Optimizing Spectrum Use in Wireless Networks by Learning Agents

Artur Popławski[1,2]
[1]*NOKIA Kraków Technology Center*
[2]*AGH University of Science and Technology*
Kraków, Poland
artur.poplawski@nokia.com

Szymon Szott
AGH University of Science and Technology
Kraków, Poland
szott@kt.agh.edu.pl

***Abstract*—The paper examines several distributed mechanisms that can be used in wireless networks consisting of base stations transmitting onto a population of receivers. The overall goal of the algorithms is to optimize a global measure: the sum of capacities of the channels formed between transmitters and receivers in the presence of interference. We introduce a mathematical model of the operation of an OFDM-based wireless network and on this premise we pose the problem of interference in a game theoretic setting. We propose a simple but expressive way of modeling "smart devices" as learning and executing agents and introduce three types of such agents. This part of the work is carried out in the scope of the multi-armed bandit framework. One of the three algorithms we propose is well known and widely studied and two others are new variants of known algorithms seemingly not yet studied. By numerical simulations we show that these mechanisms improve network performance in the considered model. We offer some basic heuristic explanations of this improvement and identify future work.**

***Index Terms*—game theory, wireless networks, spectrum allocation**

I. Introduction

Recent OFDM-based technologies like LTE or NR use advanced mechanisms (such as effective modulation and coding techniques) to increase the spectral efficiency of wireless transmissions. Nevertheless, not all aspects of high efficiency transmissions are currently standardized and there is the potential for improving network efficiency by *optimizing network operation*. This can be achieved by having smart devices perform self-tuning based on the observation of their current performance.

The behavior of such smart devices can be analyzed within the framework of game theory. In particular the multi-armed bandit approach has been studied intensively in the context of machine learning and the stochastic decision processes [1]. It differs from the classical game-theoretic approach to learning (cf. [2]) by the lack of necessity of tracking of other players' decisions in the algorithm governing a given player's behavior. This eliminates the need for direct cooperation and simplifies design. In this work we adopt the multi-armed bandit approach to study the problem of interference management [3].

We first explain the basic mathematical notation and terms used in the paper (Section II) and then proceed to introduce a mathematical model of the operation of an OFDM-based wireless network (Section III). On this premise, we pose the problem of interference in a game-theoretic setting. In Section IV we propose a simple but effective way of modeling smart devices as learning agents and introduce three types of such agents. The work is carried out in the scope of the multi-armed bandit framework. One of the proposed algorithms seems to be not discussed in the literature and another one, based on Thompson sampling, is proposed in a variant slightly different compared with what is known from the literature (cf. [1]) and inspired by the Kolmogorov-Smirnov test. In Section V, we present the assumptions and results of a simulation analysis of the game described in Section III where players behave according to algorithms introduced in Section IV. Section VI discusses the theoretical justification of the simulation results using heuristic terms. Finally, the last section discusses open questions and future research.

This work was partially supported by the Polish Ministry of Science and Higher Education with the subvention funds of the Faculty of Computer Science, Electronics and Telecommunications of AGH University.

Annex to ISBN 978-3-903176-28-7

II. Notation

Before proceeding with a description of our model, we briefly outline the mathematical notation used. First, by $\prod_{i\in I} X_i$ we denote a Cartesian product of the family of sets indexed by I. For $x \in \prod_{i\in I} X_i$ and $j \in I$, x_j is a projection of x on X_j and x_{-j} is a projection of x on $\prod_{i\in(I-\{j\})} X_i$. We denote the set of all functions from X to Y as Y^X. When $X \subset Y$ the $\mathbb{I}_X : Y \to \{0,1\}$ is an indicator function, i.e., $\mathbb{I}_X(y) = 1 \iff y \in X$. Finally, with any given X and probability distribution ds on X, we will associate a "pseudo function" $sample_{ds}$ which, whenever called returns an element of X drawn randomly according to distribution ds. If we know X and want to sample from this set according to a uniform distribution we simply write $sample(X)$.

III. Game-theoretic setting

We address the basic problem of downlink performance in wireless networks within the framework of game theory. Ultimately, we are interested in effective algorithms that can be applied locally in each of the transmitting entities in

order to increase network performance. We limit ourselves to algorithms that are based on choosing strategies in games that naturally arise when a system of interacting transmitters is considered.

