Joint Channel Selection and Power Control in Infrastructureless Wireless Networks: A Multiplayer Multiarmed Bandit Framework

Setareh Maghsudi and Sławomir Stańczak, *Senior Member, IEEE*

***Abstract*—This paper deals with the problem of efficient resource allocation in dynamic infrastructureless wireless networks. In a reactive interference-limited scenario, at each transmission trial, every transmitter selects a frequency channel from some common pool, together with a power level. As a result, for all transmitters, not only the fading gain, but the number and the power of interfering transmissions as well, vary over time. Due to the absence of a central controller and time varying network characteristics, it is highly inefficient for transmitters to acquire the global channel and network knowledge. Therefore, given no information, each transmitter selfishly intends to maximize its average reward, which is a function of the channel quality, as well as the joint selection profile of all transmitters. This scenario is modeled as an adversarial multiplayer multiarmed bandit game, where players attempt to minimize their so-called regret, while, at the network side, achieving equilibrium in some sense. Based on this model and to solve the resource allocation problem, in this paper, we develop two joint power level and channel selection strategies. We prove that the gap between the average rewards achieved by our approaches and that based on the best fixed strategy converges to zero asymptotically. Moreover, the empirical joint frequencies of the game converge to the set of correlated equilibria, which is characterized for two relaxed versions of the designed game.**

***Index Terms*—Adversarial bandit, channel selection, equilibrium, infrastructureless wireless network, power control.**

I. Introduction

A. Bandit Theory and Wireless Communication

THE MULTIARMED bandit (MAB) is a class of sequential optimization problems, to our best knowledge originally introduced in [1]. In the most traditional form of MAB, given a set of arms (actions), a player pulls an arm at every trial of the game to receive some reward. The rewards are not known to the player in advance; however, upon pulling any arm, the instantaneous reward of that arm is revealed. In such an unknown setting, at each trial, the player may lose some reward (or incur some cost) due to not selecting the best arm instead of the played arm. This loss is referred to as *regret* and is quantified by the difference between the reward that would have been achieved had the player selected the best arm and the actual achieved reward. The player decides which arm to pull in a sequence of trials so that its accumulated regret over the game horizon is minimized. This problem is a clear instance of the intrinsic tradeoff between exploration (learning) and exploitation (control), i.e., playing the arm that has exhibited the best performance in the past, on the one hand, and playing other arms to guarantee the optimal payoff in the future, on the other hand. An important class of bandit games is *adversarial bandit*, where the series of rewards generated by an arm cannot be attributed to any specific distribution function.

In recent years, bandit theory has been used as a mathematical tool to model and solve various wireless networking problems. For instance, [2] and [3] utilize the classical bandit game to model spectrum sharing in cognitive radio networks. In [4], Di *et al.* proposed a cooperative spectrum sensing scheme based on bandit theory. Furthermore, [5]–[7] use bandit theory to model relay selection, sensor scheduling, and object tracking, respectively. In [8] and [9], the channel monitoring problem is modeled as a bandit game. Bandit games have also been used for solving distributed resource allocation problems, as discussed in the following.

B. Distributed Resource Allocation in Wireless Networks

In recent years, game theory and reinforcement learning have been widely used for solving the distributed resource allocation problem. The vast majority of game-theoretical approaches is based on cooperation (e.g., coalition formation), mechanism design (e.g., auction theory), or exchange economy (e.g., supply–demand markets). Although these methods can be implemented in a distributed manner, such an implementation in a real network environment requires that each player at least knows its own utility function a priori. In addition, these approaches are in general inefficient as they necessitate information exchange among players, giving rise to the signaling and feedback overhead. For instance, most models from cooperative game theory require coordination and/or communication among players to form coalitions [10], [11]. As another example, in most auction games, bids must be submitted to

Manuscript received October 23, 2013; revised July 21, 2014; accepted October 16, 2014. Date of publication November 11, 2014; date of current version October 13, 2015. This paper was presented in part at the IEEE Wireless Communications and Networking Conference, Shanghai, China, April 7–10, 2013. This work was supported by the German Research Foundation (DFG) under Grant STA 864/3-3. The review of this paper was coordinated by Prof. M. C. Gursoy.

S. Maghsudi is with the Communications and Information Theory Group, Technical University of Berlin, 10623 Berlin, Germany (e-mail: setareh.maghsudi@tu-berlin.de).

S. Stańczak is with the Communications and Information Theory Group, Technical University of Berlin, 10623 Berlin, Germany, and also with the Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, 10587 Berlin, Germany (e-mail: slawomir.stanczak@hhi.fraunhofer.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2014.2369425

some central controller (auctioneer) that performs necessary computations and makes decisions [12], [13]. Finally, in supply–demand market models, prices and demands are exchanged among buyers and sellers [14], [15].

