Ballooning Multi-Armed Bandits

Ganesh Ghalme [*] Swapnil Dhamal[†] Shweta Jain [‡]

Sujit Gujar [§] Y. Narahari [¶]

Abstract

In this paper, we introduce *ballooning multi-armed bandits* (BL-MAB), a novel extension of 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 result in linear regret for the BL-MAB model. We prove that, if the best arm is equally likely to arrive at any time instant, a sub-linear regret cannot be achieved. Next, we 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. We conclude by showing that our algorithm would achieve sub-linear regret even if (a) the distributional parameters are not exactly known, but are obtained using a reasonable learning mechanism or (b) the best arm is not more likely to arrive early, but a large fraction of arms is likely to arrive relatively early.

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)

[*] Technion, Israel Institute of Technology, Israel. ganeshg@campus.technion.ac.il

[†] Chalmers University of Technology, Sweden. swapnil.dhamal@gmail.com

[‡] Indian Institute of Technology, Ropar, India. shwetajains20@gmail.com

[§] International Institute of Information Technology. sujit.gujar@iiit.ac.in

[¶] Indian Institute of Science. narahari@iisc.ac.in

[**] A part of this work was done while the first author was at Indian Institute of Science, India.

[††] An original version of this work is accepted for publication in the Journal of Artificial Intelligence (AIJ) by Elsevier. A previous, preliminary version of this paper appeared as an extended abstract in proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (2020), Auckland, New Zealand [GDJ+20].

from a fixed pool of finitely many actions (i.e., arms), a single arm at each discrete time instant upto arbitrary time horizon. 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, that is, the loss incurred in the 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 analysis of the achievable lower bound on regret [LR85], bandit algorithms [AO10; Tho33; GC11], empirical studies [CL11; DK17; RVRK$^+$18], and several extensions to the standard model [Sli19; BCB12]. Many papers show that some of the well known algorithms such as UCB1 [ACBF02] , THOMSON SAMPLING [AG12; KKM12], KL-UCB [GC11] are known to attain asymptotically optimal regret guarantee upto a problem-dependent constant. The above list is far from exhaustive. We refer the reader to [Sli19; LS20] for a book exposition on the topic.

The theoretical results in MAB are complemented by a wide variety of modern applications such as internet advertising [BSS09; NTGR18], crowdsourcing [JGB$^+$18], clinical trials [VBW15], and wireless communication [MS14], which can be modeled in the MAB setup. Due to a wide range of applications and an elegant theoretical foundation, several variants of the MAB problem have been proposed. Contributing to the long line of work that studies different variants of bandits, in this paper, we introduce a novel variant of MAB 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.

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 rewards 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 instant, thus incurring a regret of $O(T)$. Note in the above example that even while the best arm appeared at the first time instant itself, traditional algorithms end up pulling all other arms at least once, which leads to a high regret. As the set of available arms expands over time, the traditional algorithms could not sufficiently explore each of the arms to identify the best arm. Also, note 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.

It is clear from the above example that traditional algorithms cannot provide sublinear regret guarantees in BL-MAB setup as there is not enough time spared for exploitation. Further, as the number of arms increases (potentially linearly) with time, an optimal algorithm must ignore (or drop) a few arms. Hence, in addition to achieving an optimal trade-off between the number of exploratory pulls and exploitative pulls, the algorithm must also ensure that it does not drop too many (or too few) arms.

1.1 Motivation

The BL-MAB framework 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. We motivate the practical significance of BL-MAB with a few applications.

A contemporary example is provided by question and answer (Q&A) platforms such as Reddit, Stack Overflow, Quora, Yahoo! Answers, and ResearchGate, where for a given question, the platform's goal is to discover the highest quality answer that should be displayed in the most prominent slot. 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). If a user likes the answer displayed to her, then she may endorse the answer. 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, and decides whether to endorse them. Further, the user 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 [GH13; TH19; LH18]. However, these studies resort to the existing MAB variations which are not well suited for Q&A forums. For instance, in [GH13], the problem is modeled 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 [BAG+16]. In another work [LH18], 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), and so on. 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 at 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) [KNMS10; CGJ+17], 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 [KNMS10] would end up incurring a linear regret in BL-MAB setting.

Another MAB variant with some similarities to BL-MAB is the many-armed (potentially infinite) bandit [WAM09; CV15; BCZ+97], where the number of arms could be potentially equal to or greater than the time horizon. Berry et al. [BCZ+97] consider the case of an infinite arm bandit with Bernoulli reward distribution. However, they assume that the optimal arm has a quality of 1, which is seldom the case in practical applications. Other investigations considering infinitely many arms [WAM09; CV15] make certain assumptions on the distribution of the near optimal arm, in order to achieve sub-linear regret. A further difference is that in the many-armed bandit setting, researchers usually assume the existence of a smooth function relating the mean rewards of the arms (for instance, [CV15]). Here, we consider that the reward distributions are independent with arbitrary means. This rules out any side information that could be gathered from the pulls of other arms. Finally, 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 many-armed bandit model.

1.2 Our Contributions

Following are the main contributions of this paper:

- We introduce the BL-MAB model that allows the set of arms to grow over time.
- For the BL-MAB model, we show that, without any distributional assumptions on the arrival time of the highest quality arm, the regret grows linearly with time (Theorem 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 3) and sub-Pareto tail (Theorem 4). 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.
- We study the cost of parametric uncertainty, which we define to be the loss incurred due to not knowing the parameters of the best arm's arrival distribution exactly (Theorems 5 and 6). We also show that our algorithm is applicable to the setting which does not make distributional assumptions on the arrival time of the best arm, but instead, on the rate of arrival of arms with time (Theorem 7).

The paper is organized as follows. In Section 2, we present our proposed BL-MAB model. In Section 3, we show that no algorithm can achieve sub-linear regret in the most general setup. Hence, an additional assumption on the arrival of arms is warranted. We define two distributional assumptions on the arrival time of the best arm which would enable 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 study two extensions in Section 7, namely, relaxing the distributional assumption on the arrival of the best arm, and deducing the cost of parametric uncertainty. We conclude with related work (Section 8) and future directions (Section 9).

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 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 a distribution over 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 [KNMS10], 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 distribution-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 distribution-free regret given as $\mathcal{R}_A(T) = \sup_{\mathcal{I}} \mathcal{R}_A(T, \mathcal{I})$. Note that the distribution-free 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 that for the BL-MAB setting, it is not possible to achieve sublinear distribution-free 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[1]. However, it is not obvious that a different, more sophisticated algorithm (such as the one which randomly drops some arms) would not be able to achieve sub-linear regret. Our first result (Theorem 1) shows that no algorithm can attain sub-linear regret without any distributional assumption on the best arm's arrival.

[1] In particular, when $|K(t)| = t$.

Theorem 1. *There exists a BL-MAB instance $\mathcal{J}$ such that any MAB algorithm* Alg *satisfies*

$$\mathcal{R}_{\text{ALG}}(T, \mathcal{J}) = \Omega(T)$$

Proof. We prove the theorem in three steps. In the first step, we construct a BL-MAB instance $\mathcal{J}$ such that $\mathcal{J}$ has a single best arm and all the suboptimal arms have the same quality parameter. Further, we consider that each arm is equally likely to be the best arm, i.e., the probability that an arriving arm is the best arm is $\frac{1}{T}$ for all $t = \{1, 2, \ldots, T\}$. Next, in Step 2, we simulate any BL-MAB algorithm Alg by simulation algorithm Sim such that $\mathcal{R}_{\text{ALG}}(T, \mathcal{J}) = \mathcal{R}_{\text{SIM}}(T, \mathcal{J})$. Finally, in Step 3, we show $\mathcal{R}_{\text{SIM}}(T, \mathcal{J}) = O(T)$ for every simulated algorithm Sim. We begin with the construction of a BL-MAB instance $\mathcal{J}$.

Step 1: A new arm arrives at each discrete time instant t, till the predetermined time horizon T. There is a single best arm $i^\star$ with quality parameter $q_{i^\star} = 1/2 + \varepsilon$. Here, $\varepsilon > 0$ is a problem-independent constant. Each suboptimal arm $i \neq i^\star$ has $q_i = 1/2$. Further, each arm is equally likely to be the best arm, i.e., $\mathbb{P}(i = i^\star) = \frac{1}{T}$ for all $i \in [T]$. Next, we show an important property of the BL-MAB instance $\mathcal{J}$ (Claim 1). In particular, we show that for any algorithm, there exists a corresponding arm-pulling strategy which pulls the arms in the order of their arrival and has the same expected reward.

Step 2: Consider a single run of any BL-MAB algorithm Alg on instance $\mathcal{J}$ and let G denote the set of distinct arms pulled till time T, i.e., $G = \{i \in [T] | N_{i,T}^{\text{ALG}} > 0\}$, with $g = |G|$. Here, $N_{i,t}^{\text{ALG}}$ is the number of times arm i is pulled till (and excluding) time instant t by Alg. We drop the superscript when the algorithm is clear from the context. Further, let $M_n = \{1, 2, \ldots, n\}$ be the collection of the first[2] n arms. We simulate Alg on $\mathcal{J}$ as using a simulation Sim such that $\mathcal{R}_{\text{ALG}}(T, \mathcal{J}) = \mathcal{R}_{\text{SIM}}(T, \mathcal{J})$. For any arm pull i_t at time t, we pull arm $i_t^{'}$ in the simulation Sim as follows.

