Operations Research Letters 49 (2021) 728–733

Contents lists available at ScienceDirect

Operations Research Letters

www.elsevier.com/locate/orl

Multi-armed bandit with sub-exponential rewards

Huiwen Jia, Cong Shi*, Siqian Shen

Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109, United States of America

ARTICLE INFO

Article history:
Received 27 February 2021
Received in revised form 2 August 2021
Accepted 3 August 2021
Available online 12 August 2021

Keywords:
Multi-armed bandit
Unbounded reward
Sub-exponential reward
Upper confidence bound

ABSTRACT

We consider a general class of multi-armed bandits (MAB) problems with sub-exponential rewards. This is primarily motivated by service systems with exponential inter-arrival and service distributions. It is well-known that the celebrated Upper Confidence Bound (UCB) algorithm can achieve tight regret bound for MAB under sub-Gaussian rewards. There has been subsequent work by Bubeck et al. (2013) [4] extending this tightness result to any reward distributions with finite variance by leveraging robust mean estimators. In this paper, we present three alternative UCB based algorithms, termed UCB-Rad, UCB-Warm, and UCB-Hybrid, specifically for MAB with sub-exponential rewards. While not being the first to achieve tight regret bounds, these algorithms are conceptually simpler and provide a more explicit analysis for this problem. Moreover, we present a rental bike revenue management application and conduct numerical experiments. We find that UCB-Warm and UCB-Hybrid outperform UCB-Rad in our computational experiments.

1. Introduction

Consider multi-armed bandit (MAB) problems introduced by Robbins [16], where an agent faces a set of actions associated with unknown reward distributions. The goal of the agent is to collect as much reward as possible within a fixed number of rounds. A typical performance measure for MAB is cumulative regret, defined as the difference between the collected rewards by the policy of interest and by choosing the arm with the highest expected reward in each round under the full information scenario. The literature predominantly assumes that rewards follow sub-Gaussian (including bounded) distributions. Indeed, employing concentration inequalities of the light-tailed sub-Gaussian distributions [13,2], several classes of algorithms, including Upper Confidence Bound (UCB) and Thompson Sampling (TS), have been analyzed with tight regret bounds. We refer readers to textbooks by Slivkins [21], Lattimore and Szepesvári [12], Bubeck and Cesa-Bianchi [3] for an extensive overview of MAB.

Many real-world decision-making problems require the reward distributions to be more general. For instance, any service systems with exponential inter-arrival and service time do not enjoy the sub-Gaussian property of classical MAB. To this end, we study a general class of MAB with sub-exponential reward distributions (an important family of unbounded and heavy-tailed distributions). Bubeck et al. [4] was the first to thoroughly study MAB with heavy-tailed reward distributions. They considered rewards with moments of order $1+\epsilon$ for some $\epsilon \in (0,1]$. Their general approach is to replace the empirical mean in the upper confidence index with other robust estimators, such as the truncated empirical mean, Catoni's M-estimator, and the median-of-means estimator. Their results also imply that moments of order 2 (i.e., finite variance) are sufficient to obtain regret bounds of the same order as under sub-Gaussian reward distributions.

In this paper, we present three alternative UCB-based algorithms, termed UCB-Rad, UCB-Warm, and UCB-Hybrid, specifically for MAB with sub-exponential rewards. We prove that the cumulative regret of the proposed algorithms is $O(\sqrt{MT\log(T)} + M\log(T))$, where T is the total number of rounds and M is the number of arms, and this result matches the lower bound of general MAB up to a logarithmic factor. Although not being the first to achieve tight regret bounds, these algorithms are conceptually simpler and provide a more explicit treatment (in lieu of complex robust estimators used in Bubeck et al. [4]) for this problem. Another contribution of our study is to present a rental bike revenue management application and conduct extensive numerical studies to evaluate the empirical performance of the proposed algorithms. Both UCB-Warm and UCB-Hybrid outperform UCB-Rad numerically in our experimental results.

Besides Bubeck et al. [4], there have been other studies devoted to MAB with heavy-tailed reward distributions. Vakili et al. [22] proposed sequential phases of exploration and exploitation. Ko-

* Corresponding author.
E-mail addresses: hwjia@umich.edu (H. Jia), shicong@umich.edu (C. Shi), siqian@umich.edu (S. Shen).

https://doi.org/10.1016/j.orl.2021.08.004
0167-6377/

rda et al. [11] and Dubey and Pentland [6] modified Thompson Sampling (TS) algorithms for heavy-tailed rewards. Xia et al. [23] studied heavy-tailed rewards under the multi-play setting and Dubey and Pentland [7] considered multi-agent setting. One could potentially extend our proposed algorithms to these general settings. Moreover, we believe that our algorithms find a wide range of online learning applications, including revenue management (e.g., [9,24,5]), online advertisement assortment (e.g., [18,1]), and health-care applications (e.g., [25,10]).

