A UNIFIED APPROACH TO TRANSLATE CLASSICAL BANDIT ALGORITHMS TO STRUCTURED BANDITS

Samarth Gupta[*] *, Shreyas Chaudhari*[*]*, Subhojyoti Mukherjee*[+] *, Gauri Joshi*[*] *and Osman Yağan*[*]

ABSTRACT

We consider a finite-armed structured bandit problem in which mean rewards of different arms are known functions of a common hidden parameter θ^*. This problem setting subsumes several previously studied frameworks that assume linear or invertible reward functions. We propose a novel approach to gradually estimate the hidden θ^* and use the estimate together with the mean reward functions to substantially reduce exploration of sub-optimal arms. This approach enables us to fundamentally generalize any classic bandit algorithm including UCB and Thompson Sampling to the structured bandit setting. We prove via regret analysis that our proposed UCB-C and TS-C algorithms (structured bandit versions of UCB and Thompson Sampling, respectively) pull only a *subset* of the sub-optimal arms $\mathrm{O}(\log T)$ times while the other sub-optimal arms (referred to as *non-competitive* arms) are pulled $\mathrm{O}(1)$ times. As a result, in cases where all sub-optimal arms are non-competitive, which can happen in many practical scenarios, the proposed algorithms achieve bounded regret.

Index Terms— Multi-Armed Bandits, Sequential decision making, Online learning, Statistical learning, Regret bounds

A full version of this paper with full proofs and more details is accessible at:
https://ieeexplore.ieee.org/abstract/document/9276444 [1].

1. INTRODUCTION

The Multi-armed bandit problem [2] (MAB) falls under the umbrella of sequential decision-making problems. It has numerous applications in medical diagnosis, system testing, scheduling in computing systems, and web optimization, to name a few. In the classical K-armed bandit formulation, a player is presented with K arms. At each time step t, she decides to pull an arm $k \in \mathcal{K}$ and receives a random reward R_k with unknown mean μ_k. The goal of the player is to maximize their cumulative reward (or equivalently, minimize cumulative regret). In order to do so, the player must strike a balance between estimating the unknown rewards by pulling all the arms (exploration) and always pulling the current best arm (exploitation). The seminal work [2] proposed the UCB (upper confidence bound) algorithm that balances the exploration-exploitation tradeoff in the MAB problem. Subsequently, several algorithms such as UCB1 [3], Thompson Sampling (TS) [4] and KL-UCB [5] were proposed and analyzed for the classical MAB setting. In this work, we study a setting in which rewards corresponding to different arms are related to each other through a hidden parameter θ; see Section 2 and Figure 2. We develop an approach that leverages this reward structure and reduces exploration in UCB, TS, KL-UCB, etc., and consequently achieves a significantly lower cumulative regret.

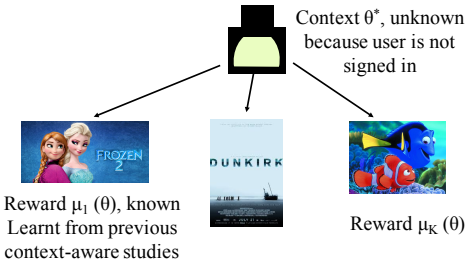

Fig. 1: *Movie recommendation application of the structured bandit framework studied in this paper. The context θ (for example, the age of the user) is unknown because the user is not signed in. But if a user gives a high rating the first movie (Frozen) one could infer that the age θ is small, which in turn implies that the user will give a high rating to the third movie.*

There are many practical applications where multi-armed bandit algorithms can be useful. For instance, let us consider the example of ad selection, where a company needs to decide which version of the ad it needs to display to the user. It has different versions for the same ad and depending on which ad is displayed the user engagement (in terms of click probability and time spent looking at the ad) is affected. In order to maximize user engagement, the company needs to identify the most appealing ad for the user in an online manner and this is where multi-armed bandit algorithms can be helpful. However, classical MAB algorithms are typically based on the (implicit) assumption that rewards from different arms are independent of each other. This assumption is unlikely to hold in reality since the user choices corresponding to different versions of an ad are likely to be related to each other; e.g., the choices corresponding to different versions may depend on the age/occupation/income of the user.