Sim

$$i_t^{'} = \begin{cases} i_t & \text{if } i_t \in M_g \\ \min\{i \in M_g | N_{i,t}^{\text{SIM}} = 0\} & \text{if } i_t \in G \setminus M_g \text{ and } N_{i_t,t}^{\text{ALG}} = 0 \\ i_\ell^{'} & \text{if } i_t \in G \setminus M_g \text{ and } N_{i_t,t}^{\text{ALG}} > 0 \\ (\ell = \min\{m < t : N_{i_t,m+1}^{\text{SIM}} = 1\}) & \end{cases}$$

Whenever Alg pulls an arm $i_t \in G \setminus M_g$ for the first time, Sim assigns a corresponding $i_t^{'} \in M_g$. Let us say that Alg pulls 3 distinct arms in its run (i.e., $G = \{1, 4, 6\}$) and the sequence of arms pulled is given by $(1, 1, 1, 4, 4, 6, 6, 4, 1, 6, \ldots, 1)$. In this case, Sim will pull arms 1, 2 and 3 in sequence, $(1, 1, 1, 2, 2, 3, 3, 2, 1, 3, \ldots, 1)$. That is, all the arm pulls of arms from set $\{1, 2, 3\}$ are retained and all the arms from outside this set that are pulled are replaced by the arms from this set as follows: the least index arm is assigned to the first arm encountered from outside the set, i.e., whenever Alg pulls

[2] In the order of their arrival.

arm 4, SIM pulls arm 2 (similarly, whenever ALG pulls arm 6, SIM pulls arm 3). We now prove that both ALG and SIM have the same expected rewards.

Claim 1. $\mathcal{R}_{\text{ALG}}(T,\mathcal{J}) = \mathcal{R}_{\text{SIM}}(T,\mathcal{J})$.

Proof.

$$\begin{aligned}\mathcal{R}_{\text{ALG}}(T,\mathcal{J}) = \mathbb{E}\big[\sum_{t=1}^{T} X^{\star}_{i_t} - X_{i_t}\big] &= \sum_{t=1}^{T}\mathbb{E}_{\text{ALG}}\big[q_{i^{\star}_t} - q_{i_t}\big] \\ &= \sum_{t=1}^{T}\mathbb{E}_{\text{ALG}}\Delta(i^{\star}_t, i_t)\end{aligned}$$

If $i_t^{'} = i_t$, we immediately have that $\Delta_{i^\star_t,i_t} = \Delta_{i^\star_t,i_t^{'}}$. Hence, without loss of generality, let $i_t^{'} \neq i_t$. We have

$$\begin{aligned}\Delta_{i^\star_t,i_t} &= \big[(1/2+\varepsilon)\cdot\mathbb{P}(i^{*}\text{ has arrived before }t) + 1/2\cdot\mathbb{P}(i^{*}\text{ arrive after time }t)\big] \\ &\quad - \big[(1/2+\varepsilon)\cdot\mathbb{P}(i_t = i^{\star}) + 1/2\cdot\mathbb{P}(i_t \neq i^{*})\big] \\ &= \big[(1/2+\varepsilon)\cdot\big(\sum_{\ell=1}^{t}\mathbb{P}(i_\ell = i^{\star})\big) + 1/2\cdot\big(1-\sum_{\ell=1}^{t}\mathbb{P}(i_\ell = i^{\star})\big)\big] \\ &\quad - \big[(1/2+\varepsilon)\cdot\mathbb{P}(i_t = i^{\star}) + 1/2\cdot(1-\mathbb{P}(i_t = i^{\star}))\big] \\ &= \big[(1/2+\varepsilon)\cdot\big(\sum_{\ell=1}^{t}\mathbb{P}(i_\ell^{'} = i^{\star})\big) + 1/2\cdot\big(1-\sum_{\ell=1}^{t}\mathbb{P}(i_\ell^{'} = i^{\star})\big)\big] \\ &\quad - \big[(1/2+\varepsilon)\cdot\mathbb{P}(i_t^{'} = i^{\star}) + 1/2\cdot(1-\mathbb{P}(i_t^{'} = i^{\star}))\big] \\ &\qquad\qquad (\text{As }\mathbb{P}(i_\ell = i^{\star}) = \mathbb{P}((i_\ell^{'} = i^{\star})) = 1/T\text{ for all }\ell\in[T]) \\ &= \Delta_{i^\star_t,i_t^{'}}\end{aligned}$$

This completes the proof. □

Henceforth, we will focus only on the simulation algorithms.

Step 3: Let G denote the set of arms pulled by the algorithm SIM in its run. The regret of SIM can be written as

$$\mathcal{R}_{\text{SIM}}(T) = \mathbb{E}_G\Big[\sum_{i=1}^{|G|}\mathbb{P}(i^{\star}\in G, i\neq i^{\star})\mathbb{E}[N_{i,T}]\cdot\varepsilon + \sum_{i=|G|+1}^{T}\mathbb{P}(i = i^{\star})(T-i)\cdot\varepsilon\Big].$$

The outer expectation is with respect to the number of arms pulled by any (possibly randomized) algorithm SIM. For a fixed value of G, the inner expectation represents the number of times arms $i \in G$ are pulled till the time horizon T. Using a classical result from [LR85], we have $\mathbb{E}[N_{i,T}] \geq \eta\log(T)$

for some $\eta > 0$ that depends on the suboptimality of the arm i. Using this, we have

$$
\begin{aligned}
\mathcal{R}_{\text{SIM}}(T) &\geq \mathbb{E}\Big[\sum_{i=1}^{|G|} \mathbb{P}(i \neq i^\star | i^\star \in G) \cdot \mathbb{P}(i^\star \in G)\eta \log(T) + \sum_{i=|G|+1}^{T} \mathbb{P}(i = i^\star)(T-i)\Big] \cdot \varepsilon \\
&= \mathbb{E}\Big[\sum_{i=1}^{|G|} \frac{|G|-1}{|G|} \cdot \frac{|G|}{T}\eta \log(T) + \sum_{i=|G|+1}^{T} \frac{T-i}{T}\Big] \cdot \varepsilon \\
&= \mathbb{E}\Big[\sum |G|(|G|-1) \cdot \eta \log(T) + \frac{(T-|G|-1)(T-|G|)}{2}\Big] \cdot \frac{\varepsilon}{T} \\
&= \mathbb{E}\Big[(1+2\eta\log(T))|G|^2 - (2(T-\eta\log(T))-1)|G| + T^2 - T\Big] \cdot \frac{\varepsilon}{2T} \\
&\geq \min_{|G| \in [0,T]} \Big[(1+2\eta\log(T))|G|^2 - (2(T-\eta\log(T))-1)|G| + T^2 - T\Big] \cdot \frac{\varepsilon}{2T}.
\end{aligned}
$$

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 1. In this case, the regret is $\Omega(T)$. For $T > 1/2 + \eta\log(T)$, the minimum occurs when $|G| = \frac{2(T-\eta\log(T))-1}{2(1+2\eta\log(T))}$. For this case, we have

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

□

Theorem 1 provides a strong impossibility result on the achievable distribution-free 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.

3.1 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., $F_X(t) > 1 - t^{-\beta}$.*

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. [AHKL12] 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 scores. 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 will show that, while the uniform distribution results in linear regret (Theorem 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 (Section 5).

4 Preliminaries

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

4.1 Lambert W Function

Definition 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., $W(x)e^{W(x)} = x$.

It is easy to check that in the non-negative domain, Lambert W function satisfies the following regularity properties [HH08]. The detailed proofs are provided in Appendix B.

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

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

Property 3. For any $x \in [0, \infty)$, the Lambert W function is unique, non-negative, and strictly increasing.

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 [CGH$^+$96].

4.2 MAB Algorithms

We now review some of the MAB algorithms, starting with UCB1, perhaps the most famous stochastic MAB algorithm. We then review THOMPSON SAMPLING [Tho33] which presents an alternate Bayesian approach to the MAB problem. In this paper, we use MOSS (Minimax Optimal Strategy in the Stochastic case) [AB10] as an underlying learning algorithm; the MOSS algorithm uses UCB-style indexing of the arms. In principle, one could use any underlying learning algorithm in a BL-MAB setup. However, as we shall discuss, one needs to carefully tune the thresholding parameter for the learning algorithm in question.

UCB1

UCB1, proposed in [ACBF02], is perhaps the most famous stochastic MAB algorithm. At each time t, UCB1 maintains a UCB index for each arm and pulls an arm with the highest UCB index. In particular, UCB1 pulls an arm i_t such that

$$i_t \in \arg\max_{i \in K} \left[\hat{q}_{i,N_{i,t}} + \sqrt{\frac{2\log(t)}{N_{i,t}}} \right].$$

Here, $K = \{1, 2, \ldots, k\}$ 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.

THOMSON SAMPLING

First proposed in [Tho33], the theoretical regret guarantee of THOMPSON SAMPLING remained an open problem for over 80 years before [AG12] and [KKM12] independently showed that THOMPSON SAMPLING achieves asymptotically optimal regret guarantee (upto problem-dependent constant). This Bayesian approach maintains a conjugate prior distribution for each arm. We refer the reader to [AG12] for the detailed algorithm as well as a regret analysis of THOMPSON SAMPLING.

MOSS

For a fixed number of k arms, the MOSS algorithm pulls an arm i_t at time t where

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

Each arm is pulled once in the beginning, and ties are broken arbitrarily.