When the utility functions are not known in advance, the resource allocation problem is often solved by using learning approaches, including bandit models. A large body of literature, such as [16]–[18], analyzes single-agent stochastic learning problems. Another example is [19]. In this work, network optimization is modeled as a stochastic bandit game, where at each trial, multiple arms are selected by the player, and the reward is a linear combination of the rewards of the selected arms. An application of this formulation might be a downlink user selection. In single-agent settings, the agent learns from its previous experiences, and no information flow is required. However, in general, single-agent learning models cannot be used in wireless networks, where multiple players selfishly act by responding to each other, and the utility of each player is influenced by the actions of other players. Moreover, similar to games with complete information, it is desired that players achieve equilibrium in some sense. As far as multiagent settings are concerned, most studies assume that players are able to observe the actions of each other. This assumption, despite being acceptable for some spectrum sharing scenarios, is not applicable to the general resource allocation problems, particularly power control games, where it is difficult to identify the transmit power level of players. In essence, the assumption that each player announces its actions (e.g., its transmit power) is not always incentive compatible. As a result, the vast majority of previous works focuses on spectrum sharing and/or sensing, as well as on channel monitoring. On the other hand, most of the previous research studies assume that the generated rewards of any given action can be attributed to some density distribution. This assumption is however highly restrictive particularly for dynamic networks. In [20], the multiagent bandit problem is investigated. This research assumes that in case of collision, no reward is paid to the colliding users. In the context of wireless transmission, such rule corresponds to the elimination of channel sharing, which is clearly suboptimal. In addition, the approach proposed in [20] requires information exchange. Moreover, no equilibrium analysis is performed. Another example is the work in [21], where opportunistic spectrum access is formulated as a multiagent learning game. In this work, upon availability, each channel pays equal rewards to all users. Such formulation simplifies the analysis; nonetheless, it is strictly restrictive as it neglects different channel qualities experienced by different users. Moreover, in [21], if a channel is selected by multiple users, an orthogonal spectrum access scheme is applied, which is known to be suboptimal in general. References [22] and [23] consider graphical games for an interference minimization problem with partially overlapping channels, where the interference exists only between neighboring users. The convergence of the proposed learning approaches is established for exact potential games. In [24], cooperative rate maximization in cognitive radio networks is modeled as a bandit game, and two learning approaches are proposed for different levels of information availability. The stability of the solutions is however not investigated. In [25], two learning algorithms that converge to Nash equilibrium in a multiplayer cognitive environment are developed. System verification, however, is only based on numerical analysis. References [26] and [27] propose two selection schemes that achieve logarithmic regret; however, no equilibrium analysis is performed. All of the aforementioned works assume that the generated rewards of any given action are independent and identically distributed.

C. Our Contribution

In this paper, our focus is on a resource allocation problem in an infrastructureless network. First, we model this problem as an adversarial multiplayer MAB (MP-MAB) game. With the aim of efficient resource management and interference mitigation, we follow an approach proposed in [28] to develop two joint power control and channel selection schemes that are adapted versions of *exponential-based weighted average* [29] and *follow the leader* [30] strategies. Both proposed strategies not only result in small (i.e., with sublinear growth in time) regret for each individual player but guarantee that the empirical joint frequencies of the game converge to the set of correlated equilibria as well. In addition, we implement the *experimental regret-testing procedure* that is known to converge to the set of Nash equilibria [31].

Our work extends the state of the art in this area significantly since it differs from the existing studies in the following crucial aspects.

- Unlike many previous works including [16] and [17] that study the single-agent learning problem, we analyze the multiagent setting while taking the selfishness of players into account.
- Our model and algorithms do not rely on the assumption that the arms' reward-generating functions are time invariant. In fact, reward functions might arbitrarily vary, which captures the dynamic nature of wireless channels and distributed networks. This is in contrast to a great majority of previous works, including [20]–[22], where the reward process of every arm is time invariant.
- In contrast to [20] and [21], we neither allow pairwise information exchange nor use a control channel, thereby minimizing the overhead.
- Moreover, players do not observe the actions of each other. As a result, the algorithms are incentive compatible and do not require any cooperation; hence, they are widely applicable. An exemplary application is a power control problem with unknown power levels used by other transmitters.
- In our system model, channel qualities are taken into account so that channels pay different rewards to different users; that is, contrary to [21] and [22], the reward-generating functions are user specific. In addition, we impose no limitation on the interference pattern.
- The convergence analysis is valid for a wide range of games. This is in contrast to many previous works where the game should be necessarily potential for the convergence analysis to hold. Examples of such works include [21]–[23], among many others.