Contextual bandits [6] consider that the player also observes the context feature (e.g., their age, occupation, income

[*]Dept. of ECE, Carnegie Mellon University
[+]Dept. of ECE, University of Wisconsin-Madison

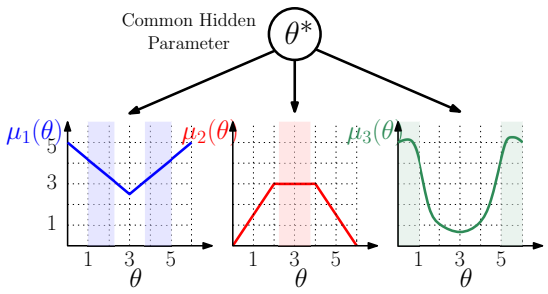

Fig. 2: *Structured bandit setup: mean rewards of different arms share a common hidden parameter. This example illustrates a 3-armed bandit problem with shaded regions indicating the values of θ for which the particular arm is optimal.*

information) of the user to whom ad is displayed. By trying to learn a mapping from feature information to the most appealing arm, contextual bandit algorithms prove useful for the application of targeted advertising. However in several use cases, observing contextual features leads to privacy concerns and contextual features may not be visible for the users who are signed in anonymously to protect their identity. In other cases, contextual information may be costly to obtain. The structured bandit setting considered in this paper can be viewed as a hidden context problem, where the objective is to do targeted advertising for a user without observing their features, as illustrated in Figure 1. Apart from ad selection, the structured bandit model also has applications in dynamic pricing (described in [7]), cellular coverage optimization (by [8]), drug dosage optimization (discussed in [9]).

2. PROBLEM FORMULATION

Consider a multi-armed bandit setting with K arms $\mathcal{K} = \{1, 2, \ldots, K\}$. At each round t, the player pulls arm $k_t \in \mathcal{K}$ and observes the reward R_{k_t}. The reward R_{k_t} is a random variable with mean $\mu_{k_t}(\theta) = \mathbb{E}[R_{k_t}|\theta, k_t]$, where θ is a *fixed, but unknown parameter* which lies in a known set Θ, as illustrated in Figure 2.

We denote the unknown true value of θ by θ^*. There are no restrictions on the set Θ; it can be countable or uncountable. Although we focus on scalar θ in this paper for brevity, the proposed algorithms and regret analysis can be generalized to the case where we have a hidden parameter *vector* $\vec{\theta} = [\theta_1, \theta_2, \ldots \theta_m]$. The mean reward functions $\mu_k(\theta) = \mathbb{E}[R_k|\theta]$ for $k \in \mathcal{K}$ can be arbitrary (no linearity or continuity constraints) functions of θ. While $\mu_k(\theta)$ are known to the player, the conditional distribution of rewards, i.e., $p(R_k|\theta)$ is not known (which are assumed to be known in the work of [10,11]). Instead, we only assume that the rewards R_k are sub-Gaussian with variance proxy σ^2, i.e., $\mathbb{E}[\exp(s(R_k - \mathbb{E}[R_k]))] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right)$ $\forall s \in \mathbb{R}$, and σ is known to the player. Both assumptions are common in the MAB literature [12].

The objective of the player is to select arm k_t in round t so as to maximize her cumulative reward $\sum_{t=1}^{T} R_{k_t}$ after

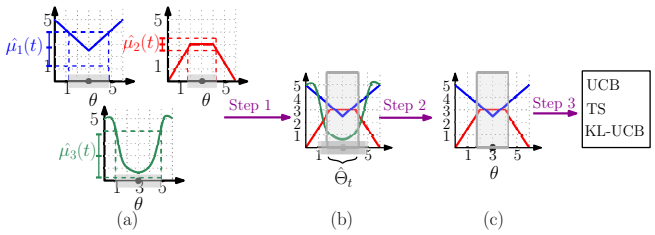

Fig. 3: Illustration of the three steps of our proposed algorithm.

