CORRELATED MULTI-ARMED BANDITS WITH A LATENT RANDOM SOURCE

Samarth Gupta, Gauri Joshi and Osman Yağan

Dept of ECE, Carnegie Mellon Unviersity, Pittsburgh PA, USA

ABSTRACT

Multi-armed bandit models are widely studied sequential decision-making problems that exemplify the exploration-exploitation trade-off. We study a novel *correlated* multi-armed bandit model where the rewards obtained from the arms are functions of a common latent random variable. We propose and analyze the performance of the *C-UCB algorithm* that leverages the correlations between arms to reduce the cumulative *regret* (i.e., to increase the total reward obtained after T rounds). Unlike the standard UCB algorithm that pulls all sub-optimal arms $\mathrm{O}(\log T)$ times, the C-UCB algorithm takes only $\mathrm{O}(1)$ times to identify that some arms, which we refer to as *non-competitive arms*, are optimal. Thus, we effectively reduce a K-armed bandit problem to a $C+1-$armed bandit problem with $C < K$ denoting the number of *competitive*, where C can be computed from the reward functions. A key consequence is that when $C = 0$, our algorithm achieves a constant (i.e., $O(1)$) regret instead of the standard $O(\log T)$ scaling with the number of rounds T. Establishing lower bounds for the regret, we show that the C-UCB algorithm is *order-wise optimal* and demonstrate its superiority against other algorithms via numerical simulations.

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

A full version of this paper with additional examples and proofs is accessible at: arxiv.org/abs/1808.05904 [1].

1. INTRODUCTION

The *multi-armed bandit* (MAB) framework is a special case of reinforcement learning where actions do not change the system state. At each time step we obtain a reward by pulling one of K arms which have unknown reward distributions, and the objective is to maximize the cumulative reward. The seminal work of Lai and Robbins [2] proposed the upper confidence bound (UCB) arm-selection algorithm, and studied its performance limits in terms of bounds on *regret*. Subsequently, multi-armed bandit algorithms [3] have been used in numerous applications including medical diagnosis [4], system testing [5], scheduling in computing systems [6], and web optimization [7].

Most existing works on MABs assume that rewards from arms are mutually independent at each step. However, in most of the envisioned applications of the model, rewards are expected to be correlated with each other through a common randomness. Consider the example of dynamic pricing, where a company needs to sequentially decide prices $p_t \in \mathcal{P}$ to maximize their long term revenue. Here, the revenue corresponding to different prices are correlated with each other and depend on the *unknown time-varying* market size X_t. The estimated revenue, $R(p_t, X_t)$, is known as a function $F(p_t, X_t)$ of the unknown market size X_t and the chosen price p_t as given in [8]. Similarly, consider the application to energy-efficient communication [9], where a device needs to decide the transmit power level in each round. The reward, which is the received signal strength depends on the *unknown time-varying* channel state X_t and is known as a function of the channel state. In both these applications, while the reward is known given the realization of the hidden random variable, the underlying randomness is hidden and its distribution is unknown.

Motivated by these applications and the limitations of using classic bandit algorithms in these settings, we consider a correlated multi-armed bandit framework in which rewards corresponding to different arms are correlated through a latent random variable X. The rewards from different arms are known functions $g_k(\cdot)$ of X but the realizations of X and its probability distribution are unknown, which is the case for the dynamic pricing and the energy-efficient communication applications described above. Although we focus on the case where rewards are deterministic functions $g_k(X)$ of X, the algorithms designed and regret bounds can be extended to a setting where the rewards are random and only their upper and lower bounds are known, as discussed briefly in Section 6.