D. Paper Structure

Section II briefly reviews some concepts and results that deal with adversarial bandit problems. In Section III, we describe the system model and formulate the joint power control and channel selection problem as an adversarial MP-MAB game. Two resource allocation strategies are described in Sections IV and V. Section VI is devoted to the experimental regret-testing procedure. The results of numerical analysis are presented in Section VII. Section VIII concludes this paper.

II. Multiplayer–Multiarmed Bandit Games

A. Notions of Regret

The MP-MAB problem is a class of sequential decision-making problems with limited information. In this game, there exists a set of players $\mathcal{K}=\{1,\ldots,K\}$, where each player $k\in\mathcal{K}$ is assigned N_k actions, indexed from 1 to N_k; therefore, the action set yields $\mathcal{N}_k=\{1,\ldots,N_k\}$, $1\le N_k\le N$. Every player selects an action at successive trials to receive an initially unknown reward that depends on the nature as well as on the joint action profile of players. The action set, the played action, and the reward achieved by each player are regarded as private information. The reward-generating processes of arms are independent. Let $\mathbf{I}_t=(I_t^{(1)},\ldots,I_t^{(k)},\ldots,I_t^{(K)})=(I_t^{(k)},\mathbf{I}_{k,t}^-)$ denote the joint action profile of players at time t, where $I_t^{(k)}$ and $\mathbf{I}_{k,t}^-$ are the action of player k and the joint action profile of all players except for k, respectively. Moreover, let $g_t^{(k)}(\mathbf{I}_t)\in[0,1]$ be the reward achieved by some player k at time t.[1] The instantaneous regret of any player k is defined as the difference between the reward of the optimal action[2] and that of the played action. Based on this definition, the *cumulative regret* [32] of player k is formally defined in the following.

Definition 1 (Cumulative Regret): The cumulative regret of player k up to time n is defined as

$$R_n^{(k)}=\max_{i=1,\ldots,N_k}\sum_{t=1}^{n}g_t^{(k)}\left(i,\mathbf{I}_{k,t}^-\right)-\sum_{t=1}^{n}g_t^{(k)}\left(I_t^{(k)},\mathbf{I}_{k,t}^-\right). \quad (1)$$

As previously described, every player aims at minimizing its accumulated regret, which is an instance of the well-known exploitation–exploration dilemma: Find a balance between exploiting actions that have exhibited a well performance in the past (control), on the one hand, and exploring actions that might lead to a better performance in the future (learning), on the other hand.

Now, suppose that players use mixed strategies. This means that, at each trial t, player k selects a probability distribution $\mathbf{P}_t^{(k)}=(p_{1,t}^{(k)},\ldots,p_{i,t}^{(k)},\ldots,p_{N_k,t}^{(k)})$ over arms and plays arm i with probability $p_{i,t}^{(k)}$. In this case, we resort to the expected regret, which is also called *external regret* [32], and is defined as follows.

[1]Note that all results can be also expressed in terms of loss (d), provided that the loss is related to the gain by $d=1-g$, $g\in[0,1]$.

[2]Optimality is defined in the sense of the highest reward.

Definition 2 (External Regret): The external cumulative regret of player k up to time n is defined as

$$\begin{aligned}R_{\mathrm{Ext}}^{(k)}&:=R_{\mathrm{Ext}}^{(k)}(n)\\&=\max_{i=1,\ldots,N_k}\sum_{t=1}^{n}g_t^{(k)}\left(i,\mathbf{I}_{k,t}^-\right)-\sum_{t=1}^{n}\bar{g}_t^{(k)}\left(\mathbf{P}_t^{(k)},\mathbf{I}_{k,t}^-\right)\\&=\max_{i=1,\ldots,N_k}\sum_{t=1}^{n}\sum_{j=1}^{N_k}p_{j,t}^{(k)}\left(g_t^{(k)}\left(i,\mathbf{I}_{k,t}^-\right)-g_t^{(k)}\left(j,\mathbf{I}_{k,t}^-\right)\right)\end{aligned} \quad (2)$$

where $\bar{g}_t^{(k)}(\cdot)$ is the expected reward at round t by using mixed strategy $\mathbf{P}_t^{(k)}$, which is defined as $\bar{g}_t^{(k)}(\cdot)=\sum_{j=1}^{N_k}g_t^{(k)}(j,\cdot)p_{j,t}^{(k)}$.

By definition, external regret compares the expected reward of the current mixed strategy with that of the best fixed action in hindsight but fails to compare the rewards achieved by changing actions in a pairwise manner. To compare actions in pairs, *internal regret* [32] is introduced, which is closely related to the concept of equilibrium in games.