T rounds. If the player had known the hidden θ^*, then she would always pull arm $k^* = \arg\max_{k \in \mathcal{K}} \mu_k(\theta^*)$ that yields the highest mean reward at $\theta = \theta^*$. We refer to this arm k^*as the optimal arm. Maximizing the cumulative reward is equivalent to minimizing the *cumulative regret*, defined as,

$$Reg(T) \triangleq \sum_{t=1}^{T} (\mu_{k^*}(\theta^*) - \mu_{k_t}(\theta^*)) = \sum_{k \neq k^*} n_k(T)\Delta_k, \quad (1)$$

where $n_k(T)$ is the number of times arm k is pulled in T slots and $\Delta_k \triangleq \mu_{k^*}(\theta^*) - \mu_k(\theta^*)$ is the *sub-optimality gap* of arm k. cumulative regret is in turn equivalent to minimizing $n_k(T)$, the number of times each sub-optimal arm $k \neq k^*$ is pulled.

Remark 1 (Connection to Classic Multi-Armed Bandits)**.** *The classic multi-armed bandit setting, which does not explicitly consider a* structure *among the mean rewards of different arms, is a special case of the proposed structured bandit framework. It corresponds to having a hidden parameter vector $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ and the mean reward of each arm being $\mu_k = \theta_k$. In fact, our proposed algorithm described in Section 3 reduces to standard UCB or Thompson sampling ([2, 3]) in this special case.*

The proposed structured bandit subsumes several previously considered models as we make no assumption on the functions $\mu_k(\theta)$. In the scenario, where $\mu_k(\theta)$ is a linear function, our framework covers the setup of [13, 14]. When $\mu_k(\theta)$ is known to be invertible and Hölder continuous, our framework subsumes the model of Global and Regional bandits [7,9]. For a situation where $\mu_k(\theta) = g(x_k^\intercal \theta)$ with known g and x_k for each arm k, our framework captures the generalized linear bandit [15] and linear bandit [16] setup. See the full paper for a detailed comparison with these works.

3. PROPOSED ALGORITHM: ALGORITHM-C

We now propose the following three-step algorithm called ALGORITHM-C. At each round $t+1$, the algorithm performs the following three steps:

Step 1: Constructing a confidence set, $\hat{\Theta}_t$. From the samples observed till round t, we define the confidence set as:

$$\hat{\Theta}_t = \left\{\theta : \forall k \in \mathcal{K}, \quad |\mu_k(\theta) - \hat{\mu}_k(t)| < \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}}\right\}. \quad (2)$$

Here, $\hat{\mu}_k(t)$ is the empirical mean of rewards obtained from the $n_k(t)$ pulls of arm k. For each arm k, we construct a confidence set of θ such that the true mean $\mu_k(\theta)$ is within an interval of size $\sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}}$ from $\hat{\mu}_k(t)$. This is illustrated by the error bars along the y-axis in Figure 3(a), with the corresponding confidence sets shown in grey for each arm. Taking the intersection of these K confidence sets gives us $\hat{\Theta}_t$, wherein θ lies with high probability, as shown in Figure 3(b).