Another common application of multi-armed bandits is advertisement recommendation systems, where responses corresponding to different ads depend on the age/occupation/income features of the user to whom the ad is displayed. Contextual bandit algorithms [10], observe the *context* (age/occupation information) of the user arriving into the system and find best recommendation for that user. Our work differs from contextual bandits in that the context is the *unknown and random* X but the reward mappings as a function of this hidden context are known. In contrast to contextual bandits, which are designed to provide personalized recommendation in the context of ads, our model can be applied in a scenario where a company needs to decide a single product recommendation for a user population X with an unknown feature demographic

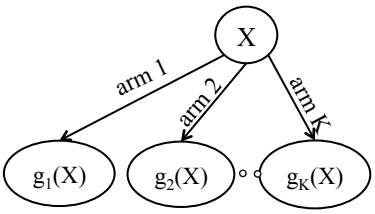


Fig. 1: *The correlated multi-armed bandit framework. The reward of arm k at round t is $g_k(x_t)$, where x_t is an i.i.d. realization of the latent random variable X.*

p_X. Another class of related models is the structured bandit framework [11–17] (i.e., linear bandit [18], generalized linear bandit [19], Lipschitz bandits [20] etc.). Structured bandit models assume that the means $\mu_k(\theta)$ of the rewards of different arms depend on a common fixed parameter θ. But the reward realizations are not necessarily correlated – they are typically independent given θ. In this work we explicitly model such correlations through the presence of a latent random source X instead of a fixed parameter θ.

2. PROBLEM FORMULATION

Consider a latent random variable X whose probability distribution is unknown. The random variable can be either discrete or continuous. For discrete X, we denote the sample space by $\mathcal{W} = \{x_1, x_2, \dots x_J\}$, and use p_j to denote the probability $\Pr(X = x_j)$ such that $\sum_{j=1}^{J} p_j = 1$. For continuous X, $f_X(x)$ denotes the pdf of X over $x \in \mathbb{R}$. As shown in Figure 1, the *reward* obtained by pulling one of the K arms in rounf t is a function of the realization X_t of the latent random variable. In other words, taking action $k_t \in \mathcal{K} = \{1, 2, \dots, K\}$ in round t yields reward $g_{k_t}(X_t)$ and X_t is an i.i.d. realization of X. Our objective is to choose the optimal sequence of arm pulls $k_1, \dots, k_T$ so as to *maximize* the cumulative reward $\sum_{t=1}^{T} g_{k_t}(x_t)$. This is equivalent to *minimizing* the cumulative *regret* which is defined as

$$Reg(T) \triangleq \mathbb{E}\left[\sum_{t=1}^{T}(g_{k^*}(X_t) - g_{k_t}(X_t))\right] \tag{1}$$

where k^* is the *optimal* arm. Put differently, $Reg(T)$ quantifies the total reward lost until time T as a consequence of making some *sub-optimal* decisions. The optimal arm k^* satisfies

$$k^* = \underset{k\in\{1,2,\dots,K\}}{\arg\max}\ \mathbb{E}\left[g_k(X)\right] = \arg\max_{k\in\{1,2,\dots K\}} \mu_k, \tag{2}$$

where μ_k denotes the mean reward of arm k. Let $\Delta_k \triangleq \mu_{k^*} - \mu_k$ be defined as the sub-optimality gap of arm k with respect to the optimal arm k^*. We also assume that the reward functions are bounded within an interval of size B, that is, $(\max_{x\in\mathcal{W}} g_k(x) - \min_{x\in\mathcal{W}} g_k(x)) \leq B$ for all arms $k \in \mathcal{K}$. We do not make any other assumptions such as the functions $g_1, \dots g_K$ being invertible.

Remark 1 (Connection to Classical Multi-armed Bandits)**.** Although we consider a scalar random variable X for brevity, our framework and algorithm can be generalized to a latent random vector $\mathbf{X} = (X_1, X_2, \dots X_m)$, as we explain in [1]. The classical multi-armed bandit framework with independent arms is a special case of this generalized model when $\mathbf{X} = (X_1, X_2, \dots X_K)$ where X_i are independent random variables and $g_k(X) = X_k$ for $k \in \mathcal{K}$.[1]

3. THE PROPOSED C-UCB ALGORITHM

Our proposed generalizes the UCB Algorithm [21] to the correlated bandit setting. The UCB algorithm at each round $t+1$ pulls the arm with the highest UCB index,

$$I_k(t) = \hat{\mu}_k + B\sqrt{\frac{2\log t}{n_k(t)}}. \tag{3}$$

Here $\hat{\mu}_k$ denotes the empirical mean of arm k and $n_k(t)$ denotes the number of times arm k has been pulled till round t. Before describing our proposed algorithm, we define the notion of pseudo-rewards and empirical competitiveness.