Definition 3 (Internal Regret): The internal cumulative regret of player k up to time n is defined as

$$\begin{aligned}R_{\mathrm{Int}}^{(k)}&:=R_{\mathrm{Int}}^{(k)}(n)=\max_{i,j=1,\ldots,N_k}R_{(i\to j),n}^{(k)}\\&=\max_{i,j=1,\ldots,N_k}\sum_{t=1}^{n}p_{i,t}^{(k)}\left(g_t^{(k)}\left(j,\mathbf{I}_{k,t}^-\right)-g_t^{(k)}\left(i,\mathbf{I}_{k,t}^-\right)\right).\end{aligned} \quad (3)$$

Notice that on the right-hand side of (3), $r_{(i\to j),t}^{(k)}=p_{i,t}^{(k)}(g_t^{(k)}(j,\cdot)-g_t^{(k)}(i,\cdot))$ denotes the expected regret caused by pulling arm i instead of arm j. By comparing (2) and (3), external regret can be bounded above by internal regret as [33]

$$\begin{aligned}R_{\mathrm{Ext}}^{(k)}&=\max_{i=1,\ldots,N_k}\sum_{j=1}^{N_k}R_{(i\to j),n}^{(k)}\\&\le N_k\max_{i,j=1,\ldots,N_k}R_{(i\to j),n}^{(k)}=N_kR_{\mathrm{Int}}^{(k)}.\end{aligned} \quad (4)$$

Remark 1: Throughout this paper, vanishing (zero average) external and internal regret means that $\lim_{n\to\infty}(1/n)R_{\mathrm{Ext}}=0$, and $\lim_{n\to\infty}(1/n)R_{\mathrm{Int}}=0$, respectively. In other words, we have $R_{\mathrm{Ext}}\in o(n)$ and $R_{\mathrm{Int}}\in o(n)$. Note that by (4), $R_{\mathrm{Int}}\in o(n)$ implies $R_{\mathrm{Ext}}\in o(n)$. Throughout this paper, we call any strategy with $R_{\mathrm{Int}}\in o(n)$ as "no-regret strategy."

B. Equilibrium

From the viewpoint of each player k, an MP-MAB is a game with two agents: player k itself and the *set* of all other $K-1$ players (referred to as the opponent), whose joint action profile affects the reward achieved by player k. We consider here the most general framework, where the opponent is nonoblivious, i.e., its series of actions depends on the actions of player k. It is known that a game against a nonoblivious opponent can be modeled by adversarial bandit games [34], where, similar to other game-theoretical formulations, the solution of this

game is equilibrium: most importantly Nash and correlated equilibria.[3]

In the context of game-theoretical bandits, an important result is the following theorem.

Theorem 1 [32]: Consider a K-player bandit game, where each player k is provided with N_k actions. Denote the internal regret of player k by $R_{\text{Int}}^{(k)}$, and the set of correlated equilibria by $\mathcal{C}$. At time n, define the empirical joint distribution of the game as

$$\hat{\pi}_n(\mathbf{J}) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}_{\{\mathbf{I}_t=\mathbf{J}\}} \tag{5}$$

where $\mathbf{J} = (J^{(1)}, \ldots, J^{(K)}) \in \bigotimes_{k=1}^{K}\{1, \ldots, N_K\}$ and with $\mathbf{1}_{\{x\}}$ being the indicator function that returns one if x holds and zero otherwise. If all players $k \in \{1, \ldots, K\}$ play according to any strategy so that

$$\lim_{n\to\infty} \frac{1}{n} R_{\text{Int}}^{(k)} = 0 \tag{6}$$

then the distance $\inf_{\pi\in\mathcal{C}} \sum_{\mathbf{J}} |\hat{\pi}_n(\mathbf{J}) - \pi(\mathbf{J})|$ between the empirical joint distribution of plays and the set of correlated equilibria converges to 0 almost surely.

Theorem 1 simply states that in an MP-MAB game, if all players play according to a strategy with vanishing internal regret (no-regret), then the empirical joint distribution of plays converges to the set of correlated equilibria. Note that the strategies used by players are not required to be identical. Since a rational player is interested in minimizing its regret, the assumption that every player plays according to a no-regret strategy is reasonable.

C. From Vanishing External Regret to Vanishing Internal Regret

In [33], an approach is proposed for converting any selection strategy with vanishing external regret to another version with vanishing internal regret. We describe this approach briefly. The player index k is omitted for brevity.

