Resource Allocation for Multi-source Multi-relay Wireless Networks: A Multi-Armed Bandit Approach

Ali Al Khansa[1,2(✉)], Raphael Visoz[1], Yezekael Hayel[2], and Samson Lasaulce[3]

[1] Orange Labs, Chatillon, France
{ali.alkhansa,raphael.visoz}@orange.com
[2] LIA, Avignon University, Avignon, France
yezekael.hayel@univ-avignon.fr
[3] CRAN, CNRS, Nancy, France
samson.lasaulce@univ-lorraine.fr

Abstract. In this paper, we consider the problem of link adaptation (rate allocation) of Orthogonal Multiple Access Multiple Relay Channel (OMAMRC) using the Multi-Armed Bandit (MAB) online learning framework. The cooperative system is composed of a transmission phase where sources transmit in a round robin manner, and a retransmission phase where a scheduled node sends redundancies. We assume that we have no knowledge of the Channel State Information (CSI) nor of the Channel Distributed Information (CDI). Accordingly, rate allocation must be learned online following a sequential learning algorithm. We adapt to one variant of the MAB framework algorithms, the Upper Confidence Bound (UCB) family, and specifically the UCB1 algorithm. The UCB1 algorithm achieves a logarithmic regret uniformly over time, without any preliminary knowledge about the reward distributions. Due to the exponential growth of the number of arms, following the multiple sources included in the rate allocation, the UCB1 algorithm features a complexity problem. Thus, we propose a sequential UCB1 (SUCB1) algorithm which solves the complexity issue, and outperforms the UCB1 algorithm.

Keywords: Link Adaptation · Multi-Armed Bandit · Upper confidence bound · Multi-source multi-relay wireless network · Spectral efficiency

1 Introduction

One of the main objectives for 5G and 5G-beyond cellular networks is to allow heterogeneous services to coexist within the same network architecture. Some of these services need a very high peak data rates and a fast adaptation of the channel state, as in enhanced Mobile Broadband (eMBB). In order to meet those needs, we aim at improving the spectral efficiency. Cooperative communication [1] represents one of the key physical layer technologies which aims to optimize the spectral efficiency. The concept is to use the shared resources and

H. Elbiaze et al. (Eds.): UNet 2021, LNCS 12845, pp. 62–75, 2021.
https://doi.org/10.1007/978-3-030-86356-2_6

information of the users to improve the transmission and reception processes. The cooperation process can be performed by sources themselves (user cooperation), or by using some dedicated relay nodes. The difference between a relay node and a source node which implements user cooperation is the fact that the latter has its own message whereas the relay node does not.

Cooperative models have been analyzed extensively in the prior literature. These models depend on the number of source nodes, relay nodes, and destination nodes included. For example, we call the system Multiple Relay Channel (MRC) when we have multiple relays helping a single source to communicate with a single destination [2]. Other two examples are the Relay Broadcast Channel (RBC) [2] and the Multiple Access Relay Channel (MARC) [3]. In RBC, the system is composed of a single source, a single relay, and multiple destination nodes, whereas in the MARC, we have a single relay node helping multiple users to communicate with a single destination.

Here, we consider the Multiple Access Multiple Relay Channel (MAMRC), where we have multiple relay nodes, helping multiple users to communicate with a single destination. In addition, we consider user cooperation, where users that have no message to send, act as relays. Specifically, we consider a slow-fading half-duplex Orthogonal Multiple Access Multiple Relay Channel (OMAMRC), where orthogonality is achieved using Time Division Multiplexing (TDM) (check Fig. 1).

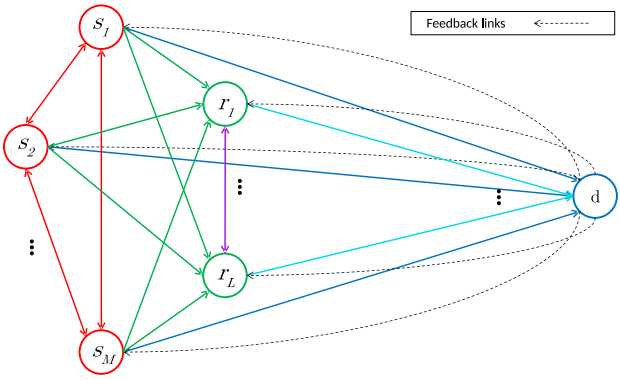

Fig. 1. Cooperative Orthogonal Multiple Access Multiple Relay Channel (OMAMRC) with feedback broadcast control channel to indicate the destination decoding set.

There are several relaying protocols widely used in cooperative communication. One category of these protocols is the linear relaying protocols, and its famous example the Amplify-and-Forward (AF) protocol [4]. Another category is the non-linear (regenerative) relaying protocols. Some examples of regenerative protocols are the Compress-and-Forward [5] and the Quantize-Map-and-Forward [6]. In our work, we use the Selective Decode-and-Forward (SDF) relaying protocol, where relays can forward only a signal representative of successfully decoded source messages. The error detection is based on Cyclic Redundancy

Check (CRC) bits that are appended to each source message. The used protocol is an updated version of the well known Decode-and-Forward (DF) protocol [7]. In DF, cooperative nodes are obliged to wait to successfully decode all the source messages before starting to cooperate, whereas in SDF, they can start cooperating before.

