Ballooning Multi-Armed Bandits

Ganesh Ghalme
Indian Institute of Science
Bangalore, India

Swapnil Dhamal
Chalmers University of Technology
Gothenburg, Sweden

Shweta Jain
Indian Institute of Technology
Ropar, India

Sujit Gujar
International Institute of Information Technology
Hyderabad, India

Y. Narahari
Indian Institute of Science
Bangalore, India

ABSTRACT

In this paper, we introduce *ballooning multi-armed bandits* (BL-MAB), a novel extension to the classical stochastic MAB model. In the BL-MAB model, the set of available arms grows (or balloons) over time. In contrast to the classical MAB setting where the regret is computed with respect to the best arm overall, the regret in a BL-MAB setting is computed with respect to the best available arm at each time. We first observe that the existing stochastic MAB algorithms are not regret-optimal for the BL-MAB model. We show that if the best arm is equally likely to arrive at any time, a sub-linear regret cannot be achieved, irrespective of the arrival of other arms. We further show that if the best arm is more likely to arrive in the early rounds, one can achieve sub-linear regret. Our proposed algorithm determines (1) the fraction of the time horizon for which the newly arriving arms should be explored and (2) the sequence of arm pulls in the exploitation phase from among the explored arms. Making reasonable assumptions on the arrival distribution of the best arm in terms of the thinness of the distribution's tail, we prove that the proposed algorithm achieves sub-linear instance-independent regret. We further quantify explicit dependence of regret on the arrival distribution parameters. We reinforce our theoretical findings with extensive simulation results.

1 INTRODUCTION

The classical stochastic multi-armed bandit (MAB) problem provides an elegant abstraction to a number of important sequential decision making problems. In this setting, the planner chooses (or pulls) a single arm in each discrete time instant from a fixed pool of finitely many arms for a finite number of time instants. Each arm, when pulled, generates a reward from a fixed but a priori unknown stochastic distribution corresponding to the pulled arm. The planner's goal is to minimize the regret, i.e., the loss incurred in expected cumulative reward due to not knowing the reward distribution of the arms beforehand. The MAB problem encapsulates the classical exploration versus exploitation dilemma, in that the planner's algorithm has to arrive at an optimal trade-off between exploration (pulling relatively unexplored arms) and exploitation (pulling the best arms according to the history of pulls thus far). This problem has been extensively studied in the literature. These studies include analyzing the lower bound on regret [21], analysis of asymptotically optimal algorithms [1, 5, 7, 29], empirical studies [13, 16, 26], and several extensions [10, 27]. We provide a detailed review of the relevant literature in Section 7.

The theoretical results in MAB are complemented by a wide variety of modern applications which can be seamlessly modelled in the MAB setup. Internet advertising [8, 25], crowdsourcing [19], clinical trials [30], wireless communication [23] represent a few of the many applications. Due to its wide applications and an elegant theoretical foundation, many variants of the MAB problem have been proposed. In this paper, we propose a novel variant which we call Ballooning multi-armed bandits (BL-MAB). In contrast to the classical MAB where the set of available arms is fixed throughout the run of an algorithm, the set of arms in BL-MAB grows (or balloons) over time. As the number of arms increases (potentially linearly) with time, it is clear that an optimal algorithm has to ignore (or drop) a few arms. Hence, in addition to achieving an optimal trade-off between the number of exploratory pulls and exploitation pulls, the algorithm must also ensure that it does not drop too many or too few arms.

To see that the traditional algorithms are not regret-optimal in the BL-MAB setting, consider the following thought experiment. Let a new arm arrive at each time instant in decreasing order of mean reward, and let the MAB algorithm run for a total of T time instants. The traditional MAB algorithms (such as UCB1, Moss etc.) would pull the newly arrived arm at each time and thus would incur a regret of $O(T)$. Note here that the best arm appeared at the first time instant itself, however, as the set of arms is monotonically expanding over time, the algorithm could not sufficiently explore the arms. Observe that the regret in BL-MAB depends not only on the mean reward of the arms, but also on when they arrive. Hence, any BL-MAB algorithm ought to be aware of the arrival of the arms.

Motivation

We motivate the practical significance of BL-MAB with a few applications. In general, BL-MAB is directly applicable in any scenario where the set of options grows over time, and, the objective is to choose the best option available at any given time.

A contemporary example is provided by question and answer (Q&A) platforms such as Reddit, Stack Overflow, Quora, Yahoo! Answers, and ResearchGate, where the platform's goal is to discover the highest quality answer that should be displayed in the most prominent slot, for a given question. Each answer post is modeled as a distinct arm of a BL-MAB instance, and the rewards are distributed according to a Bernoulli distribution parameterized by the quality of the posted answer. Note that this quality is a priori unknown to the platform and hence needs to be learnt. For this, the platform employs certain endorsement mechanisms with indicators such as upvotes, likes, and shares (or re-posts). A user endorses the answer that is displayed to her, if she likes it. Each display of a posted answer corresponds to a pull of the corresponding arm. At each time instant, a new user observes the existing answer posts shown by the platform, decides whether to endorse them, and may

also choose to post her own answer, thus increasing the number of available arms. Hence, the number of available arms (answers) monotonically increases over time.

The problem of learning qualities of the answers on Q&A forums has been modeled under the MAB framework in various studies [17, 22, 28]. However, these studies resort to the existing MAB variations which are not well suited for Q&A forums. For instance, Ghosh and Hummel [17] model the problem with a classical MAB framework by limiting the number of arms via strategic choice of an agent, by assuming that a user incurs a certain cost for posting an answer and hence posts it only if she derives a positive utility by doing so. However, a user's behavior on the platform may be driven by simple cognitive heuristics rather than a well calibrated strategic decision [11]. In another work [22], the number of arms is limited by randomly dropping some of the arms from consideration. The regret is then computed with respect to only the considered arms. That is, they do not account for the regret incurred due to the randomly dropped arms.

Some of the other applications of BL-MAB framework are in various websites that feature user reviews, such as Amazon and Flipkart (product reviews), Tripadvisor (hotel reviews), IMDB (movie reviews). As time progresses, the reviews for a product (or a hotel or a movie) keep arriving, and the website aims to display the most useful reviews for that product (or hotel or movie) at the top. The usefulness of a review is estimated using users' endorsements for that review, similar to that in Q&A forums. BL-MAB is also applicable in scenarios where users comment on a video or news article on a video or news hosting website, where the website's objective is to display the most popular or interesting comment on the top.

The BL-MAB setting thus provides a natural framework to be considered in such type of applications. It needs an independent investigation owing to a number of reasons. For instance, one of the MAB variants that holds some similarity with BL-MAB is sleeping multi-armed bandit (S-MAB) [14, 20], where a subset of a fixed set of base arms is available at each time instant. Though the S-MAB framework captures the availability of a small subset of arms at each time, it assumes that the set of base arms is fixed and is small as compared to the time horizon. In contrast, the BL-MAB framework allows for the number of available arms to increase, potentially linearly with time. Hence, an optimal sleeping bandits algorithm such as Auer would end up giving a linear regret in BL-MAB setting.

Another MAB variant which is somewhat similar is the many-armed (potentially infinite) bandit [9, 12, 31], where the number of arms could be potentially equal to or greater than the time horizon. Berry et al. [9] consider the case of an infinite arm bandit with Bernoulli reward distribution. However, they consider that the optimal arm has a quality of 1, which is seldom the case in practical applications. Other investigations considering infinitely many arms [12, 31] make certain assumptions on the distribution of the near optimal arm to achieve sub-linear regret. Further, all the above works consider that all the arms are available in all time instants, and hence use the traditional notion of regret. In our case, the regret incurred by an algorithm in a given time instant is the difference between the quality of the best available arm during that time and the quality of the arm pulled by the algorithm (same as the notion of regret considered in sleeping bandits). The BL-MAB framework is thus an interesting blend of both the sleeping bandit model and the infinite arms bandit model.