Consider some selection strategy κ that at each time t selects one of the N actions according to some probability distribution $\mathbf{P}_t$. Let $\mathbf{P}_1$ be the uniform distribution. To calculate $\mathbf{P}_t$ for some $t > 1$, κ constructs a metastrategy κ' with $N(N-1)$ virtual actions $(i \to j)$, $(i, j \in \{1, \ldots, N\}, i \neq j)$. Assume that κ' uses some mixed strategy $\underline{\delta}_t$ over $N(N-1)$ virtual actions, where the probability of the virtual action $(i \to j)$, i.e., $\delta_{(i\to j),t}$, depends on its past performance in some way.[4] Given $\underline{\delta}_t$ and $\mathbf{P}_{t-1} = (p_{1,t-1}, \ldots, p_{i,t-1}, \ldots, p_{j,t-1}, \ldots, p_{N,t-1})$, κ defines $\mathbf{P}_{(i\to j),t-1} = (p_{1,t-1}, \ldots, 0, \ldots, p_{j,t-1} + p_{i,t-1}, \ldots, p_{N,t-1})$, which has 0 and $p_{j,t-1} + p_{i,t-1}$ at the place of $p_{i,t-1}$ and $p_{j,t-1}$, respectively, and all other elements remain unchanged. Then, $\mathbf{P}_t = \sum_{(i,j):i\neq j} \mathbf{P}_{(i\to j),t}\delta_{(i\to j),t}$. As a result, κ has the characteristic that its internal regret is upper bounded by the external regret of κ'. Thus, if κ' exhibits vanishing external regret, then κ results in vanishing internal regret. In Sections IV and V, we use this property to design no-regret selection strategies.

III. Bandit-Theoretical Model of Infrastructureless Wireless Networks

We consider a network consisting of a set $\mathcal{K} = \{1, \ldots, K\}$ of transmitter–receiver pairs, referred to as D2D users. Each pair is denoted either by just k or by the pair (k, k') and can access a set of mutually orthogonal channels, i.e., $\mathcal{N}'_k = \{1, \ldots, N'_k\}$. The transmission power can also be selected from a set of quantized power levels $\mathcal{N}''_k = \{1, \ldots, N''_k\}$. This implies that the strategy set includes $N_k = N'_k \times N''_k$ actions, where at time t, each action $I_t^{(k)} = (I_t'^{(k)}, I_t''^{(k)})$ consists of one channel index $I_t'^{(k)}$ (which corresponds to some channel quality) and one power level $I_t''^{(k)}$. Therefore, the joint action profile of users, i.e., $\mathbf{I}_t$, is to be understood here as the pair $(\mathbf{I}'_t, \mathbf{I}''_t)$, where $\mathbf{I}'_t = (I_t'^{(1)}, \ldots, I_t'^{(K)})$ and $\mathbf{I}''_t = (I_t''^{(1)}, \ldots, I_t''^{(K)})$. As each channel might be accessible by multiple users, cochannel interference (collision, interchangeably) is likely to arise. Since users are allowed to select a new channel and to adapt their power levels at each transmission trial, the interference pattern changes over time. In addition, the distribution of fading coefficients is also time varying in general so that acquiring channel and/or network information at the level of autonomous transmitters would be extremely challenging and inefficient. Therefore, we assume the following.

Assumption A1: Throughout this paper, we have the following.

a) Transmitters have no channel knowledge or any other side information such as the number of users or their selected actions.
b) Users do not coordinate their actions that can be chosen completely asynchronously by each user.

As users do not observe the actions of each other, it might be in their interest to select their actions at the beginning of trials, thereby using the remaining time for data transmission.

In this paper, we model the joint channel and power level selection problem as a K-player adversarial bandit game, where player k decides for one of the N_k actions. We define the average utility function (reward) of player k to be[5]

$$f_t^{(k)}(\mathbf{I}) = \log\left(\frac{I''^{(k)}\left|h_{kk',t,I'^{(k)}}\right|^2}{\sum_{q\in\mathcal{Q}^{(k)}} I''^{(q)}\left|h_{qk',t,I'^{(k)}}\right|^2 + N_0}\right) - \alpha \cdot I''^{(k)} \tag{7}$$

for some given joint action profile $\mathbf{I} = (\mathbf{I}', \mathbf{I}'')$. In (7), $\mathcal{Q}^{(k)}$ is the set of players that interfere with user k in channel $I'^{(k)}$. Throughout this paper, $|h_{uv,t,c}|^2 > 0$ is used to denote the average gain of channel c between u and v at time t. N_0 is the variance of zero-mean additive white Gaussian noise, and $\alpha \geq 0$ is the constant power price factor. The last term in (7) is used

[3]These definitions are quite standard (see, e.g., [35]), and thus, we do not state them here.

[4]Note that the gains of virtual actions cannot be explicitly calculated. Later, we see that the gain achieved by any virtual action $(i \to j)$ is calculated based on the gain achieved by playing true actions i and j.