Step 2: Finding the set $\mathcal{C}_t$ of $\hat{\Theta}_t$-Competitive arms. We let $\mathcal{C}_t$ denote the set of $\hat{\Theta}_t$-Competitive arms at round t, where,

Definition 1 ($\hat{\Theta}_t$-Competitive arm)**.** *An arm k is said to be $\hat{\Theta}_t$-Competitive if its mean reward is the highest among all arms for some $\theta \in \hat{\Theta}_t$; i.e., $\exists \theta \in \hat{\Theta}_t$ such that $\mu_k(\theta) = \max_{\ell \in \mathcal{K}} \mu_\ell(\theta)$.*

Definition 2 ($\hat{\Theta}_t$-Non-competitive arm)**.** *An arm k is said to be $\hat{\Theta}_t$-Non-competitive if it is* not *$\hat{\Theta}_t$-Competitive; i.e., if $\mu_k(\theta) < \max_{\ell \in \mathcal{K}} \mu_\ell(\theta)$ for all $\theta \in \hat{\Theta}_t$.*

If an arm is $\hat{\Theta}_t$-Non-competitive, then it cannot be optimal if the true parameter lies inside the confidence set $\hat{\Theta}_t$. These $\hat{\Theta}_t$-Non-competitive arms are not considered in Step 3 of the algorithm for round $t+1$. However, these arms can be $\hat{\Theta}_t$-Competitive in subsequent rounds. For example, in Figure 3(b), the mean reward of Arm 3 (the green colored arm) is strictly lower than the two other arms for all $\theta \in \hat{\Theta}_t$. Hence, this arm is declared as $\hat{\Theta}_t$-Non-competitive and only Arms 1 and 2 are included in the competitive set $\mathcal{C}_t$. In the rare case when $\hat{\Theta}_t$ is empty, we let $\mathcal{C}_t = \{1, \ldots, K\}$ (i.e., it contains all arms) and go directly to step 3 below.

Step 3: Pull an arm from the set $\mathcal{C}_t$ using a classic bandit algorithm. At round $t+1$, we choose one of the $\hat{\Theta}_t$-Competitive arms using any classical bandit ALGORITHM (for e.g., UCB, Thompson sampling, KL-UCB). For instance UCB-C selects

$$k_{t+1} = \arg\max_{k \in \mathcal{C}_t} I_k(t),$$

where $I_k(t) = \hat{\mu}_k(t) + \sqrt{\frac{2\alpha\sigma^2 \log t}{n_k(t)}}$ is the UCB index [3]. The ability to employ any bandit algorithm in its last step is an important advantage of our algorithm. In particular, Thompson sampling has attracted a lot of attention [4, 17, 18] due to its superior empirical performance over UCB.

Remark 2 (Comparison with UCB-S proposed in [19])**.** *The paper [19] proposes an algorithm called UCB-S for the same structured bandit framework considered in this work. UCB-S constructs the confidence set $\hat{\Theta}_t$ in the same way as Step 1 described above. It then pulls the arm $k = \arg\max_{k \in \mathcal{K}} \sup_{\theta \in \hat{\Theta}_t} \mu_k(\theta)$. Taking the supremum of $\mu_k(\theta)$ over θ makes UCB-S sensitive to small changes in $\mu_k(\theta)$ and to the confidence set $\hat{\Theta}_t$. Our approach of identifying competitive arms is more robust, as observed in Section 4. Moreover, the flexibility of using Thompson Sampling in Step 3 results in a large regret improvement over UCB-S. As noted in [19], the approach used to design UCB-S cannot be directly generalized to Thompson Sampling and other bandit algorithms.*

4. REGRET ANALYSIS AND INSIGHTS

In this section, we evaluate the performance of the UCB-C and TS-C algorithms through a finite-time analysis of the expected cumulative regret defined as $\mathbb{E}[Reg(T)] = \sum_{k=1}^{K} \mathbb{E}[n_k(T)]\,\Delta_k$, (See (1)). We derive $\mathbb{E}[n_k(T)]$ separately for competitive and non-competitive arms and show that it is $\mathrm{O}(1)$ for non-competitive arms.

4.1. Competitive and Non-competitive Arms

In Section 3, we defined the notion of competitiveness of arms with respect to the confidence set $\hat{\Theta}_t$ at a fixed round t. For our regret analysis, we need asymptotic notions of competitiveness of arms, which are given below.

