Globally Informative Thompson Sampling for Structured Bandit Problems with Application to CrowdTranscoding

Xingchi Liu*, Mahsa Derakhshani*, Ziming Zhu[†] and Sangarapillai Lambotharan*

*Signal Processing and Networks Research Group, Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough, LE11 3TU, UK

[†]Toshiba Europe Limited, Bristol Research & Innovation Laboratory, Bristol, BS1 4ND, UK

Email: X.Liu3@lboro.ac.uk; M.Derakhshani@lboro.ac.uk; ziming.zhu@toshiba-bril.com; S.Lambotharan@lboro.ac.uk

Abstract—Multi-armed bandit is a widely-studied model for sequential decision-making problems. The most studied model in the literature is stochastic bandits wherein the reward of each arm follows an independent distribution. However, there is a wide range of applications where the rewards of different alternatives are correlated to some extent. In this paper, a class of structured bandit problems is studied in which rewards of different arms are functions of the same unknown parameter vector. To minimize the cumulative learning regret, we propose a globally-informative Thompson sampling algorithm to learn and leverage the correlation among arms, which can deal with unknown multi-dimensional parameter and non-monotonic reward functions. Our studies demonstrate that the proposed algorithm achieves significant improvement in the learning speed. In particular, the designed algorithm is used to solve an edge transcoder selection problem in crowdsourced live video streaming systems and shows superior performance as compared to the existing schemes.

Index Terms—Multi-armed bandit, Thompson sampling, Structured bandit, Edge computing

I. INTRODUCTION

Multi-armed bandit (MAB) is used to model the decision-making problem, which can be applied into a wide range of applications including cognitive radio networks [1], crowdsourcing systems [2], and online recommendation [3]. In an MAB problem, a machine (which is referred to as an arm) can instantly generate a random reward if it is played and the agent's aim is to maximize the received cumulative reward by sequentially pulling a number of arms. The challenge of solving the MAB problem is how to balance the exploration-exploitation (EE) dilemma. If the agent only exploits the learned knowledge to select a relatively good arm, the opportunity of gaining higher cumulative reward is missed. However, if the agent keeps exploring every arm, obviously the arms with lower rewards can be played frequently.

In the classical MAB problem, the reward of an arm is assumed to follow an independent probability distribution. Therefore, the observations from one arm cannot reveal any

information of other arms. This type of MAB can be categorized as the *non-informative bandit* and many algorithms are designed to solve it including upper confidence bound (UCB) [4] and Thompson sampling (TS) [5]. However, this assumption does not hold in many applications. For example, in a news recommendation system, a company needs to decide which news to push to the users so that the click-through rate can be maximized. To identify the most appealing news for different users [6], the correlation of user preferences shall be considered since users with similar ages or occupations may have similar preferences.

Considering the correlation among different arms in decision-making problems, the structured bandit model [7] is introduced. One model of structured bandits is contextual MAB in which the correlation of arms is modelled under assumption that the reward distributions are conditioned on the context information and hence the arms with similar context information can generate close rewards. In [8], a contextual MAB was modelled which assumes the expected reward of an arm is a linear function of contextual information and unknown parameters. In [9], a TS-based algorithm was proposed to solve the contextual MAB problem with the same reward function assumption. Moreover, in [10], instead of explicitly assuming the mapping from the context to the reward to be a linear function, Gaussian process regression was used to model the relationship between the context and the reward.

Contextual MAB relies on knowledge of context information of the environment. However, this information might be unavailable in some problems due to the privacy concerns [11]. In this paper, we study a different model of structured bandit problems assuming the context information is hidden and the reward mappings are functions of one or multiple unknown parameters shared by all the arms. In [12], a structured bandit problem is modelled which assumes each group of arms shares the same unknown parameter. In [13] and [14], the unknown parameter is estimated using the knowledge of the reward functions and empirical mean rewards. However, the proposed algorithms in [13] and [14] are limited to considering monotonic expected reward functions.

This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/R006385/1 and in part by the Royal Academy of Engineering under the Leverhulme Trust Research Fellowship scheme under Grant Derakhshani-LTRF1920\16\67.