Our Contributions

- We introduce the BL-MAB model that allows the set of arms to grow over time.
- We show for the BL-MAB model that the regret will grow linearly with time, in the absence of any distributional assumption on the arrival time of the highest quality arm (Theorem 3.1).
- We propose an algorithm (BL-Moss) which determines: (1) the fraction of the time horizon until which the newly arriving arms should be explored at least once and (2) the sequence of arm pulls during the exploitation phase. Our key finding is that BL-Moss achieves sub-linear regret under practical and minimal assumptions on the arrival distribution of the best arm, namely, sub-exponential tail (Theorem 5.3) and sub-Pareto tail (Theorem 5.5). Note that we make no assumption on the arrival of the other arms. As the regret depends on the qualities of the arms and the sequence of their arrivals, it is interesting that with sub-exponential and sub-Pareto assumption on only the best arm's arrival pattern, we can achieve sub-linear regret.
- We carry out a pertinent simulation study to empirically observe how the expected regret varies with the time horizon. We find a strong validation for our theoretically derived regret bounds.

The paper is organized as follows. In Section 2, we present the BL-MAB model. In Section 3, we first show that if the best arm arrives uniformly at random, one cannot achieve sub-linear regret. We hence define two distributions on the arrival time of the best arm which enables us to achieve sub-linear regret. Next, we present some preliminaries in Section 4, followed by our proposed algorithm and its theoretical analysis in Section 5. Section 6 presents our simulation results. We conclude the paper with related work (Section 7) and future directions (Section 8).

2 THE MODEL

A classical MAB instance is given by the tuple $\langle K, (\mathcal{D}_i)_{i \in K} \rangle$. Here, K is a fixed set of arms and $\mathcal{D}_i$ is the reward distribution corresponding to an arm i. Denote by q_i, the mean of distribution $\mathcal{D}_i$. Consider that each of the distributions $\mathcal{D}_i$ is supported over a finite interval and is unknown to the algorithm. Throughout the paper, without loss of generality, we consider that $\mathcal{D}_i$ is supported over $[0, 1]$. Further, we will refer to q_i as the quality of the arm i. A MAB algorithm is run in discrete time instants, and the total number of time instants is denoted by time horizon T. In each time instant aka round, the algorithm selects a single arm and observes the reward corresponding to the selected arm. The arms which are not selected, do not give any reward. More precisely, a MAB algorithm is a mapping from the history of arm pulls and obtained rewards, to the set of arms.

At each time instant, a BL-MAB algorithm chooses a single arm from the set of available arms and receives a reward generated randomly according to the reward distribution $\mathcal{D}_i$ of the chosen arm i. New arms may spring up at each time instant. Throughout the paper, we consider that at most one new arm arrives at each time, and the arms are never dropped. Let $K(t)$ denote the set of arms available at round t. In the BL-MAB model, this set of available arms grows by at most one arm per round, i.e., $K(t) \subseteq K(t+1)$ and

$|K(t)| \leq |K(t+1)| \leq |K(t)|+1$. A BL-MAB instance, therefore, is given by $\langle T, (K(t), (\mathcal{D}_i)_{i\in K(t)})_{t=1}^{T}\rangle$.

Similar to the notion of regret in the sleeping stochastic MAB model, we introduce the notion of regret in BL-MAB setting that takes into account the availability of the arms at each time t. Let i_t denote the arm pulled by the algorithm and $i_t^\star$ be the best available arm at time t, i.e., $i_t^\star = \arg\max_{i\in K(t)} q_i$. Further, let $\mathcal{I}$ denote a BL-MAB instance and A be a BL-MAB algorithm. The instance-dependent regret of A is given by

$$\mathcal{R}_A(T,\mathcal{I}) = \mathbb{E}\Big[\sum_{t=1}^{T}(q_{i_t^\star} - q_{i_t})\Big].$$

Throughout the paper, we consider instance-independent regret given as $\mathcal{R}_A(T) = \sup_{\mathcal{I}} \mathcal{R}_A(T,\mathcal{I})$. Note that the instance-independent regret bound is a worst case regret bound over all the arrival sequences of the arms and all possible reward distributions. In the next section, we show, for the BL-MAB setting, that it is not possible to achieve sublinear instance-independent regret bound.

3 LOWER BOUND ON REGRET

As pointed out in Section 1, it is clear that UCB-style algorithms (which pull arms based on uncertainty) would pull each incoming arm at least once, leaving no rounds for exploitation. Hence, they incur linear regret in the ballooning bandit setup (in particular, when $K(t) = t$). However, it is not obvious that a different, more sophisticated algorithm (such as the one which randomly drops some arms) may not be able to achieve sub-linear regret. Our first result (Theorem 3.1) shows that no algorithm can attain sub-linear regret under a general BL-MAB setting.

Consider the following BL-MAB instance $\mathcal{I}$. Let there be a unique best arm $i^\star$ with quality $q_{i^\star} = 1/2+\varepsilon$ and all other arms $j \neq i^\star$ have quality $q_j = 1/2$. A new arm arrives at each time, i.e., $|K(t)| = t$. Further, let the arrival of the best arm be uniformly distributed over time, i.e., $\mathbb{P}(t = i^\star) = 1/T$ for all $t = 1, 2, \ldots, T$. Let i_t^* denote the optimal arm till time t. Further, let G be the set of arms pulled by the algorithm, i.e., $G = \{i : N_{i,T} \geq 1\}$. We will show that for any fixed $G \subset \{1, 2, \cdots, T\}$, the expected regret is lower bounded by $\Omega(T)$. Here, expectation is taken over randomness in arrival of arms as well as in the algorithm (if any). We first observe that an algorithm that pulls $|G|$ number of arms achieves the minimum regret when it pulls the first $|G|$ arms (see Claim 1 in Appendix).

Theorem 3.1. *For the BL-MAB instance $\mathcal{I}$, the expected regret of any algorithm A is lower bounded by $\Omega(T)$.*

Proof. The expected regret of a BL-MAB algorithm is given by

$$\begin{aligned}\mathcal{R}_A(T) \geq \mathcal{R}_A(T,\mathcal{I}) &= \mathbb{E}_A\Big[\sum_{t=1}^{T}(q_{i_t^\star} - q_{i_t})\Big] \\ &\geq P(i^\star \in G)\sum_{i\in G}\mathbb{E}[N_{i,T}]\Delta(i^\star,i) + \sum_{i\notin G}P(i^\star = i)(T-i)\Delta(i^\star,i)\end{aligned}$$

In the above expression, the first term represents the internal regret of the learning algorithm and the second quantity is the external regret. Here, $\Delta(i^*, i) = q_{i^\star} - q_i = \varepsilon$ if $i^\star < i$ and 0 otherwise. From Claim 1, we have that an algorithm will incur least regret if it pulls first $|G|$ arms. Further, from the classical result in [21], in order to separate the quality of the arms, we should have $\mathbb{E}[N_{i,T}] \geq \eta\cdot\log(T)$ for some positive problem dependent constant η, for all $i \in G$ and $i \neq i^\star$. Hence, we have

$$\begin{aligned}\mathcal{R}_A(T) &\geq P(i^\star \in G)\sum_{i=1}^{|G|-1}\eta\log(T)\varepsilon + \sum_{i=|G|+1}^{T}P(i=i^\star)(T-i)\varepsilon \\ &= \Big[\frac{\eta|G|(|G|-1)\log(T)}{T} + \frac{(T-|G|-1)(T-|G|)}{2T}\Big]\cdot\varepsilon \\ &= \Big[(1+2\eta\log(T))|G|^2 - (2(T-\eta\log(T))-1)|G| + T^2 - T\Big]\cdot\frac{\varepsilon}{2T}\end{aligned}$$