In contrast to other popular MAB algorithms such as THOMPSON SAMPLING [Tho33], UCB1 [ACBF02] and KL-UCB [GC11], MOSS simultaneously achieves the optimal instance-dependent as well as optimal instance-independent regret guarantee [AB10]. However, the time horizon is assumed to be known to the algorithm a priori. The problem of achieving simultaneous optimal anytime regret guarantees had remained open until recently, when modified versions of KL-UCB algorithms, namely, KL-UCB++ [MG17] and KL-UCB-SWITCH [GHMS18], were proven to be simultaneously optimal. However, the instance-independent regret bound of these algorithms still depends linearly on the number of available arms (Theorem 1 in [MG17] and Theorem 4 in [GHMS18], respectively).

Algorithm 1: BL-Moss

```
Input: Time horizon $T$, Distributional parameter $\lambda > 0$ or $\beta > 0$
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($K(\lceil \alpha T \rceil)$)
    end
end
```

As we shall see in Algorithm 1 in the next section, we need a threshold parameter α which signifies the fraction of arms that a learning algorithm should explore. This parameter must be tuned based on the regret guarantees of the learning algorithm, i.e., the internal regret in the BL-MAB framework and the external regret. We choose Moss for its simplicity and optimality (both in terms of number of arms and the time horizon). For instance, Moss achieves an optimal instance-independent regret guarantee of $O(\sqrt{kT})$. Other algorithms such as Thompson Sampling, UCB1 or KL-UCB may also be employed as underlying learning algorithms, albeit with a slight ($O(\sqrt{\log(T)})$) increase in internal regret. We leave the determination of the threshold parameter and the corresponding regret analysis of BL-MAB using other algorithms as an interesting direction for future work.

5 The BL-MOSS Algorithm and Regret Analysis

5.1 The BL-Moss Algorithm

We now present our algorithm, BL-Moss (Algorithm 1), that uses Moss as the underlying learning algorithm. 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 algorithm is a simple extension of Moss and 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.

5.2 Regret Analysis of BL-Moss

We begin with an upper bound on the expected regret of the Moss algorithm. Note that Moss achieves optimal (up to a constant factor) regret bound. Throughout the paper, we use the notation Moss(k) to denote that the Moss algorithm is run with k arms.

Theorem 2. *[AB10] For any time horizon $T \geq 1$, the expected regret of* Moss *is given by $\mathcal{R}_{\text{Moss}(k)}(T) \leq 6\sqrt{kT}$.*

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 distribution-dependent regret $\mathcal{R}_{\text{BL-MOSS}}(T, \mathcal{I})$ is given as

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

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 $L \subseteq K$ and for any any time horizon T. From Theorem 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 $\mathcal{R}^{\text{int}}_{\text{BL-MOSS}}(T) \leq \mathcal{R}_{\text{MOSS}(\alpha T)}(T) \leq 6\sqrt{\alpha}T$.

The first inequality in Observation 1 follows from the fact that in a classical MAB setting, all the arms are available at all times, whereas in a BL-MAB setting, arms arrive online. Hence, the best arm is available at all times in MAB, whereas in BL-MAB, the arrival of the best arm is delayed.

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 1. *The upper bound on the expected regret for any BL-MAB instance is given by* $\mathcal{R}_{\text{BL-MOSS}}(T) \leq T(1 - (1 - 6 \cdot \sqrt{\alpha}) F_X(\alpha T))$, *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 $q_{i^\star} \geq q_{j^\star}$.

$$\begin{aligned}
\mathcal{R}_{\text{BL-MOSS}}(T, \mathcal{I}) &\leq \mathbb{E}\Big[\sum_{t=1}^{\alpha T}(q_{j^\star} - q_{i_t}) + \sum_{t=\alpha T+1}^{T}(q_{i^\star} - q_{i_t})\Big] && (\because q_{i^\star} > q_{j^\star}) \\
&= \mathbb{P}(i^\star = j^\star)\Big[\sum_{t=1}^{T}(q_{j^\star} - q_{i_t})\Big] \\
&\quad + \mathbb{P}(i^\star \neq j^\star)\Big[\sum_{t=1}^{\alpha T}(q_{j^\star} - q_{i_t}) + \sum_{t=\alpha T+1}^{T}(q_{i^\star} - q_{i_t})\Big]
\end{aligned}$$

$$
\begin{aligned}
&\leq 6\mathbb{P}(i^\star = j^\star)\sqrt{\alpha T \cdot T} + \sum_{t=1}^{T}(q_{i^\star} - q_{i_t})\mathbb{P}(i^\star \neq j^\star) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(From Observation 1 and since } q_{i^\star} \geq q_{j^\star}\text{)} \\
&\leq 6T\sqrt{\alpha}\cdot\mathbb{P}(i^\star = j^\star) + \mathbb{P}(i^\star \neq j^\star)T \qquad (\because \textstyle\sum_{t=1}^{T}(q_{i^\star} - q_{i_t}) \leq T) \\
&= 6T\sqrt{\alpha}\cdot\mathbb{P}(t_{i^\star} \leq \alpha T) + (1 - \mathbb{P}(t_{i^\star} \leq \alpha\cdot T))T \\
&= T(1 - (1 - 6\cdot\sqrt{\alpha})\mathbb{P}(t_{i^\star} \leq \alpha T)) \\
&= T(1 - (1 - 6\cdot\sqrt{\alpha})F_X(\alpha T))
\end{aligned}
$$

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

5.2.1 Sub-exponential tail distribution

We now 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 2. *Let the arm arrival distribution of the best arm satisfy sub-exponential tail property for some $\lambda \geq 0$. 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. Note that from the Property 2 of the Lambert W function we have $\frac{\log(x)}{2} \leq W(x) \leq \log(x)$ for $x \geq e$. We have,

$$
\begin{aligned}
&\alpha \geq \frac{W(\lambda T/c)}{\lambda T/c} \implies \frac{\alpha\lambda T}{c} \geq W(\lambda T/c) \implies W\big(\frac{\alpha\lambda T}{c}\cdot e^{\alpha\lambda T/c}\big) \geq W(\lambda T/c) \qquad \text{(by Property 1)} \\
&\implies \frac{\alpha\lambda T}{c}\cdot e^{\alpha\lambda T/c} \geq \lambda T/c
\end{aligned}
$$

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 3. *Let the arrival distribution of the best arm satisfy the sub-exponential tail property for some $\lambda \geq 0$, 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*, $\mathcal{R}_{\text{BL-Moss}}(T)$, 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 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 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 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 A).

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 equivalent definition of Lambert W function (Property 1).

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 A). For Case 2, we have from Property 3 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 2, since $2\lambda T \geq e$ ($\because T > \frac{36c \log(36)}{\lambda}$ where $c = 1/2$). □

If we absorb λ (which is a constant with respect to T) in order notation, we have $R_{\text{BL-Moss}} = O(\sqrt{T \log(T)})$.

5.2.2 Sub-Pareto tail distribution

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

Lemma 3. *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 4. *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*, $\mathcal{R}_{\text{BL-Moss}}(T)$, 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 1 and 3, 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 3, 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 $c \in (1/2, \infty)$. This shows that the minimum regret is achieved when $c = 1/2$. □

5.3 Important Observations

We conclude the section with some key observations and remarks.

Observation 2. If the best arm arrival satisfies sub-exponential tail property with parameter λ, then

1. $\mathcal{R}_{\text{BL-Moss}}(T) \to 0$ as $\lambda \to \infty$
2. $\mathcal{R}_{\text{BL-Moss}}(T) \to O(T)$ as $\lambda \to 0$

Proof. Recall that, from Equation (1), we have

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

Note that, for large enough λ such that $\frac{W(2\lambda T)}{2\lambda T} \leq \frac{1}{T}$, we have that $\lceil \alpha T \rceil = 1$ and the algorithm pulls arm 1 at all times i.e. $i_t = 1$ for all $t \leq T$. Furthermore, as $\lceil \alpha T \rceil = 1$, we have $j^\star = 1$. Hence, the internal regret is zero. The total regret of the algorithm is, hence,

$$\begin{aligned}
\mathcal{R}_{\text{BL-Moss}}(T) = \mathcal{R}^{\text{ext}}_{\text{BL-Moss}}(T) & \\
&= \mathbb{P}(i^\star \neq 1)\Big[\sum_{t=1}^{t_{i^\star}-1} \Delta(i_t^\star, 1) + \sum_{t=t_{i^\star}}^{T} \Delta(i^\star, 1)\Big] \\
&\leq \mathbb{P}(i^\star \neq 1)\sum_{t=1}^{T} \Delta(i^\star, 1) && (\because \Delta(i_t^\star, 1) \leq \Delta(i^\star, 1)) \\
&\leq T(1 - F_X(1)) \\
&\leq Te^{-\lambda} \approx 0 && (\because \lambda \to \infty)
\end{aligned}$$

The last inequality follows from the sub-exponential tail assumption.

If $\lambda \to 0$, we have that $F_X(t) > 1 - e^{-\lambda t} \to 0$. Hence, the arrival distribution of the best arm captures the uniform distribution, i.e., $F_X(t) = \frac{1}{T}$ as well. Hence, from Theorem 1, we have that a regret of $O(T)$ is unavoidable. □

Observation 3. If the best arm arrival satisfies sub-Pareto tail property with parameter β, then

1. $\mathcal{R}_{\text{BL-MOSS}}(T) \to O(\sqrt{T})$ as $\beta \to \infty$
2. $\mathcal{R}_{\text{BL-MOSS}}(T) \to O(T)$ as $\beta \to 0$

Proof. Note that under sub-Pareto tail assumption, we have $1 - F_X(2) < 2^{-\beta}$. Hence, as $\beta \to \infty$, we have that $F_X(2) \to 1$. The internal regret of Moss is upper bounded by $6\sqrt{2T}$, whereas the external regret is given as

$$\mathcal{R}^{\text{ext}}_{\text{BL-MOSS}}(T) \leq \sum_{t=3}^{T} \mathbb{P}(i_t = i^\star)(T-t) \leq (T-3)(1-(1-2^{-\beta})) \approx 0.$$

Hence, the total regret is $O(\sqrt{T})$.