3.1. Pseudo-Reward of Arm ℓ with respect to Arm k

The pseudo-reward of arm ℓ with respect to arm k is an artificial sample of arm ℓ's reward generated using the reward observed from arm k. It is defined as follows.

Definition 1 (Pseudo-Reward)**.** *Suppose we pull arm k and observe reward r. Then the pseudo-reward of arm ℓ with respect to arm k is*

$$s_{\ell,k}(r) \triangleq \max_{x:g_k(x)=r} g_\ell(x). \tag{4}$$

The pseudo-reward $s_{\ell,k}(r)$ gives the maximum possible reward that could have been obtained from arm ℓ, given the reward observed from arm k.

Definition 2 (Empirical and Expected Pseudo-Reward)**.** *After t rounds, arm k is pulled $n_k(t)$ times. Using these $n_k(t)$ reward realizations, we can construct the empirical pseudo-reward $\hat{\phi}_{\ell,k}(t)$ for each arm ℓ with respect to arm k as follows.*

$$\hat{\phi}_{\ell,k}(t) \triangleq \frac{\sum_{\tau=1}^{t} \mathbb{1}_{k_\tau=k} s_{\ell,k}(r_t)}{n_k(t)}, \quad \ell \in \mathcal{K}\setminus\{k\}. \tag{5}$$

The expected pseudo-reward of arm ℓ with respect to arm k is defined as

$$\phi_{\ell,k} \triangleq \mathbb{E}\left[s_{\ell,k}(g_k(X))\right]. \tag{6}$$

[1]The framework we study leads to rewards from different arms being correlated for general functions $g_1, ..g_K$ and general pdf of X, i.e., $f_X(x)$. However, there are specific reward functions and distributions $f_X(x)$, where rewards across (some) arms might be uncorrelated or independent.

3.2. Competitive and Non-competitive arms

Using the pseudo-reward estimates defined above, we can classify each arm $\ell \neq k$ as *competitive* or *non-competitive* with respect the arm k. To this end, we first define the notion of the pseudo-gap.

Definition 3 (Pseudo-Gap)**.** *The pseudo-gap $\tilde{\Delta}_{\ell,k}$ of arm ℓ with respect to arm k is defined as*

$$\tilde{\Delta}_{\ell,k} \triangleq \mu_k - \phi_{\ell,k}, \tag{7}$$

i.e., the difference between expected reward of arm k and the expected pseudo-reward of arm ℓ with respect to arm k.

From the definition of pseudo-reward, it follows that the expected pseudo-reward $\phi_{\ell,k}$ is greater than or equal to the expected reward μ_ℓ from arm ℓ. Thus, a positive pseudo-gap $\tilde{\Delta}_{\ell,k} > 0$ indicates that it is possible to classify arm ℓ as sub-optimal using only the rewards observed from arm k (with *high* probability as the number of pulls for arm k gets *large*); thus, arm ℓ needs not be explored. Such arms are called non-competitive, as we define below.

Definition 4 (Competitive and Non-Competitive arms)**.** *An arm ℓ is said to be non-competitive if its pseudo-gap with respect to the optimal arm k^* is positive, that is, $\tilde{\Delta}_{\ell,k^*} > 0$. Similarly, an arm ℓ is said to be competitive if $\tilde{\Delta}_{\ell,k^*} < 0$. The unique best arm k^* has $\tilde{\Delta}_{k^*,k^*} = 0$ and is not counted in the set of competitive arms.*

Since the distribution of X is unknown, we can not find the pseudo-gap of each arm and thus have to resort to empirical estimates based on observed rewards. In our algorithm, we use a noisy notion of the competitiveness of an arm defined as follows. Note that since the optimal arm k^* is also not known, empirical competitiveness of an arm ℓ is defined with respect to each of the other arms $k \neq \ell$.