In this paper, we investigate the problem of Link Adaptation (LA) (rate allocation), where the destination is considered as the centralized node which allocates the rates for the multiple sources. In the prior art, several heuristic algorithms were presented. In [8], a Slow-Link Adaptation (SLA) algorithm was proposed. The algorithms proposed were heuristic, and based on the information available at the destination. When Channel State Information (CSI) is available, Fast Link Adaptation (FLA) algorithm is used, where allocation is performed once there is a change in the CSI. On the contrary, if CSI is not available (in high mobility scenarios), the SLA algorithm is used, where allocation is based on the Channel Distribution Information (CDI) (for example: the average Signal to Noise Ratio (SNR)) of the links.

In this work, we aim to solve the LA problem using a different perspective. First, we aim to use an algorithm which is not heuristic, and where the regret is bounded and tractable. Next, we want to solve the problem when no information is given at the destination. In other words, we aim to perform rate allocation using a learning algorithm, where the probability of transmission success at a certain rate is unknown (since the channel state is unknown) and rather needed to be learned. We adapt to the well known framework called Multi-Armed Bandit (MAB), where it addresses the exploration-exploitation dilemma.

The main contributions of the paper can be summarized as the following:

- To our knowledge, this work is the first which tackles LA for MAMRC with online learning perspective, using the MAB framework.
- We state the MAB problem based on the utility metric (spectral efficiency per frame), following the definitions of the common and individual outage events.
- We implement the UCB1 algorithm in the presented framework, and then, an approximated UCB1 algorithm was presented aiming to reduce the initialization step of the algorithm.
- We finally propose a sequential UCB1 algorithm which solves the problem of exponential dimension of the number of arms.

The rest of the paper is organized as follows: In the next section, the related work of the MAB literature is presented. In Sect. 3, the system model is presented. In Sect. 4, outage events are given, followed by the MAB problem formulation. In Sect. 5, the LA algorithms are described. Finally, numerical results and main conclusions are presented in Sects. 6 and 7 respectively.

2 Related Work

First, we state the main issue which MAB framework tackles, i.e., the exploration-exploitation dilemma. In scenarios where multiple choices are possible (multiple arms), each with an unknown average reward, MAB algorithms give sequential steps to decide whether we need to learn more (exploration), or to stay with the option that gave best rewards in the past (exploitation). There are different types of MAB problems, each based on the assumptions of the problem. In the survey [9], three different fundamental types of MAB problems were mentioned, stochastic, adversarial, and Markovian. In this paper, we are interested in the stochastic MAB problem, as it aligns with the case of rate allocation problem (the reward is stochastic). From a historical point of view, Lai and Robbins [10] introduced the first analysis of stochastic bandits with asymptotic analysis of regret. There, the principle of *optimism in the face of uncertainty* (to be optimistic while thinking about the not well explored choices) was used and the Upper Confidence Bound (UCB) algorithm was proposed. This concept is widely used in most of the MAB literature.

In UCB-like algorithms, we favor the exploration of actions with a strong potential to have an optimal value [11], and UCB measures this potential by an upper confidence bound of the reward value. Based on this type of literature, a lot of algorithms have been further proposed [12] (Sect. 2.2) and [13]. In [13], the authors proved that the proposed KL-UCB algorithm attains the optimal rate in finite-time. In addition, they proved that this algorithm is optimal for Bernoulli distributions (problems with reward of Bernoulli distribution).

Another type of algorithms widely used is based on Thompson Sampling (TS) (also known as posterior sampling and probability matching) [14]. Contrary to UCB-type algorithms, the TS algorithms are based on the assumption of posterior distribution for the unknown metric we are trying to learn. The algorithm chooses the arm which maximize the expected reward based on the current distribution. Then, after each iteration, the posterior distribution is updated. Although this type of algorithms was ignored in the academic literature until recently, several nowadays problems are using these strategies [15]. For interested readers, [16] gives a detailed discussion on when, why, and how to apply TS.

Besides UCB-type and TS-type algorithms, there are also different approaches tackling the MAB problem. In [17], rather than using the concept of *optimism in the face of uncertainty*, a new general algorithm is proposed aiming at matching the minimal exploration rates of sub-optimal arms as characterized in the derivation of the regret lower bound. In this algorithm, rather than only performing exploration and exploitation, a third process is taken into consideration as well: estimation. For simpler algorithms, ϵ-greedy is a well-known algorithm, where a fixed value $\epsilon \in [0, 1]$, decides the percentage of time you spend on exploration and exploitation. Since, with a fixed value of ϵ, we will reach a linear (not logarithmic) regret, decreasing ϵ-greedy algorithms are used, taking ϵ as a decreasing variable with time (usually it is in the form of a fraction between a constant and time). We find in literature several papers comparing the

previously mentioned algorithms, as in [18], where the power allocation problem was solved using several algorithms, as the UCB, TS and ϵ-greedy.