In this paper, we study a model which removes the restrictions on the properties of the expected reward function and the number of parameters in state-of-the-art works. In order to solve the structured bandit problem, we first build a confidence set which is a set of parameters whose expected reward is close to the empirical mean rewards of all arms, and then a novel technique is designed to estimate the true value of the unknown parameter based on the established confidence set. In addition, a novel TS-based algorithm is proposed to handle the EE dilemma. Simulation results demonstrate that the proposed globally-informative Thompson sampling (GI-TS) algorithm can solve the structured bandit problem with a noteworthy improvement of the learning regret compared with the existing benchmarks.

We apply the proposed GI-TS algorithm to tackle the transcoder selection problem for crowdsourced live streaming (CLS) services as a case study. Recently, the edge transcoding has been introduced for CLS to utilize the abundant computational resources at the user end for quality of experience enhancement [15]. However, the optimal transcoder selection for seamless video transcoding remains to be a challenge due to performance uncertainties of devices. Since the edge devices with similar specifications and attributes shall offer similar transcoding performance, the transcoder selection problem can be modelled as a structured bandit problem. Numerical results confirm the effectiveness of the proposed GI-TS approach for the case study of transcoder selection for crowdtranscoding.

The rest of this paper is organized as follows. Section II describes the system model. The detailed GI-TS algorithm is presented in Section III. The simulation setup and results are provided in Section IV, followed by conclusions in Section V.

II. SYSTEM MODEL

Consider a structured bandit problem with K arms, i.e., arm $k \in \mathcal{K} = \{1, 2, \dots, K\}$. At each round, one of the K alternatives must be selected by an agent to play, and the reward of the played arm can be observed. For example, in the edge transcoder selection problem, at each round, a transcoder need to be selected to do transcoding and each edge transcoder can be treated as an arm. Define the arm played at round t as k_t where $k_t \in \mathcal{K}$, the observed reward of playing arm k_t can be defined as $R_{k_t}(t)$, which is a random variable following an unknown probability distribution. As a structured bandit problem, the expectation of $R_{k_t}(t)$ can be defined as

$$\mu_k(\theta^*) = \mathbb{E}[R_{k_t}(t) \mid \theta^*], \quad (1)$$

where $\theta^* = [\theta_1, \theta_2, \dots, \theta_N]$ is a fixed but unknown parameter vector, which belongs to a parameter set Θ and is shared by all the arms. In addition, the expected reward function of each arm which is known can be of any form. The observed rewards are assumed to be sub-Gaussian with variance σ^2 and the variance is assumed to be known to the agent. This assumption is common in MAB literature ([7], [15]) and enables us to utilize Hoeffding's inequality which is the foundation of the proposed algorithm (discussed in the next section).

The goal of the agent is to select arm k_t at each round t to maximize the total reward in any T rounds. Assume there is an oracle policy that knows the true value of θ^* and always selects the optimal arm which is defined as $k^* = \operatorname{argmax}_{k \in \mathcal{K}} \mu_k(\theta^*)$. The cumulative reward following this policy can be represented as $\sum_{t=1}^T \mu_{k^*}(\theta^*)$. If the agent misses choosing the optimal arm at round t , a reward loss is incurred, which is defined as $\mu_{k^*}(\theta^*) - \mu_{k_t}(\theta^*)$. Therefore, the expected cumulative learning regret can be defined as

$$\operatorname{Reg}(T) := \mathbb{E} \left[\sum_{t=1}^T \mu_{k^*}(\theta^*) - \mu_{k_t}(\theta^*) \right]. \quad (2)$$

It should be noticed that minimizing the cumulative learning regret equals to maximizing the cumulative reward.

III. PROPOSED ALGORITHM

In the structured bandit, the agent knows the forms of reward functions. Therefore, if the unknown parameter vector can be accurately estimated, the optimal arm can be determined surely. Moreover, since all arms share θ^* , no matter which arm is played, the observed reward can be used to refine the estimation of θ^* . In this regard, enlightened by [7], a TS-based algorithm called GI-TS is designed to utilize the feature of the setting to solve the structured bandit problem.