Definition 5 (Empirically Competitive and Non-Competitive arms)**.** *An arm ℓ is said to be "empirically non-competitive with respect to arm k at round t" if its empirical pseudo-reward is less than the empirical reward of arm k, that is, $\hat{\mu}_k(t) - \hat{\phi}_{\ell,k}(t) > 0$. Similarly, an arm $\ell \neq k$ is deemed empirically competitive with respect to arm k at round t, if $\hat{\mu}_k(t) - \hat{\phi}_{\ell,k}(t) \leq 0$.*

3.3. The C-UCB Algorithm

The central idea in our correlated UCB algorithm is that after pulling the optimal arm k^* sufficiently many times, the non-competitive (and thus sub-optimal) arms can be classified as empirically non-competitive with increasing confidence, and thus need not be explored. However, the competitive arms cannot be discerned as sub-optimal with high confidence just by using the rewards observed from the optimal arm. Motivated by this, in each round t, the proposed C-UCB algorithm performs the following steps:

1. Select arm $k^{max} = \arg\max_k n_k(t-1)$, that has been pulled the most until round $t-1$.
2. Identify the set $\mathcal{A}_t$ of arms that are empirically competitive with respect to arm k^{max} at round t.
3. Pull the arm $k_t \in \{\mathcal{A}_t \cup k^{max}\}$ with the highest UCB1 index $I_k(t-1)$ (defined in (3)).
4. Update the empirical pseudo-rewards s_{ℓ,k_t} for all ℓ, the empirical reward $\hat{\phi}_{\ell,k_t}(t)$, and the UCB1 indices of all arms based on the observed reward r_t.

Note that the set of competitive arms is not known beforehand, due to which we identify the set of empirically competitive arms based on the samples of $k^{\max}$, as it is likely to provide the best estimate. Moreover, an empirically non-competitive arm at round t can be empirically competitive in subsequent rounds as empirical competitiveness is a noisy notion.

4. REGRET ANALYSIS AND BOUNDS

We now characterize the performance of the C-UCB algorithm by analyzing the expected value of the cumulative regret ((1)). The expected regret can be expressed as

$$\mathbb{E}\left[Reg(T)\right] = \sum_{k=1}^{K} \mathbb{E}\left[n_k(T)\right] \Delta_k, \tag{8}$$

where $\Delta_k = \mathbb{E}\left[g_{k^*}(X)\right] - \mathbb{E}\left[g_k(X)\right] = \mu_{k^*} - \mu_k$ and $n_k(T)$ is the number of times arm k is pulled in T slots. For the regret analysis, we assume without loss of generality that the reward functions $g_k(X)$ satisfy $0 \leq g_k(X) \leq 1$ for all $k \in \{1, 2, \ldots K\}$. Theorem 1 and Theorem 2 below show that $\mathbb{E}\left[n_k(T)\right]$ scales as $O(1)$ and $O(\log T)$ for non-competitive and competitive arms respectively.

Theorem 1 (Expected Pulls of a Non-competitive Arm)**.** *The expected number of times C-UCB pulls a non-competitive arm is bounded as,*

$$\mathbb{E}\left[n_k(T)\right] \leq Kt_0 + K^2 \sum_{t=Kt_0}^{T} 3\left(\frac{t}{K}\right)^{-2} + \sum_{t=1}^{T} t^{-3}, \tag{9}$$

$$= \mathrm{O}(1), \tag{10}$$

where $t_0 = \inf\left\{\tau \geq 2 : \Delta_{min}, \tilde{\Delta}_{k,k^*} \geq 4\sqrt{\frac{K \log \tau}{\tau}}\right\}$.

Theorem 2 (Expected Pulls of a Competitive Arm)**.** *The expected number of times a competitive arm is pulled by C-UCB is bounded as*

$$\mathbb{E}\left[n_k(T)\right] \leq 8\frac{\log(T)}{\Delta_k^2} + 2 + \sum_{t=1}^{T} t \exp\left(-\frac{t\Delta_{min}^2}{2K}\right), \tag{11}$$

$$= \mathrm{O}(\log T) \quad \textit{if } \Delta_{min} = \min_k \Delta_k > 0. \tag{12}$$

Substituting the bounds on $\mathbb{E}\left[n_k(T)\right]$ derived in Theorem 1 and Theorem 2 into (8), we get the following regret bound

