Ballooning Multi-Armed Bandits

Extended Abstract

Ganesh Ghalme
Indian Institute of Science
Bangalore, India
ganeshg@iisc.ac.in

Swapnil Dhamal
Chalmers University of Technology
Gothenburg, Sweden
swapnil.dhamal@gmail.com

Shweta Jain
Indian Institute of Technology
Ropar, India
shwetajains20@gmail.com

Sujit Gujar
International Institute of Information Technology
Hyderabad, India
sujit.gujar@iiit.ac.in

Y. Narahari
Indian Institute of Science
Bangalore, India
narahari@iisc.ac.in

ABSTRACT

We introduce *ballooning multi-armed bandits* (BL-MAB), a novel extension to the classical stochastic MAB model. In the BL-MAB model, the set of available arms grows (or balloons) over time. The regret in a BL-MAB setting is computed with respect to the best available arm at each time. We first observe that the existing stochastic MAB algorithms are not regret-optimal for the BL-MAB model. We show that if the best arm is equally likely to arrive at any time, a sub-linear regret cannot be achieved, irrespective of the arrival of the other arms. We further show that if the best arm is more likely to arrive in the early rounds, one can achieve sub-linear regret. 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.

ACM Reference Format:
Ganesh Ghalme, Swapnil Dhamal, Shweta Jain, Sujit Gujar, and Y. Narahari. 2020. Ballooning Multi-Armed Bandits. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

The classical stochastic multi-armed bandit (MAB) problem encapsulates the classical exploration versus exploitation dilemma, in that the planner's algorithm has to arrive at an optimal trade-off between pulling relatively unexplored arms and pulling the best arms according to the history of pulls thus far. This problem has been extensively studied in the literature. These studies include analyzing the lower bound on regret [14], analysis of asymptotically optimal algorithms [1, 4, 5, 19], empirical studies [8, 10, 17], and several extensions [7, 18].

The theoretical results in MAB are complemented by a wide variety of modern applications which can be seamlessly modelled in the MAB setup. Internet advertising [6, 16], crowdsourcing [12], clinical trials [20], wireless communication [15] represent a few of the many applications. In this paper, we propose 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. As the number of arms increases (potentially linearly) with time, it is clear that an optimal algorithm has to ignore (or drop) a few arms. Hence, in addition to achieving an optimal trade-off between the number of exploratory pulls and exploitation pulls, the algorithm must also ensure that it does not drop too many or too few arms.

BL-MAB, in general, is directly applicable in any scenario, where the set of options grows over time and the objective is to choose the best option available at any given time. A contemporary example is provided by question and answer (Q&A) platforms such as Reddit, Stack Overflow, Quora, Yahoo! Answers, and ResearchGate, where the platform's goal is to discover the highest quality answer that should be displayed in the most prominent slot, for a given question. Each display of a posted answer corresponds to a pull of the corresponding arm (answer post). At each time instant, a new user observes the existing answer posts shown by the platform, decides whether to endorse them, and may also choose to post her own answer, thus increasing the number of available arms.

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

Our Contributions: We introduce the BL-MAB model that allows the set of arms to grow over time, and show that in the absence of any distributional assumption on the arrival time of the best (or highest quality) arm, the regret will grow linearly with time (Theorem 2.1). 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

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

on the arrival distribution of the best arm, namely, sub-exponential tail (Theorem 2.2) and sub-Pareto tail (Theorem 2.3). Note that we make no assumption on the arrival of the other arms.

2 THE MODEL AND MAIN RESULTS

A BL-MAB instance is given by $\langle T, (K(t), (\mathcal{D}_i)_{i \in K(t)})_{t=1}^{T} \rangle$. Here, $K(t)$ is the set of arms available at time t 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 BL-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. 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 $|K(t)| \leq |K(t+1)| \leq |K(t)+1|$.

Similar to the notion of regret in the sleeping stochastic MAB model [9, 13], the notion of regret in BL-MAB setting 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., $i_t^\star = \arg\max_{i \in K(t)} q_i$. The instance-dependent regret of a BL-MAB algorithm A is given by $\mathcal{R}_A(T, \mathcal{I}) = \mathbb{E}\big[\sum_{t=1}^{T} (q_{i_t^\star} - q_{i_t})\big]$. Throughout the paper, we consider instance-independent regret, given as $\mathcal{R}_A(T) = \sup_{\mathcal{I}} \mathcal{R}_A(T, \mathcal{I})$. Our first result shows that, without any side information about the arms, it is impossible to achieve sub-linear regret.[1]

Theorem 2.1. *There exists a BL-MAB instance $\mathcal{I}$ such that the expected regret of any algorithm A is lower bounded by $\Omega(T)$.*

Theorem 2.1 provides a strong impossibility result on the achievable instance-independent regret bound under BL-MAB setting. We hence impose a restriction on the arrival of the best arm $i^\star = \arg\max_{i \in K(T)} q_i$, that the probability of $i^\star$ arriving early is large enough. This would allow a learning algorithm to explore the best arm enough to estimate its true quality with high probability.

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.

Sub-exponential tail property: 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., $F_X(t) > 1 - e^{-\lambda t}$.

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

The aforementioned assumptions naturally arise in the context of Q&A forums as observed in extensive empirical studies on the nature of answering as well as voting behavior of the users. Anderson et al. [2] observe that high reputation users hasten to post their answers early.