Definition 3 (Non-competitive and Competitive Arms)**.** *For any $\epsilon > 0$, let*

$$\Theta^{*(\epsilon)} = \{\theta : |\mu_{k^*}(\theta^*) - \mu_{k^*}(\theta)| < \epsilon\}. \tag{3}$$

An arm k is said to be non-competitive if there exists an $\epsilon > 0$ such that k is not the optimal arm for any $\theta \in \Theta^{*(\epsilon)}$; i.e., if $\mu_k(\theta) < \max_{\ell \in \mathcal{K}} \mu_\ell(\theta)$ for all $\theta \in \Theta^{*(\epsilon)}$. Otherwise, the arm is said to be competitive; i.e., if for all $\epsilon > 0$, $\exists \theta \in \Theta^{*(\epsilon)}$ such that $\mu_k(\theta) = \max_{\ell \in \mathcal{K}} \mu_\ell(\theta)$. The number of competitive arms is denoted by $C(\theta^*)$.

Since the optimal arm k^* is competitive by definition, we have $1 \leq C(\theta^*) \leq K$. We can think of $\Theta^{*(\epsilon)}$ as a confidence set for θ obtained from the samples of the best arm k^*. We note that the number of competitive $C(\theta^*)$ arms is a function of the unknown parameter θ^* and the mean reward functions $\mu_k(\theta)$.

4.2. Upper Bounds on Regret

Definition 4 (Degree of Non-competitiveness, ϵ_k)**.** *The degree of non-competitiveness ϵ_k of a non-competitive arm k is the largest ϵ for which $\mu_k(\theta) < \max_{\ell \in \mathcal{K}} \mu_\ell(\theta)$ for all $\theta \in \Theta^{*(\epsilon)}$, where $\Theta^{*(\epsilon)} = \{\theta : |\mu_{k^*}(\theta^*) - \mu_{k^*}(\theta)| < \epsilon\}$. In other words, ϵ_k is the largest ϵ for which arm k is $\Theta^{*(\epsilon)}$-non-competitive.*

Our first result shows that the expected pulls for non-competitive arms are bounded with respect to time T.

Theorem 1. *If an arm k is non-competitive with degree ϵ_k, then the number of times it is pulled by UCB-C is upper bounded as*

$$\mathbb{E}[n_k(T)] \leq Kt_0 + \sum_{t=1}^{T} 2Kt^{1-\alpha} + K^3 \sum_{t=Kt_0}^{T} 6\left(\frac{t}{K}\right)^{2-\alpha}$$
$$= \mathrm{O}(1) \quad \textit{for } \alpha > 3, \qquad \textit{where,} \tag{4}$$

$$t_0 = \inf\left\{\tau \geq 2 : \Delta_{min}, \epsilon_k \geq 4\sqrt{\frac{K\alpha\sigma^2 \log \tau}{\tau}}\right\}; \Delta_{min} = \min_{k \in \mathcal{K}} \Delta_k$$

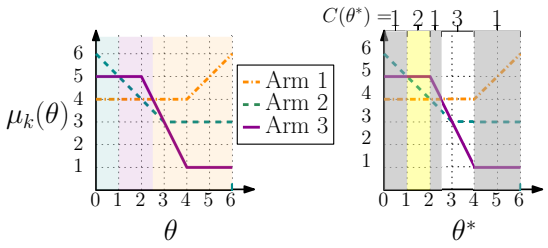

Fig. 4: *Arm 2 is optimal for $\theta^* \in [0, 1]$, Arm 3 is optimal for $\theta^* \in [1, 2.5]$ and Arm 1 is optimal for $\theta^* \in [2.5, 6]$. The figure on the right illustrates the number of competitive arms for different values of θ through shaded regions grey ($C(\theta) = 1$), yellow ($C(\theta) = 2$) and white ($C(\theta) = 3$).*

We now show that $\mathbb{E}[n_k(T)]$ is $\mathrm{O}(\log T)$ for competitive arms.