In our framework, there is a fixed set of Modulation and Coding Scheme (MCS) representing the available set of rates. These rates represent the possible choices of the MAB problem. Since we are considering MAMRC framework, at each frame transmission, the destination will allocate a rate for each given source. In other words, rather than selecting a single arm of the MAB, we need to select multiple arms, each corresponding to each of the multiple source nodes. Such kind of MAB problems is given under the name of Combinatorial MAB (CMAB), where a subset of arms is selected at each step, forming a *Super Arm*. In the literature, CMAB was investigated in several applications. In [19], the problem of beam selection in a vehicular network was solved using CMAB algorithms, based on TS. In [20], CMAB was also presented but this time using UCB-type algorithms. There, two applications were selected, online advertising and social influence maximization for viral marketing. In [21], Combinatorial Sleeping MAB model with Fairness constraints (CSMAB-F) was presented. The concept of sleeping arm is when some arms are not always available. In the next section, we present the system model, describing how a frame is transmitted. Then, based on this model, we present the CMAB rate allocation problem.

3 System Model

The system model is a slow-fading half-duplex OMAMRC. There are M source nodes, a single destination node, and L dedicated relay nodes. The source nodes belong to the set $\mathcal{S} = \{1, \ldots, M\}$, the relay nodes belong to the set $\mathcal{R} = \{M + 1, \ldots, M+L\}$, and we define the set of all source and relay nodes as $\mathcal{N} = \mathcal{S} \cup \mathcal{R} = \{1, \ldots, M + L\}$. In other words, a source s_i will be the node i in set $\mathcal{N}$, and a relay r_i will be the node $i+M$ in set $\mathcal{N}$. In order to explain a frame transmission in a cooperative system, the two steps of transmission should be explained, i.e., the transmission phase and the retransmission phase.

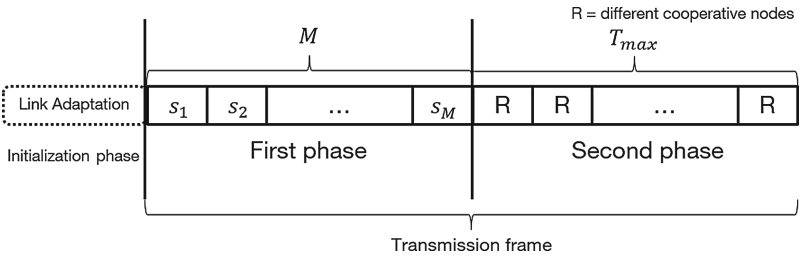

Fig. 2. Transmission of a frame: initialization, first and second phases.

As seen in Fig. 2, the frame transmission is composed of two steps, M time slots for transmission (each source of the M sources sends in one time slot), and

several time slots for retransmission (up to $T_{\max}$ time slots). At the beginning of each time slot of the retransmission phase, the scheduler (at the destination) selects a relay node to send redundancies based on its correctly decoded source messages. We call the set of correctly decoded source messages the decoding set. The selection strategy used is based on maximizing the number of correctly decoded messages at the destination. This is done by choosing the node with the highest mutual information with the destination. Note that only the nodes which were able to decode at least one source from the set of non-successfully decoded sources at the destination are possible choices for selection (nodes which do not satisfy this condition cannot help no matter what the mutual information is). In [22], it is shown that this low-complexity strategy can reach the upper bound selection strategy (based on exhaustive search approach). For that reason, we retain to this selection strategy in the paper, although other strategies might also be used. Nevertheless, the LA problem and the proposed algorithms will not change.

The messages of all sources are mutually independent. A message $\mathbf{u}_s \in \mathbb{F}_2^{K_s}$ of a source s has a length of K_s information bits, where $\mathbb{F}_2$ represents the binary Galois field. In addition, the length K_s depends on the selected Modulation and Coding Schemes for that source. We assume that we have a finite set of possible rates of size n_{MCS}. These possible rates represent the arms of the MAB problem. In the initialization phase, the phase which comes before transmission and retransmission (check Fig. 2), the destination allocates M rates for the M sources. In other words, the destination chooses a set of M arms of the n_{MCS} arms, forming a *super arm*. As a results, our MAB problem is now a CMAB problem, with n_{MCS}^{M} arms.

In the prior art, the selection of the rates was based on the information available at the destination, i.e., CSI or CDI. Here, no information is available, and rather needed to be learned. Finally, for a given transmitting node $a \in \mathcal{S} \cup \mathcal{R}$, and a receiving node $b \in \mathcal{S} \cup \mathcal{R} \cup \{d\}$, at a given channel use k, the received signal $y_{a,b,k}$ can be written as:

$$y_{a,b,k} = h_{a,b} x_{a,k} + n_{a,b,k}, \tag{1}$$

where $x_{a,k} \in \mathbb{C}$ is the coded modulated symbol whose power is normalized to unity, $h_{a,b}$ are the channel fading gains, which are independent and follow a zero-mean circularly symmetric complex Gaussian distribution with variance $\gamma_{a,b}$, and $n_{a,b,k}$ represents the independent and identically distributed AWGN samples, which follow a zero-mean circularly-symmetric complex Gaussian distribution with unit variance. During the first phase, a given channel use k belong to $\{1, \ldots, U\}$, while during the second phase, k belongs to $\{1, \ldots, Q\}$, where U and Q represent the number of channel uses in each phase respectively.