Corollary 1 (Upper Bound on Expected Regret)**.** *The expected cumulative regret of the C-UCB algorithm is,*

$$\mathbb{E}\left[Reg(T)\right] \leq \sum_{k\in\mathcal{C}} \Delta_k U_k^{(c)}(T) + \sum_{k'\in\mathcal{K}\setminus\{\mathcal{C}\cup k^*\}} \Delta_{k'} U_{k'}^{(nc)}(T),$$
$$= C\cdot \mathrm{O}(\log T) + \mathrm{O}(1),$$

where $\mathcal{C} \subseteq \{1,\ldots,K\}\setminus\{k^*\}$ is set of competitive arms with cardinality C, $U_k^{(c)}(T)$ is the upper bound on $\mathbb{E}\left[n_k(T)\right]$ for competitive arms given in (11)*, and $U_k^{(nc)}(T)$ is the upper bound for non-competitive arms given in* (9)*.*

Reducing dimension of Bandit problem. For the UCB1 algorithm [21], a regret bound similar to (1) is known to hold, but with the first sum taken over *all* arms. In this sense, we can say that our C-UCB algorithm reduces the K-armed bandit problem to a $C+1$-armed bandit problem.

Achieving Bounded Regret. If the set of competitive arms $\mathcal{C}$ is empty (i.e., the number of competitive arms $C=0$), then our algorithm will lead to (see Corollary 1) an expected regret of $\mathrm{O}(1)$, instead of the typical $\mathrm{O}(\log T)$ regret scaling in classic multi-armed bandits. A simple case where $\mathcal{C}$ is empty is when the reward function $g_{k^*}(X)$ corresponding to the arm k^* is invertible. This is because, for all arms $\ell \neq k^*$, the pseudo-gap satisfies $\tilde{\Delta}_{\ell,k^*} = \Delta_\ell > 0$ resulting in them being non-competitive. The set $\mathcal{C}$ can be empty in more general cases where none of the arms are invertible, in which case our algorithm still achieves an expected regret of $\mathrm{O}(1)$.

Next, we present a lower bound on the expected regret $\mathbb{E}\left[Reg(T)\right]$. Intuitively, if an arm ℓ is *competitive*, it can not be deemed sub-optimal just by using the samples from the optimal arm k^*. Theorem 3 shows that $\mathrm{O}(\log T)$ regret is unavoidable in such cases.

Theorem 3 (Lower Bound on Expected Regret)**.** *For any algorithm that achieves a sub-polynomial regret,*

$$\mathbb{E}\left[Reg(T)\right] = \begin{cases} \Omega(\log T) & \text{if } C>0, \\ \Omega(1) & \text{if } C=0. \end{cases}$$

Bounded regret whenever possible. From Corollary 1, we see that when $C>0$, our algorithm achieves a regret of $\mathrm{O}(\log T)$, matching the lower bound given in Theorem 3 orderwise. When $C=0$, our algorithm achieves $\mathrm{O}(1)$ regret, meaning that it achieves bounded regret whenever possible.

5. SIMULATION RESULTS

We now present simulation results for the case where X is a discrete random variable (simulations for continuous X and random vector $\mathbf{X}$ can be found in [1]). We consider the reward functions $g_1(X), g_2(X)$ and $g_3(X)$ shown in Figure 2 for all simulation plots. However, the probability distribution $P_X = (p_{x_1}, p_{x_2}, \ldots p_{x_5})$ of X is different for each of the following cases given below. For each case, Figure 3 shows the cumulative regret versus the number of rounds. The cumulative

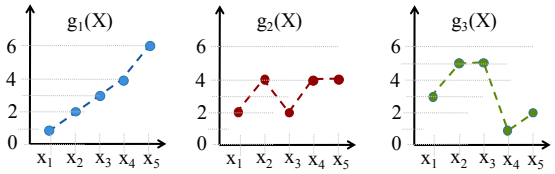


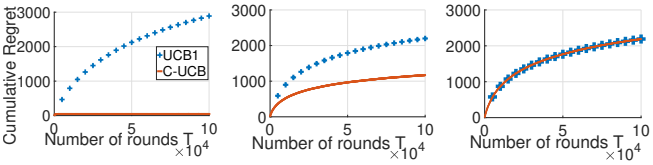
Fig. 2: *Reward Functions used for the simulation results presented in Figure 3.*

regret is averaged over 500 simulation runs, and for each run we use the same reward realizations for both the algorithms.