A. Confidence Set Construction

Define the empirical mean reward of arm k at round t as $\bar{\mu}_k(t)$ and define the number of times that arm k has been played till round t as $n_k(t)$. GI-TS starts to estimate the unknown parameter vector by constructing a confidence set based on the following fact:

Fact 1: (Hoeffding inequality) Let Z be a random variable which follows a σ^2 sub-Gaussian distribution with mean $\mu(\theta^*)$, then

$$\Pr(|\mu(\theta^*) - \bar{\mu}| \geq \delta) \leq 2 \exp \left(-\frac{\delta^2 n}{2\sigma^2} \right), \quad (3)$$

where $\bar{\mu}$ represents the empirical mean of Z , and n denotes the number of samples. When $\delta = \sqrt{\frac{2\alpha\sigma^2 \log t}{n}}$, Hoeffding inequality can be bounded as

$$\Pr \left(|\mu(\theta^*) - \bar{\mu}| \geq \sqrt{\frac{2\alpha\sigma^2 \log t}{n}} \right) \leq 2t^{-\alpha}, \quad (4)$$

which implies that the gap between the empirical mean and the real mean being larger than a threshold is increasing unlikely to happen with t increasing. Here α is introduced to control the variation rate of the bound.

According to Fact 1, for each arm $k \in \mathcal{K}$, we choose a set of parameters whose mean rewards are close to the empirical mean reward. This set of parameters is defined as $\tilde{\Theta}_k(t)$, which can be written as

$$\tilde{\Theta}_k(t) = \left\{ \theta : |\mu_k(\theta) - \bar{\mu}_k(t)| < \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}} \right\}. \quad (5)$$

Since θ^* is shared by all the arms in their reward functions, the constructed $\tilde{\Theta}_k(t)$ of each arm are utilized to estimate θ^* more precisely. We define a set of parameters satisfying the specified condition in (5) for every arm as the confidence set. The confidence set $\tilde{\Theta}_t$ is calculated as the intersection of $\tilde{\Theta}_k(t)$'s of all the arms, which can be written as

$$\tilde{\Theta}_t = \bigcap_{k \in \mathcal{K}} \tilde{\Theta}_k(t). \quad (6)$$

An example of confidence set construction is illustrated in Fig. 1.

B. Confidence Set Analysis

In this section, we prove that the parameter vector θ^* falls into the constructed confidence set with a high probability.

Theorem 1: The probability that θ^* belongs to the confidence set is bounded as

$$\Pr(\theta^* \in \tilde{\Theta}_t) \geq 1 - 2Kt^{1-\alpha}, \quad (7)$$

Proof: Based on (5), define $\tilde{\Theta}_k^c(t)$ as the complement set of $\tilde{\Theta}_k(t)$. The probability that θ^* does not belong to the confidence set can be written as

$$\Pr(\theta^* \notin \tilde{\Theta}_t) = \Pr\left(\bigcup_{k \in \mathcal{K}} \theta^* \in \tilde{\Theta}_k^c(t)\right). \quad (8)$$

Following the union bound, we have

$$\begin{aligned} & \Pr\left(\bigcup_{k \in \mathcal{K}} \theta^* \in \tilde{\Theta}_k^c(t)\right) \\ & \leq \sum_{k=1}^K \Pr\left(|\mu_k(\theta^*) - \bar{\mu}_k(t)| \geq \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}}\right), \end{aligned} \quad (9)$$

$$\leq \sum_{k=1}^K \sum_{m=1}^t \Pr\left(|\mu_k(\theta^*) - \bar{\mu}_{k,m}(t)| \geq \sqrt{\frac{2\alpha\sigma^2 \log t}{m}}\right). \quad (10)$$

Applying Fact 1, (10) can be upper bounded as

$$\Pr(\theta^* \notin \tilde{\Theta}_t) \leq K \sum_{m=1}^t 2t^{-\alpha} = 2Kt^{1-\alpha}, \quad (11)$$

which completes the proof.