The remainder of this paper is organized as follows. In §2, we define the problem and present three modified UCB algorithms. In §3, we study a stylized revenue management example with exponential reward distributions (one family of sub-exponential distributions) and present numerical results of proposed algorithms. In §4, we conclude the paper and point out future research directions.

2. Three UCB based algorithms for sub-exponential rewards

Consider M arms, denoted by a set $\mathcal{M}$ and $|\mathcal{M}| = M$. The agent can pull one arm $m \in \mathcal{M}$ in each round t, $\forall t = 1, \dots T$ and receive a random reward r_m. The uncertain reward r_m of each arm follows a (τ_m^2, b_m)-sub-exponential distribution with mean μ_m. We assume that these three distribution parameters of each arm m, i.e., (μ_m, τ_m^2, b_m), are unknown to the agent. The goal of the agent is to find an effective policy π, where the agent pulls arm $m_t(\pi) \in \mathcal{M}$ in round t, to maximize the expected cumulative reward by the end of round T, denoted by $J^\pi = \mathbb{E}[\sum_{t=1}^T r_{m_t(\pi)}]$. Examples of sub-exponential variables include (i) exponential random variables and (ii) χ^2 random variables. We refer readers to Foss et al. [8] for more properties of sub-exponential distributions.

Definition 1. A random variable X with mean $\mathbb{E}[X]$ is (τ^2, b)-sub-exponential if

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \le \exp\left(\frac{\lambda^2\tau^2}{2}\right) \text{ for } |\lambda| \le \frac{1}{b}.$$

Regret. The notion cumulative regret or simply regret is commonly used in online learning (see, e.g., [20]) to evaluate the performance of a policy if the decision maker has limited information of the system against the optimal performance under full information. In this problem, the optimal policy is a static policy where the agent always pulls the arm with the highest expected reward. Denote the arm with the highest expected reward as m^*, and thus $m^* = \text{argmax}_{m\in\mathcal{M}}\mu_m$. Denote the expected cumulative reward of the optimal policy as J^* and thus we have $J^* = T\mu_{m^*}$. Therefore, we define the regret of this problem as follows.

Definition 2. The regret by the end of round T of policy π is defined as $R(\pi, T) = J^* - J^\pi$.

We introduce three algorithms: UCB-Rad, UCB-Warm, and UCB-Hybrid. UCB-Rad is a direct extension from general UCB to handle sub-exponential rewards, by enlarging the confidence radius and substituting Lemma 1 for Hoeffding's inequality. A potential pitfall of UCB-Rad is that the resulting confidence radius could be too large, which slows down learning. To reduce the confidence radius of UCB-Rad, we develop another two modified UCB algorithms. UCB-Warm includes a warm-up phase (uniform sampling) before starting active exploration. UCB-Hybrid utilized hybrid confidence radii along the learning process. We present an algorithm overview in Table 1.

To provide a basis for our proposed algorithms, we now invoke a standard concentration inequality for sub-exponentials from Bernstein's inequality (see, e.g., Rigollet and Hütter [15]).

Lemma 1 *(Concentration of sub-exponentials from Bernstein's Inequality). Let a random variable X be (τ^2, b)-sub-exponential. Then for a non-negative number $t \ge 0$:*

$$\mathbb{P}(|X - \mathbb{E}[X]| \ge t) \le \left\{2\exp\left(-\frac{t^2}{2\tau^2}\right) \text{ if } 0 \le t \le \frac{\tau^2}{b};\ 2\exp\left(-\frac{t}{2b}\right) \text{ if } t \ge \frac{\tau^2}{b}\right\}.$$

Assumption 1. The agent knows valid upper bounds $\tilde{\tau}_m^2$ and η_m for arm $m \in \mathcal{M}$ such that

$$\tilde{\tau}_m^2 \ge \tau_m^2 \text{ and } \eta_m \ge (b_m^2/\tau_m^2).$$

Assumption 1 requires the agent to only know the upper bounds of boundary/shape information of the distribution parameters. Technically, this assumption is used in Proposition 1 to mitigate the heavy-tailed effects. Note that UCB-Rad and UCB-Hybrid only require Assumption 1.