Note that the above expression is quadratic in $|G|$. For $T \leq 1/2 + \eta\log(T)$, the minimum occurs when the value of $|G|$ is the least (in the positive domain), which is 1, for which the above expression cannot be sub-linear in T. For the case where $T > 1/2 + \eta\log(T)$, the minimum occurs when $|G| = \frac{2(T-\eta\log(T))-1}{2(1+2\eta\log(T))}$. Hence,

$$\begin{aligned}\mathcal{R}_A(T) &\geq \Big[\frac{(2(T-\eta\log(T))-1)^2}{4(1+2\eta\log(T))} - \frac{(2(T-\eta\log(T))-1)^2}{2(1+2\eta\log(T))} + T^2 - T\Big]\cdot\frac{\varepsilon}{2T} \\ &= \Big[T^2 - T - \frac{(T-\eta\log(T)-1/2)^2}{(1+2\eta\log(T))}\Big]\cdot\frac{\varepsilon}{2T} \\ &> \Big[\frac{(T-1/2)}{2}\frac{2\eta\log(T)}{1+2\eta\log(T)} - 1/4\Big]\cdot\varepsilon \;=\; \Omega(T)\end{aligned}$$ □

Theorem 3.1 provides a strong impossibility result on the achievable instance-independent regret bound under BL-MAB setting. However, one can still achieve sub-linear regret by imposing appropriate structure on the BL-MAB instances. Observe that the regret depends on the arrival of arms, i.e., $(K(t))_{t=1}^{T}$, and their reward distributions $(\mathcal{D}_i)_{i\in K(t)}$. We impose restrictions on the arrival of the best arm $i^\star = \arg\max_{i\in K(T)} q_i$ so that the probability that $i^\star$ arrives early is large enough; this would allow a learning algorithm to explore the best arm enough to estimate the true quality of that arm with high probability. As noted previously, the other arms may arrive arbitrarily. Further note that we make no assumption on the qualities of individual arms.

Arrival of the Best Arm

Let X be the random variable denoting the time at which the best arm arrives. Further, let $F_X(t)$ denote the cumulative distribution function of X. In our first result, we use the following Sub-exponential tail assumption on the arrival time of the best arm.

Sub-exponential tail. There exists a constant $\lambda > 0$ such that the probability of the best arm arriving later than t rounds, is upper bounded by $e^{-\lambda t}$, i.e., $F_X(t) > 1 - e^{-\lambda t}$.

Next, we consider a relaxed condition on the tail probabilities, i.e., when the tail does not shrink as fast as in the sub-exponential case. We consider the family of distributions whose tail is thinner than that of Pareto distribution.

Sub-Pareto tail. There exists a constant $\beta > 0$ such that the probability of the best arm arriving later than t rounds, is upper bounded by $t^{-\beta}$, i.e., $F(t) > 1 - t^{-\beta}$.

The aforementioned assumptions naturally arise in the context of Q&A forums as observed in extensive empirical studies on the nature of answering as well as voting behavior of the users. Anderson et al. [2] observe that high reputation users hasten to post their answers early. One possible explanation for this phenomenon could be that the users who are motivated by the visibility that

their answers receive, tend to be more active on the platform and also provide high quality answers early on, which explains their reputation score. Thus, it is reasonable to assume that the best answer arrives, with high probability, in early rounds.

Note that the uniform distribution is the limiting case of the sub-exponential case, when $\lambda = 0$. We show that, while the uniform distribution results in linear regret (Theorem 3.1), a sub-linear regret can be achieved for BL-MAB instances having the best arm arrival distribution with even slightly thinner tail than that of uniform distribution.

4 PRELIMINARIES

We now present some essentials which will be useful for our analysis in the remainder of the paper.

Lambert W Function

Definition 4.1. For any $x > -e^{-1}$, the Lambert W function, $W(x)$, is defined as the solution to the equation $we^w = x$, i.e., $W(x)e^{W(x)} = x$.

Lambert W function satisfies the following properties [18]:

P 1. *The Lambert W function can be equivalently written as the inverse of the function $f(x) := xe^x$, i.e., $W(xe^x) = x$.*

P 2. *For any $x \in [0, \infty)$, the Lambert W function is unique, non-negative, and strictly increasing.*

P 3. *For any $x \geq e$, we have $\log(x)/2 < W(x) \leq \log(x)$.*

It can be noted that it is easy to numerically approximate $W(x)$ for a given x, using Newton-Raphson's or Halley's method. Moreover, there exist efficient numerical methods for evaluating it to arbitrary precision [15].

The Moss Algorithm

We use Moss (Minimax Optimal Strategy in the Stochastic case) [3] as a black box learning algorithm. For a fixed number of k arms, the Moss algorithm pulls an arm i_t at time t such that

$$i_t \in \arg\max_{i \in K} \left[\hat{q}_{i,N_{i,t}} + \sqrt{\frac{\max(\log(\frac{T}{k \cdot N_{i,t}}), 0)}{N_{i,t}}} \right].$$

Here, $K = \{1, 2, \cdots, k\}$ denotes the set of arms and $N_{i,t}$ is the number of times arm i was pulled before (and excluding) round t and $\hat{q}_{i,N_{i,t}}$ are the empirical estimates of the arm i from $N_{i,t}$ samples. Each arm is pulled once in the beginning, and ties are broken arbitrarily. The following result gives an upper bound on the expected regret of Moss which is optimal up to a constant factor (it achieves the lower bound on regret given by [6]). Throughout the paper, we use the notation Moss (K) to denote that the Moss algorithm is run with set of arms K.

Theorem 4.2. *[Audibert and Bubeck [3]] For any time horizon $T \geq 1$, the expected regret of Moss is given by $\mathcal{R}_{Moss}(T) \leq 6\sqrt{kT}$.*

5 THE BL-MOSS ALGORITHM AND REGRET ANALYSIS

The BL-Moss Algorithm

We now present our algorithm, BL-Moss (Algorithm 1), that uses Moss as a black-box. The number of arms explored by BL-Moss is dependent on the distribution of arrival of the best arm. In particular, BL-Moss considers only the first $\lceil \alpha T \rceil$ arms in its execution ($\alpha \in (0, 1]$). Later in this section, we show how to derive the value of α for distributions with sub-exponential and sub-Pareto tails. Observe that the proposed BL-Moss is a simple extension of Moss and this algorithm is practically easy to implement. Further, Moss does not assume any structure on the arrival of suboptimal arms. Thus we are able to obtain sub-linear regret with minimal assumptions.