Theorem 1 indicates that, as t increases, the probability of θ^* falling into the confidence set $\tilde{\Theta}_t$ superlinearly converges to 1 with $\alpha > 2$.

C. Unknown Parameter Estimation

After establishing the confidence set, we need to estimate the unknown parameter vector θ^* . Since the confidence set maintains a group of parameters whose corresponding expected rewards are close to the empirical mean rewards of each arm, the confidence set can be used to estimate θ^* . Because the more times an agent plays, the higher is the probability that the true value of θ^* falls into this set as presented in (7).

To utilize the knowledge of the reward functions, we do not simply estimate θ^* by directly solving the equation $\bar{\mu}_k(t) = \mu_k(\theta)$, which can incur high computation complexity

Fig. 1: Confidence set construction: An illustration of how the confidence set is built. The marked curves represent the reward functions of the two arms. For each arm, a set of parameters can be found based on (5), so the gap between $\bar{\mu}_k$ and $\mu_k(\theta)$ is smaller than $\sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}}$. By taking the intersection of the two sets, the confidence set is established. In this case, we can find that θ^* falls into the confidence set.

and hardly guarantee a sufficiently accurate solution. Besides, the reward function can be non-monotonic and hence directly solving the equation may lead to multiple solutions.

In GI-TS, we define a parameter in the confidence set as θ_n , the estimated parameter $\hat{\theta}_t$ can be calculated as

$$\hat{\theta}_t = \sum_{\theta_n \in \tilde{\Theta}_t} \theta_n w_{\theta_n}(t), \quad (12)$$

where $w_{\theta_n}(t)$ represents the weight of θ_n at round t . The weight can be calculate as

$$w_{\theta_n}(t) = \frac{G_{\theta_n}(t)^{-1}}{\sum_{\theta_n \in \tilde{\Theta}_t} G_{\theta_n}(t)^{-1}}, \quad (13)$$

where $G_{\theta_n}(t) = \sum_{k \in \mathcal{K}} |\mu_k(\theta_n) - \bar{\mu}_k(t)|$. To estimate θ^* , θ_n is uniformly selected within the confidence set. The logic behind is that, instead of calculating the mean value of the confidence set, the parameter whose expected reward is closer to the empirical mean reward has a greater impact on the estimated unknown parameter vector. Therefore, the estimation of θ^* is accomplished by calculating the weighted summation of the parameters inside the confidence set.

D. Arm Selection

After calculating $\hat{\theta}_t$, we need to utilize it for the arm selection to minimize the learning regret. However, we still need to balance the tradeoff between exploring the reward of each arm and exploiting the learned knowledge to play the best arm, because the empirical mean reward can be inaccurate especially at the early stage. Therefore, a TS-based scheme namely GI-TS is designed to handle the EE dilemma.

Consider (5), (6), and (12), if $\theta^* \in \tilde{\Theta}_t$, given $\hat{\theta}_t \in \tilde{\Theta}_t$, for each arm, it can be inferred that $|\mu_k(\theta^*) - \mu_k(\hat{\theta}_t)| \leq \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}}$. Since the confidence set is the intersection among $\tilde{\Theta}_k(t)$ of different arms, the worst case occurs if $n_k(t) = t/K$, which leads to the possibility of having the largest gaps

between $\mu_k(\theta^*)$ and $\mu_k(\hat{\theta}_t)$. Therefore, in the worst case, the gap between $\mu_k(\hat{\theta}_t)$ and $\mu_k(\theta^*)$ can be upper bounded as

$$|\mu_k(\theta^*) - \mu_k(\hat{\theta}_t)| \leq \sqrt{\frac{2K\alpha\sigma^2 \log t}{t}}, \quad (14)$$

which means the estimated reward $\mu_k(\hat{\theta}_t)$ will converge to its real reward as t increases.