The proof of the second part follows on similar line as the second part of Observation 2. □

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.

6.1 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 $\lceil \alpha T \rceil$ depends 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 that the arm arriving at time t is 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

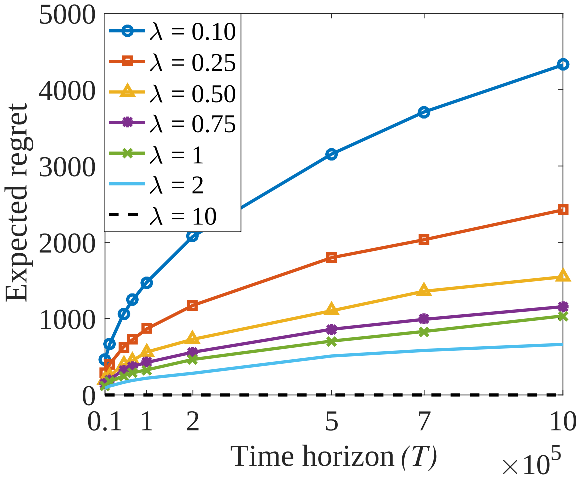

Figure 1
Expected regret as a function of time horizon for various sub-exponential tails

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 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).

6.2 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

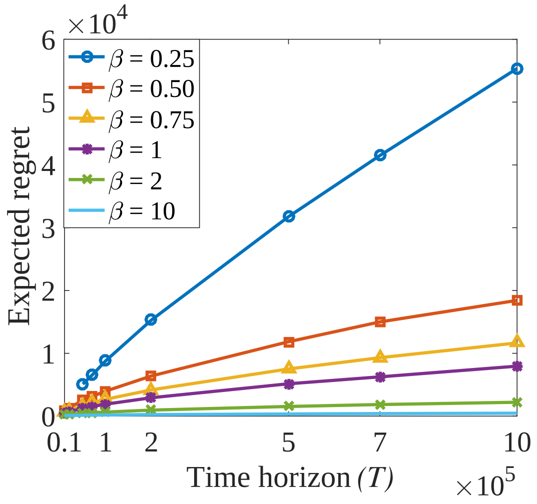

Figure 2

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

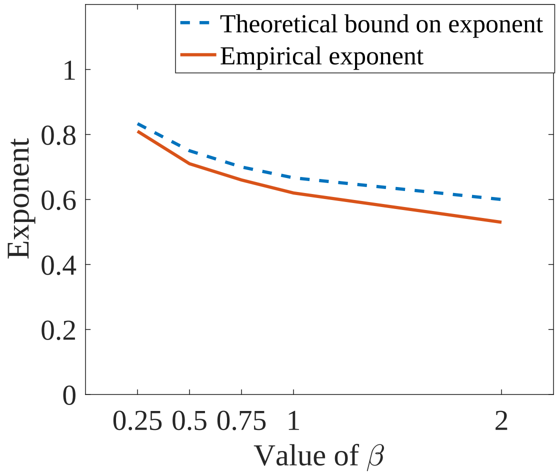

Figure 3

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

Figure 1). 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 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 2, 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 2 follow a far less sub-linear nature and exhibit a much higher expected regret than those in Figure 1. 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 3 that the upper bound on the theoretical exponent (which is $\frac{1+\beta}{1+2\beta}$ from Theorem 4) 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 distribution-free) 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 5.2) are, in all probability, tight.

6.3 Additional Notes on Simulation

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 regret is averaged over 1000 random instances with same arrival distribution of the best arm. 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 Extensions

In this section, we discuss possible extensions and relaxations of the BL-MAB setting studied in the paper. Thus far, we assumed that the true parameter of the arrival distribution of the best arm (i.e., λ or β) is known a priori to the BL-MOSS algorithm. We relax this assumption and consider that the parameter (λ or β) is known only approximately correctly. We show that the proposed algorithm achieves sublinear regret guarantees even with this relaxation. Next, we relax the assumption that the best arm arrives early, and instead consider that a large fraction of arms arrive early. We show that our algorithm is applicable even with this alternative assumption; we validate this assumption with real-world datasets.

7.1 Unknown Distributional Parameters of Best Arm's Arrival

So far, we have assumed that the distributional parameters (β and λ) are known to the algorithm designer. In most practical settings, these parameters are not known but can be learnt using previous data. In this section, we show the effect on the regret bound if learned parameters are used instead of the true parameters.

7.1.1 Sub-exponential tail distribution

Let $(X_i)_{i=1}^n$ be a collection of i.i.d. random variables sampled from an exponential distribution with parameter λ truncated at T. Further, let T be large enough such that $F_X(T) = 1 - e^{-\lambda T} \approx 1$. Let $X = \frac{1}{n}\sum_{i=1}^n X_i$ denote the empirical average over n questions posted. Let $\hat{\lambda}$ be the estimated parameter of the n i.i.d. exponential random variables; then $X \approx \frac{1}{\hat{\lambda}}$. Further, let there be two parameters μ and δ such that the Hoeffding's inequality [H$^+$56] gives us the following:

$$\mathbb{P}\left(\left|\frac{1}{\hat{\lambda}} - \frac{1}{\lambda}\right| \geq \mu\right) \leq \delta \tag{2}$$

Here, μ and δ denote how close the learned parameter and the true parameter are. The value of these parameters depend on the number of samples that we select. For example, in order to achieve confidence of $1-\delta$, we need $\delta \leq 2e^{-\frac{2n\mu^2}{T^2}} \implies n \geq \frac{T^2 \ln\left(\frac{2}{\delta}\right)}{2\mu^2}$, where n is the number of samples used to learn parameter $\hat{\lambda}$.

In Algorithm 1, we will now choose $\alpha = \frac{W(2\hat{\lambda}T)}{2\hat{\lambda}T}$ instead of $\frac{W(2\lambda T)}{2\lambda T}$ (which is the regret optimal α, from Theorem 3). Choosing this α would not change Lemma 1 and we would still have $R_{\text{BL-Moss}}(T) \leq T(1-(1-6\sqrt{\alpha})F_X(\alpha T))$. We now have the following theorem on regret.

Theorem 5. *If* BL-Moss *(Algorithm 1) is run with a learned parameter $\hat{\lambda}$ that satisfies Inequality (2), then with probability at least $1-\delta$, the regret under sub-exponential tail distribution assumption is upper bounded by $O\left(T^{\frac{1+\mu\lambda}{2}}\left(\sqrt{\frac{W(2\hat{\lambda}T)}{2\hat{\lambda}}}\right)^{1-\mu\lambda}\right)$ if $\mu\lambda < 1$ and $O\left(\sqrt{\frac{T\cdot W(2\hat{\lambda}T)}{2\hat{\lambda}}}\right)$ otherwise. The regret is sub-linear in both the cases.*

Proof. Recall that α is the fraction of arms explored by BL-Moss. The following is true with probability at least $1-\delta$.

$$\begin{aligned}
\alpha &\geq \frac{W(\hat{\lambda}T/c)}{\hat{\lambda}T/c} = e^{-W(2\hat{\lambda}T)} && (\because\ c = 1/2 \text{ is regret optimal (Theorem 3)})\\
\implies \log\alpha &\geq -W\left(2T\hat{\lambda}\right) && \\
\implies W\left(\log\left(\frac{1}{\alpha}\right)e^{\log\left(\frac{1}{\alpha}\right)}\right) &\leq W(2T\hat{\lambda}) && \text{(Property 1 of Lambert W function)}\\
\implies \log\left(\frac{1}{\alpha}\right) &\leq 2\alpha T\hat{\lambda} && \text{(Property 3 of Lambert W function)}\\
\implies \alpha T &\geq \frac{\log(1/\alpha)}{2\hat{\lambda}} && \\
\implies F_X(\alpha T) &> 1 - e^{-\frac{\lambda\log(1/\alpha)}{2\hat{\lambda}}} = 1 - \alpha^{\frac{\lambda}{2\hat{\lambda}}} && \text{(sub-exponential tail assumption)}
\end{aligned}$$

Thus, by Lemma 1, we have

$$\begin{aligned} R_{\text{BL-Moss}}(T) &\leq T\left(1-(1-6\sqrt{\alpha})F_X(\alpha T)\right) \\ &< T\left(1-(1-6\sqrt{\alpha})(1-\alpha^{\frac{\lambda}{2\hat{\lambda}}})\right) \\ &= T\left(6\sqrt{\alpha}+\alpha^{\frac{\lambda}{2\hat{\lambda}}}-6\alpha^{\frac{1}{2}+\frac{\lambda}{2\hat{\lambda}}}\right) \\ &< T\left(6\sqrt{\alpha}+\alpha^{\frac{\lambda}{2\hat{\lambda}}}\right) \end{aligned}$$