Assumption 2. $T \gg \sum_{m\in\mathcal{M}} 8\eta_m \log(T)$.

Assumption 2 is further assumed to apply UCB-Warm. It requires the playing round horizon to be sufficiently long, since it takes time for the agent to eliminate the heavy-tailed effects and learn the underlying reward distributions through dynamic interactions with the environment.

2.1. Modified UCB algorithm with enlarged radius: UCB-Rad

By the end of round t, denote the observed reward of arm m as $\hat{r}_m^i$ for $i = 1, \dots, n_t(m)$, where $n_t(m)$ is the number of observed rewards of arm m, i.e., the number of rounds that the agent pulls arm m. The UCB-Rad algorithm is an extension of UCB1 (see, e.g., Slivkins [21]). In each round t, UCB-Rad chooses the arm with the maximum upper confidence bound of the mean of the reward.

Definition 3. The upper confidence bound of the mean of the reward associated with arm m in UCB-Rad is defined as:

$$U_t^R(m) = \frac{\sum_{i=1}^{n_t(m)} \hat{r}_m^i}{n_t(m)} + \sqrt{\frac{8\tilde{\tau}_m^2 \log(T)}{n_t(m)}} + \frac{8\sqrt{\eta_m}\tilde{\tau}_m \log(T)}{n_t(m)}, \tag{1}$$

and $\mathbf{Rad}_t^R(m) = \sqrt{\frac{8\tilde{\tau}_m^2 \log(T)}{n_t(m)}} + \frac{8\sqrt{\eta_m}\tilde{\tau}_m \log(T)}{n_t(m)}$ is the confidence radius of arm m by end of round t.

We present UCB-Rad in Algorithm 1. Note that compared with UCB1, we enlarge the confidence radius from $\sqrt{\frac{2\log(T)}{n_m}}$ to $\mathbf{Rad}_t^R(m)$ to counteract the heavy-tailed effects of sub-exponentials.

Algorithm 1 Modified UCB algorithm with enlarged radius: UCB-Rad.

1: Input: $\mathcal{M}$, T, η_m, $\tilde{\tau}_m$, $\forall m \in \mathcal{M}$.
2: Initialize: $U_0^R(m) \leftarrow +\infty$, $\forall m \in \mathcal{M}$.
3: **for** $t = 1, \dots, T$ **do**
4: Pull arm $m_t = \text{argmax}_{m\in\mathcal{M}} U_{t-1}^R(m)$ and receive reward $\hat{r}_{m_t}^i$.
5: Update $U_t^R(m_t)$ based on (1), $t \leftarrow t + 1$.
6: **end for**

Proposition 1 *(Sub-exponential concentration bound for UCB-Rad). For any arm $m \in \mathcal{M}$ by the end of round $t = 1, \dots, T$, we have*

$$\mathbb{P}\left(\left|\mu_m - \frac{\sum_{i=1}^{n_t(m)} \hat{r}_m^i}{n_t(m)}\right| \le \mathbf{Rad}_t^R(m)\right) \ge 1 - \frac{2}{T^4}.$$

Table 1
Overview and comparison of the three algorithms in this paper.

	UCB-Rad	UCB-Warm	UCB-Hybrid
Assumptions	Assumption 1	Assumptions 1 and 2	Assumption 1
Needs a Warm-up Phase?	No	Yes	No
Estimated Mean	$\sum_{i=1}^{n_t(m)} \hat{r}_m^i / n_t(m)$		
Confidence Radius	$\sqrt{\frac{8\tilde{\tau}_m^2 \log(T)}{n_t(m)}} + \frac{8\sqrt{\eta_m}\tilde{\tau}_m \log(T)}{n_t(m)}$	$\sqrt{\frac{8\tilde{\tau}_m^2 \log(T)}{n_t(m)}}$	Hybrid of $\sqrt{\frac{8\tilde{\tau}_m^2 \log(T)}{n_t(m)}}$ and $\frac{8\sqrt{\eta_m}\tilde{\tau}_m \log(T)}{n_t(m)}$
Regret Bound		$O\left(M\log(T) + \sqrt{MT\log(T)}\right)$	

Theorem 1. *The cumulative regret of UCB-Rad is bounded by* $O\left(\sqrt{MT\log(T)} + M\log(T)\right)$.

Note that the regret lower bound for nominal MAB is $\Omega(\sqrt{MT} + M)$ (see, e.g., [21]), which suggests that our regret bound is tight, up to a logarithmic factor.