In the absence of prior and likelihood knowledge, we assume the posterior distribution of the arm k 's reward to be Gaussian distribution. According to the previous discussion, we choose the estimated reward $\mu_k(\hat{\theta}_t)$ to be the mean of the distribution since it can converge to the real mean as more samples are observed. The variance of the posterior distribution is set to be $\frac{\sigma^2}{t}$ given the fact that any play is informative for all arms and hence can help to refine the estimation of θ^* . Therefore, we will generate an independent sample $\tilde{R}_k(t)$ from the distribution $\mathcal{N}(\mu_k(\hat{\theta}_t), \frac{\sigma^2}{t})$ for each arm. The agent will then play the arm with the highest $\tilde{R}_k(t)$. The pseudocode of GI-TS is presented in Algorithm 1.

Algorithm 1 GI-TS

Require: Reward functions $\{\mu_1, \mu_2, \dots, \mu_k\}$, $n_k \leftarrow 0$, t_h

for $t \leftarrow 1$ **to** T **do**

if $t < t_h$ **then**

 Randomly play an arm $k_t \in \mathcal{K}$, observe $R_{k_t}(t)$

$n_{k_t}(t) = n_{k_t}(t-1) + 1$

$\bar{\mu}_{k_t}(t) = \frac{\bar{\mu}_{k_t}(t-1) \cdot n_{k_t}(t-1) + R_{k_t}(t)}{n_{k_t}(t)}$

else

 Construct confidence set $\tilde{\Theta}_t$ based on (5) and (6)

 Calculate $\hat{\theta}_t$ based on (12), and (13)

for $k \in \mathcal{K}$ **do**

 Sample $\tilde{R}_k(t) \sim \mathcal{N}(\mu_k(\hat{\theta}_t), \frac{\sigma^2}{t})$

end for

 Play arm $k_t = \operatorname{argmax}_{k \in \mathcal{K}} \tilde{R}_k(t)$, observe $R_{k_t}(t)$

$n_{k_t}(t) = n_{k_t}(t-1) + 1$

$\bar{\mu}_{k_t}(t) = \frac{\bar{\mu}_{k_t}(t-1) \cdot n_{k_t}(t-1) + R_{k_t}(t)}{n_{k_t}(t)}$

end if

end for

IV. SIMULATION RESULTS

A. Simulation Setup

In this section, we evaluate the performance of the GI-TS. We first generate a range of different reward functions to demonstrate the robustness and performance of the proposed algorithm, and three scenarios are set by changing the number of arms and the length of the unknown parameter vector. The setting is depicted in Table I. In addition, we apply GI-TS to solve a transcoder selection problem in CLS systems which can be modelled as a structured bandit problem.

In the simulations, α is set to be 3. We choose the variance of each arm in all simulations as $\sigma^2 = 4$, and R_k is drawn from $\mathcal{N}(\mu_k(\theta^*), 4)$. t_h is set to be 10, which is the number of initialization rounds for arm selection. In addition, each of

	Number of arms	Length of θ^*
1 st scenario	3	1
2 nd scenario	10	1
3 rd scenario	3	2

TABLE I: Simulation setting

Fig. 2: Reward function versus θ

simulations runs for 10000 rounds and the learning regret is presented as the average of 100 independent experiments.

B. Benchmarks

Several existing algorithms are simulated as benchmarks. First, the classical scheme UCB and TS are simulated. In addition, we also simulate the UCBc and TSc schemes which are specially designed for the structured bandit problem [15]. In these two schemes, first a confidence set of parameters is built with the knowledge of reward functions, and some of the arms are identified to be the competitive arms with the help of the confidence set. After that, either UCB or TS is applied to select only one of the competitive arms.

C. Numerical Results

1) *1st scenario:* In Fig. 2, the reward functions of arms versus θ are presented. We can find that arm 1 is optimal when $\theta^* \in [2.5, 5]$, arm 2 is optimal when $\theta^* \in [0, 1]$, and arm 3 is optimal when $\theta^* \in [1, 2.5]$. With this setting, we test the proposed algorithm and the benchmarks with varying θ^* . In Fig. 3, it is apparent that GI-TS and TSc outperform classical schemes by considering the correlation among the arms and hence the learning regret is greatly reduced. This is because no matter which arm is played, the observed reward can help to find θ^* by refining the confidence set each round and to improve the learning speed. In addition, our proposed GI-TS algorithm reaches the lowest regret compared with all the benchmarks. This is because GI-TS directly estimates the value of θ^* to find the optimal arm with the help of reward function information.