Algorithm 1: BL-Moss

Input: Time horizon T, Distributional parameter λ or β

$$\text{Set } \alpha := \begin{cases} \frac{W(2\lambda T)}{2\lambda T} & \text{under sub-exponential tail property} \\ T^{\frac{-2\beta}{1+2\beta}} & \text{under sub-Pareto tail property} \end{cases}$$

for $t = 1, 2, \ldots T$ **do**
  Input: A newly arriving arm at time t
  if $|K(t)| \leq \lceil \alpha T \rceil$ **then**
    Moss $(K(t))$
  else
    Moss $(K(\lceil \alpha T \rceil))$

Regret Analysis of BL-Moss

For a given BL-MAB instance $\mathcal{I}$, let $j^\star = \arg\max_{i \in K(\lceil \alpha T \rceil)} q_i$ and $i^\star = \arg\max_{i \in K(T)} q_i$. Clearly, we have that $q_{i^\star} \geq q_{j^\star}$. As stated earlier, the regret of the algorithm can be decomposed into internal regret, i.e., the regret incurred by the learning algorithm considering only $\lceil \alpha T \rceil$ arms and external regret, i.e., the regret incurred by BL-Moss due to the fact that BL-Moss might have ignored the best arm. Write $\Delta(i, j) = q_i - q_j$ and let t_i be the time of arrival of arm i. Further, let $i_t^\star$ denote the best arm till time t. The instance-dependent regret $\mathcal{R}_{\text{BL-Moss}}(T, \mathcal{I})$ is given as

$$\mathbb{P}(i^\star = j^\star) \underbrace{\left[\sum_{t=1}^{t_{j^\star}-1} \Delta(i_t^\star, i_t) + \sum_{t=t_{j^\star}}^{T} \Delta(j^\star, i_t) \right]}_{\mathcal{R}^{\text{int}}_{\text{BL-Moss}}(T)}$$
$$+ \mathbb{P}(i^\star \neq j^\star) \underbrace{\left[\sum_{t=1}^{t_{i^\star}-1} \Delta(i_t^\star, i_t) + \sum_{t=t_{i^\star}}^{T} \Delta(i^\star, i_t) \right]}_{\mathcal{R}^{\text{ext}}_{\text{BL-Moss}}(T)}$$

The first and the second terms respectively denote the internal regret and the external regret of BL-Moss. We ignore the ceiling in $\lceil \alpha T \rceil$ throughout this section to avoid notation clutter.

Note that $\mathcal{R}_{\text{Moss}(L)}(T) \leq \mathcal{R}_{\text{Moss}(K)}(T)$ for all $L \subset K$. This is true for any time horizon T. From Theorem 4.2, we have the following observation about the internal regret of BL-Moss.

Observation 1. *For the value of α computed by BL-Moss , we have $\mathcal{R}^{int}_{BL\text{-}Moss}(T) \leq \mathcal{R}_{Moss\,(\alpha T)}(T) \leq 6\sqrt{\alpha}T$.*

In order to bound the overall regret, we begin with the following lemma which explicitly shows the relation between the expected regret of the algorithm and $F_X(\cdot)$. Recall that the random variable X denotes the time of arrival of the best arm.

Lemma 5.1. *The upper bound on the expected regret for any BL-MAB instance is given by $\mathcal{R}_{BL\text{-}Moss}(T) \leq T(1-(1-6\cdot\sqrt{\alpha})F_X(\alpha T))$, with BL-Moss exploring only the first αT arrived arms.*

Proof. For a given BL-MAB instance $\mathcal{I}$, let t_i denote the time at which arm i becomes available for the first time. Let $i^\star$ denote the best arm till T rounds, i.e., $i^\star = \arg\max_{i\in K(T)} q_i$. Further, let $j^\star$ be the best arm among the arms considered by BL-Moss, i.e., $j^\star = \arg\max_{j\in K(\alpha T)} q_i$. Notice that $K(\alpha T) \subseteq K(T)$. This implies $q_{i^\star} \geq q_{j^\star}$.

$$\begin{aligned}
\mathcal{R}_{\text{BL-Moss}}(T,\mathcal{I}) &\leq \mathbb{E}\Big[\sum_{t=1}^{\alpha T}(q_{j^\star}-q_{i_t}) + \sum_{t=\alpha T+1}^{T}(q_{i^\star}-q_{i_t})\Big] \\
&\qquad (\because q_{i^\star} > q_{j^\star}) \\
&= \mathbb{P}(i^\star = j^\star)\Big[\sum_{t=1}^{T}(q_{j^\star}-q_{i_t})\Big] \\
&\quad + \mathbb{P}(i^\star \neq j^\star)\Big[\sum_{t=1}^{\alpha T}(q_{j^\star}-q_{i_t}) + \sum_{t=\alpha T+1}^{T}(q_{i^\star}-q_{i_t})\Big] \\
&\leq 6\mathbb{P}(i^\star = j^\star)\sqrt{\alpha T\cdot T} + \sum_{t=1}^{T}(q_{i^\star}-q_{i_t})\mathbb{P}(i^\star \neq j^\star) \\
&\qquad \text{(From Observation 1 and since } q_{i^\star} \geq q_{j^\star}) \\
&\leq 6T\sqrt{\alpha}\cdot\mathbb{P}(i^\star = j^\star) + \mathbb{P}(i^\star \neq j^\star)T \\
&\qquad (\because \textstyle\sum_{t=1}^{T}(q_{i^\star}-q_{i_t}) \leq T) \\
&= 6T\sqrt{\alpha}\cdot\mathbb{P}(t_{i^\star} \leq \alpha T) + (1-\mathbb{P}(t_{i^\star} \leq \alpha\cdot T))T \\
&= T(1-(1-6\cdot\sqrt{\alpha})\mathbb{P}(t_{i^\star} \leq \alpha T)) \\
&= T(1-(1-6\cdot\sqrt{\alpha})F_X(\alpha T))
\end{aligned}$$

Note that the above inequality holds for any BL-MAB instance and hence we have $\mathcal{R}_{\text{BL-Moss}}(T) = \sup_{\mathcal{I}} \mathcal{R}_{\text{BL-Moss}}(T,\mathcal{I}) \leq T(1-(1-6\cdot\sqrt{\alpha})F_X(\alpha T))$. □

Next, we show that under the sub-exponential tail property on X, BL-Moss achieves sub-linear regret. We begin with the following lemma that lower bounds the probability of the arrival of the best quality arm in the initial αT rounds.

Lemma 5.2. *Let the arm arrival distribution of the best arm satisfy sub-exponential tail property for some $\lambda > 0$. Then for any $c > 0$ and $\alpha \geq \frac{W(\lambda T/c)}{\lambda T/c}$, we have that $F_X(\alpha T) > (1-\alpha^c)$.*

Proof. We have that $\alpha \geq \frac{W(\lambda T/c)}{\lambda T/c} \implies \frac{\alpha\lambda T}{c} \geq W(\lambda T/c) \implies W\big(\frac{\alpha\lambda T}{c}\cdot e^{\alpha\lambda T/c}\big) \geq W(\lambda T/c)$ (by Property **P1**) $\implies \frac{\alpha\lambda T}{c}\cdot e^{\alpha\lambda T/c} \geq \lambda T/c$ (by Property **P2**) $\implies \alpha \geq e^{-\alpha\lambda T/c}$. So, we have $1-\alpha^c \leq 1-e^{-\lambda(\alpha T)} < F_X(\alpha T)$. The last inequality follows from the sub-exponential tail property. □

Theorem 5.3. *Let the arrival distribution of the best arm satisfy the sub-exponential tail property for some $\lambda > 0$, and let T be large enough such that $T > \frac{36c\log(36)}{\lambda}$ for some $c > 0$. Then with $\alpha = \frac{W(\lambda T/c)}{\lambda T/c}$, the upper bound on the expected regret of BL-Moss, $\mathcal{R}_{BL\text{-}Moss}(T)$, is $O\Big(T\cdot\max\big(e^{-cW(\lambda T/c)}, e^{-\frac{W(\lambda T/c)}{2}}\big)\Big)$. The upper bound on the expected regret is minimized when $c = 1/2$ and is given by $O\Big(\sqrt{\frac{T\log(2\lambda T)}{2\lambda}}\Big)$.*

Proof. From Lemma 5.2, we have $F_X(\alpha T) > 1-\alpha^c$ for all $\alpha \geq \frac{W(\lambda T/c)}{\lambda T/c}$. Thus, from Lemma 5.1, we have $\mathcal{R}_{\text{BL-Moss}}(T) < T(1-(1-6\cdot\sqrt{\alpha})(1-\alpha^c))$.