2.2. Modified UCB algorithm with Warm-up phase: UCB-Warm

The UCB-Warm algorithm includes a Warm-up phase, which allows to use the similar tight concentration bound as that for sub-Gaussian tail. In the *Warm-up* Phase, the algorithm pulls each arm for $8\eta_m \log(T)$ number of rounds and collect the reward. The second phase is *Learning* Phase and UCB-Warm chooses the arm with the maximum upper confidence bound $U_t^W(m)$, which is smaller than $U_t^R(m)$ and is defined in Definition 4. We present UCB-Warm in Algorithm 2.

Definition 4. The upper confidence bound of the mean of the reward associated with arm m in UCB-Warm is defined as:

$$U_t^W(m) = \frac{\sum_{i=1}^{n_t(m)} \hat{r}_m^i}{n_t(m)} + \sqrt{\frac{8\tilde{\tau}_m^2 \log(T)}{n_t(m)}} \tag{2}$$

where $\mathbf{Rad}_t^W(m) = \sqrt{\frac{8\tilde{\tau}_m^2 \log(T)}{n_t(m)}}$ is the confidence radius of arm m by end of round t of UCB-Warm.

Algorithm 2 Modified UCB algorithm with Warm-up phase: UCB-Warm.

1: Input: $\mathcal{M}$, T, η_m, $\tilde{\tau}_m$, $\forall m \in \mathcal{M}$.
2: Warm-up Phase:
3: **for** $m \in \mathcal{M}$ **do**
4: Pull arm m for $8\eta_m \log(T)$ rounds and collect reward $\hat{r}_m^i$, $i = 1, \ldots, 8\eta_m \log(T)$.
5: Compute $U_t(m)$.
6: **end for**
7: Learning Phase:
8: **for** $t = t_L, \ldots, T$ **do**
9: Pull arm $m_t = \text{argmax}_{m\in\mathcal{M}} U_{t-1}^W(m)$ and receive reward $\hat{r}_{m_t}^i$.
10: Update $U_t^W(m_t)$ by (2), $t \leftarrow t+1$.
11: **end for**

Corollary 1 *(Sub-exponential concentration bound for learning phase of UCB-Warm). For any arm $m \in \mathcal{M}$, if the number of reward observations $n_t(m) \geq 8\eta_m \log(T)$, then*

$$\mathbb{P}\left(\left|\mu_m - \frac{\sum_{i=1}^{n_t(m)} \hat{r}_m^i}{n_t(m)}\right| \leq \sqrt{\frac{8\tilde{\tau}_m^2 \log(T)}{n_t(m)}}\right) \geq 1 - \frac{2}{T^4}.$$

Corollary 1 shows that after observing a certain amount of observations, the negative effect of the heavy-tailed sub-exponentials has been eliminated and the concentration inequality is as tight as that for sub-Gaussians. The proof is similar to that of Proposition 1 and hence omitted.

Theorem 2. *The cumulative regret of UCB-Rad is bounded by* $O(\sqrt{MT\log(T)} + M\log(T))$.

The regret during the warm-up phase is linear on the length of the operation time and thus one has $R_W(\pi_{\text{UCB-Warm}}, T) = O(M\log(T))$. The regret during the learning phase can be bounded by using the same techniques for Theorem 1 and thus one has $R_L(\pi_{\text{UCB-Warm}}, T) \leq O(\sqrt{MT\log(T)})$. By adding these two regret terms together, $R(\pi_{\text{UCB-Warm}}, T) \leq (\sqrt{MT\log(T)} + M\log(T))$.

2.3. Modified UCB algorithm with hybrid radii: UCB-Hybrid

The UCB-Hybrid algorithm is a hybrid algorithm combining the merits of UCB-Rad and UCB-Warm. Compared to UCB-Rad, UCB-Hybrid utilizes a smaller confidence radius while pursuing the same level of confidence, i.e., $1 - 2/T^4$. Compared to UCB-Warm, UCB-Hybrid algorithm makes adaptive decisions even when the number of trials is less than the threshold $8\eta_m \log(T)$ while pursuing the same level of confidence. The UCB-Hybrid works similarly as UCB1 and UCB-Rad, but with two types of upper confidence bounds. We present UCB-Hybrid in Algorithm 3.

Algorithm 3 Modified UCB algorithm with hybrid radii: UCB-Hybrid.

