Combinatorial Multi-Armed Bandits in Cognitive Radio Networks: A Brief Overview

Sunjung Kang
School of Electrical and Computer Engineering
UNIST
Ulsan, Korea
sky12382@unist.ac.kr

Changhee Joo
School of Electrical and Computer Engineering
UNIST
Ulsan, Korea
cjoo@unist.ac.kr

***Abstract*—Combinatorial multi-armed bandit (MAB) problem can be used to formulate sequential decision problems with exploration-exploitation tradeoff. Dynamic spectrum access (DSA) in cognitive radio (CR) networks is one of important applications. In this work, we briefly overview combinatorial MAB problems with its possible applications to CR networks. We first investigate the standard MAB problems where a single player either explores an arm to gather information to improve its decision strategy, or exploits the arm based on the information that it has collected at each round. Then, we study the taxonomy of combinatorial MAB problems, in particular for multi-player scenarios with independent and identically distributed (i.i.d.) rewards. Finally, we discuss limitations of existing works and interesting open problems.**

***Index Terms*—Multi-armed bandits, Combinatorial multi-armed bandits, Cognitive radio networks**

I. Introduction

It has been observed that license-based spectrum management has suffered from low spectrum utilization [1]. As an alternative solution, cognitive radio (CR) networks have attracted much attention as a promising approach to improve spectrum utility. In CR networks, dynamic spectrum access (DSA) allows unlicensed users (or secondary users) to identify unused channels that are licensed to primary users and to access them opportunistically. DSA is known to successfully improve spectrum utility [2].

A (secondary) user who has no prior knowledge about channel characteristics should either explore a channel to gather information to improve its decision strategy, or exploit the channel based on the information that it has collected. Hence, a user faces the well-known exploration-exploitation tradeoff.

The sequential decision problems with exploration-exploitation tradeoff have been studied in the literature on multi-armed bandit (MAB) problem [3]–[7]. In the standard K-armed bandit problems, a player (or a user) chooses an arm (i.e., a channel) $x(t) \in \{1, ..., K\}$ at each round (or time) t, and then given the choice, the player receives reward $X_{x(t)}(t)$.

Multi-armed bandit problem:
For each round $t = 1, 2, ..., T$

This work was supported by the research fund of the Signal Intelligence Research Center supervised by the Defense Acquisition Program Administration and Agency for Defense Development of Korea.

(i) the player chooses $x(t) \in \{1, 2, .., K\}$ based on its past observations,
(ii) the environment generates reward $X_{x(t)}(t)$, and sends it back to the player.

A policy decides which arm to play with the objective of maximizing the total reward $\sum_{t=1}^{T} X_{x(t)}(t)$. As the performance metric for evaluating a policy, *regret* is often used. It is the difference between the total expected reward by playing an arm with the highest total expected reward and that achieved under the policy of interest. Maximizing the reward is equivalent to minimizing the regret.

In stochastic MAB problems, the rewards are assumed to be i.i.d. processes with unknown distributions with bounded support, without loss of generality, of $[0, 1]$. In [3], the authors have shown that the regret of stochastic MAB grows at least logarithmically with respect to time. In [4], an index-based policy for stochastic MAB is proposed using upper confidence bound (UCB) called UCB1, and is shown to achieve the logarithmic growth of the regret with respect to time. In [8], the authors have shown that in stochastic MAB problem with multiple plays, the regret grows at least logarithmically with respect to time. In restless MAB problems, the rewards follow Markov chains which are known to the player, and the player learns the state of channels instead of expected rewards of channels. An index-based policy called Whittle index is prosed in [5], and restless MAB problem with multiple plays is considered in [6]. On the other hand, adversarial MAB problems consider non-stochastic rewards, and an index-based policy called EXP3 whose regret is sub-linear has been proposed in [7].

In multi-user multi-channel CR networks where each user can access at most one channel and each channel can be accessed by at most one user, multiple users access channels at the same time, and a collision occurs if more than one user access the same channel. This problem can be formulated as combinatorial MAB problem. In combinatorial MAB problem, multiple arms are played at the same time, and the total reward received by playing multiple arms is either the reward sum of played arms (linear rewards) or a function of rewards vector (non-linear rewards). Combinatorial MAB problems with non-linear rewards have been studied in [9], and linear rewards problems have been studied in [10]–[17]. In [10] and [11],

combinatorial MAB problems with restless Markovian rewards and with non-stochastic rewards are studied, respectively. In [12]–[17], combinatorial MAB problems with i.i.d. rewards are studied, and we investigate these papers detailedly in the rest of this paper.

The paper is organized as follows. In section, we provide a survey of combinatorial MAB problems with i.i.d. rewards. In Section III, we discuss limitations of existing works and interesting open problems.