Since $\alpha < 1$ and $\frac{1}{\hat{\lambda}} \geq \frac{1}{\lambda} - \mu$, we have

$$R_{\text{BL-Moss}}(T) = \begin{cases} O(T\sqrt{\alpha}), & \text{if } \mu\lambda \geq 1 \\ O(T\alpha^{\frac{1-\mu\lambda}{2}}), & \text{if } \mu\lambda < 1 \end{cases}$$

Case 1 ($\mu\lambda < 1$):

$$\begin{aligned} R_{\text{BL-Moss}}(T) &= O(T\alpha^{\frac{1-\mu\lambda}{2}}) \\ &= O\left(T\left(\frac{W(2\hat{\lambda}T)}{2\hat{\lambda}T}\right)^{\frac{1-\mu\lambda}{2}}\right) \\ &= O\left(T^{\frac{1+\mu\lambda}{2}}\left(\sqrt{\frac{W(2\hat{\lambda}T)}{2\hat{\lambda}}}\right)^{1-\mu\lambda}\right) \end{aligned}$$

Since $\mu\lambda < 1$ in this case, the above expression is sub-linear in T. If we absorb the constant parameters ($\hat{\lambda}$, λ, and μ), we have $R_{\text{BL-Moss}}(T) = O\left(T^{\frac{1+\mu\lambda}{2}}\left(\sqrt{\log T}\right)^{1-\mu\lambda}\right) \leq O\left(T^{\frac{1+\mu\lambda}{2}}\left(\sqrt{\log T}\right)\right)$ (since $\mu > 0$). Note also that if the learned mean parameter $\frac{1}{\hat{\lambda}}$ is close to the true mean parameter $\frac{1}{\lambda}$ (i.e., μ is close to zero), we recover the original regret guarantee which is $O(\sqrt{T\log(T)})$, with probability at least $1-\delta$.

Case 2 ($\mu\lambda \geq 1$): Here, $\alpha^{\frac{\lambda}{2\hat{\lambda}}} \leq \sqrt{\alpha}$ ($\because \alpha \leq 1$), and hence,

$$\begin{aligned} R_{\text{BL-Moss}}(T) &\leq O(T\sqrt{\alpha}) \\ &= O\left(T\sqrt{\frac{W(2\hat{\lambda}T)}{2\hat{\lambda}T}}\right) \\ &= O\left(\sqrt{\frac{T\cdot W(2\hat{\lambda}T)}{2\hat{\lambda}}}\right) \end{aligned}$$

If we absorb the constant parameter $\hat{\lambda}$ in order notation, we have $R_{\text{BL-Moss}}(T) = O(\sqrt{T\log(T)}$, that is, we recover the original regret guarantee with probability at least $1-\delta$.

□

7.1.2 Sub-Pareto tail distribution

Let $\hat{\beta}$ be the learned parameter of the sub-Pareto tail distribution. Like in the sub-exponential case, let there be two parameters μ and δ such that

$$\mathbb{P}\left(\left|\frac{\hat{\beta}}{\hat{\beta}-1} - \frac{\beta}{\beta-1}\right| \geq \mu\right) \leq \delta \tag{3}$$

Note that for sub-Pareto tail distribution, mean is defined only for $\beta > 1$, and thus we assume that $\beta > 1$ in the rest of the analysis. We can further derive the number of samples required as in the sub-exponential case, which will turn out to be the same for the given values of μ and δ. In Algorithm 1, we will now choose $\alpha = T^{\frac{-2\hat{\beta}}{1+2\hat{\beta}}}$ instead of $T^{\frac{-2\beta}{1+2\beta}}$ (which is the regret optimal α, from Theorem 4). We now have the following theorem on regret.

Theorem 6. *Let $\hat{\beta}$ be a learned distributional parameter such that it satisfies Inequality (3). If* BL-MOSS *(Algorithm 1) is run with $\hat{\beta}$, then with probability at least $1-\delta$, the regret under sub-Pareto tail distribution assumption is upper bounded by $O\left(T^{1-\frac{\beta(1-\mu\beta+\mu)}{1+2\beta-3\mu\beta+3\mu}}\right)$ if $\hat{\beta} > \beta$ and by $O\left(T^{\frac{1+\beta+2\mu(\beta-1)}{1+2\beta+3\mu(\beta-1)}}\right)$ if $\hat{\beta} \leq \beta$. In both the cases, the regret is sub-linear.*

Proof. For $c = 1/2$ (regret optimal value in Theorem 4), the following is true with probability at least $1-\delta$.

$$\begin{aligned}
&\alpha \geq T^{\frac{-\hat{\beta}}{\hat{\beta}+1/2}} \\
\implies &\alpha^{1-\frac{\hat{\beta}}{\hat{\beta}+1/2}} \geq (\alpha T)^{\frac{-\hat{\beta}}{\hat{\beta}+1/2}} \\
\implies &\alpha^{\frac{1/2}{\hat{\beta}+1/2}} \geq (\alpha T)^{\frac{-\hat{\beta}}{\hat{\beta}+1/2}} \\
\implies &\alpha^{\frac{\beta}{2\hat{\beta}}} \geq (\alpha T)^{-\beta} \\
\implies &F_X(\alpha T) > 1-(\alpha T)^{-\beta} \geq 1-\alpha^{\frac{\beta}{2\hat{\beta}}} && \text{(sub-Pareto tail assumption)}
\end{aligned}$$

Thus, by Lemma 1, we have $R_{\text{BL-MOSS}}(T) \leq T(1-(1-6\sqrt{\alpha})F_X(\alpha T)) < T\left(1-(1-6\sqrt{\alpha})(1-\alpha^{\frac{\beta}{2\hat{\beta}}})\right)$. Hence,

$$R_{\text{BL-MOSS}} < T\Big(6\sqrt{\alpha} + \alpha^{\frac{\beta}{2\hat{\beta}}} - 6\alpha^{\frac{\beta}{2\hat{\beta}}+\frac{1}{2}}\Big) < T\Big(6\sqrt{\alpha} + \alpha^{\frac{\beta}{2\hat{\beta}}}\Big)$$

Case 1 ($\hat{\beta} > \beta$): Note that, $\frac{\beta}{\beta-1} - \mu \leq \frac{\hat{\beta}}{\hat{\beta}-1}$ is equivalent to

$$\hat{\beta} \leq \frac{\beta - \mu\beta + \mu}{1 - \mu\beta + \mu} \tag{4}$$

We further have $\frac{\hat{\beta}}{\hat{\beta}-1} > 1$ and hence, the lower bound estimate on the mean should also be greater than 1. The lower bound estimate of $\frac{\hat{\beta}}{\hat{\beta}-1}$ with probability $1-\delta$ is: $\frac{\beta}{\beta-1} - \mu$. Thus, $\frac{\beta}{\beta-1} - \mu > 1$, which gives us that for the case $\hat{\beta} > \beta$:

$$\mu < \frac{1}{\beta - 1} \tag{5}$$

For this case, $\alpha^{\frac{\beta}{2\hat{\beta}}} > \sqrt{\alpha}$ $(\because \alpha \le 1)$, and hence,

$$\begin{aligned}
R_{\text{BL-Moss}} &\le O\Big(T\alpha^{\frac{\beta}{2\hat{\beta}}}\Big) \\
&= O\left(T^{1-\frac{\beta}{2\hat{\beta}}\left(\frac{2\hat{\beta}}{2\hat{\beta}+1}\right)}\right) && (\because \alpha = T^{\frac{-2\hat{\beta}}{2\hat{\beta}+1}} \text{ minimizes regret (Theorem 4)}) \\
&= O\left(T^{1-\frac{\beta}{2\hat{\beta}+1}}\right) \\
&\le O\left(T^{1-\frac{\beta}{2\left(\frac{\beta-\mu\beta+\mu}{1-\mu\beta+\mu}\right)+1}}\right) \\
& && (\because \text{the upper bound on } \hat{\beta} \text{ is } \tfrac{\beta-\mu\beta+\mu}{1-\mu\beta+\mu} \text{ w.p. at least } 1-\delta \text{ (Inequality (4))}) \\
&= O\left(T^{1-\frac{\beta(1-\mu\beta+\mu)}{1+2\beta-3\mu\beta+3\mu}}\right)
\end{aligned}$$

Note that the regret is sub-linear in T since $\frac{\beta(1-\mu\beta+\mu)}{1+2\beta-3\mu\beta+3\mu} > 0$ ($\because \mu < \frac{1}{\beta-1}$ for this case, from Inequality (5)).

Case 2 ($\hat{\beta} \le \beta$): We have $\frac{\hat{\beta}}{\hat{\beta}-1} \le \frac{\beta}{\beta-1} + \mu$ with probability at least $1-\delta$, which is equivalent to

$$\hat{\beta} \ge \frac{\beta + \mu\beta - \mu}{1 + \mu\beta - \mu} \tag{6}$$

For this case, $\alpha^{\frac{\beta}{2\hat{\beta}}} \le \sqrt{\alpha}$ $(\because \alpha \le 1)$, and hence,

$$\begin{aligned}
R_{\text{BL-Moss}} &\le O\Big(T\sqrt{\alpha}\Big) \\
&= O\left(T^{1-\frac{1}{2+\frac{1}{\hat{\beta}}}}\right) && (\because \alpha = T^{\frac{-2\hat{\beta}}{2\hat{\beta}+1}} = T^{\frac{-2}{2+\frac{1}{\hat{\beta}}}} \text{ (Theorem 4)}) \\
&\le O\left(T^{1-\frac{1}{2+\frac{1+\mu\beta-\mu}{\beta+\mu\beta-\mu}}}\right) \\
& && (\because \text{the lower bound on } \hat{\beta} \text{ is } \tfrac{\beta+\mu\beta-\mu}{1+\mu\beta-\mu} \text{ w.p. at least } 1-\delta \text{ (Inequality (6))}) \\
&= O\left(T^{1-\frac{\beta+\mu\beta-\mu}{1+2\beta+3\mu\beta-3\mu}}\right) \\
&= O\left(T^{\frac{1+\beta+2\mu(\beta-1)}{1+2\beta+3\mu(\beta-1)}}\right)
\end{aligned}$$

Note that the regret is sub-linear in T.