The Proposed BL-Moss Algorithm: We now present our algorithm, BL-Moss, that uses Moss [3] as a black-box. The number of arms explored by BL-Moss is dependent on the distribution of arrival of the best arm. In particular, BL-Moss considers only the first $\lceil \alpha T \rceil$ arms in its execution ($\alpha \in (0, 1]$). The Moss algorithm is run with $\lceil \alpha T \rceil$ arms whereas the later arms are ignored.[1]

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, $q_{i^\star} \geq q_{j^\star}$. As BL-Moss does not consider all the arms, the regret of BL-Moss can be decomposed into the following two parts. The external regret of BL-Moss (say $\mathcal{R}^{\text{ext}}_{\text{BL-Moss}}(T)$) is the regret incurred due to considering only a subset consisting of the first $\lceil \alpha T \rceil$ of the available arms, whereas the internal regret (say $\mathcal{R}^{\text{int}}_{\text{BL-Moss}}(T)$) is due to not apriori knowing the qualities of each of the $\lceil \alpha T \rceil$ arms it considers. Write $\Delta(i, j) = q_i - q_j$ and let t_i be the time of arrival of arm i. Let $i_t^\star$ denote the best arm till time t. The instance-dependent regret is given as

$$\mathcal{R}_{\text{BL-Moss}}(T, \mathcal{I}) = \mathbb{P}(i^\star = j^\star) \cdot \mathcal{R}^{\text{int}}_{\text{BL-Moss}}(T) + \mathbb{P}(i^\star \neq j^\star) \cdot \mathcal{R}^{\text{ext}}_{\text{BL-Moss}}(T),$$

where $\mathcal{R}^{\text{int}}_{\text{BL-Moss}}(T) = \sum_{t=1}^{t_{j^\star}-1} \Delta(i_t^\star, i_t) + \sum_{t=t_{j^\star}}^{T} \Delta(j^\star, i_t)$

and $\mathcal{R}^{\text{ext}}_{\text{BL-Moss}}(T) = \sum_{t=1}^{t_{i^\star}-1} \Delta(i_t^\star, i_t) + \sum_{t=t_{i^\star}}^{T} \Delta(i^\star, i_t)$.

We ignore the ceiling in $\lceil \alpha T \rceil$ to avoid notation clutter. We now present the main results of the paper under sub-exponential and sub-Pareto tail assumptions.[1]

Theorem 2.2. *Let the arrival distribution of the best arm satisfy the sub-exponential tail property for some $\lambda > 0$, and let T be large enough such that $T > 36c\log(36)/\lambda$ 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 $O\Big(T \cdot \max\big(e^{-cW(\lambda T/c)}, e^{-W(\lambda T/c)/2}\big)\Big)$. The upper bound on the expected regret is minimized when $c = 1/2$ and is given by $O(\sqrt{T \cdot \log(2\lambda T)/2\lambda})$.*

Theorem 2.3. *Let the arrival distribution of the best arm satisfy the sub-Pareto tail property for some $\beta > 0$, and let T be large enough such that $T > (36)^{(c+\beta)/\beta}$ for some $c > 0$. Then with $\alpha = T^{-\beta/(\beta+c)}$, 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 $O(T^{(1+\beta)/(1+2\beta)})$.*

3 DISCUSSION AND FUTURE WORK

We presented the Ballooning multi-armed bandits (BL-MAB) model, a novel extension to the classical MAB model. We showed that, in the absence of any side information, it is impossible to achieve a sub-linear regret in the BL-MAB setting. We provided sufficient conditions under which the proposed algorithm (BL-Moss) achieves sub-linear regret. When the arrival distribution of the best quality arm has a sub-exponential or sub-Pareto tail, our algorithm BL-Moss achieves sub-linear regret by restricting the number of arms to be explored in an intelligent way. Our results indicate that, the number of arms to be explored depends on the distributional parameters, namely, λ (for sub-exponential case) and β (for sub-Pareto case), which must be known to the algorithm. It would be interesting to see how a learning algorithm could be designed to learn these parameters as well. As noted earlier, we consider a structure on only the arrival of the best arm. One may also consider a more sophisticated arrival process of the arms, for obtaining better regret guarantees.

[1] The details and proofs are available in the extended version of the paper [11].

REFERENCES

[1] Shipra Agrawal and Navin Goyal. 2012. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Conference on Learning Theory*. 1–39.

[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Discovering value from community activity on focused question answering sites: A case study of Stack Overflow. In *Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 850–858.

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

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

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

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

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

[8] Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*. 2249–2257.

[9] Aritra Chatterjee, Ganesh Ghalme, Shweta Jain, Rohit Vaish, and Y Narahari. 2017. Analysis of Thompson sampling for stochastic sleeping bandits. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

[10] R Devanand and P Kumar. 2017. Empirical study of Thompson sampling: Tuning the posterior parameters. *American Institute of Physics (AIP) Conference Proceedings* 1853, 1 (2017), 050001.

[11] Ganesh Ghalme, Swapnil Dhamal, Shweta Jain, Sujit Gujar, and Y Narahari. 2020. Ballooning Multi-Armed Bandits. *arXiv preprint arXiv:2001.10055* (2020).

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

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

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

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

[16] Alessandro Nuara, Francesco Trovo, Nicola Gatti, and Marcello Restelli. 2018. A combinatorial-bandit algorithm for the online joint bid/budget optimization of pay-per-click advertising campaigns. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 2379–2386.

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

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

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

[20] Sofía S Villar, Jack Bowden, and James Wason. 2015. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 30, 2 (2015), 199–215.