II. Multi-Player Multi-Armed Bandit Problem with i.i.d. Rewards

We start with the description of multi-player multi-armed bandit problem. In N-player K-armed bandit problem where $K \geq N$, each player can play at most one arm, and each arm can be played by at most one user at each round (or time slot). If player i plays arm k at time slot t, then it gets reward $X_{i,k}(t)$ which is an i.i.d. random variable drawn from distribution $f_{i,k}$ with bound support. Without loss of generality, $X_{i,k}(t)$ lies in $[0,1]$ with mean $\mu_{i,k}$. If a channel is played more than one user, then a collision occurs and all conflicting players get no reward. Let $Y_{i,k}(t)$ denote the returned reward that player i receives from arm k at time slot t. If player i plays arm k without a collision, then $Y_{i,k}(t) = X_{i,k}(t)$, and otherwise $Y_{i,k}(t) = 0$. Each player does not have prior knowledge about channel rewards, and can only observe the returned reward that is used to choose an arm for future decisions.

Let $\mathcal{K} = \{1, .., K\}$ denote the set of arms (i.e., the set of actions of players), $x_i(t) \in \mathcal{K}$ denote an action of player i at time slot t. Its vector $\mathbf{x}(t)$ is denoted as the schedule at round t. Then, the history of player i by round t is $\mathcal{H}_i(t) = \{(x_i(1), Y_{i,x_i(1)}(1)), ..., (x_i(t), Y_{i,x_i(t)}(t))\}$ with $\mathcal{H}_i(0) = \emptyset$. A policy $\pi_i = \{\pi_i(t)\}_{t\geq 1}$ for player i is a sequence of maps $\pi_i(t) : \mathcal{H}_i(t-1) \to \mathcal{K}$. Let $\mathcal{M}$ be the set of feasible schedules such that $\mathcal{M} = \{\mathbf{a} = (a_1, ..., a_N) \ : \ a_i \in \mathcal{K}, a_i \neq a_j \text{ for } i \neq j\}$, which is equivalent to the set of all maximal matchings in complete bipartite graph $\mathcal{G} = (\mathcal{N} \cup \mathcal{K}, E)$, where $\mathcal{N}$ and $\mathcal{K}$ are the sets of users and channels, respectively, and E denotes the set of possible edges (user, channel). Let $\mathbf{a}^*$ denote as an optimal matching (i.e., an maximum weighted matching in $\mathcal{G}$) such that

$$\mathbf{a}^* \in \arg\max_{\mathbf{a}\in\mathcal{M}} \sum_{i=1}^{N} \mu_{i,a_i}.$$

Without prior knowledge of $\mu_{i,k}$, a policy π cannot schedule an optimal matching every time. Thus, whenever π schedules non-optimal matching $\mathbf{a} \neq \mathbf{a}^*$, the loss of a reward occurs, and the total expected regret by time T under π is denoted as

$$\mathcal{R}_\pi(T) := T \sum_{i=1}^{N} \mu_{i,a_i^*} - \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}[X_{i,\pi_i(t)}(t)].$$

It is known that the logarithmic growth of the total expected regret is order-optimal [12].

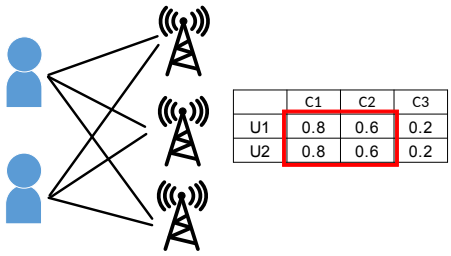


Fig. 1: Cognitive radio networks with identical channel statistics for all the users. Combinations in marked rectangles, i.e., $\{(U1, C1), (U2, C2)\}$ and $\{(U2, C1), (U1, C2)\}$, are an optimal matching.

A. Identical arms for each player

In [12]–[14], it is assumed that the distribution of a channel reward is identical for all players (i.e., $\mu_{i,k} = \mu_k$ for all i and k). Thus, an optimal matching is combinations of arms with N highest expected rewards. Fig. 1 illustrates an example of cognitive radio networks with identical channels for each user. There are three identical channels C1, C2, and C3 for two users U1 and U2. The matrix shows the expected rewards of channels, and the optimal matchings are $(a_1^*, a_2^*) \in \{(1, 2), (2, 1)\}$.

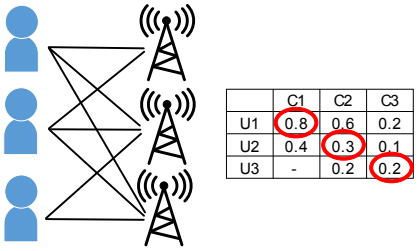
In [12], the authors have established a lower bound on the rate of the regret growth for a general class of distributed algorithms. They have proposed a distributed algorithm based on time division fair sharing (TDFS), which achieves order-optimal of the regret and ensures fairness among players (i.e., the players achieve the same time-average reward). The TDFS policy can employ any order-optimal single-player MAB policy.

The authors of [13] consider two scenarios: when the players have a prioritized ranking, and when the fairness among players is requested. The authors have proposed distributed algorithms for those two scenarios, and shown that they achieve order-optimal performance. However, they need for the players to know the number of players in the system, and to have pre-allocated identifications (or ranks) since the fairness is achieved in the round-robin manner.