2) *2nd scenario:* In this section, we further test the proposed algorithm in a more complex setting with 10 arms whose reward functions are similar to the functions in the 1st scenario.

Fig. 3: Cumulative regret versus θ^*

Fig. 6: Cumulative regret when $\theta^* = [0.8, 0.4]$

Fig. 4: Cumulative regret when $\theta^* = 3.8$

Fig. 7: Cumulative regret when $\theta^* = [-0.9, -0.2]$

Fig. 5: Cumulative regret when $\theta^* = 4.4$

According to Figs. 4 and 5, we can find that GI-TS outperforms all the benchmarks given different θ^* , and the learning regret of GI-TS is only around one-fifth of the regret incurred by TS_c. Compared with the 1st scenario, GI-TS scales well when more arms are involved. This is because, instead of using the number of plays to estimate the variance of the reward distribution of an arm, which is used in TS_c, the total number of rounds t is utilized to estimate the variance of the reward distribution. This is based on the fact that the confidence set is updated in each round no matter which arm is played. This novel method improves the convergence rate and helps to achieve a competitively small learning regret.

3) *3rd scenario*: In this section, we test the proposed algorithm when the unknown parameter θ^* is a 2-dimensional vector. Noted that this setting cannot be considered when linear or generalized linear models are used. The reward functions of the three arms are $\mu_1(\theta) = \theta_1 + \theta_2$, $\mu_2(\theta) = \theta_1^2 - \theta_2 + 0.5$, and $\mu_3(\theta) = 0.5 \max\{\theta_1, \theta_2\} + 0.1$.

In Figs. 6 and 7, we choose different pairs of θ^* . The result demonstrates that GI-TS scales well when one more parameter is added as compared to the results in the 1st scenario, and GI-TS still outperforms all the benchmarks by reaching smaller learning regrets with a higher convergence speed.

D. Case Study: Transcoder Selection Problem

In this section, we study a decision-making problem of edge transcoder selection for CLS services, which can be modelled as a structured bandit problem.

The revolution of the Internet driven by the powerful mobile devices and social networks has greatly enriched the sources of live video and the CLS has emerged as a new type of video services that both serve tremendous viewers and receive videos from various sources. To provide viewers with adaptive bit rate videos, it is necessary to transcode the heterogeneous original live videos into industrial standard representations. To meet the massive transcoding demands, the edge computing is proposed to assist CLS by incentivizing end viewers' devices to be candidate transcoders [16]. However, as discussed in Section I, the candidate devices can be offline during transcoding which

Fig. 8: Cumulative regret when $\theta^* = 22.5$

can severely deteriorate the transcoding performance, and hence more stable devices should be selected as transcoders.

According to [17], the distribution of candidate devices' online durations follows the Pareto distribution, which implies that the longer a device is online, the more likely this device will continue to be online. The cumulative distribution function of a Pareto random variable X with parameters β and x_m is

$$F_X(x) = \begin{cases} 1 - (x_m/x)^\beta & x \geq x_m \\ 0 & x < x_m \end{cases} \quad (15)$$

For simplicity, assume $x \geq x_m$ always holds, define $S_f(x_m) = 1 - (x_m/x_f)^{\beta_f}$ as the transcoding stability of edge device f to show the probability that the transcoder will be online after the selection. In addition, x_f and β_f are predefined parameters which vary due to the heterogeneity of transcoders.

Based on above settings, each edge device can be regarded as an arm, and the expected reward function of arm f can be written as $\mu_f = S_f(x_m)$. Moreover, the form of each reward function is known with $\theta^* = x_m$, which is shared by all the arms. Therefore, this transcoder selection problem can be modelled as a structured bandit problem and can be solved by the proposed GI-TS algorithm. More specifically, in each round, the transcoder k with the highest sampled reward $\hat{R}_k(t)$ is selected for live video transcoding.