4 Problem Formulation

4.1 Objective Function

Here, we define the utility function as the spectral efficiency per frame defined as the ratio between the total number of successfully received bits and the total

number of channel uses in a given frame. The utility function depends on the vector of selected rates $\{R_i\}, i \in \{1, \ldots, M\}$ (rates we are allocating) chosen from a fixed set of possible rates. It also depends on the number of channel uses used at each transmission and retransmission phase. We define also the outage events $\mathcal{O}_{i,t}$ which occur when source i is not decoded correctly after the transmission phase ($t = 0$) and at each retransmission l up to t ($l = 1, ..., t$). We define, accordingly, the outage event indication $\mathrm{O}_{i,t}$ which takes value 1 if the event $\mathcal{O}_{i,t}$ happens, and 0 otherwise. Or, in mathematical term, for any elementary event w, $\mathrm{O}_{i,t}(w) = [w \in \mathcal{O}_{i,t}]$ where [P] denotes the Inverson bracket which takes the value 1 if P is true, and 0 otherwise. The spectral efficiency per frame can be written as

$$\begin{aligned} \eta^{frame}(\{R_i\}, \alpha) &= \frac{\text{nb bits successfully received}}{\text{nb channel uses}} \\ &= \frac{\sum_{i=1}^{M} K_i(1 - \mathrm{O}_{i,T_{max}})}{MU + QT_{used}} \\ &= \frac{\sum_{i=1}^{M} R_i(1 - \mathrm{O}_{i,T_{max}})}{M + \alpha T_{used}} \end{aligned} \tag{2}$$

where

- $R_i = K_i/U$ represents the rate of a source i,
- $\alpha = Q/U$ denotes the ratio of the channel uses of retransmission phase Q and the channel uses of the transmission phase U,
- $\mathrm{O}_{i,T_{max}}$ is the outage indication as defined above, i.e., $\mathrm{O}_{i,T_{max}} = 1$ means that source i is not decoded correctly during a frame (since the maximum number of retransmissions is T_{max}),
- $T_{used} \in \{1, \ldots, T_{max}\}$ is the number of retransmissions activated for a frame.

The value of T_{uesd} depends on the number of retransmission rounds needed for the destination to decode all the source nodes, or to state an outage event (after T_{max} retransmissions). The outage indication $\mathrm{O}_{i,T_{max}}$ is obtained from an information theory perspective. This means that we don't use practical decoding schemes (e.g., LDPC or Turbo codes), but we follow the ideal information theory assumptions. In other words, we use the knowledge of the links' states, assume as infinite codeword length with mutual information achieving (spatially distributed) channel coding codebooks, and a Maximum Likelihood (ML) decoding. Also, we assume that two successive frames are received with correlation time separation (to ensure no correlation). Finally, we formulate the analytical expression of the outage events. The detailed expressions of the common and individual outage, based on the results of [23], are presented in the Appendix *A*.

4.2 MAB Problem

After defining the utility metric, as well as the outage events, we can now formulate the considered rate adaptation problem as a MAB problem. We consider a finite set of possible arms of size n_{MCS}. At each step, a *super arm* of size M is

selected for the M source nodes included. This leads us to an equivalent CMAB of arms size n_{MCS}^{M}. The reward of each arm is a stochastic random variable, with an unknown distribution and unknown average. We define the random variable $X_i(t)$ as the reward given when we select the *super arm* i at the t^{th} transmission frame. The reward was defined before as the spectral efficiency per frame, and the randomness is within the variables T_{used} which varies between zero and T_{max}, and the outage event indications of each source node. We define the expected value of the reward of the *super arm* i as $\theta_i = \mathbb{E}[X_i(t)]$.

For a given online sequential algorithm π, where at each frame j, a decision I_j of a *super arm* i is selected ($I_j = i$), we define the regret as the difference between the rewards of the optimal algorithm (Oracle algorithm selecting the optimal arm each round) and the given algorithm. The regret of algorithm π up to transmission frame t can be written as:

$$\text{Reg}^{\pi}(t) = \theta^* t - \sum_{i=1}^{n_{\text{MCS}}^{M}} \theta_i \mathbb{E}[n_i^{\pi}(t)], \tag{3}$$

where θ^* represents the expected value of the optimal reward (i.e., the reward of the optimal *super arm* i^*), and $\mathbb{E}[n_i^{\pi}(t)]$ represents the expected value of the number of times arm i was selected after t rounds when using algorithm π. We aim to propose a rate allocation algorithm which performs exploration and exploitation in a way that minimizes this regret.