1: Input: $\mathcal{M}$, T, $\eta_m, \tilde{\tau}_m$, $\forall m \in \mathcal{M}$.
2: Initialize: $U_0^H(m) \leftarrow +\infty$, $\forall m \in \mathcal{M}$.
3: **for** $t = 1, \ldots, T$ **do**
4: Pull arm $m_t = \text{argmax}_{m\in\mathcal{M}} U_{t-1}^H(m)$ and receive reward $\hat{r}_{m_t}^i$.
5: **if** $n_t(m_t) < 8\eta_{m_t} \log(T)$ **then**
6: Update $U_t^H(m_t) = \frac{\sum_{i=1}^{n_t(m_t)} \hat{r}_{m_t}^i}{n_t(m_t)} + \frac{8\sqrt{\eta_{m_t}}\tilde{\tau}_{m_t} \log(T)}{n_t(m_t)}$, $t \leftarrow t+1$;
7: **else**
8: Update $U_t^H(m_t) = \frac{\sum_{i=1}^{n_t(m_t)} \hat{r}_{m_t}^i}{n_t(m_t)} + \sqrt{\frac{8\tilde{\tau}_{m_t}^2 \log(T)}{n_t(m_t)}}$, $t \leftarrow t+1$.
9: **end if**
10: **end for**

Corollary 2 *(Sub-exponential concentration bound for UCB-Hybrid). For any arm $m \in \mathcal{M}$, then one has*

$$\mathbb{P}\left(\left|\mu_m - \frac{\sum_{i=1}^{n_t(m)} \hat{r}_m^i}{n_t(m)}\right| \leq \mathbf{Rad}_t^H(m)\right) \geq 1 - \frac{2}{T^4},$$

where $\mathbf{Rad}_t^H(m) = \frac{8\sqrt{\eta_m}\tilde{\tau}_m \log(T)}{n_t(m)}$ *if* $n_t(m) < 8\eta_m \log(T)$ *and* $\mathbf{Rad}_t^H(m) = \sqrt{\frac{8\tilde{\tau}_m^2 \log(T)}{n_t(m)}}$ *otherwise.*

The proof of Corollary 2 is similar to that of Proposition 1 and hence omitted.

Theorem 3. *The cumulative regret of UCB-Hybrid is bounded by* $O(\sqrt{MT\log(T)} + M\log(T))$.

Table 2
Instance settings.

Index	Mean of the rewards associated with each price (ν_m)	Size	Difference between prices
1	5, 6, 7	Small	Median
2	5, 10, 15	Small	Large
3	5, 6, 7, 8, 9, 10	Median	Median
4	5, 10, 15, 20, 25, 30	Median	Large
5	20, 21, 22, 23, 24, 25	Median	Small
6	20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30	Large	Small
7	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	Large	Median
8	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55	Large	Large
9	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25	Large	Small

The proof of Theorem 3 is omitted because it is similar to that of Theorem 1 since the concentration inequalities are in the same order. Note that because UCB-Hybrid always utilizes a smaller confidence radius versus UCB-Rad, the theoretical regret of UCB-Hybrid is also smaller.

3. Rental bike pricing: an application of sub-exponential MAB

We study a pricing-based revenue management problem (see, e.g., [19]) and provide specific algorithmic steps for three algorithms, UCB-Rad, UCB-Warm, and UCB-Hybrid. Consider a bike rental company that sequentially serves T homogeneous customers with dynamically adjusted unit rental prices $m \in \mathcal{M}$. Under each price m, $\forall m \in \mathcal{M}$, the usage duration of the customer follows an exponential distribution with mean $1/\gamma_m$ and thus the rent paid to the rental company follows an exponential distribution with mean m/γ_m. This assumption of exponential service time has been empirically justified in the literature (see, e.g., [14,19]). Letting $\nu_m = \gamma_m/m$, we have the rent that the company can collect by posting price m, $\forall m \in \mathcal{M}$ follows an exponential distribution with mean $1/\nu_m$. The parameter ν_m, $\forall m \in \mathcal{M}$ is unknown to the company in the beginning, and the company aims to collect as much as rental revenue by dynamically adjusting the offered rental price based on the information of the historical rents. Denote the price for customer $t = 1, \ldots, T$ as m_t. The goal of the company is to maximize the expected cumulative revenue $\sum_{t=1}^{T} 1/\nu_{m_t}$.