[5]Throughout this paper, logarithms are base 2, unless otherwise stated.

to penalize the use of excessive power. According to Section II, let $g_t^{(k)}(\mathbf{I}_t) \in [0,1]$ denote the achieved reward of player k at time t, as a function of joint action profile $\mathbf{I}_t$. We consider a game with noisy rewards, where $g_t^{(k)}(\mathbf{I}) = f_t^{(k)}(\mathbf{I}) + \epsilon_t$, with ϵ being some zero-mean random variable with bounded variance, independent from all other random variables. As is well known, in a noncooperative game, the primary goal of each selfish player is to maximize its own accumulated reward. This can be written as

$$\underset{\{I_t'^{(k)}, I_t''^{(k)}\}_{t=1}^n}{\text{maximize}} \sum_{t=1}^{n} g_t^{(k)}\left(\mathbf{I}_t', \mathbf{I}_t''\right) \tag{8}$$

where $I_t'^{(k)} \in \mathcal{N}_K'$ and $I_t''^{(k)} \in \mathcal{N}_K''$. From Assumptions A1, however, it can be concluded that the objective function in (8) is not available. For this reason, we argue for a less ambitious goal, which is known as *regret minimization*. Formally, every player k attempts to achieve vanishing external regret in the sense that

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} R_{\text{Ext}}^{(k)} &= \lim_{n\to\infty} \frac{1}{n}\left(\max_{i\in\mathcal{N}_k} \sum_{t=1}^{n} g_t^{(k)}\left(i, \mathbf{I}_{k,t}^-\right) - \sum_{t=1}^{n} \bar{g}_t^{(k)}\left(\mathbf{P}_t^{(k)}, \mathbf{I}_{k,t}^-\right)\right) \\ &= 0. \end{aligned} \tag{9}$$

In addition to the individual strategy of each user aiming at satisfying (9), at the network level, it is desired to achieve some steady state, i.e., equilibrium. Therefore, in this paper, we develop algorithmic solutions to the resource allocation problem with a twofold objective in mind: 1) External regret of each user should asymptotically vanish according to (9), and 2) the players' actions should converge to equilibrium.

By (4), the external regret of each user is upper bounded by its internal regret. As a result, if all users select their actions according to some no-regret strategy, not only (9) is achieved by all of them (see also Remark 1), but the corresponding game converges to equilibrium in some sense as well, which immediately follows from Theorem 1. In Sections IV and V, we present two internal regret minimizing strategies that are shown to solve the game and, with it, to achieve the two objectives mentioned above. Both algorithms can be applied in a fully decentralized manner by each player, since at each time, they only require the set of past rewards of the respective player.

Finally, it is worth noting that the set of correlated equilibria for the discrete time varying repeated game defined by (7) cannot be characterized. Nevertheless, in what follows, we characterize the set of equilibria for two relaxed versions of this game. In doing so, we assume that the strategy sets of all players are convex and compact subsets of $\mathbb{R}^2$. With this assumption in mind, consider a game where for all players, the mean reward process is time invariant, i.e.,

$$f^{(k)}(\mathbf{I}) = \log\left(\frac{I''^{(k)}\left|h_{kk',I'^{(k)}}\right|^2}{\sum_{q\in\mathcal{Q}^{(k)}} I''^{(q)}\left|h_{qk',I'^{(k)}}\right|^2 + N_0}\right) - \alpha\cdot I''^{(k)} \tag{10}$$

which implies that the average channel gains are time invariant. By the following proposition, this game has a unique correlated equilibrium.

Proposition 1: Consider a K-player game where the mean reward function of each player k is defined by (10). This game has a unique correlated equilibrium that places probability one on its unique pure strategy Nash equilibrium.

Proof: See Appendix B. ■

Now, let the mean reward function be defined as follows:

$$f^{(k)}(\mathbf{I}) = \log\left(I''^{(k)}\frac{\left|h_{kk',I'^{(k)}}\right|^2}{N_0}\right) - \alpha^{(k)} I''^{(k)} \tag{11}$$

which is more restricted but simpler than (10). Note that here the pricing factor is user specific and can be used by an authority to control the interferences generated by each user. With this choice of average reward function, the game can be shown to have a unique correlated equilibrium that maximizes the aggregate utility of all players. This result is formally stated in the following proposition.

Proposition 2: Consider a K-player game where each player k has the mean reward function $f^{(k)}(\cdot)$ given by (11). This game has a unique correlated equilibrium that places probability one on a unique pure strategy Nash equilibrium that maximizes $\sum_{k=1}^{K} f^{(k)}(\cdot)$.

Proof: See Appendix C. ■

IV. No-Regret Bandit Exponential-Based Weighted Average Strategy

The basic idea of an exponential-based weighted average strategy is to assign each action, at every trial, some selection probability that is inversely proportional to the exponentially weighted accumulated regret (or directly proportional to the exponentially weighted accumulated reward) caused by that action in the past [36]. Roughly speaking, if playing an action has resulted in large regret in the past, its future selection probability is small, and *vice versa*.