The considered telecommunication problem is the following: we have many transmitters which share the same resources (radio spectrum) and send data to receivers assigned to them. We assume that each agent possesses knowledge only about their own performance and past behavior.

First, let us define the game Γ as a triple:

$$\Gamma = (P, \{S_p\}_{p\in P}, \{u_p\}_{p\in P}),$$

where P is a finite set of players (agents), S_p is a finite set of strategies available for each player $p \in P$. The individual payoff function for each player is $u_p : \prod_{l\in P} S_l \to \mathbb{R}$.

We model the network as an interference game with the set P of the downlink transmitters using OFDM transmissions in a common band. We also have a population of terminals which are the receivers of these transmissions. We assume that both transmitters and receivers are immobile and the environment is static (i.e., the radio channel between each transmitter and receiver can be considered as constant in time). We also assume that each receiver is assigned to a single transmitter and that this does not change in time.

We divide the band designated for transmission into K disjoint parts:

$$B = \bigcup_{i=0,\ldots,K-1} b_i.$$

Subsets of $\{b_0, \ldots, b_{K-1}\}$ are strategies available for transmitters. The interaction between transmitters occurs through interference and the payoffs can be expressed in terms of the capacity of the channels between the transmitter and receivers assigned to it. Thus, for each $p \in P$:

$$S_p \subset \{s | s \subset \{b_0, \ldots, b_{K-1}\}\}$$

Furthermore, we denote by $Home(p)$ the set of receivers assigned to transmitter p. For transmitter p and receiver ue and $m \in B$, the *channel* between p and ue (assuming a sufficiently long channel coherence time and sufficiently wide coherence bandwidth) can be expressed by a complex number $h_{p,ue,m}$. Using this notation and following the standard formula for the capacity of a Gaussian channel with interference, we define the payoff of player $p \in P$ as:

$$u_p(s) = \frac{1}{|Home(p)|} \sum_{ue\in Home(p)} \left(\frac{BW}{|B|} \times \sum_{m\in s_p} log\left(1 + \frac{|h_{p,ue,m}|^2}{n + \sum_{p'\in P-\{p\}} \mathbb{I}_{s_{p'}}(m)|h_{p',ue,m}|^2}\right)\right), \quad (1)$$

where we have assumed the same normalized power for each of the transmitters, the same level of Gaussian noise n and the same bandwidth $b_i = \frac{BW}{|B|}$ for each $b_i \in B$. We also assume that there is at least one receiver connected to each transmitter ($Home(p) \neq \emptyset$ for all p), so we do not need to worry about the denominator in (1). Intuitively, the payoff is an average of the instant capacities of the channels to all receivers handled by the transmitter taking into account interference from all of the other transmitters using the same part of the spectrum.

We assume that $K = 2$, the bandwidth of b_0 is the same as bandwidth of b_1, and there is a flat radio channel between transmitter and receiver across the whole bandwidth, i.e., $h_{p,ue,b_0} = h_{p,ue,b_1}$ for each transmitter p and receiver ue. Furthermore, we assume the following variant of the game where, for each $p \in P$, $S_p = \{\{b_0\}, \{b_1\}, \{b_0, b_1\}\}$.

An important observation is that assuming the presence of receivers, playing $\{b_0, b_1\}$ is always a dominating strategy. Using all of the spectrum under the assumption that other transmitters stay with their strategies is always better than using only part of it.

To measure total network performance we consider the sum of payoffs of all the players. Thus, we define the welfare function as a mapping $w : \prod_{p\in P} S_p \to \mathbb{R}$, given by the equation

$$w(s) = \sum_{p\in P} u_p(s).$$

Finally, we assume games are played in sequence. This, in the context of wireless games, is a natural assumption. For example, in LTE networks time is quantified and divided into TTIs (Transmission Time Intervals) and the decision about allocating spectral resources in each TTI corresponds to a single play.

IV. Learning Players as Local Optimizers

A. Model

Our general objective to find methods that would allow the optimization of the welfare function in the network over the players' strategy space. We limit ourselves to the class of algorithms that can be executed by each player independently and where the decision about the strategy is based on the observation of player performance and the "local environment".