3.1. Algorithmic steps: three UCB algorithms

We assume that (i) the bike capacity of the company is sufficiently large (e.g., larger than or equal to T), (ii) the usage time of each customer is known to the company when the service begins, (iii) the company knows a valid lower bound of the exponential rates of all arms based on the information of the bike rental industry, i.e., the company knows $\nu_{\min}$ such that $\nu_{\min} \le \nu_m$, $\forall m \in \mathcal{M}$, and (iv) $T >> 8M\log(T)$. The assumption (i) says that there are enough capacity such that for each customer, she can immediately start renting a bike. The assumption (ii) says that after posting a price, the company can know the usage, as well as the rent, before serving the next customer. The assumption (iii) says that though the performance of each candidate price is unknown, the company has information of the boundary case of all possible performance scenarios. Following the definition of ν_m, we have $1/\nu_{\min} \ge \max_m\{m/\gamma_m\}$, where m is the unit price and $1/\gamma_m$ is the average usage time. Therefore, the company can estimate a valid lower bound for $\nu_{\min}$ by analyzing the maximum revenue it can gain from one customer in the local market. Based on the known lower bound of ν_m, $\forall m \in \mathcal{M}$, we employ a uniform upper bound where $\tilde{\tau}^2 = 4/\nu_{\min}^2 \ge \tilde{\tau}_m^2 \ge \tau_m^2$ for all $m \in \mathcal{M}$. The specific property of exponential distributions results in an arm-independent value of η_m where $\eta_m = b_m^2/\tau_m^2 = 1$ for all $m \in \mathcal{M}$. Thus, by plugging in these values, we conclude that this problem satisfies Assumption 1. The assumption (iv) requires the number of customers is large enough for the company to find the best candidate price through interactions, which satisfies Assumption 2. Therefore, this pricing-based revenue management problem can be formulated as an MAB problem with sub-exponential rewards and solve with proposed algorithms.

Lemma 2. ***(Sub-exponential property.)*** *If a random variable X follows an exponential distribution with mean $1/\nu$, then the random variable X is $(\frac{4}{\nu^2}, \frac{2}{\nu})$-sub-exponential.*

UCB-Rad, UCB-Warm, and UCB-Hybrid. In UCB-Rad, the algorithm utilizes a confidence radius as $\sqrt{32\log(T)/\nu_{\min}^2 n_t(m)} + 16\log(T)/\nu_{\min} n_t(m)$. In UCB-Warm, the algorithm firstly posts each price for serving $8\log(T)$ customers, and then utilizes a confidence radius as $\sqrt{32\log(T)/\nu_{\min}^2 n_t(m)}$. In UCB-Hybrid, the algorithm uses $16\log(T)/\nu_{\min} n_t(m)$ as the confidence radius when the number of served customers under this specific price is less than $8\log(T)$ and otherwise uses $\sqrt{32\log(T)/\nu_{\min}^2 n_t(m)}$.

3.2. Numerical experiments

We present the numerical results of different instances settings, including various size of candidate price sets and distributional parameters, see instance settings in Table 2. For each price set, we consider the total time horizon $T \in \{M \times (400 + 200 \times i) : i = 0, 1, \ldots, 10\}$. For each price set under a specific T, we run 10 replicates of the instances. We present the empirical average regret over T and 2 $\times$ empirical standard deviation in Fig. 1. The cumulative regret of three UCB algorithms shows the same pattern, where the cumulative regret over time converges to zero. Moreover, the regret of UCB-Warm and UCB-Hybrid is smaller than that of UCB-Rad. The number of rounds for choosing the optimal price by UCB-Warm and UCB-Hybrid is significantly larger than that of UCB-Rad and the selected prices of UCB-Warm and UCB-Hybrid are stabilized more quickly than that of UCB-Rad (see Figure 2 and Figure 3 in e-companion). These results show that UCB-Warm and UCB-Hybrid have better numerical performance than UCB-Rad in most of the scenarios.

4. Conclusion

We developed three provably tight learning algorithms, namely, UCB-Rad, UCB-Warm, and UCB-Hybrid, for a general class of MAB problems with sub-exponential rewards. The proposed algorithms can be readily extended to Thompson Sampling (TS) based approaches (with the notion of Bayesian regret) if the reward has a specific conjugate distribution (see [17]). We conducted extensive numerical studies on a revenue management example with exponential reward distributions, to analyze and compare the three algorithms.