Note also that in both the cases, when μ tends to zero, the regret tends to go to the original regret of $O(T^{\frac{1+\beta}{1+2\beta}})$ (Theorem 4), with probability at least $1-\delta$.

□

Cost of Parametric Uncertainty (CPU): We now quantify the robustness of the regret guarantee provided by BL-Moss towards uncertainty in the tail distribution of the best arm's arrival. We define CPU to be the ratio of the regret achieved with the learned parameters and the regret achieved with the actual parameters. From the above theorems, we have the following:

1. CPU for sub-exponential tail distribution is given as (on absorbing the constant parameters $\hat{\lambda}$, λ, and μ in order notation):

$$CPU(\lambda, \mu) = \begin{cases} \dfrac{T^{\frac{1+\mu\lambda}{2}}\left(\sqrt{\log(T)}\right)^{1-\mu\lambda}}{\sqrt{T\log(T)}} = \left(\sqrt{\frac{T}{\log(T)}}\right)^{\mu\lambda} & \text{if } \mu\lambda < 1 \\ \dfrac{\sqrt{T\log(T)}}{\sqrt{T\log(T)}} = 1 & \text{if } \mu\lambda \geq 1 \end{cases}$$

2. CPU for sub-Pareto tail distribution is given as:

$$CPU(\beta, \mu) = \begin{cases} \dfrac{T^{1-\frac{\beta(1-\mu\beta+\mu)}{1+2\beta-3\mu\beta+3\mu}}}{T^{\frac{1+\beta}{1+2\beta}}} = T^{\frac{2\mu\beta(\beta-1)}{(1+2\beta)(1+2\beta-3\mu(\beta-1))}} & \text{if } \hat{\beta} > \beta \\ \dfrac{T^{\frac{1+\beta+2\mu(\beta-1)}{1+2\beta+3\mu(\beta-1)}}}{T^{\frac{1+\beta}{1+2\beta}}} = T^{\frac{\mu(\beta-1)^2}{(1+2\beta)(1+2\beta+3\mu(\beta-1))}} & \text{if } \hat{\beta} \leq \beta \end{cases}$$

7.2 Arm Arrival Distribution

Throughout the paper, we considered a setting that assumed certain distributions on the arrival of the best arm, namely, the best arm is more likely to arrive in early rounds. In this section, we discuss another practically relevant setting, which assumes distribution on the arrival rate of arms with time. If arms arrive at a faster rate in early rounds, that is, if a large fraction of arms arrive relatively early, it can be shown that one can use our proposed BL-MOSS algorithm. The following result establishes the equivalence between the distributional assumptions in the two settings.

Theorem 7. *Let $f(t)$ denote the fraction of arms arrived till time t. Further, let the quality of each arriving arm be an i.i.d. sample from unif[0, 1]. Then, the best arm's arrival distribution $F_X(t)$ satisfies $F_X(t) = f(t)$.*

Proof. Let $M \geq 1$ be the total number of arms arrived till time T. First, observe that $\mathbb{P}(i = i^\star) = \frac{1}{M}$. Here, $i^\star = \arg\max_{i \in [M]} q_i$. We have

$$\begin{aligned} F_X(t) &= \sum_{\ell=1}^{t} M \cdot (f(\ell) - f(\ell-1))\mathbb{P}(i_\ell = i^\star) \\ &= M \cdot (f(t) - f(0))\frac{1}{M} = f(t) && (\because f(0) = 0) \end{aligned}$$

□

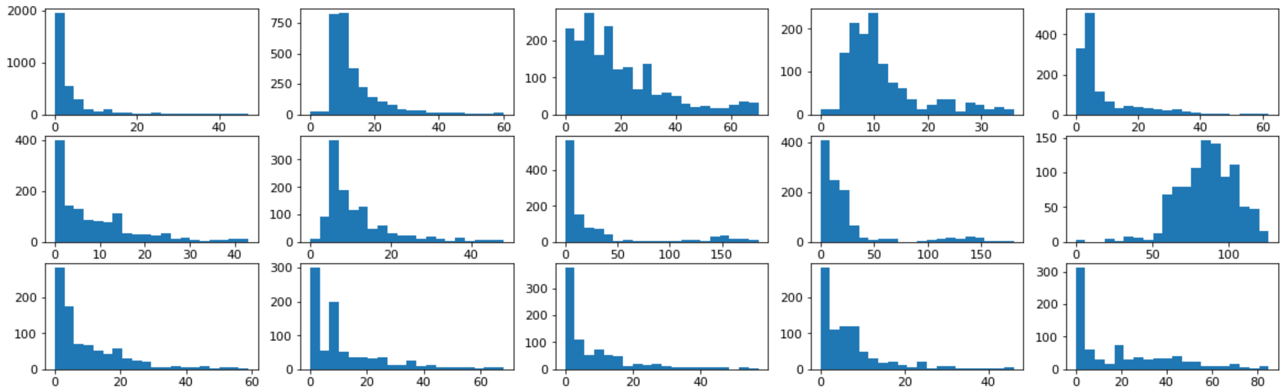

Figure 4

Arrival distribution of reviews for representative Digital Music products on Amazon (X-axis: number of months elapsed since the first review, Y-axis: number of reviews)

A fundamental difference between the two settings is that, in the first setting, we consider that a new arm arrives at each time instant; whereas in the second setting, we consider that arms follow an arrival process having a sharp tail. In the first setting, our proposed algorithm achieves sublinear regret due to early arrival of the best arm. In the second setting, even if each arm is equally likely to be the best arm, our algorithm achieves sublinear regret owing to most of the arms arriving relatively early.

We now validate our distributional assumption on the arrival of arms, with real-world data such as posting times of answers for questions on StackExchange[3] and posting times of reviews for products on Amazon[4] and Steam[5]. Figure 4 presents the arrival distribution of reviews over time for a representative set of the most popular digital music products on Amazon. Each subfigure shows the arrival of reviews for a particular product. In each subfigure, the X-axis represents the number of months elapsed since the first posted review for that product, and the Y-axis represents the number of reviews. It is to be noted that in MAB applications, X-axis usually represents the number of opportunities to pull the arms (here, the cumulative number of views to the reviews on a given product page) and not the wall-clock time (here, the number of months). We assume that the number of such views does not change significantly across different time intervals, and hence consider the wall-clock time as a proxy for the number of opportunities for arm pulls.

We observe that for most products, the number of reviews follows a decreasing trend over time (i.e., a large fraction of the reviews arrive early), which is aptly captured by sub-exponential and sub-Pareto tail distributions. A very similar trend was observed for questions on various sub-domains of StackExchange, reviews on Amazon products belonging to other categories like CDs & vinyls, video games, software, movies and TV, etc., as well as reviews on video games on Steam. Figure 5 presents the arrival distribution of answers over time for a representative set of the most answered questions on the Mathematics StackExchange platform. Likewise, the arrival distribution of reviews for representative products belonging to categories of CDs & Vinyls and video games are respectively presented in Figures 6 and 7 in Appendix C.

[3] StackExchange data dump is available publicly at [SE220].

[4] Amazon review data is available publicly at [Ni18].

[5] Steam video game and bundle data is available publicly at [ST117].

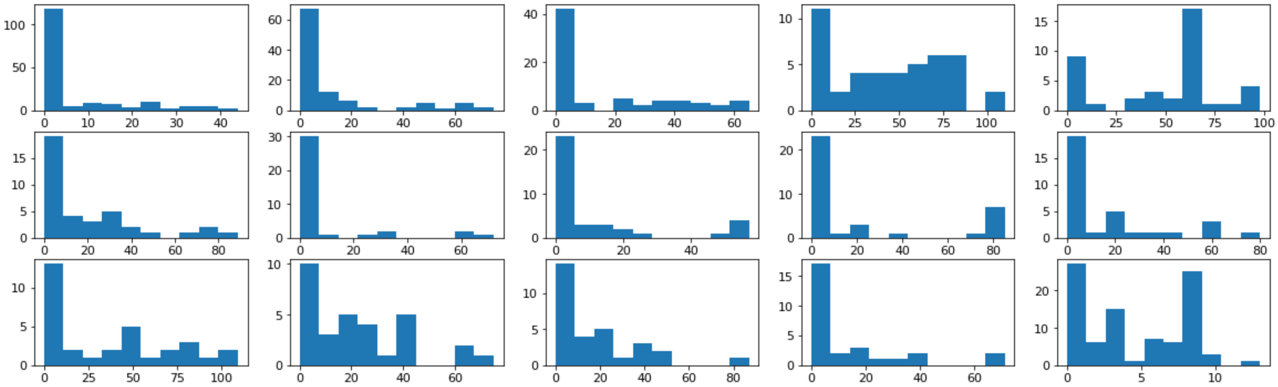

Figure 5

Arrival distribution of answers for representative questions on Mathematics StackExchange (X-axis: number of months since the first posted answer, Y-axis: number of answers)

It is relevant to note here that this trend is not necessarily followed across all product categories. For instance, certain products would follow upticks or a gradual increase in the number of posted reviews, either due to marketing campaigns (through media advertising or word-of-mouth publicity) or spike in demand during festive seasons. Amazon gift cards, cell phones and accessories, etc. are popular examples of such products (Figure 8 in Appendix C presents the distribution for Amazon gift cards).