5 Algorithm

We retain here to a well-known algorithm in the literature, specifically, a UCB-like algorithm. Several types of UCB algorithms are seen in the prior art, each depending on the problem considered, the reward type, and the way we choose the upper bound. In this paper, we use the UCB1 algorithm [24], where it is known that it achieves a logarithmic regret uniformly over t and without any preliminary knowledge about the reward distributions. The only condition is to assume that the rewards are bounded in [0, 1], and this normalization can be assumed easily with no loss of generality. The sketch of the algorithm is presented in Algorithm 1.

Algorithm 1. UCB1

Initialization: For $t = 0, \ldots, n_{\text{MCS}}^{M} - 1$, for the $(t+1)^{th}$ transmission, select the super arm t+1 (play each super arm once).

UCB: For $t \geq n_{\text{MCS}}^{M}$, for the $(t+1)^{th}$ transmission, select the super arm i which maximizes $\overline{X}_i + \sqrt{\frac{2 \ln t}{n_i}}$.

After the initialization step, where each arm is explored once, we start choosing the next arms based on the information collected. We see next that the choice is based on two terms summed together, $\overline{X}_i$ representing the average reward obtained from *super arm* i up to transmission t, and the upper confidence term represented by $\sqrt{\frac{2\ln t}{n_i}}$, where n_i represents the number of times *super arm* i was selected up to transmission t. The first term, i.e., $\overline{X}_i$, gives the exploitation term, where the history rewards of the arms are taken into consideration. On the other hand, the second term, i.e., $\sqrt{\frac{2\ln t}{n_i}}$, gives the exploration term. The ratio can be understood as, when a given arm i is not selected for enough time, compared to other arms, the fraction increases, and then the index of this arm composed of the sum of the two terms increases. In this way, we tend to compromise between the history of the rewards of each arm and the number of times this arm was selected. One final comment, about the logarithmic in the expression: In UCB1, we try to decrease the exploration coefficient as time increases, trying to set a limit to the exploration phase when enough information is collected through previously selected arms. The mathematical aspect of this result is based on the Hoeffding's Inequality, a theorem applicable to any bounded distribution. In theorem 1, the expected regret of the UCB1 algorithm when played t times is presented.

Theorem 1. *For all* $n_{\mathrm{MCS}}^{M} > 1$*, if policy UCB1 is run on* n_{MCS}^{M} *machines having arbitrary reward distributions* $P_1, \ldots, P_{n_{\mathrm{MCS}}^{M}}$ *with support in* $[0, 1]$*, then its expected regret after any number* t *of plays is at most:*

$$8 \sum_{i:\theta_i<\theta^*} \left(\frac{\ln t}{\Delta_i}\right) + \left(1 + \frac{\pi^2}{3}\right) \left(\sum_{j=1}^{n_{\mathrm{MCS}}^{M}} \Delta_j\right),$$

where $\theta_1, \ldots, \theta_{n_{\mathrm{MCS}}^{M}}$ *are the expected values of* $P_1, \ldots, P_{n_{\mathrm{MCS}}^{M}}$*, and* Δ_i *is defined as:*

$$\Delta_i = \theta^* - \theta_i.$$

Proof. The proof of the above upper bound is omitted and can be found in [24].

In practice, the proposed algorithm suffers mainly from the exponential growth of arms. Specifically, the initialization phase (pure exploration phase) will take too much time before reaching the exploitation-exploration phase. Thus, we propose an Approximated UCB1 (AUCB1) algorithm, which reduces the complexity of the initialization phase.

The goal of the initialization phase is to explore each *super arm* once, and to set its index, we call index the sum $\overline{X}_i + \sqrt{\frac{2\ln t}{n_i}}$. We propose here setting an approximated initial index in order to decrease the complexity of the initialization phase. One way in doing so is by removing the exponential relationship between the sources forming the *super arm*. In other words, rather than taking *super arms* initially, we take each arm by itself (each possible rate), and we test this arm with all the possible sources. In this case, when a source is sending with

a given rate, other sources send nothing. We repeat this process for a given arm with all the given sources. Finally, we average for this arm the number of transmitted bits (Rate × success or failure), and we save the highest T_{used} needed with all the sources. We repeat this process for all arms (rates). Finally, for each *super arm* composed of M subset of arms, we calculate the reward (index) as the average of transmitted bits divided by the number of channel uses while using the highest T_{used} of the considered subset of arms (rates). Following these steps, we approximate the reward (recall Eq. 2). The complexity of the initialization phase is reduced from $O(n_{\text{MCS}}^{M})$ to $O(n_{\text{MCS}} \times M)$. For brevity, we omit here the step-by-step AUCB1 algorithm, as it will only be an initialization step in the proposed Sequential UCB1 (SUCB1) algorithm presented next.

In SUCB1, the idea is to generalize the AUCB1 algorithm for all iterations rather than only the initialization step. After setting the indices of each arm using AUCB1, SUCB1 chooses each *super arms* successively, arm by arm. In other words, instead of choosing the *super arm* directly, we choose for each source of the M sources the arm with the highest index. After each selection, we update the indices' counter. Finally, we update the indices based on the cumulative reward, each based on decoding the signal of the related source. In SUCB1, we have n_{MCS} arms, rather than n_{MCS}^{M} arms, and this reduction will decrease the regret as we will see in the numerical results section. The sketch of the algorithm is presented in Algorithm 2.