Note that for achieving sub-linear regret, it is necessary that $(1-6\cdot\sqrt{\alpha})$ is strictly positive, for which it is necessary that $\alpha < 1/36$. From Lemma 5.2, we also have $\alpha \geq \frac{W(\lambda T/c)}{\lambda T/c}$. Since such a feasible α may not exist for small values of T, we consider that T is large enough. It can be easily shown that $\frac{W(\lambda T/c)}{\lambda T/c} < 1/36 \iff T > \frac{36c\log(36)}{\lambda} \approx \frac{129c}{\lambda}$ (see Claim 2 in Appendix).

Thus, for $1/36 > \alpha \geq \frac{W(\lambda T/c)}{\lambda T/c}$, we have: $\mathcal{R}_{\text{BL-Moss}}(T) < T(6\cdot\sqrt{\alpha}+\alpha^c-6\cdot\alpha^{c+1/2})$. Recall that by definition, we have $\alpha \leq 1$. Thus when $c \in (0,1/2]$, the term α^c dominates the other terms in the regret expression, whereas when $c > 1/2$, the term $\sqrt{\alpha}$ dominates. We analyze these cases separately.

Case 1 ($c \in (0,1/2]$): In this case, the regret is given by $\mathcal{R}_{\text{BL-Moss}}(T) = O(\alpha^c T)$. Note that the regret is minimized for the lowest feasible value of α, i.e., $\alpha = \frac{W(\lambda T/c)}{\lambda T/c}$, resulting in $\mathcal{R}_{\text{BL-Moss}}(T) = O\Big(T\big(\frac{W(\lambda T/c)}{\lambda T/c}\big)^c\Big) = O(T\cdot e^{-cW(\lambda T/c)})$. The last equality follows from the definition of Lambert W function.

Case 2 ($c \in [1/2,\infty)$): In this case, the regret is given by $\mathcal{R}_{\text{BL-Moss}}(T) = O(\sqrt{\alpha}T)$. Again, the regret is minimized when $\alpha = \frac{W(\lambda T/c)}{\lambda T/c}$. The regret in this case is given by $\mathcal{R}_{\text{BL-Moss}}(T) = O\Big(T\cdot\sqrt{\frac{W(\lambda T/c)}{\lambda T/c}}\Big) = O\big(T\cdot e^{\frac{-W(\lambda T/c)}{2}}\big)$.

Further, we have that in Case 1, $e^{-cW(\lambda T/c)} > e^{\frac{-W(2\lambda T)}{2}}$ for any $c \in (0,1/2)$ (see Claim 3 in Appendix). For Case 2, we have from Property **P2** that, $W(\lambda T/c)$ is decreasing in c, which gives us that $e^{\frac{-W(2\lambda T)}{2}} < e^{\frac{-W(\lambda T/c)}{2}}$ for any $c \in (1/2,\infty)$. This shows that the minimum regret is achieved when $c = 1/2$, and the regret is given by $\mathcal{R}_{\text{BL-Moss}}(T) = O\Big(\sqrt{\frac{T\cdot W(2\lambda T)}{2\lambda}}\Big) = O\Big(\sqrt{\frac{T\log(2\lambda T)}{2\lambda}}\Big)$. The last inequality follows from Property **P3**, since $2\lambda T \geq e$ ($\because T > \frac{36c\log(36)}{\lambda}$ where $c = 1/2$). □

We now prove the sub-linear regret of BL-Moss under the sub-Pareto tail property.

Lemma 5.4. *Let the arm arrival distribution of the best arm satisfy sub-Pareto tail property for some $\beta > 0$. Then for any $c > 0$ and $\alpha \geq T^{\frac{-\beta}{c+\beta}}$, we have that $F_X(\alpha T) > (1-\alpha^c)$.*

Proof. First note that $\alpha \geq T^{\frac{-\beta}{c+\beta}} \iff \alpha^c \geq (\alpha T)^{-\beta}$. This implies that $(1-\alpha^c) \leq 1-(\alpha T)^{-\beta}$. Further, from the sub-Pareto tail property, we have that $1-(\alpha T)^{-\beta} < F_X(\alpha T)$. □

Theorem 5.5. *Let the arrival distribution of arms satisfy the sub-Pareto tail property for some $\beta > 0$, and let T be large enough such that $T > (36)^{\frac{c+\beta}{\beta}}$ for some $c > 0$. Then with $\alpha = T^{\frac{-\beta}{\beta+c}}$, the upper bound on the expected regret of BL-Moss, $\mathcal{R}_{BL\text{-}Moss}(T)$, is $O(\max(T^{\frac{c+\beta(1-c)}{c+\beta}}, T^{\frac{2c+\beta}{2(c+\beta)}}))$. The upper bound on the expected regret is minimized when $c = 1/2$ and is given by $O(T^{\frac{1+\beta}{1+2\beta}})$.*

Proof. From Lemmas 5.1 and 5.4, we have $\mathcal{R}_{\text{BL-Moss}}(T) < T(1-(1-6\cdot\sqrt{\alpha})(1-\alpha^c))$. For achieving sub-linear regret, it is necessary that $(1-6\cdot\sqrt{\alpha})$ is strictly positive. So, we should have $\alpha < 1/36$. Further, from Lemma 5.4, we have $\alpha \geq T^{\frac{-\beta}{c+\beta}}$. So, for a feasible α to exist, it is necessary that $T^{\frac{-\beta}{c+\beta}} < 1/36 \iff T > (36)^{\frac{c+\beta}{\beta}}$, i.e., T is large enough. Thus, for $1/36 > \alpha \geq T^{\frac{-\beta}{c+\beta}}$, we have $\mathcal{R}_{\text{BL-Moss}}(T) < T(6\cdot\sqrt{\alpha}+\alpha^c-6\cdot\alpha^{c+1/2})$. As earlier, we analyze two cases.

Case 1 ($c \in (0, 1/2]$): In this case, the regret is given by $\mathcal{R}_{\text{BL-Moss}}(T) = O(\alpha^c T)$. The minimum regret is obtained when $\alpha = T^{\frac{-\beta}{\beta+c}}$ and is given by $O(T^{1-\frac{c\beta}{c+\beta}})$.

Case 2 ($c \in [1/2, \infty)$): In this case, the regret is given by $\mathcal{R}_{\text{BL-Moss}}(T) = O(\sqrt{\alpha}T)$. Again, the regret is minimum when $\alpha = T^{\frac{-\beta}{\beta+c}}$ and is given by $O(T^{\frac{2c+\beta}{2(c+\beta)}})$.

Furthermore, it is easy to see that in Case 1, $T^{\frac{1+\beta}{1+2\beta}} > T^{\frac{\beta+c(1-\beta)}{c+\beta}}$ for any $c \in (0, 1/2)$. Similarly, in Case 2, $T^{\frac{1+\beta}{1+2\beta}} > T^{\frac{2c+\beta}{2(c+\beta)}}$ for any $c \in (1/2, \infty)$. Thus, the minimum regret is achieved when $c = 1/2$. □

Important Observations

We conclude the section with some key observations.