A note to practitioners: In practice, the arrival distribution of reviews for any product would depend on the type of the product, marketing strategy of the product manufacturer, true quality of the product, and so on. Though the proposed BL-MAB framework provides curation strategy to identify the best arm (user generated content) from ever increasing choices, one must use expert knowledge about arrival rate of the arms, qualities of arriving arms, total expected number of arms, and so on, so as to design an optimal learning BL-MAB algorithm. We leave mathematical modeling and design of specialized BL-MAB algorithms which take into account the specific arrival of arms, as an interesting future direction to our work.

A Remark on Unknown Distributional Parameters of Arm Arrival Distribution

Note that if we have uncertainty with respect to the distributional parameters (discussed in Section 7.1) in the setting which makes distributional assumption on the rate of arrival of arms (discussed in Section 7.2), we could transform it into the setting which makes distributional assumption on the arrival time of the best arm using Theorem 7, and then show that the sub-linearity of regret is preserved even if we have uncertainty with respect to the distributional parameters (using Theorem 5 or 6). Thus, though the distributional parameters signify very different things in the two settings, one can prove that our algorithm would achieve sub-linear regret in such a combined case.

8 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 [LR85], the authors showed that any MAB algorithm in such a setting must incur a regret of $\Omega(\frac{\log T}{D_{\mathrm{KL}}})$ where D_{KL} is the Kullback-Leibler divergence between the best arm and the second best arm. Auer [Aue02] proposed the UCB1 algorithm which attains a matching upper bound on the expected regret. However, the distribution-free (i.e., in adversarial case) regret of the variant of UCB1, (α, ψ)-UCB, is given by $O(\sqrt{kT \log T})$ [BCB12]. The MOSS algorithm proposed by Audibert and Bubeck [AB10] achieves the distribution-free regret of $O(\sqrt{kT})$. Bubeck and Cesa-Bianchi [BCB12] present a detailed survey on regret bounds of these algorithms.

A setting similar to ballooning bandits is studied under Markovian bandits framework; where each arm is characterized by a known MDP. This setting, known as *arm-acquiring bandits* [Whi81] was first studied by [Nas73]. In arm acquiring bandits framework the goal is to maximize the discounted, infinite time cumulative reward whereas in ballooning bandits goal is to minimize the finite time cumulative regret. The difference in the two models 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 [GH13] 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 [TH19] 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 [LH18] limits the growth of the bandit arms by randomly dropping some of the arms from consideration, and computing the regret with respect to only the considered arms. That is, they do not account for the regret incurred due to the randomly dropped arms.

9 Discussion and Future Work

In this paper, we presented a novel extension to the classical MAB model, which we call the Ballooning bandits model (BL-MAB). We showed that, it is impossible to attain a sub-linear regret guarantee without any distributional assumption on the best arm's arrival. 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. However, in practice, these parameters may not be known exactly. We studied the increase in regret when one must use approximations of these values. It will be interesting to see how a learning algorithm can be designed to learn these parameters. We also studied the effect of a varying rate of arrival of arms (instead of the arrival time of the best arm). Owing to our equivalence theorem, our algorithm and results are directly applicable to cases wherein arm arrivals follow a sub-exponential or sub-Pareto structure. However, a general result with arbitrary (albeit sublinear) arrival of arms is still an open question. One could also consider other arrival processes for arms, in order to obtain tighter, arrival specific regret guarantees.

In this work, we employed MOSS as the underlying learning algorithm owing to its simplicity and optimality, in terms of both the number of arms and the time horizon. It is an interesting future direction to determine the threshold parameter α under other learning algorithms such as THOMPSON SAMPLING, UCB1, KL-UCB, and analyze the corresponding regret bounds. We assumed the knowledge of time horizon, as is the case with several works on MAB. Note that even if the time horizon is not known, one could always work with its approximate value which is typically known from past experiences. Extending our algorithm to the case of unknown time horizon using techniques such as MOSS-anytime [DP16] or doubling trick [BK18], is a promising direction for future work.

Acknowledgement

GG gratefully acknowledges the support from MHRD fellowship, Govt. of India and the Israeli Ministry of Science and Technology Grant 19400214.

References

[AB10] Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(Oct):2785–2836, 2010. (Cited on pages 10, 11, and 27)

[ACBF02] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine learning*, 47(2-3):235–256, 2002. (Cited on pages 2 and 10)

[AG12] Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 1–39, 2012. (Cited on pages 2 and 10)

[AHKL12] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Discovering value from community activity on focused question answering sites: A case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD*, pages 850–858, 2012. (Cited on page 9)

[AO10] Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010. (Cited on page 2)

[Aue02] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002. (Cited on page 27)

[BAG$^+$16] Keith Burghardt, Emanuel Alsina, Michelle Girvan, William Rand, and Kristina Lerman.

The myopia of crowds: A study of collective evaluation on stack exchange. *Robert H. Smith School Research Paper No. RHS*, 2736568, 2016. (Cited on page 3)

[BCB12] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. (Cited on pages 2 and 27)

[BCZ$^+$97] Donald A Berry, Robert W Chen, Alan Zame, David C Heath, and Larry A Shepp. Bandit problems with infinitely many arms. *The Annals of Statistics*, 25(5):2103–2116, 1997. (Cited on page 3)

[BK18] Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018. (Cited on page 28)

[BSS09] Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 79–88. ACM, 2009. (Cited on page 2)

[CGH$^+$96] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the Lambert W function. *Advances in Computational Mathematics*, 5(1):329–359, 1996. (Cited on page 9)

[CGJ$^+$17] Aritra Chatterjee, Ganesh Ghalme, Shweta Jain, Rohit Vaish, and Y Narahari. Analysis of Thompson sampling for stochastic sleeping bandits. In *Uncertainty in Artificial Intelligence, UAI 2017*, 2017. (Cited on page 3)

[CL11] Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011. (Cited on page 2)

[CV15] Alexandra Carpentier and Michal Valko. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning*, pages 1133–1141, 2015. (Cited on pages 3 and 4)

[DK17] R Devanand and P Kumar. Empirical study of Thompson sampling: Tuning the posterior parameters. In *AIP Conference Proceedings*, volume 1853, 2017. (Cited on page 2)

[DP16] Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 1587–1595, 2016. (Cited on page 28)

[GC11] Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011. (Cited on pages 2 and 10)

[GDJ$^+$20] Ganesh Ghalme, Swapnil Dhamal, Shweta Jain, Sujit Gujar, and Y. Narahari. Ballooning multi-armed bandits. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, page 1849–1851. IFAAMAS, 2020. (Cited on page 1)

[GH13] Arpita Ghosh and Patrick Hummel. 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*, pages 233–246, 2013. (Cited on pages 3 and 27)

[GHMS18] Aurélien Garivier, Hédi Hadiji, Pierre Menard, and Gilles Stoltz. KL-UCB-switch: Optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints, 2018. (Cited on page 10)

[H⁺56] Wassily Hoeffding et al. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 27(3):713–721, 1956. (Cited on page 20)

[HH08] Abdolhossein Hoorfar and Mehdi Hassani. Inequalities on the Lambert W function and hyperpower function. *J. Inequal. Pure and Appl. Math*, 9(2):5–9, 2008. (Cited on page 9)

[JGB⁺18] Shweta Jain, Sujit Gujar, Satyanath Bhat, Onno Zoeter, and Y. Narahari. A quality assuring, cost optimal multi-armed bandit mechanism for expertsourcing. *Artificial Intelligence*, 254:44 – 63, 2018. (Cited on page 2)

[KKM12] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012. (Cited on pages 2 and 10)

[KNMS10] Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2-3):245–272, 2010. (Cited on pages 3 and 5)

[LH18] Yang Liu and Chien-Ju Ho. Incentivizing high quality user contributions: New arm generation in bandit learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. (Cited on pages 3 and 27)

[LR85] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985. (Cited on pages 2, 7, and 26)

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. (Cited on page 2)

[MG17] Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pages 223–237, 2017. (Cited on page 10)

[MS14] Setareh Maghsudi and Sławomir Stańczak. Joint channel selection and power control in infrastructureless wireless networks: A multiplayer multiarmed bandit framework. *IEEE Transactions on Vehicular Technology*, 64(10):4565–4578, 2014. (Cited on page 2)

[Nas73] Peter Nash. *Optimal allocation of Resources Between Research Projects.* PhD thesis, University of Cambridge, 1973. (Cited on page 27)

[Ni18] Jianmo Ni. Amazon review data. "`https://nijianmo.github.io/amazon/index.html`", 2018. (Cited on page 25)

[NTGR18] Alessandro Nuara, Francesco Trovo, Nicola Gatti, and Marcello Restelli. 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*, 2018. (Cited on page 2)

[RVRK$^+$18] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018. (Cited on page 2)

[SE220] Stack exchange data dump. "https://archive.org/details/stackexchange", 2020. (Cited on page 25)

[Sli19] Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 12(1-2):1–286, 2019. (Cited on page 2)

[ST117] Steam video game and bundle data. "https://cseweb.ucsd.edu/~jmcauley/datasets.html#steam_data", 2017. (Cited on page 25)