We will define algorithms according to the following scheme. We assume that with each player there is associated some set of states representing "memory" M_p. Gathering experience or "learning" will be modeled by the function

$$learnM_p : M_p \times S \to M_p,$$

where $S = \prod_{i\in P} S_i$.

The decision of a player is a function of the memory and is modeled by

$$select_p : M_p \to S_p.$$

Since we want to consider also algorithms that use randomness, we slightly abuse the function definition here. In such cases "functions" $learnM$ and $select$ should be understood as sampling from some probability measure on the appropriate sets.

Now, if we denote by $M = \prod_{p \in P} M_p$ we can define $ev : M \to M$ as

$$ev(m) = \left(\prod_{p \in P} learnM_p \circ \prod_{p \in P} select_p \right)(m).$$

We can consider the history or trajectory of the game, i.e., a sequence of iterations $(m_i)_{i \in \mathbb{N}}$ so that $m_{i+1} = ev(m_i)$. We can also consider the history of the game in terms of choices as $(s_i)_{i \in \mathbb{N}} = ((select_p)_{p \in P}(m_i))_{i \in \mathbb{N}}$ where m_i are as above.

With these definitions one can define the average performance of the process at the nth step in terms of welfare w as

$$q(n) = \frac{\sum_{i=0}^{n-1} w(s_i)}{n}.$$

This value depends on the initial condition s_0 or, since we consider the process to be driven by the evolution in the memory space, on the initial condition m_0.

B. Players

Every time we describe the algorithm for a player, we define appropriate memory spaces and "functions" $learnM$ and $select$.

1) ϵ-greedy player: In this case $M_p = \mathbb{R}^{S_p}$ and the parameters are the probability of exploration $\epsilon \in [0,1]$ and $\alpha \in [0,1]$. We assume only knowledge of outcome of the game and define:

$$learnM_p(m_p, s) = m_p + (\alpha - 1) m_p \mathbb{I}_{\{s_p\}} + (1-\alpha) u_p(s) \mathbb{I}_{\{s_p\}},$$

which amounts to updating the remembered function by modifying the value at the element s_p to be $\alpha m_p(s_p) + (1-\alpha) u_p(s)$. The $select$ function is probabilistic and is defined as

$$select_p(m_p) = \begin{cases} sample(\arg\max_{x \in S_i} h(x)), & \text{for } sample([0,1]) \geq \epsilon, \\ sample(S_p), & \text{for } sample([0,1]) < \epsilon. \end{cases}$$

2) Greedy with auto cooling player: This is a variant of ϵ-greedy in which the probability of exploration is adaptive. We define these internal parameters: $\epsilon \in [0,1]$ being the initial value for the parameter controlling the probability of exploration, $\alpha \in [0,1]$ controlling how fast estimation of the value of the strategy changes, $\gamma \in [0,1]$ being the maximal allowed probability of exploration, and $\eta \in [0,1]$ controlling how fast the probability of the exploration changes. We have:

$$M_p = \mathbb{R}^{S_p} \times \mathbb{R} \times [0,1],$$

$$learnM_p((e_p, l, \pi), s) = (e'_p, u_p(s), \pi'),$$

where

$$e'_p = e_p + (\alpha - 1) e_p \mathbb{I}_{\{s_p\}} + (1-\alpha) u_p(s) \mathbb{I}_{\{e_p\}}$$

is the update known from ϵ-greedy and

$$\pi' = \begin{cases} \eta\pi + (1-\eta)\gamma, & \text{if } (s_p \notin \arg\max_{x \in S_p} e_p(x) \text{ and} \\ & u_p(s) > e_p(s_p)) \text{ or} \\ & (s_p \in \arg\max_{x \in S_p} e_p(x) \text{ and} \\ & u_p(s) < e_p(s_p)) \\ \eta\pi, & \text{otherwise} \end{cases}$$

Now we define:

$$select_p((e_p, l, \pi)) = \begin{cases} sample(\arg\max_{x \in S_p} e_p(x)), & \text{for } sample([0,1]) \geq \pi \\ sample(S_p), & \text{for } sample([0,1]) < \pi \end{cases}$$

Instead of a fixed value of ϵ as in ϵ-greedy, in this case we maintain a variable π describing the probability of deviating in the choice of the strategy from what is currently considered the best choice. π is modified according to the "level of correctness" of the current optimal choice. i.e., if the currently selected strategy is different then current optimal but payoff is higher than estimated for the current optimal the probability of exploration is increased. Similarly, it is increased when for the currently considered optimal strategy one obtains a value below the estimation. In any other case the probability of exploration is decreased.