Theorem 2. *The expected number of times any sub-optimal arm is pulled by UCB-C Algorithm is upper bounded as*

$$\mathbb{E}[n_k(T)] \leq 8\alpha\sigma^2 \frac{\log T}{\Delta_k^2} + \frac{2\alpha}{\alpha - 2} + \sum_{t=1}^{T} 2Kt^{1-\alpha}$$
$$= \mathrm{O}(\log T) \quad \textit{for } \alpha > 2, \tag{5}$$

Plugging the results of Theorem 1 and Theorem 2 in (1) yields the bound on the expected regret in Theorem 3 [1].

Theorem 3 (Expected Regret Scaling)**.** *The expected regret of the UCB-C and TS-C algorithms has the following scaling with respect to the number of rounds T:*

$$\mathbb{E}[Reg(T)] \leq (C(\theta^*) - 1) \cdot \mathrm{O}(\log T) + \mathrm{O}(1) \tag{6}$$

where $C(\theta^*)$ is the number of competitive and θ^* is the true value of the common unknown parameter.

4.3. Discussion on Regret Bounds

Reduction in the Effective Number of Arms. The classic multi-armed bandit algorithms, which are agnostic to the structure of the problem, pull each of the $(K - 1)$ sub-optimal arms $\mathrm{O}(\log T)$ times. In contrast, our algorithms UCB-C and TS-C pull only a *subset* of the sub-optimal arms $\mathrm{O}(\log T)$ times, with the rest (i.e., non-competitive arms) being pulled only $O(1)$ times. When $C(\theta^*) = 1$, all sub-optimal arms are pulled only $\mathrm{O}(1)$ times, leading to a bounded regret. Cases with $C(\theta^*) = 1$ can arise quite often in practical settings. For example, when functions are continuous or Θ is countable, this occurs when the optimal arm k^* is *invertible*, or has a unique maximum at $\mu_{k^*}(\theta^*)$, or any case where the set $\Theta^* = \{\theta : \mu_{k^*}(\theta) = \mu_{k^*}(\theta^*)\}$ is a *singleton*. These cases lead to having all sub-optimal arms non-competitive, whence both UCB-C and TS-C achieve bounded (i.e., $O(1)$) regret.

Empirical performance of ALGORITHM-C. In Figure 5 we compare the regret of ALGORITHM-C against the regret of ALGORITHM (UCB/TS/KL-UCB). We plot the corresponding cumulative regret attained under ALGORITHM-C vs. ALGORITHM of the example shown in Figure 4 for three different values of θ^* : 0.5, 1.5 and 2.6. Refer to Figure 4 to see that $C = 1$, 2 and 3 for $\theta^* = 0.5$,1.5 and 2.6, respectively. Due to this, we see that ALGORITHM-C achieves bounded regret for $\theta^* = 0.5$, and reduced regret relative to ALGORITHM for $\theta^* = 1.5$ as only one arm is pulled $\mathrm{O}(\log T)$ times. For $\theta^* = 2.6$, even though $C = 3$ (i.e., all arms are competitive), ALGORITHM-C achieves empirically smaller regret than ALGORITHM. We also see the advantage of using TS-C and KL-UCB-C over UCB-C in Figure 5 as Thompson Sampling and KL-UCB are known to outperform UCB empirically. For all the simulations, we set $\alpha = 3, \beta = 1$. Rewards are drawn from the distribution $\mathcal{N}(\mu_k(\theta^*), 4)$, i.e, $\sigma = 2$. We report average regret after conducting 100 independent experiments.