In Fig. 8, the transcoder selection problem is simulated assuming there are 10 transcoders. The simulation runs for 20000 rounds and the learning regret is calculated as the average of 100 independent experiments. To build a challenging scenario that the expected rewards of arms are close to each other, we set $\theta^* = 22.5$. The result demonstrates that the proposed GI-TS algorithm outperforms all the benchmarks and achieves the lowest learning regret.

V. CONCLUSIONS

In this paper, a new model of MAB called structured bandit is studied which assumes the arms are correlated by sharing the same unknown parameter vector in their reward functions. In order to maximize the cumulative reward in this new bandit problem, we design an algorithm to accurately estimate the true value of the unknown parameter with the

knowledge of the reward functions and the historical reward observations. Simulation results demonstrate that the proposed GI-TS algorithm outperforms all the benchmarks by reaching a low cumulative learning regret. The GI-TS algorithm is robust and scales well when more arms are added or multi-dimensional unknown parameter is considered. As a future work, we envision the extension of the proposed structured bandit model to consider the risk awareness [18] in the decision making aiming to identify the most stable arm with relatively higher mean reward and lower reward variation.

REFERENCES

- [1] Z. Tian, J. Wang, J. Wang, and J. Song, "Distributed noma-based multi-armed bandit approach for channel access in cognitive radio networks," *IEEE Wireless Communications Letters*, vol. 8, no. 4, pp. 1112–1115, Aug 2019.
- [2] L. Tran-Thanh, S. Stein, A. Rogers, and N. R. Jennings, "Efficient crowdsourcing of unknown experts using bounded multi-armed bandits," *Artificial Intelligence*, vol. 214, pp. 89–111, 2014.
- [3] Q. Wang, C. Zeng, W. Zhou, T. Li, S. S. Iyengar, L. Shwartz, and G. Y. Grabarnik, "Online interactive collaborative filtering using multi-armed bandit with dependent arms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 8, pp. 1569–1580, Aug 2019.
- [4] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.
- [5] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.
- [6] X. Wang, S. C. Hoi, C. Liu, and M. Ester, "Interactive social recommendation," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Singapore, 2017.
- [7] T. Lattimore and R. Munos, "Bounded regret for finite-armed structured bandits," in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014.
- [8] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th International Conference on World Wide Web*, 2010.
- [9] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, no. 3, 2013, pp. 127–135.
- [10] N. Srinivas, A. Krause, S. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, Haifa, 2010.
- [11] S. Gupta, G. Joshi, and O. Yağan, "Correlated multi-armed bandits with a latent random source," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.
- [12] Z. Wang, R. Zhou, and C. Shen, "Regional multi-armed bandits with partial informativeness," *IEEE Transactions on Signal Processing*, vol. 66, no. 21, pp. 5705–5717, Nov 2018.
- [13] O. Atan, C. Tekin, and M. van der Schaar, "Global bandits," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 12, pp. 5798–5811, 2018.
- [14] C. Shen, R. Zhou, C. Tekin, and M. van der Schaar, "Generalized global bandit and its application in cellular coverage optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 218–232, Feb 2018.
- [15] S. Gupta, S. Chaudhari, S. Mukherjee, G. Joshi, and O. Yağan, "A unified approach to translate classical bandit algorithms to the structured bandit setting," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 840–853, 2020.
- [16] Q. He, C. Zhang, X. Ma, and J. Liu, "Fog-based transcoding for crowdsourced video livecast," *IEEE Communications Magazine*, vol. 55, no. 4, pp. 28–33, 2017.
- [17] Q. He, C. Zhang, and J. Liu, "Crowdtranscoding: Online video transcoding with massive viewers," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1365–1375, 2017.
- [18] X. Liu, M. Derakhshani, S. Lambotaran, and M. van der Schaar, "Risk-aware multi-armed bandits with refined upper confidence bounds," *IEEE Signal Processing Letters*, vol. 28, pp. 269–273, 2021.