3) Thompson sampling player: The last considered algorithm is a Thompson sampler. The idea behind Thompson sampling is to treat the payoff associated with each strategy as a random variable. The distribution of this variable is unknown, however, each time players choose a strategy and receive a payoff, they update their empirical distribution associated with this strategy. The choice of strategy is done first by sampling for each strategy from the empirical distribution associated with each strategy and choosing the strategy with the highest value of the sample (or break the tie by sampling uniformly if more then one is highest). Typically, in Thompson sampling empirical distributions are chosen to be from some parametrized family of distributions (e.g., the beta distribution [1]). We use a different approach, which we will call a non-parametric Thompson sampler.

For any set X we denote $X^* = \bigcup_{n \in \mathbb{N}} X^n$. Now, we define $M_p = (\mathbb{R}^*)^{S_p}$ and $learnM_p(f, (s_l)_{l \in P})) = g$, where

$$g(r) = \begin{cases} (f(r), u_p(s)) & \text{for } r = s_p, \\ f(r) & \text{for } r \neq s_p \end{cases}$$

and we use the natural identification $X^n \times X \sim X^{n+1}$. So, any time a player plays some strategy it notes the payoff associated with the strategy.

Now let us assign to $x \in \mathbb{R}^*$ the distribution $d(x)$ by the following formula:

$$d((x_0, \ldots, x_{n-1})) = \frac{1}{n} \sum_{i<n} \mathbb{I}_{\{x_i\}}$$

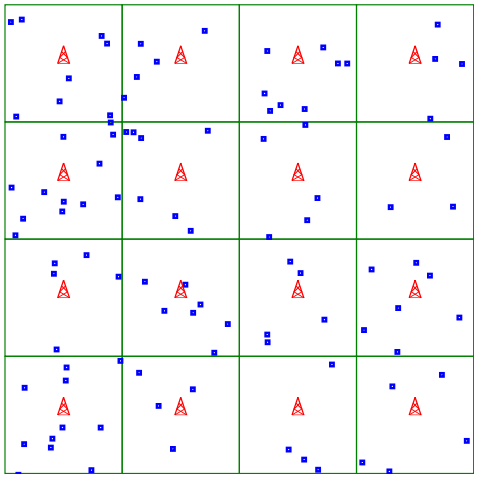

Fig. 1. The simulated topology. Icons of towers represent transmitters, small rectangles represent receivers, and lines represent division of the grid which determines the assignment of receivers to transmitters.

This is the probability distribution on $\mathbb{R}$ concentrated only in the set $\{x_0, \ldots, x_{n-1}\}$.

Now we can define $select_p$ for the player:

$$select_p(f) = sample\left(\max \arg_{s_p \in S_p} sample_{d(f(s_p))}\right)$$

The outermost sampling comes from the fact that the $\arg\max$ operator returns a set (finite in this case). When a singleton is returned, sampling is trivial and returns the only element in the set.

V. Results

In this section we present results of a simulation of the game where all players are of one of the types described in Section IV-B. The game is be the interference game $\Gamma_{\{b_0\},\{b_1\},\{b_0,b_1\}}$ as defined in Section III. To fully specify the game one needs to specify the position and power of transmitters (that constitute the player set P) and the set of receivers as well as the assignment of receivers to transmitters. The last component defines the channels between the transmitters and receivers.

In all simulations we assumed transmitters using equal power, regardless of selected band. 16 transmitters are placed in the centers of squares constituting a rectangular 4×4 grid. There are 100 receivers in positions chosen randomly with a uniform distribution over the whole grid. We assume a natural rule of assignment of receiver to transmitter by selecting the nearest one (Fig. 1). We do not assume any channel effects except white noise of a constant power and the attenuation being proportional to the square of the distance between transmitter (or source of interference) and receiver.