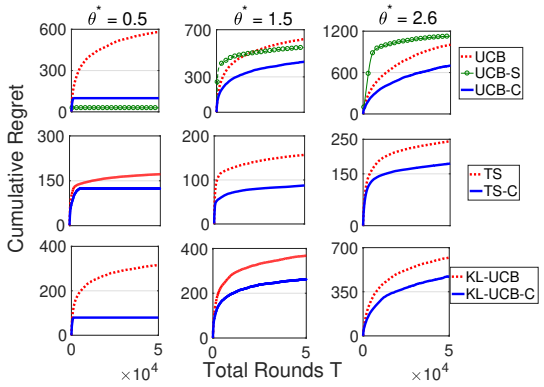

Fig. 5: *Cumulative regret of* ALGORITHM-C *vs.* ALGORITHM *(UCB in row 1, TS in row 2 and KL-UCB in row 3) for the setting in Figure 4. The value of $C(\theta^*) = 1$ in the 1st column, $C(\theta^*) = 2$ in 2nd and $C(\theta^*) = 3$ in 3rd column.*

Performance comparison with UCB-S. In the first row of Figure 5, we also plot the performance of the UCB-S algorithm proposed in [19], alongside UCB and UCB-C. UCB-S tends to favor pulling arms that have the largest mean reward for $\theta \in \Theta^{*(\epsilon)}$ (Remark 2). This bias renders the performance of UCB-S to depend heavily on θ^*. When $\theta^* = 0.5$, UCB-S has least regret among the three algorithms compared in Figure 5, but when $\theta^* = 2.6$ it gives even worse regret than UCB.

5. CONCLUDING REMARKS

We studied a structured bandit problem in which the mean rewards of different arms are related through a common hidden parameter. Our problem setting makes no assumptions on mean reward functions, due to which it subsumes several previously studied frameworks [7, 9, 13]. Our proposed approach extends *any* classical bandit algorithm to the structured bandit setting. Rigorous evaluation of UCB-C and TS-C reveal that the designed algorithm pulls only $C(\theta^*) - 1$ of the $K - 1$ sub-optimal arms $\mathrm{O}(\log T)$ times and all other arms are pulled only $\mathrm{O}(1)$ times. An open direction is to study the problem of best-arm identification in the considered problem setting.

[1] A corresponding result of C-TS is available in our full paper. Additional Simulations and Experiment on real-world Movielens recommendation dataset are available in our full paper

6. REFERENCES

[1] 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*, 2020.

[2] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[3] 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.

[4] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *COLT*, 2012, pp. 39–1.

[5] A. Garivier and O. Cappé, "The kl-ucb algorithm for bounded stochastic bandits and beyond," in *Proceedings of the 24th annual Conference On Learning Theory*, 2011, pp. 359–376.

[6] 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*. ACM, 2010, pp. 661–670.

[7] O. Atan, C. Tekin, and M. van der Schaar, "Global multi-armed bandits with Hölder continuity," in *AISTATS*, 2015.

[8] 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, 2018.

[9] Z. Wang, R. Zhou, and C. Shen, "Regional multi-armed bandits," in *AISTATS*, 2018.

[10] R. Combes, S. Magureanu, and A. Proutière, "Minimal exploration in structured stochastic bandits," in *NIPS*, 2017.

[11] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, pp. 1221–1243, 2014.

[12] S. Bubeck, N. Cesa-Bianchi *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[13] A. J. Mersereau, P. Rusmevichientong, and J. N. Tsitsiklis, "A structured multi-armed bandit problem and the greedy policy," *IEEE Transactions on Automatic Control*, vol. 54, no. 12, pp. 2787–2802, Dec 2009.

[14] T. Lattimore and C. Szepesvari, "The end of optimism? an asymptotic analysis of finite-armed linear bandits," *arXiv preprint arXiv:1610.04491*, 2016.

[15] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári, "Parametric bandits: The generalized linear case," in *Advances in Neural Information Processing Systems*, 2010, pp. 586–594.

[16] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, 2011, pp. 2312–2320.

[17] 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, Dec. 1933.

[18] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds., 2011, pp. 2249–2257.

[19] T. Lattimore and R. Munos, "Bounded regret for finite-armed structured bandits," in *Advances in Neural Information Processing Systems*, 2014, pp. 550–558.