- In the sub-exponential tail case, as $\lambda \to \infty$, we have $\frac{W(2\lambda T)}{2\lambda T} \to 0$. This implies that the upper bound on the expected regret goes to 0. Note that in this case, $\lceil \alpha T \rceil = 1$. Since BL-Moss considers a single arm, the internal regret is zero. Further, we have $F_X(1) \to 1$, i.e., the first arm is optimal with probability approaching 1, the external regret is also zero. As $\lambda \to 0$, the tail bounds are trivial and are satisfied by uniform distribution. From Theorem 3.1, we have that the regret in this case cannot be sub-linear.
- In the sub-Pareto tail case, following the similar argument as in the sub-exponential tail case, we have that as $\beta \to 0$, the regret $\mathcal{R}_{\text{BL-Moss}}(T) \to O(T)$. On the other hand, as $\beta \to \infty$, the regret goes to $O(\sqrt{T})$. The larger value of β implies that the probability that the optimal arm arrives by $t = 2$ is close to 1; then we have that the regret of BL-Moss is asymptotically optimal. Further, it asymptotically achieves the information theoretic lower bound (which is $O(\sqrt{T})$).
- One could also consider UCB1 instead of Moss. While UCB1 is any-time algorithm, Moss needs the time horizon as an input. However, the important distinction between the two algorithms is that the instance-independent regret of UCB1, which is $O(\sqrt{kT\log(T)})$, is greater than that of Moss; hence we use Moss in BL-Moss. One can similarly use UCB1 to get any-time version of BL-Moss with slightly more (up to $\sqrt{\log(T)}$) regret guarantee.

6 SIMULATION STUDY

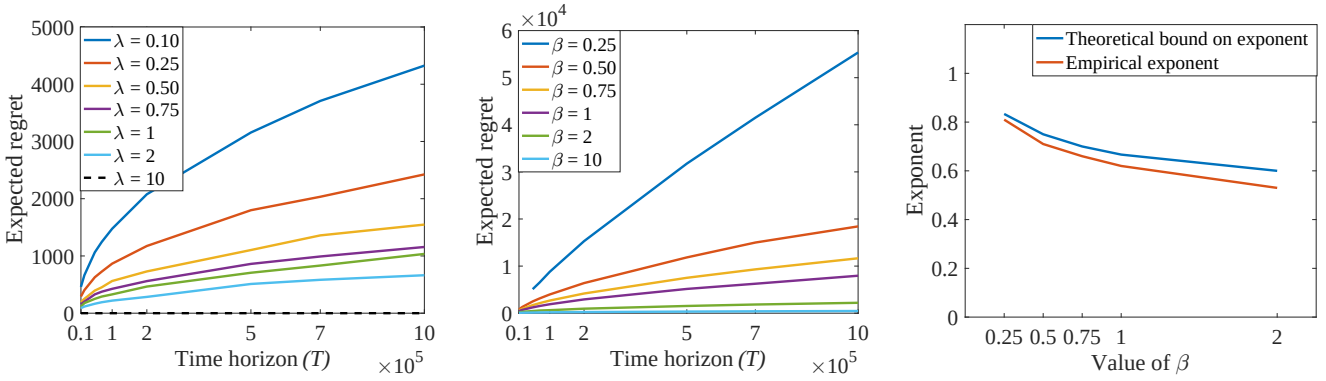
So far, we focused on deriving upper bounds on regret for distributions (on the arrival time of the best arm) having sub-exponential and sub-Pareto tail with different values of λ and β, respectively. In particular, for the case of sub-Pareto tail, we deduced that the extent of sublinearity of the regret (the exponent of T in the order of the regret) depends on the value of β. On the other hand, the upper bound on regret for the case of sub-exponential tail had the same order with respect to T for any reasonable value of λ, albeit with different multiplicative and additive terms for different values of λ. In this section, we aim to illustrate how the expected regret varies with the time horizon T, and how the empirical exponents compare with their theoretical bounds for different values of β and λ, for time horizons up to 10^6 rounds.

Simulation Setup

Note that in a traditional MAB setup, a simulation for a larger time horizon T'' could be conducted as an extension of a simulation for a smaller time horizon $T' < T''$. In other words, after obtaining the results for time horizon T', the results for time horizon T'' can be obtained by running simulations for an additional $T'' - T'$ rounds. However, in the BL-MAB setup where new arms continue arriving with time and the desired time horizon is known, we have seen that the optimal value of α and hence $\lceil \alpha T \rceil$ depend on the time horizon. Owing to different values of $\lceil \alpha T \rceil$ for different time horizons T, the simulation for a time horizon T' are not extendable to time horizon $T'' > T'$. So even if we have simulation results for time horizon T', it is necessary to run a fresh set of simulations for obtaining results for time horizon $T'' > T'$. In our simulation study, we consider the following values of time horizon: $\{1, 2, 5, 7\} \times 10^4$, $\{1, 2, 5, 7\} \times 10^5$, 10^6.

We consider that a new arm arrives in each round, and the probability of an arm arriving at time t being the best arm is determined by the distribution function $F_X(t)$. Thereafter, this best arm ($i^\star$) is assigned a quality ($q_{i^\star}$) between 0 and 1 uniformly at random, and the rest of the arms are assigned quality parameters between 0 and $q_{i^\star}$ uniformly at random. Given a time horizon T, the value of α and hence $\lceil \alpha T \rceil$ are obtained based on our theoretical analysis. The arm to be pulled in a round is determined by Algorithm 1, wherein the pulled arm generates unit reward with probability equal to its quality, and no reward otherwise (i.e., as per Bernoulli distribution). The regret in each round is computed as the difference between the quality of the best arm available in that round and the quality of the pulled arm. The overall regret is the sum of the regrets over all rounds from 1 till T. Note that we are concerned with the regret irrespective of the numerical values of the arms' qualities. So, for a given instance of the arrival of the best arm, we consider the worst-case regret over 50 sub-instances, where the quality parameters assigned to the arms in different sub-instances are independent of each other. Also, since different instances would have the best arm arriving in different rounds, the expected regret is obtained by simulating over 1000 such random instances and averaging over the corresponding worst-case regret values.

Our primary objective is to observe how the expected regret varies with the time horizon T. In order to observe the influence of various sub-exponential and sub-Pareto tail distributions over the arrival time of the best arm, we conduct simulations for different values of parameters λ and β: $\{0.10, 0.25, 0.50, 0.75, 1, 2, 10\}$. The other objective is to determine the empirical exponent of the plots (i.e., the value of γ such that the expected regret is approximately a constant multiple of T^γ). To achieve this, we first estimate the constant factor ξ by dividing the expected regret for $T = 10^6$ by T^γ, for a given value of γ. We then compute the squared error when attempting to fit the expected regret with ξT^γ. Considering candidate values of γ to be between 0 and 1 with intervals of 0.01, we deduce the empirical exponent to be the value of γ which results in the least squared error. We also consider another method for determining the empirical exponent: we produce the line of best fit



(a) Expected regret as a function of time horizon for various sub-exponential tails

(b) Expected regret as a function of time horizon for various sub-Pareto tails

(c) Empirical exponents vs. theoretical bounds for time horizons up to 10^6

Figure 1: Simulation results

for the scatter plot of $\log(T)$ versus the log of the expected regret for that T; the slope of this line gives the empirical exponent. The empirical exponents obtained using the two methods are almost identical (differing by less than 0.01).