As described in Section II-A, in bandit formulation, players only observe the reward of the played action, and not those of other actions. Therefore, the reward of any action i is estimated as [32]

$$\tilde{g}_t^{(k)}(i) = \begin{cases} \frac{g_t^{(k)}\left(I_t^{(k)}\right)}{p_{i,t}^{(k)}}, & i = I_t^{(k)} \\ 0, & o.w. \end{cases} \tag{12}$$

which is an unbiased estimate of the true reward of action i. Estimated rewards are then used to calculate regrets. For example, the regret of *not* playing action j instead of action i yields

$$\tilde{R}_{(i\to j),t-1}^{(k)} = \sum_{s=1}^{t-1} \tilde{r}_{(i\to j),s}^{(k)} = \sum_{s=1}^{t-1} p_{i,s}^{(k)}\left(\tilde{g}_s^{(k)}(j) - \tilde{g}_s^{(k)}(i)\right). \tag{13}$$

Despite exhibiting vanishing external regret, weighted average strategies yield in general large internal regret; as a result, even if all players play according to such strategies, the game does not converge to equilibrium. In the following, we utilize the bandit version of exponentially weighted average strategy [37] and convert it into an improved version that yields small internal regret, using the approach in Section II-C. The strategy is called the no-regret bandit exponentially weighted average strategy (NR-BEWAS) and is described in Algorithm 1.

Algorithm 1 NR-BEWAS

1: If the game horizon, n, is known, define γ_t and η_t as given in Proposition 3, otherwise as given in Proposition 4.
2: Define $\Phi(\mathbf{U}) = (1/\eta_t)\ln(\sum_{i=1}^{N_k}\exp(\eta_t u_i))$, where $\mathbf{U} = (u_1,\ldots,u_{N_k}) \in \mathbb{R}^{N_k}$.
3: Let $\mathbf{P}_1^{(k)} = (1/N_k,\ldots,1/N_k)$ (uniform distribution).
4: Select an action using $\mathbf{P}_1^{(k)}$.
5: Play and observe the reward.
6: **for** $t = 2,\ldots,n$ **do**
7: Let $\mathbf{P}_{t-1}^{(k)} = (p_{1,t-1}^{(k)},\ldots,p_{i,t-1}^{(k)},\ldots,p_{j,t-1}^{(k)},\ldots,p_{N_k,t-1}^{(k)})$ be the mixed strategy at time $t-1$.
8: Construct $\mathbf{P}_{(i\to j),t-1}^{(k)}$ as follows: replace $p_{i,t-1}^{(k)}$ in $\mathbf{P}_{t-1}^{(k)}$ by zero, and instead increase $p_{j,t-1}^{(k)}$ to $p_{j,t-1}^{(k)} + p_{i,t-1}^{(k)}$. Other elements remain unchanged. We obtain $\mathbf{P}_{(i\to j),t-1}^{(k)} = (p_{1,t-1}^{(k)},\ldots,0,\ldots,p_{j,t-1}^{(k)}+p_{i,t-1}^{(k)},\ldots,p_{N_k,t-1}^{(k)})$.
9: Define

$$\delta_{(i\to j),t}^{(k)} = \frac{\exp\left(\eta_t \tilde{R}_{(i\to j),t-1}^{(k)}\right)}{\sum_{(m\to l):m\neq l}\exp\left(\eta_t \tilde{R}_{(m\to l),t-1}^{(k)}\right)} \tag{14}$$

where $\tilde{R}_{(i\to j),t-1}^{(k)}$ is calculated by using (12) and (13).
10: Given $\delta_{(i\to j),t}^{(k)}$, solve the following fixed-point equation to find $\mathbf{P}_t^{(k)}$:

$$\mathbf{P}_t^{(k)} = \sum_{(i\to j):i\neq j} \mathbf{P}_{(i\to j),t}^{(k)}\delta_{(i\to j),t}^{(k)}. \tag{15}$$

11: The final probability distribution yields

$$\mathbf{P}_t^{(k)} = (1-\gamma_t)\mathbf{P}_t^{(k)} + \frac{\gamma_t}{N_k}. \tag{16}$$

12: Using the final $\mathbf{P}_t^{(k)}$, given by (16), select an action.
13: Play and observe the reward.
14: **end for**

From Algorithm 1, NR-BEWAS has two parameters, namely, γ_t and η_t. In the event that the game horizon, i.e., n, is known in advance, these two parameters are constant over time ($\eta_t = \eta$ and $\gamma_t = \gamma$), and the growth rate of regret can be precisely bounded, mainly based on the results in [32]. Otherwise, they vary with time. In this case, vanishing (sublinear in time) internal regret can be guaranteed; nevertheless, this bound might be loose. This discussion is formalized by the following propositions.