There are two plausible future research avenues. First, the algorithmic structures, with a Warm-up Phase or using different radius

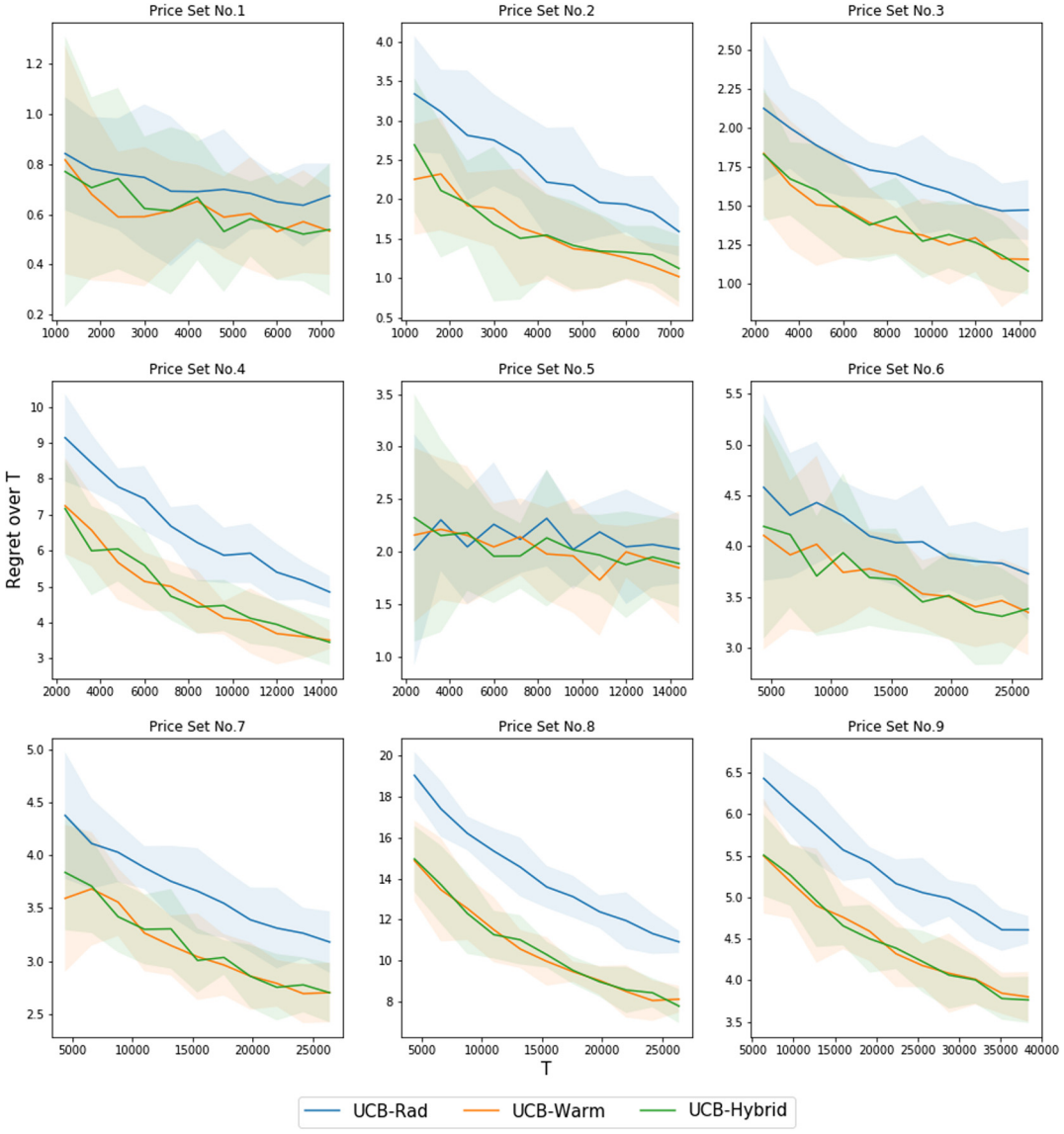

Fig. 1. Regret over T: empirical mean $\pm$ 2 $\times$ empirical std. (For interpretation of the colors in the figure, the reader is referred to the web version of this article.)

along the learning process, can also be extended to other algorithms when dealing with heavy-tailed effects. Second, one may consider exploring other specific heavy-tailed families to propose algorithms with tight regret bounds.

Acknowledgements

The authors are grateful to the department editor Professor Mahesh Nagarajan, the anonymous associate editor, and the two anonymous referees for their detailed comments and suggestions, which have helped significantly improve both the content and the exposition of this paper. This research is partially supported by a University of Michigan Mcubed Award, an Amazon Research Award, U.S. National Science Foundation award #CMMI-1727618, and Department of Energy award #DE-SC0018018.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.orl.2021.08.004.