Algorithm 2. SUCB1

Initialization: For $t = 0, \ldots, n_{\text{MCS}} \times (M-1)$, for the $(t+1)^{th}$ transmission, initialize the arms indices following the steps of *AUCB1*

SUCB: For $t \geq n_{\text{MCS}} \times M$, for the $(t+1)^{th}$ transmission, select the super arm i successively, arm by arm, for each of the M sources as:

- select the arm which maximizes $\overline{X}_i + \sqrt{\frac{2 \ln t}{n_i}}$.
- update n_i
- repeat for all sources within M to reach the super arm i

6 Numerical Results

In this section, we validate the learning algorithms with three source nodes and three relay nodes, while using 4 possible retransmissions in the second phase ($T_{max} = 4$) and $\alpha = 1/2$. We assume independent Gaussian distributed channel inputs (with zero mean and unit variance), with $I_{a,b} = \log_2(1 + |h_{a,b}|^2)$. Note that some other formulas could be also used for calculating $I_{a,b}$ but they would not have any impact on the basic concepts of this work. There are many factors to investigate: links configuration, SNR levels, and different MCS families.

Due to brevity, and after carefully checking different possible scenarios, we present the results of symmetric link configuration (SNR of all channel links

is symmetric). Three different levels of SNR will be considered, specifically, SNR = $\{-4, 6, 21\}$ dB. The importance of choosing the different SNR links, is that the optimal rate allocation (the Oracle allocation) is different at each SNR level. Following the discrete MCS family whose rates belong to the set {0.5;1;1.5;2;2.5;3;3.5} [b.c.u], the Oracle rate allocation of sources $\{s_1, s_2, s_3\}$ will be $\{1, 1, 1\}$, $\{3, 3, 2.5\}$, and $\{3.5, 3.5, 3.5\}$ respectively to the SNR set investigated.

In Figs. 3, 4, and 5, we see the regret analysis of the three different SNR levels. For clarity of the results, we present the regret in the form of percentage loss with respect to the optimal efficiency. In other words, we compare the efficiency of the algorithms as a ratio of the rewards of the algorithms and the Oracle. In Fig. 3, for SNR = -4 dB, we see that the three algorithms are featuring a close regret level (up to 25% loss after 1000 samples). Next, in Fig. 4, for SNR = 6 dB, we see a great improvement with using SUCB1 (reaching 90% of the optimal reward), as compared to UCB1 and AUCB1 which act closely as in the case when $\gamma = -4$ dB. In Fig. 5, the same result is seen for SNR = 21 dB, where SUCB1 is outperforms other algorithms, while AUCB1 is slightly better than UCB1. Finally, in Fig. 6, we present the Average Spectral Efficiency (ASE), for the different SNR levels between -5 and 15 dB after 500 samples (larger numbers of samples were investigated and gave the same results). We see that the proposed SUCB1 algorithm approaches the upper bound (the Oracle) while outperforming UCB1 and AUCB1.

7 Conclusion

In this paper, we investigate the LA of OMAMRC using an online learning framework, MAB. First, we formulate the system model as a MAB problem. Then, we adapt to the UCB-type family, specifically the UCB1 algorithm. In order to solve the problem of complexity of exponential number of arms included in the MAMRC system, a sequential algorithm SUCB1 is proposed. Within SUCB1, we use an approximated initialization phase AUCB1, then, we choose arms sequentially for the considered set of sources. The numerical results show that the proposed algorithm outperforms the traditional UCB1 algorithm in terms of regret and average spectral efficiency.

Appendix

A: Outage Events

Based on [23] proposition 1, we see a direct relation between the individual outage and the common outage. The individual outage is defined as the event that an individual source is not decoded correctly at the destination after T_{max} rounds. Similarly, common outage is defined for a set of sources, and it is declared when at least one of the sources within this set is not decoded correctly at the