Simulation Results

As mentioned at the end of our theoretical analysis, for the sub-exponential tail case when $\lambda \to \infty$, the upper bound on the expected regret goes to 0. In our simulations with the maximum observed time horizon of 10^6, the expected regret was observed to be uniformly zero, even for $\lambda = 10$ (see Figure 1(a)). Further, for other considered values of λ, the plots exhibit a prominent sub-linear nature. In particular, considering the maximum observed time horizon of 10^6, the empirical exponents for different values of λ were consistently observed to be between 0.45 and 0.5 (Theorem 5.3 showed the order of the regret with respect to T, for reasonable values of λ, to be bounded by $\sqrt{T\log(T)}$, which is an exponent close to 0.5).

For the sub-Pareto tail case illustrated in Figure 1(b), note that we have no result for $\beta = 0.10$ because the value of T for obtaining a feasible α should be greater than 36^6, which is beyond our maximum observed time horizon of 10^6. Moreover, we have partial results for $\beta = 0.25$ because the value of T for obtaining a feasible α should be greater than 36^3; so the plot starts with $T = 0.5 \times 10^5$. It can be seen, in general, that the plots in Figure 1(b) follow a far less sub-linear nature and exhibit a much higher expected regret than those in Figure 1(a). This is intuitive from our analysis that the sub-exponential tail case is likely to result in a much lower regret than the sub-Pareto tail case. In particular, the empirical exponent for $\beta = 0.25$ was deduced to be 0.8, which is close to linear (its theoretical upper bound as per our analysis is 0.83). In general, considering the maximum observed time horizon of 10^6, it can be seen from Figure 1(c) that the upper bound on the theoretical exponent (which is $\frac{1+\beta}{1+2\beta}$ from Theorem 5.5) and the empirical exponent are close to each other.

Note that the gap between the empirical exponents and the corresponding theoretical upper bounds could be attributed to the fact that it is difficult to find the worst-case distribution over the reward parameters of the arms. Hence, it is unlikely that the worst-case (or instance-independent) expected regret could be attained in the simulations with a random reward structure. Since the gap is not very significant, the simulation results suggest that the bounds derived in our regret analysis of BL-Moss (in Section 5) are, in all probability, tight.

Additional Notes on Simulations

It is to be noted that our theoretical analysis holds for any arbitrary time horizon as long as the time horizon is known to BL-Moss. In our simulations, we considered time horizons up to 10^6 for computational reasons. The expected regret for a given arrival distribution of the best arm is computed using 50000 random instances (by averaging over 1000 instances for different arrival times of the best arm, where in each instance, the worst case is taken over 50 sub-instances for different quality parameters). In practice, as only one instance is realized, the computational overhead is not an impediment in the real world applicability of the proposed algorithm.

Note also that the standard MAB algorithms (e.g., the UCB family) which are oblivious to the structure on the arrival of arms, would incur linear regret. Also, since these algorithms explore each incoming arm at least once, they would incur linear regret even with sub-exponential or sub-Pareto assumption, when the number of arms grows linearly with time. Our simulations aimed to observe the order of sublinearity of regret (exponent of T). Since existing algorithms would give linear regret, the exponent of T is trivially 1.

7 ADDITIONAL RELATED WORK

A standard stochastic MAB framework considers that the number of available arms is fixed (say k) and typically much less than the time horizon (say T). In the seminal work of Lai and Robbins [21], the authors showed that any MAB algorithm in such a setting must incur a regret of $\Omega(\frac{\log T}{D_{\text{KL}}})$ where D_{KL} is the Kullback-Leibler divergence between the best arm and the second best arm. Auer [4] proposed the UCB1 algorithm which attains a matching upper bound on the expected regret. However, the instance-independent (i.e., in adversarial case) regret of the variant of UCB1, (α, ψ)-UCB, is given by $O(\sqrt{kT\log T})$ [10]. The Moss algorithm proposed by Audibert and Bubeck [3] achieves the instance-independent regret of $O(\sqrt{kT})$. Bubeck and Cesa-Bianchi [10] present a detailed survey on regret bounds of these algorithms. A similar setting, known as *arm-acquiring bandits* is studied under Markovian bandits framework [24, 32]. Here, the goal is to maximize the discounted, infinite

time cumulative reward whereas in ballooning bandits goal is to minimize the finite time cumulative regret. This difference is further highlighted by the fact that ballooning bandits is a *learning* problem whereas arm-acquiring bandits is a planning problem.

The problem of learning qualities of the answers on Q&A forums was first modeled under MAB framework by Ghosh and Hummel [17] where generation of a new arm was considered as a consequence of strategic choice of an agent. Though this model captures strategic aspects of the contributors, there is an important practical issue with such modelling. Each agent, being a strategic attention seeker, is assumed to produce the effort just enough to satisfy incentive compatibility in the equilibrium. We do not assume an efforts-and-costs model and show that, even when the number of answers grows linearly with time if the qualities of arriving answers follow certain mild distributional assumption, the proposed algorithm achieves sub-linear regret. Tang and Ho [28] consider a model with fixed number of arms but with a platform where agents provide biased feedback. On such Q&A forums, it is more relevant to consider the problem with increasing number of arms. A recent work by Liu and Ho [22] limits the growth of the bandit arms by randomly dropping some arms from consideration, and computing the regret with respect to only the considered arms. They do not account for the regret incurred due to the randomly dropped arms.

8 DISCUSSION AND FUTURE WORK

In this paper, we presented Ballooning bandits model (BL-MAB) and showed that, in the absence of any distributional assumption on the arrival of the best quality arm, it is impossible to achieve sub-linear regret. We proposed an algorithm for the BL-MAB model and provided sufficient conditions under which the proposed algorithm achieves sub-linear regret. In particular, when the arrival distribution of the best quality arm has a sub-exponential or sub-Pareto tail, our algorithm BL-Moss achieves sub-linear regret by restricting the number of arms to be explored in an intelligent way.

Our results indicate that, the number of arms to be explored depends on the distributional parameters, namely, λ (for sub-exponential case) and β (for sub-Pareto case), which must be known to the algorithm. It will be interesting to see how a learning algorithm can be designed to learn these parameters as well. For the worst case analysis, we considered that a new arm arrives at every time instant (similar to [17, 22]), hence the number of arms equals the time horizon T. The case where the number of arms is $\sqrt{T}$ or $\log(T)$ can be easily analyzed using a sleeping bandit algorithm, and sublinear regret can be achieved. With additional distributional assumptions on the best arm's arrival, the regret will be correspondingly lower; analyzing this regret bound is an interesting future direction. In this paper, we only consider a structure on the arrival of the best arm. One could also consider a more sophisticated arrival process of the arms, for obtaining better regret guarantees.