Case 1: No competitive arms. Here, we set $P_X = (0.1, 0.2, 0.25, 0.25, 0.2)$. In this setting, arm 1 is optimal, and arms 2 and 3 are *non-competitive*. We see in Figure 3a that the proposed C-UCB algorithm achieves a constant regret and is significantly superior to the UCB1 algorithm as it is able to exploit the correlation of rewards between the arms.

Case 2: One competitive arm. Let $P_X = (0.25, 0.17, 0.25, 0.17, 0.16)$ which results in arm 3 being optimal, while arm 1 is *non-competitive* and arm 2 is *competitive*. We see in Figure 3b that C-UCB performs less exploration than UCB as only one arm is *competitive*.

Case 3: Two competitive arms. Here, we set $P_X = (0.05, 0.3, 0.3, 0.05, 0.3)$, under which arm 3 is optimal and arms 1 and 2 are *competitive*. Since both arms are competitive, exploration is necessary for both arms and hence in Figure 3c the regret obtained under C-UCB and UCB1 is similar.



(a) *Case 1* (b) *Case 2* (c) *Case 3*

Fig. 3: *For the reward functions in Figure 2, the cumulative regret of C-UCB is smaller than UCB1 in all the three cases.*

6. DISCUSSION AND FUTURE WORK

While this work focuses on a setting where the rewards are deterministic functions $g_k(X)$ of the latent random source X. The algorithm, regret analysis and bounds can be extended to a setup where rewards are random variables R_k that correlated with the common X. If upper and lower bounds on R_k denoted by $\bar{g}_k(x)$ and $\underline{g}_k(x)$ are known, then the current algorithm and regret bounds can be directly extended by redefining pseudo-reward $s_{\ell,k}(r)$ as

$$s_{\ell,k}(r) = \max_{\underline{g}_k(x)<r<\bar{g}_k(x)} \bar{g}_\ell(x).$$

Open future directions include improving the current algorithm by using correlation information from all arms instead of just considering the arm that is pulled the most times, and designing best-arm identification algorithms for the correlated bandit setting.

7. REFERENCES

[1] S. Gupta, G. Joshi, and O. Yağan, "Correlated multi-armed bandits with a latent random source," *arXiv preprint arXiv:1808.05904*, 2018.

[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] 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.

[4] S. S. Villar, J. Bowden, and J. Wason, "Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges," *Statistical science*, vol. 30, no. 2, p. 199, 2015.

[5] C. Tekin and E. Turgay, "Multi-objective contextual multi-armed bandit problem with a dominant objective," *arXiv:1708.05655*, 2017.

[6] S. Krishnasamy, R. Sen, R. Johari, and S. Shakkottai, "Regret of queueing bandits," *CoRR*, vol. abs/1604.06377, 2016.

[7] D. Agarwal, B.-C. Chen, and P. Elango, "Explore/exploit schemes for web content optimization," in *ICDM'09*. IEEE, 2009, pp. 1–10.

[8] J. Huang, M. Leng, and M. Parlar, "Demand functions in decision modeling: A comprehensive survey and research directions," *Decision Sciences*, vol. 44, no. 3, pp. 557–609, 2013.

[9] N. Mastronarde and M. van der Schaar, "Reinforcement learning for energy-efficient wireless transmission," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 3452–3455.

[10] L. Zhou, "A survey on contextual multi-armed bandits," *CoRR*, vol. 1508.03326 [cs.LG], 2016.

[11] S. Pandey, D. Chakrabarti, and D. Agarwal, "Multi-armed bandit problems with dependent arms," in *ICML*, 2007, pp. 721–728.

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

[13] V. Srivastava, P. Reverdy, and N. E. Leonard, "Correlated multiarmed bandit problem: Bayesian algorithms and regret analysis," *CoRR*, vol. arXiv:1507.01160 [math.OC], Jul. 2015.

[14] 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.

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

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

[17] S. Gupta, G. Joshi, and O. Yağan, "Exploiting correlation in finite-armed structured bandits," *arXiv preprint arXiv:1810.08164*, 2018.

[18] 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.

[19] 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.

[20] S. Magureanu, R. Combes, and A. Proutiere, "Lipschitz bandits: Regret lower bounds and optimal algorithms," *arXiv preprint arXiv:1405.4758*, 2014.

[21] 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.