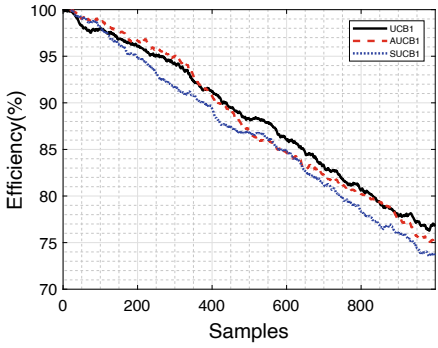

Fig. 3. Efficiency for $\gamma = -4$ dB

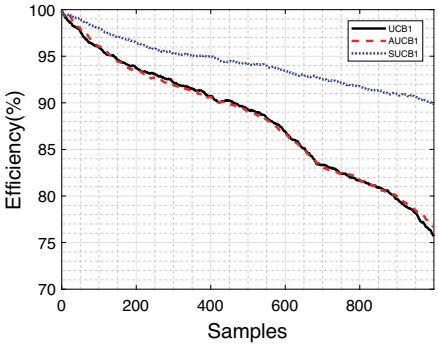

Fig. 4. Efficiency for $\gamma = 6$ dB

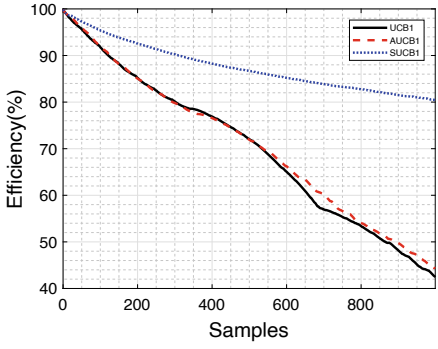

Fig. 5. Efficiency for $\gamma = 21$ dB

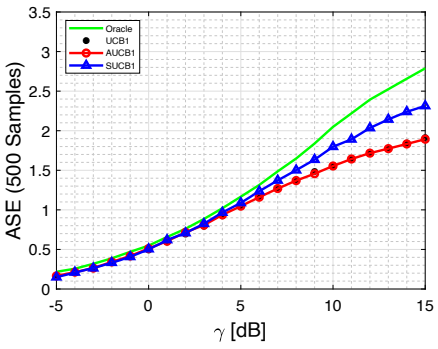

Fig. 6. ASE vs γ after 500 samples

destination. In other words, common outage of a set occurs when one or more of its source nodes are in an individual outage.

Both, the individual outage event $\mathcal{O}_{s,t}(a_t, \mathcal{S}_{a_t,t-1}|\mathcal{P}_{t-1})$ of a source s after round t, and the common outage event $\mathcal{E}_t(a_t, \mathcal{S}_{a_t,t-1}|\mathcal{P}_{t-1})$ after round t, depend directly on the rate being scheduled. In addition, they depend on the selected node $a_t \in \mathcal{N}$ and its associated decoding set $\mathcal{S}_{a_t,t-1}$. They are conditional on the knowledge of $\mathcal{P}_{t-1}$, where $\mathcal{P}_{t-1}$ denotes the set collecting the nodes $\widehat{a}_l$ which were selected in rounds $l \in \{1, \ldots, t-1\}$ prior to round t together with their associated decoding sets $\mathcal{S}_{\widehat{a}_l,l-1}$, and the decoding set of the destination $\mathcal{S}_{d,t-1}$ ($\mathcal{S}_{d,0}$ is the destination's decoding set after the first phase).

Analytically, the common outage event of a given subset of sources is declared if the vector of their rates lies outside of the corresponding MAC capacity region. For some subset of sources $\mathcal{B} \subseteq \overline{\mathcal{S}}_{d,t-1}$, where $\overline{\mathcal{S}}_{d,t-1} = \mathcal{S} \setminus \mathcal{S}_{d,t-1}$ is the set of non-successfully decoded sources at the destination after round $t-1$, and for a candidate node a_t this event can be expressed as:

$$\mathcal{E}_{t,\mathcal{B}}(a_t, \mathcal{S}_{a_t,t-1}) = \bigcup_{\mathcal{U}\subseteq\mathcal{B}} \Big\{ \sum_{i\in\mathcal{U}} R_i > \sum_{i\in\mathcal{U}} I_{i,d} + \alpha \sum_{l=1}^{t-1} I_{\widehat{a}_l,d}\mathcal{C}_{\widehat{a}_l}(\mathcal{U}) + \alpha I_{a_t,d}\mathcal{C}_{a_t}(\mathcal{U}) \Big\}, \tag{4}$$

where $I_{a,b}$ denotes the mutual information between the nodes a and b (the mutual information is defined based on the channel inputs, check Sect. 6 for Gaussian inputs example), and where $\mathcal{C}_{\widehat{a}_l}$ and $\mathcal{C}_{a_t}$ have the following definitions:

$$\begin{aligned} \mathcal{C}_{\widehat{a}_l}(\mathcal{U}) &= \Big[(\mathcal{S}_{\widehat{a}_l,l-1} \cap \mathcal{U} \neq \emptyset) \wedge (\mathcal{S}_{\widehat{a}_l,l-1} \cap \mathcal{I} = \emptyset) \Big], \\ \mathcal{C}_{a_t}(\mathcal{U}) &= \Big[(\mathcal{S}_{a_t,t-1} \cap \mathcal{U} \neq \emptyset) \wedge (\mathcal{S}_{a_t,t-1} \cap \mathcal{I} = \emptyset) \Big]. \end{aligned} \tag{5}$$