The results of the simulations are presented in Fig. 2. Besides the average welfare as a function of the number of steps for the considered algorithms, we present the average welfare for the case when all the players play a completely random strategy and when all players play the dominant strategy.

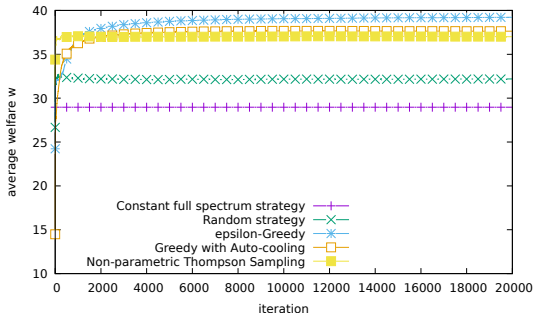

Fig. 2. Comparison of the performance of players of various types for the $\Gamma_{\{b_0\},\{b_1\},\{b_0,b_1\}}$ game.

In such a simple setting one can observe significantly higher welfare when players are learning players of any of the discussed types comparing not only to the set of players using a random strategy, but also to the set of players playing the dominant strategy ($\{b_0, b_1\}$ in this case). This means, if the results generalize to a more realistic scenario, that with a relatively small investment in the algorithm, one can achieve good overall performance of the network (with less power consumption) even in the case when there is no direct cooperation between players.

VI. Discussion

There are significant difficulties in analyzing even such a simple setting as introduced in the previous section. From a formal point of view, the quantity we are studying here (see Section IV) can be seen as the average value of the function over a realization of the random process. Although we are using the trajectory of the game in this definition, it is better to reason in terms of the function ev defined on the global memory space. In this case, one should consider instead w evaluated on the actual selection of the strategies random function $\mathbb{E}_{\circ \prod_{p \in P} \overline{select_p}(m)}[w]$. It can be shown that the average of this function over a large number of realizations of the process defined by $\overline{ev}$ is the same as the average of w over a large number of realization of selections of s. Furthermore, in all analyzed cases we do not deal with a finite memory space. Also, although all considered players realize a Markov process given by ev, this process is not ergodic. Thus, in none of the cases can we use an "off-the-shelf" theorem to show that in the limit the process will behave according to any invariant measure.

We will argue that in the considered game with learning players, the players will be biased towards playing a dominant strategy. Obviously, by the very construction of the algorithms, players will deviate from playing only the dominant strategy.

Intuitively we may expect some long term equilibrium scenario, so, for further considerations we will assume that an invariant measure is finally reached (this is not necessarily a unique measure independent of the realization of the process). Also, one can show that, in any moment for two individual players, the selection of the strategy is independent.

Consider the case of the ϵ-greedy algorithm. For a given player $p \in P$ and $s_i \in S_p$, the value of $m_p(s_i$ is a random variable. Since the selection of strategies for all players is independent and the probability of selection is $\geq \epsilon$, in the long run, from the rule of updating $m_p(s_i)$ we should have $\mathbb{E}[m_p(s_i)] = \mathbb{E}[\alpha m_p(s_i) + (1-\alpha)u_p(s_i, s_{-i})]$, where expectation is over the joint probability of S_{-i} under the invariant measure. This means that $\mathbb{E}[m_p(s_i)] = \mathbb{E}[u_p(s_i, s_{-i})]$. It is easy to observe that this also means, for this specific game, that $\mathbb{E}[m_p(\{b_1\})] + \mathbb{E}[m_p(\{b_2\})] = \mathbb{E}[m_p(\{b_0, b_1\})]$. Since we have a symmetry between players and the same reasoning can be applied to other players, one can infer that an invariant measure corresponding to the equilibrium in the long run will prefer playing the dominant strategy. Similar reasoning can lead to the same conclusion for the greedy with auto cooling algorithm.