[TH19] Wei Tang and Chien-Ju Ho. Bandit learning with biased human feedback. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1324–1332, 2019. (Cited on pages 3 and 27)

[Tho33] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. (Cited on pages 2 and 10)

[VBW15] Sofía S Villar, Jack Bowden, and James Wason. 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):199, 2015. (Cited on page 2)

[WAM09] Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for infinitely many-armed bandits. In *Advances in Neural Information Processing Systems*, pages 1729–1736, 2009. (Cited on pages 3 and 4)

[Whi81] Peter Whittle. Arm-acquiring bandits. *The Annals of Probability*, 9(2):284–292, 1981. (Cited on page 27)

Appendices

A Omitted Proofs

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.

$$\begin{aligned}
&\frac{W(\lambda T/c)}{\lambda T/c} < \frac{1}{36} \\
\iff &e^{-W(\lambda T/c)} < \frac{1}{36} \quad (\because W(x)e^{W(x)} = x) \\
\iff &W(\lambda T/c) > \log(36) \\
\iff &\frac{\lambda T}{c} > \log(36)e^{\log(36)} \\
\iff &T > \frac{36c\log(36)}{\lambda}
\end{aligned}$$

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

$$\begin{aligned}
&\lambda T/c > \lambda T/c_1 \\
\iff &W(\lambda T/c) > W(\lambda T/c_1) \quad \text{(Property 3 of Lambert } W) \\
\iff &e^{-W(\lambda T/c)} < e^{-W(\lambda T/c_1)} \\
\iff &\frac{W(\lambda T/c)}{\lambda T/c} < \frac{W(\lambda T/c_1)}{\lambda T/c_1} \quad (\because W(x)e^{W(x)} = x) \\
\iff &cW(\lambda T/c) < c_1 W(\lambda T/c_1) \\
\iff &e^{-cW(\lambda T/c)} > e^{-c_1 W(\lambda T/c_1)}
\end{aligned}$$

□

B Properties of Lambert W function

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

Proof. The forward direction is straightforward. As $W(\cdot)$ is one to one function in the non-negative domain, we have $W(W(x)e^{W(x)}) = W(x)$. Let $y = W(x)$ then we have $W(ye^y) = y$. To show that $W(xe^x) = x$ implies $W(z)e^{W(z)} = z$, observe that $W(W(z_0)e^{W(z_0)}) = W(z_0)$. We get the required result by taking the inverse. □

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

Proof. By definition, the Lambert W function satisfies $W(x)e^{W(x)} = x$. It is easy to see that $W(e) = 1$. Further we have,

$$\begin{aligned} 1 &= \frac{dW(x)}{dx} \cdot e^{W(x)} + \frac{dW(x)}{dx} \cdot e^{W(x)} W(x) \\ &= \frac{dW(x)}{dx}(x + e^{W(x)}) \\ \implies \frac{dW(x)}{dx} &= \frac{1}{x + e^{W(x)}} \end{aligned}$$

Let $f(x) = \log(x) - W(x)$. We have that $f(e) = 0$. We have that $\frac{df(x)}{dx} = \frac{1}{x} - \frac{1}{x+e^{W(x)}} = \frac{e^{W(x)}}{x(x+e^{W(x)})} = \frac{1}{x(1+W(x))} > 0$. Hence we have that $f(\cdot)$ is increasing i.e. $f(x) > 0$ for all $x > e$. This shows that $W(x) \leq \log(x)$.

Now, let $g(x) = \frac{\log(x)}{2} - W(x)$. Here, we have $g(e) < 0$, and $\frac{dg(x)}{dx} = \frac{1}{2x} - \frac{1}{x+e^{W(x)}} = \frac{e^{W(x)} - e^{\log(x)}}{2x(x+e^{W(x)})} \leq 0$ (since $W(x) \leq \log(x)$ for $x \geq e$). So, $g(x) < 0$ for all $x \geq e$, implying that $\frac{\log(x)}{2} < W(x)$. This completes the proof. □

Property 3. For any $x \in [0, \infty)$, the Lambert W function is unique, non-negative, and strictly increasing.

Proof. Observe that $W(0) = 0$. Note that in the non-negative domain, $f(x) = xe^x$ is continuous, one to one and strictly increasing. Hence, its inverse, $W(\cdot)$, is also increasing. □

C Validating Arm Arrival Distributions

In each subfigure, X-axis represents the number of months elapsed since the first posted review for the corresponding product, and Y-axis represents the number of reviews.

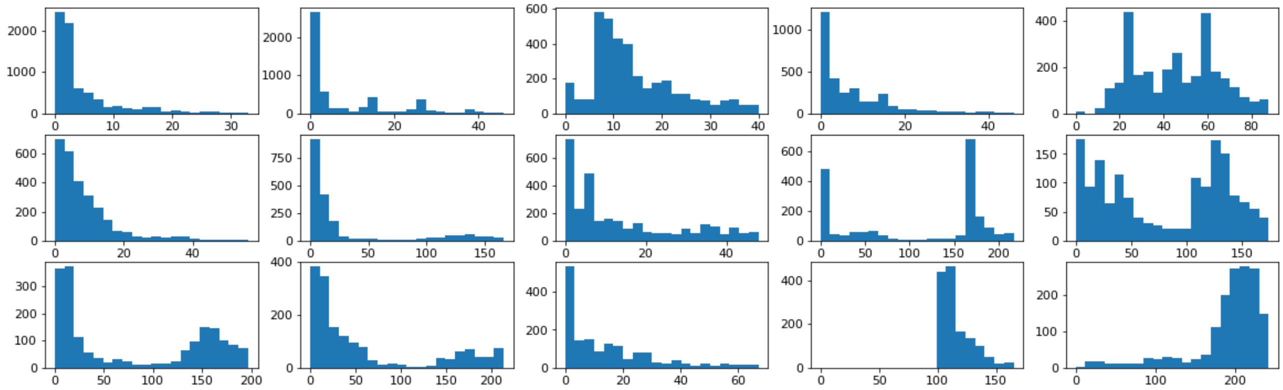

Figure 6

Arrival distribution of reviews for representative CDs & vinyl products on Amazon

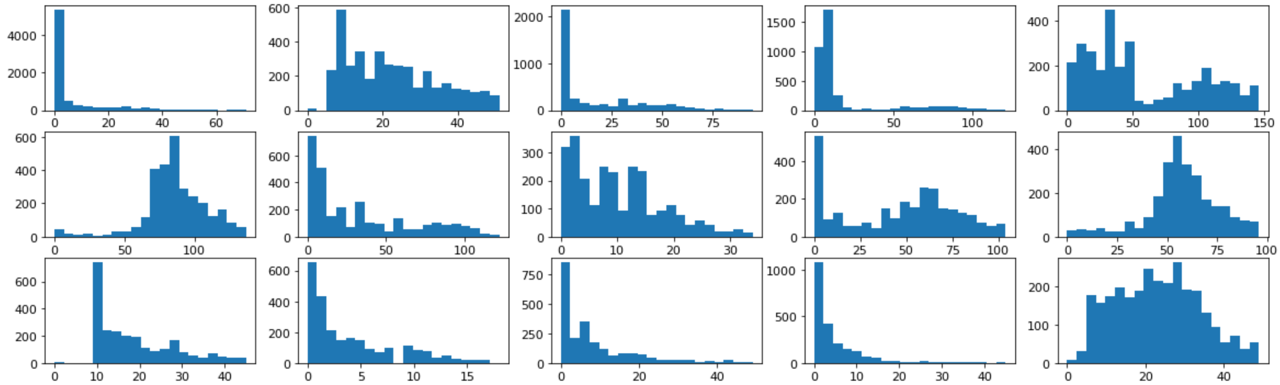

Figure 7

Arrival distribution of reviews for representative video game products on Amazon

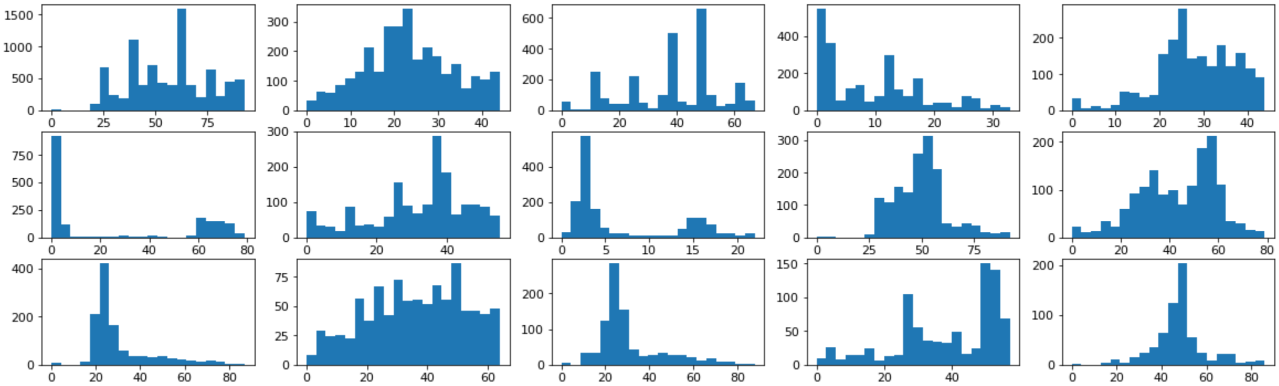

Figure 8

Arrival distribution of reviews for representative gift card products on Amazon