The individual outage event of a source s after round t for a candidate node a_t can be defined as:

$$\begin{aligned} \mathcal{O}_{s,t}(a_t, \mathcal{S}_{a_t,t-1}) &= \bigcap_{\mathcal{I}\subset\overline{\mathcal{S}}_{d,t-1}, \mathcal{B}=\overline{\mathcal{I}}, s\in\mathcal{B}} \mathcal{E}_{t,\mathcal{B}}(a_t, \mathcal{S}_{a_t,t-1}), \\ &= \bigcap_{\mathcal{I}\subset\overline{\mathcal{S}}_{d,t-1}} \bigcup_{\mathcal{U}\subseteq\overline{\mathcal{I}}: s\in\mathcal{U}} \Big\{ \sum_{i\in\mathcal{U}} R_i > \sum_{i\in\mathcal{U}} I_{i,d} + \alpha \sum_{l=1}^{t-1} I_{\widehat{a}_l,d}\mathcal{C}_{\widehat{a}_l}(\mathcal{U}) + \alpha I_{a_t,d}\mathcal{C}_{a_t}(\mathcal{U}) \Big\}, \end{aligned} \tag{6}$$

where $\overline{\mathcal{I}} = \overline{\mathcal{S}}_{d,t-1} \setminus \mathcal{I}$.

References

1. Kramer, G., Marić, I., Yates, R.D.: Cooperative communications (2007)
2. Kramer, G., Gastpar, M., Gupta, P.: Cooperative strategies and capacity theorems for relay networks. IEEE Trans. Inf. Theory **51**(9), 3037–3063 (2005)
3. Sankaranarayanan, L., Kramer, G., Mandayam, N.B.: Hierarchical sensor networks: capacity bounds and cooperative strategies using the multiple-access relay channel model. In: 2004 First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, SECON 2004, pp. 191–199. IEEE (2004)
4. Laneman, J.N.: Cooperative diversity in wireless networks: algorithms and architectures. Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA (2002)
5. Lim, S.H., Kim, Y.H., Gamal, A.E., Chung, S.Y.: Noisy network coding. IEEE Trans. Inf. Theory **57**(5), 3132–3152 (2011)
6. Avestimehr, A.S., Diggavi, S.N., Tse, D.N.C.: Wireless network information flow: a deterministic approach. IEEE Trans. Inf. Theory **57**(4), 1872–1905 (2011)
7. Cover, T., Gamal, A.E.: Capacity theorems for the relay channel. IEEE Trans. Inf. Theory **25**(5), 572–584 (1979)
8. Khansa, A.A., Cerovic, S., Visoz, R., Hayel, Y., Lasaulce, S.: Slow-link adaptation algorithm for multi-source multi-relay wireless networks using best-response dynamics. To be presented at NetGCOOP (2021)
9. Bubeck, S., Cesa-Bianchi, N.: Regret analysis of stochastic and nonstochastic multi-armed bandit problems. arXiv preprint arXiv:1204.5721 (2012)

10. Lai, T.L., Robbins, H.: Asymptotically efficient adaptive allocation rules. Adv. Appl. Math. **6**(1), 4–22 (1985)
11. Weng, L.: The multi-armed bandit problem and its solutions. lilianweng.github.io/lil-log (2018)
12. Bubeck, S.: Bandits games and clustering foundations. Ph.D. dissertation, INRIA Nord Europe (2010)
13. Garivier, A., Cappé, O.: The KL-UCB algorithm for bounded stochastic bandits and beyond. In: Proceedings of the 24th Annual Conference on Learning Theory. In: JMLR Workshop and Conference Proceedings, pp. 359–376 (2011)
14. Thompson, W.R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika **25**(3/4), 285–294 (1933)
15. Chapelle, O., Li, L.: An empirical evaluation of Thompson sampling. Adv. Neural Inf. Process. Syst. **24**, 2249–2257 (2011)
16. Russo, D., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z.: A tutorial on Thompson sampling. arXiv preprint arXiv:1707.02038 (2017)
17. Combes, R., Magureanu, S., Proutiere, A.: Minimal exploration in structured stochastic bandits. arXiv preprint arXiv:1711.00400 (2017)
18. Ameur, W.B., Mary, P., Hélard, J.-F., Dumay, M., Schwoerer, J.: Autonomous power decision for the grant free access MUSA scheme in the mMTC scenario. Sensors **21**(1), 116 (2021)
19. Nasim, I., Ibrahim, A.S., Kim, S.: Learning-based beamforming for multi-user vehicular communications: a combinatorial multi-armed bandit approach. IEEE Access **8**, 219 891–219 902 (2020)
20. Chen, W., Wang, Y., Yuan, Y.: Combinatorial multi-armed bandit: general framework and applications. In: International Conference on Machine Learning. PMLR, pp. 151–159 (2013)
21. Kuchibhotla, V., Harshitha, P., Elugoti, D.: Combinatorial sleeping bandits with fairness constraints and long-term non-availability of arms. In: 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1575–1581. IEEE (2020)
22. Cerovic, S., Visoz, R., Madier, L., Berthet, A.O.: Centralized scheduling strategies for cooperative HARQ retransmissions in multi-source multi-relay wireless networks. In: Proceedings of IEEE ICC 2018, Kansas City, MO, USA, May 2018
23. Mohamad, A., Visoz, R., Berthet, A.O.: Outage analysis of various cooperative strategies for the multiple access multiple relay channel. In: IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). IEEE 2013, pp. 1321–1326 (2013)
24. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Mach. Learn. **47**(2), 235–256 (2002)