APPENDIX

Claim 1. *For a given BL-MAB instance $\mathcal{I}$, minimum regret is achieved when algorithm pulls the first $|G|$ arms.*

Proof. We prove the result by contradiction. Without loss of generality, let $|G| \neq 0$ and $|G| \neq T$ (since the result is trivially true in both the cases). Let an optimal algorithm pull the set of arms G in its execution and that there is atleast one arm pulled after $|G|$ time instants. For contradiction, assume that the achieved regret is strictly less than when the algorithm pulls the first $|G|$ arms. As G is not the set of first $|G|$ arms, there exists an arm i such that $i \notin G$ and $i \leq |G|$. Also, there exists a corresponding arm j such that $j \in G$ and $j > |G|$. Consider all such (i, j) pairs and construct a set $G' = G \cup \{i\} \setminus \{j\}$ by swapping the arm j with arm i. Whenever an arm $j \in G$ is pulled by the algorithm, we make a pull of the corresponding arm $i \in G'$. We now prove that the expected regret guarantee with G and G' is the same. Let,

$$\Delta(i_t^\star, i_t) = \begin{cases} \varepsilon & \text{if } i_t^\star = i^\star \text{ and } i_t \neq i_t^\star \\ 0 & \text{otherwise.} \end{cases}$$

Here, $i_t^\star$ is the highest quality arm available at time t. In the given example in the main text, we have $q_{i_t^\star} = 1/2 + \varepsilon$ if the best arm $i^\star$ has arrived on or before time instant t, otherwise $q_{i_t^\star} = 1/2$. Note that, as the best arm is uniformly distributed, both i and j have equal probability of being an optimal arm. Hence, the expected t-time regret

$$\begin{aligned}
\mathbb{E}_G[\Delta(i_t^\star, i_t)] &= \varepsilon \cdot \mathbb{P}(i_t \neq i_t^\star, i_t^\star = i^\star) \\
&= \varepsilon \cdot \mathbb{P}(i_t \neq i_t^\star | i_t^\star = i^\star) \cdot \mathbb{P}(i_t^\star = i^\star) \\
&= \varepsilon \cdot \mathbb{P}(i_t^\star = i^\star) \sum_{j \in G} \mathbb{1}(i_t = j) \mathbb{P}(j \neq i_t^\star | i_t^\star = i^\star) \\
&= \varepsilon \cdot \mathbb{P}(i_t^\star = i^\star) \sum_{i \in G'} \mathbb{1}(i_t = i) \mathbb{P}(i \neq i_t^\star | i_t^\star = i^\star) \\
&= \mathbb{E}_{G'}[\Delta(i_t^\star, i_t)]
\end{aligned}$$

Note that the above equality holds for any time instant t. This contradicts the assumption that $\mathbb{E}_G[\sum_{t=1}^T \Delta(i_t^\star, i_t)] < \mathbb{E}_{G'}[\sum_{t=1}^T \Delta(i_t^\star, i_t)]$. This completes the proof. □

Claim 2. $\frac{W(\lambda T/c)}{\lambda T/c} < 1/36 \iff T > \frac{36c\log(36)}{\lambda}$

Proof. We have the following equivalent inequalities.

$$\frac{W(\lambda T/c)}{\lambda T/c} < \frac{1}{36} \iff e^{-W(\lambda T/c)} < \frac{1}{36} \ (\because W(x)e^{W(x)} = x)$$

$$\iff W(\lambda T/c) > \log(36) \iff \frac{\lambda T}{c} > \log(36) e^{\log(36)}$$

$$\iff T > \frac{36c\log(36)}{\lambda}$$

The second to last inequality is obtained by applying the monotone increasing function $f(x) := xe^x$ on both sides, and then using Definition 4.1 of Lambert W function. □

Claim 3. *$e^{-cW(\lambda T/c)}$ is decreasing in c for $c \in (0, 1/2]$.*

Proof. For $c_1 > c$, we have

$$\begin{aligned}
&\lambda T/c > \lambda T/c_1 \\
\iff & W(\lambda T/c) > W(\lambda T/c_1) \quad \text{(Property } \mathbf{P2} \text{ of Lambert } W\text{)} \\
\iff & e^{-W(\lambda T/c)} < e^{-W(\lambda T/c_1)} \\
\iff & \frac{W(\lambda T/c)}{\lambda T/c} < \frac{W(\lambda T/c_1)}{\lambda T/c_1} \quad (\because W(x)e^{W(x)} = x) \\
\iff & cW(\lambda T/c) < c_1 W(\lambda T/c_1) \\
\iff & e^{-cW(\lambda T/c)} > e^{-c_1 W(\lambda T/c_1)}
\end{aligned}$$

□

REFERENCES

[1] Shipra Agrawal and Navin Goyal. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*. 1–39.

[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD*. 850–858.

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

[4] Peter Auer. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* 3, Nov (2002), 397–422.

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

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

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

[8] Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. 2009. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*. ACM, 79–88.

[9] Donald A Berry, Robert W Chen, Alan Zame, David C Heath, Larry A Shepp, et al. 1997. Bandit problems with infinitely many arms. *The Annals of Statistics* 25, 5 (1997), 2103–2116.

[10] Sébastien Bubeck and Nicolo Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* 5, 1 (2012), 1–122.

[11] Keith Burghardt, Emanuel Alsina, Michelle Girvan, William Rand, and Kristina Lerman. 2016. The myopia of crowds: A study of collective evaluation on stack exchange. *Robert H. Smith School Research Paper No. RHS* 2736568 (2016).

[12] Alexandra Carpentier and Michal Valko. 2015. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning*. 1133–1141.

[13] Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*. 2249–2257.

[14] Aritra Chatterjee, Ganesh Ghalme, Shweta Jain, Rohit Vaish, and Y Narahari. 2017. Analysis of Thompson Sampling for Stochastic Sleeping Bandits.. In *Uncertainty in Artificial Intelligence, UAI 2017*.

[15] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. 1996. On the Lambert *W* function. *Advances in Computational Mathematics* 5, 1 (1996), 329–359.

[16] R Devanand and P Kumar. 2017. Empirical study of Thompson sampling: Tuning the posterior parameters. In *AIP Conference Proceedings*, Vol. 1853.

[17] Arpita Ghosh and Patrick Hummel. 2013. Learning and incentives in user-generated content: Multi-armed bandits with endogenous arms. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*. 233–246.

[18] Abdolhossein Hoorfar and Mehdi Hassani. 2008. Inequalities on the Lambert *W* function and hyperpower function. *J. Inequal. Pure and Appl. Math* 9, 2 (2008), 5–9.

[19] Shweta Jain, Sujit Gujar, Satyanath Bhat, Onno Zoeter, and Y. Narahari. 2018. A quality assuring, cost optimal multi-armed bandit mechanism for expertsourcing. *Artificial Intelligence* 254 (2018), 44 – 63.

[20] Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. 2010. Regret bounds for sleeping experts and bandits. *Machine learning* 80, 2-3 (2010), 245–272.

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

[22] Yang Liu and Chien-Ju Ho. 2018. Incentivizing high quality user contributions: New arm generation in bandit learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[23] Setareh Maghsudi and Sławomir Stańczak. 2014. Joint channel selection and power control in infrastructureless wireless networks: A multiplayer multiarmed bandit framework. *IEEE Transactions on Vehicular Technology* 64, 10 (2014), 4565–4578.

[24] Peter Nash. 1973. *Optimal allocation of Resources Between Research Projects*. Ph.D. Dissertation. University of Cambridge.

[25] Alessandro Nuara, Francesco Trovo, Nicola Gatti, and Marcello Restelli. 2018. A Combinatorial-Bandit Algorithm for the Online Joint Bid/Budget Optimization of Pay-per-Click Advertising Campaigns. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[26] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. 2018. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning* 11, 1 (2018), 1–96.

[27] Aleksandrs Slivkins. 2019. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272* (2019).

[28] Wei Tang and Chien-Ju Ho. 2019. Bandit Learning with Biased Human Feedback. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 1324–1332.

[29] William R Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3/4 (1933), 285–294.

[30] Sofía S Villar, Jack Bowden, and James Wason. 2015. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics* 30, 2 (2015), 199.

[31] Yizao Wang, Jean-Yves Audibert, and Rémi Munos. 2009. Algorithms for infinitely many-armed bandits. In *Advances in Neural Information Processing Systems*. 1729–1736.

[32] Peter Whittle et al. 1981. Arm-acquiring bandits. *The Annals of Probability* 9, 2 (1981), 284–292.