References

[1] S. Agrawal, V. Avadhanula, V. Goyal, A. Zeevi, A near-optimal exploration-exploitation approach for assortment selection, in: V. Conitzer (Ed.), Proceedings of the 2016 ACM Conference on Economics and Computation, in: EC '16, Association for Computing Machinery, New York, NY, 2016, pp. 599–600.
[2] S. Boucheron, G. Lugosi, P. Massart, Concentration Inequalities: A Nonasymptotic Theory of Independence, Oxford University Press, Oxford, UK, 2013.
[3] S. Bubeck, N. Cesa-Bianchi, Regret analysis of stochastic and nonstochastic multi-armed bandit problems, Found. Trends Mach. Learn. 5 (1) (2012) 1–122.
[4] S. Bubeck, N. Cesa-Bianchi, G. Lugosi, Bandits with heavy tail, IEEE Trans. Inf. Theory 59 (11) (2013) 7711–7717.
[5] X.V. Doan, X. Lei, S. Shen, Pricing of reusable resources under ambiguous distributions of demand and service time with emerging applications, Eur. J. Oper. Res. 282 (1) (2020) 235–251.
[6] A. Dubey, A. Pentland, Thompson sampling on symmetric α-stable bandits, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI'19, AAAI Press, Macao, China, 2019, pp. 5715–5721.
[7] A. Dubey, A. Pentland, Cooperative multi-agent bandits with heavy tails, in: H. Daumé III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, PMLR, Held Virtually, 2020, pp. 2730–2739.
[8] S. Foss, D. Korshunov, S. Zachary, et al., An Introduction to Heavy-Tailed and Subexponential Distributions, Springer, New York, NY, 2011.
[9] H. Jia, C. Shi, S. Shen, Online learning and pricing for service systems with reusable resources, working paper, University of Michigan, Ann Arbor, MI, 2020. Available at SSRN 3755902.

[10] E. Keyvanshokooh, M. Zhalechian, C. Shi, M.P. Van Oyen, P. Kazemian, Contextual learning with online convex optimization: Theory and application to chronic diseases, working Paper, University of Michigan, Ann Arbor, MI, 2019. Available at SSRN 3501316.
[11] N. Korda, E. Kaufmann, R. Munos, Thompson sampling for 1-dimensional exponential family bandits, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 26, Curran Associates, Inc., Lake Tahoe, NV, 2013.
[12] T. Lattimore, C. Szepesvári, Bandit Algorithms, Cambridge University Press, Cambridge, UK, 2020.
[13] M. Ledoux, The Concentration of Measure Phenomenon, American Mathematical Society, Providence, RI, 2001.
[14] Z. Owen, D. Simchi-Levi, Price and assortment optimization for reusable resources, working paper, Massachusetts Institute of Technology, Cambridge, MA, 2018. Available at SSRN 3070625.
[15] P. Rigollet, J.C. Hütter, High Dimensional Statistics, Lecture Notes for Course 18S997, Massachusetts Institute of Technology, Cambridge, MA, 2015.
[16] H. Robbins, Some aspects of the sequential design of experiments, Bull. Am. Math. Soc. 58 (5) (1952) 527–535.
[17] D. Russo, B. Van Roy, Learning to optimize via posterior sampling, Math. Oper. Res. 39 (4) (2014) 1221–1243.
[18] D. Sauré, A. Zeevi, Optimal dynamic assortment planning with demand learning, Manuf. Serv. Oper. Manag. 15 (3) (2013) 387–404.
[19] J. Schuijbroek, R.C. Hampshire, W.J. Van Hoeve, Inventory rebalancing and vehicle routing in bike sharing systems, Eur. J. Oper. Res. 257 (3) (2017) 992–1004.
[20] S. Shalev-Shwartz, et al., Online learning and online convex optimization, Found. Trends Mach. Learn. 4 (2) (2011) 107–194.
[21] A. Slivkins, Introduction to multi-armed bandits, Found. Trends Mach. Learn. 12 (1–2) (2019) 1–286.
[22] S. Vakili, K. Liu, Q. Zhao, Deterministic sequencing of exploration and exploitation for multi-armed bandit problems, IEEE J. Sel. Top. Signal Process. 7 (5) (2013) 759–767.
[23] Y. Xia, T. Qin, W. Ma, N. Yu, T.Y. Liu, Budgeted multi-armed bandits with multiple plays, in: S. Kambhampati (Ed.), Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, AAAI Press, New York, NY, 2016, pp. 2210–2216.
[24] H. Yuan, Q. Luo, C. Shi, Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs, working paper, University of Michigan, Ann Arbor, MI, 2019. Available at SSRN 3329611.
[25] M. Zhalechian, E. Keyvanshokooh, C. Shi, M.P. Van Oyen, Personalized hospital admission control: a contextual learning approach, working paper, University of Michigan, Ann Arbor, MI, 2020. Available at SSRN 3653433.