In [14], the authors have proposed distributed algorithms which require neither information exchange nor prior agreement (e.g. pre-allocated identification). They first assume that the number of players in the system is fixed and known to all players, and then consider the scenario where the number of players is unknown to players. The proposed algorithms guarantee order-optimal of the regret, but the fairness among players is not achieved.

B. Non-identical arms for each player

In [15]–[17], the scenarios where the distribution of a channel reward is different for each player have been considered. In these cases, it is difficult to find an optimal matching in distributed manner without any information exchange. Fig. 2 illustrates an example of system model with non-identical channels for each user with two users and three channels.



	C1	C2	C3
U1	0.8	0.6	0.2
U2	0.4	0.3	0.1
U3	-	0.2	0.2

Fig. 2: Cognitive radio networks with different channel statistics for each player. The maximum weighted matching (i.e., the optimal matching) is marked by circles.

The matrix shows the expected rewards of each user-channel pair, and the optimal matching (i.e., the maximum weighted matching) is $(a_1^*, a_2^*) = (1, 2)$.

A naive approach for solving this problem is to consider each matching as an arm, and applying UCB1 algorithm [4], which guarantees the logarithmic growth of the regret with respect to time. However, this approach is very inefficient since storage and computational complexity are $O(P(K, N))$. In [15], the authors have proposed a centralized algorithm called matching learning with polynomial storage (MLPS). The algorithm calculates UCB indices for each user-channel pair, and schedules the maximum weighted matching with UCB index as a weight on each edge (i.e., using Hungarian algorithm whose computational complexity is $O((N + K)^3)$. The authors have shown that the proposed algorithm guarantees the logarithmic growth of the regret.

The authors of [16] pay attention to computational cost of MAB algorithms, in particular, algorithms that solve combinatorial optimization problems that have high-order computational complexity. Thus, they add computational cost to the regret, i.e., the regret considers the accumulated loss of rewards by scheduling non-optimal matching and computational cost (or communication cost in distributed setting). They have proposed a distributed algorithm called dUCB_4 that achieves the regret of $O(\log^2 T)$. Under dUCB_4, when the matching recomputation is needed, the distributed players participate to the Bertsekas auction algorithm which converges to an ϵ-optimal matching with convergence time of $O(N^2 \cdot \max_{i,k} \mu_{i,k}/\epsilon)$.

In [17], the authors are interested in learning an orthogonal stable marriage configuration (i.e., matching) (SMC) rather than an optimal matching. They have proposed a distributed algorithm that converges to an orthogonal SMC, and thus does not guarantees the logarithmic growth of the regret with respect to time. The computational complexity of this algorithm is $O(K)$.

III. Discussion

Combinatorial MAB problems have many potential applications such as cognitive radio networks, multi-channel communication systems, shortest path, etc. In many cases, the multi-player MAB problem with non-identical arms is more practical than the problem with identical arms. Although there are several works that consider multi-player MAB with non-identical arms that guarantees the order-optimal of the regret, they have high-order computational complexity. Thus, developing low-complexity learning algorithm that guarantees the order-optimal of the regret in multi-player MAB with non-identical arms is very interesting open problem.

Combinatorial MAB problem can be used to formulate not only cognitive radio networks, but also many other problems such as shortest path or multi-channel communication systems. Although in cognitive radio networks application, combinatorial MAB is modeled with bipartite graph, other applications may need to be modeled with general graphs. Developing learning algorithm that can be applied to more general graphs is another interesting open problem.

References

[1] P. Kolodzy, "Spectrum Policy Task Force," *Federal Commun. Comm., Washington, DC, Rep. ET Docket*, no. 02-135, 2002.

[2] M. M. Buddhikot, "Understanding dynamic spectrum access: Models, taxonomy and challenges," in *2nd IEEE International Symposium on DySPAN 2007*.

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

[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] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of applied probability*, vol. 25, no. A, pp. 287–298, 1988.

[6] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5547–5567, 2010.

[7] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM journal on computing*, vol. 32, no. 1, pp. 48–77, 2002.

[8] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: IID rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, 1987.

[9] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *International Conference on Machine Learning*, 2013, pp. 151–159.

[10] Y. Gai, B. Krishnamachari, and M. Liu, "Online learning for combinatorial network optimization with restless markovian rewards," in *9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2012*.

[11] R. Combes, M. S. T. M. Shahi, A. Proutiere, *et al.*, "Combinatorial bandits revisited," in *Advances in Neural Information Processing Systems*, 2015, pp. 2116–2124.

[12] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, 2010.

[13] Y. Gai and B. Krishnamachari, "Decentralized online learning algorithms for opportunistic spectrum access," in *GLOBECOM*, 2011.

[14] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[15] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *IEEE Symposium on New Frontiers in Dynamic Spectrum*, 2010.

[16] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[17] O. Avner and S. Mannor, "Multi-user lax communications: a multi-armed bandit approach," in *Computer Communications, IEEE INFOCOM 2016*.