The Thompson sampler requires a different approach. We assume that the game is played sufficiently long time, not only to be close to equilibrium, but also so that the empiric distributions associated (in our variant) with each of the players' strategies be close enough to their limits. First observe that in the long term, the cumulative distribution function (CDF) F_i of the empirical distribution associated with dominating strategy s_i with high probability is less or equal point-wise comparing to other CDFs associated with the distribution functions for other strategies. This is because, taking any $j \neq i$ $\{z : u_p(s_i, z) \leq x\} \subset \{z : u_p(s_j, z) \leq x\}$. So, if we assume that the equilibrium measure restricted to the choices of another player is μ, we have: $F_i(x) \sim \mu(\{z : u_p(s_i, z) \leq x\}) \leq \mu(\{z : u_p(s_j, z) \leq x\}) \sim F_j(x)$. For the sake of simplicity of calculations we operate in the scope of distribution theory. For any i distribution corresponding to CDF, F_i is in this case F_i'. It can be shown that for $i \neq j$ the probability of drawing a larger sample according to distribution corresponding to s_i then to s_j is $\geq \frac{1}{2}$. This further implies that the probability of drawing a maximal number from the distribution associated to s_i which corresponds to the probability that the stochastic function $select$ returns s_i is highest among all strategies.

On the other hand, numerical experiments reveal that solutions, where players play a dominant strategy most of the time but randomly and independently deviate from it (with some appropriately small probability), lead to solutions of similar overall quality as those achieved by learning players. For the game described here, under the same assumptions as in the simulation, the dependency of the average welfare w on the *perturbation*, i.e., deviation from the dominant strategy (understood as the probability of playing a different strategy by the player) is presented in Fig. 3.

As a consequence, at least in such a simple model of the network as considered here, a strategy based on just simple giving up maximal allocation randomly with some probability (a "dumb" strategy) may be comparable to more intelligent behavior based on the multi-armed bandit model. This is not an argument against the usefulness of implementing intelligent agents to manage cells in the network. It is not entirely clear,

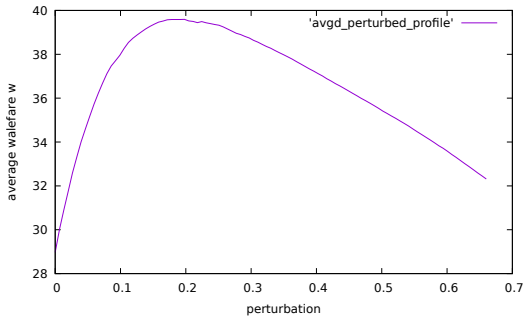

Fig. 3. Dependency of the welfare on the perturbation from dominant strategy for the $\Gamma_{\{b_0\},\{b_1\},\{b_0,b_1\}}$ game.

e.g., what should be the size of perturbation from the dominant strategy. Algorithms based on the multi-armed bandit model may find close to optimal equilibria.

VII. Conclusions and Further Research

This work should be treated as a preliminary study. Even in the considered subclass of algorithms from the area of multi-bandit solvers, there remain many algorithms in the literature to be verified from the point of view of their efficiency in selecting good solutions. Furthermore, assuming any form of cooperation (e.g., informing neighbors of current payoffs or choice of strategy) leads to a rich variety of possible designs of interactions and opens new interesting possibilities. We anticipate that full power of the intelligent learning agents' controlling the transmission strategy in cells will become more visible exposed in these, more sophisticated designs.

The optimization objective in the presented approach was done over whole network and not for a single player. So, it is not obvious that if a network composed of any single type of player would perform better then a diversified population of players. This question has rarely been asked (e.g., in [4]) partially because it imposes high difficulties on the mathematical analysis.

The problem we started from is purely practical and one should expect that the effect of the investigation would be the design of a real technical system. This requires understanding how much we lose coming from a theoretical model to the reality of technical application because of the overheads imposed by limitation of the standards, imperfections of the devices, delays, etc.

References

[1] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *JMLR: Workshop and Conference Proceedings*, vol. 23, 2012.

[2] D. Fudenberg, *The Theory of Learning in Games*. Cambridge, MA: MIT Press, 1998.

[3] J. Zheng, Y. Cai, Y. Liu, Y. Xu, B. Duan, and X. Shen, "Optimal power allocation and user scheduling in multicell networks: Base station cooperation using a game-theoretic approach," *IEEE Transactions on Wireless Communications*, vol. 13, no. 12, pp. 6928–6942, 2014.

[4] H. Tembine, *Distributed Strategic Learning for Wireless Engineers.* CRC Press, 2012.