Proposition 3: Let $\eta_t = \eta = (\ln(N_k)/2N_k n)^{2/3}$ and $\gamma_t = \gamma = (N_k^2\ln(N_k)/4n)^{1/3}$. Then, Algorithm 1 (NR-BEWAS) yields vanishing internal regret, and we have $R_{\text{Int}}^{(k)} \in ((N_k^2 n)^{2/3}(\ln(N_k))^{1/3})$.

Proof: See Appendix D. ■

Proposition 4: Let $\eta_t = (\gamma_t^3/N_k^2)$ and $\gamma_t = t^{-(1/3)}$. Then, Algorithm 1 (NR-BEWAS) yields vanishing internal regret, i.e., $R_{\text{Int}}^{(k)} \in o(n)$.

Proof: See Appendix E. ■

Corollary 1: If all players play according to NR-BEWAS, then the empirical joint frequencies of the game converge to the set of correlated equilibria.

Proof: The proof is a direct consequence of Theorem 1 and Proposition 3 or Proposition 4. ■

Corollary 2: Assume that ϵ-correlated equilibrium approximate correlated equilibrium in the sense that $\bigcap_{\epsilon>0}\mathcal{C}_\epsilon = \mathcal{C}$. If the game horizon is known and all players play according to NR-BEWAS, then the minimum required number of trials to achieve ϵ-correlated equilibrium yields $\max_{k=1,\ldots,K}\epsilon^{-(3/2)} \times O((N_k K)(N_k^2\ln(N_k) + K^2\ln(K)))$, which is proportional to $\epsilon^{-(3/2)}$ and polynomially increases in the number of actions and players.

Proof: The proof follows from the bound of Proposition 3 and [32, Remark 7.6].[6] ■

V. No-Regret Bandit Follow the Perturbed Leader Strategy

Similar to the weighted average strategy presented in the previous section, the strategy *follow the perturbed leader* is an approach to solve online decision-making problems. In the basic version of this approach, called *follow the leader* [38], the action with the minimum regret in the past is selected at each trial. This rule is however deterministic and, therefore, does not achieve vanishing regret against nonoblivious opponents. As a result, in *follow the perturbed leader*, the player adds a random perturbation to the vector of accumulated regrets, and the action with the minimum perturbed regret in the past is selected [32]. In [39], a bandit version of this algorithm is constructed, where unobserved rewards are estimated. The authors show that the developed algorithm exhibits vanishing external regret. Similar to NR-BEWAS, we here modify the algorithm in [39] to ensure vanishing internal regret. The approach is called no-regret bandit follow the perturbed leader strategy (NR-BFPLS) and is described in Algorithm 2.

Algorithm 2 NR-BFPLS

1: Define $\epsilon_t = \epsilon_n = (\sqrt{\ln(n)}/3\sqrt{N_k n})$, and $\gamma_t = \min(1, N_k\epsilon_t)$. Note that unlike NR-BEWAS, here, we know the game horizon (n) in advance.
2: Let $\mathbf{P}_1^{(k)} = (1/N_k,\ldots,1/N_k)$ (uniform distribution).
3: Select an action using $\mathbf{P}_1^{(k)}$.
4: Play and observe the reward.
5: **for** $t = 2,\ldots,n$ **do**
6: Let $\mathbf{P}_{t-1}^{(k)} = (p_{1,t-1}^{(k)},\ldots,p_{i,t-1}^{(k)},\ldots,p_{j,t-1}^{(k)},\ldots,p_{N_k,t-1}^{(k)})$ be the mixed strategy at time $t-1$.
7: Construct $\mathbf{P}_{(i\to j),t-1}^{(k)}$ as follows: replace $p_{i,t-1}^{(k)}$ in $\mathbf{P}_{t-1}^{(k)}$ by zero, and instead increase $p_{j,t-1}^{(k)}$ to $p_{j,t-1}^{(k)} + p_{i,t-1}^{(k)}$. Other elements remain unchanged. We obtain $\mathbf{P}_{(i\to j),t-1}^{(k)} = (p_{1,t-1}^{(k)},\ldots,0,\ldots,p_{j,t-1}^{(k)}+p_{i,t-1}^{(k)},\ldots,p_{N_k,t-1}^{(k)})$.
8: Calculate $\tilde{R}_{(i\to j),t-1}^{(k)}$ using (12) and (13).

[6] Details are omitted to avoid an unnecessary restatement of the existing analysis.

9: Define $\sigma_{(i\to j),t-1}=(\sum_{\tau=1}^{t-1}(1/\delta^{(k)}_{(i\to j),t}))^{1/2}$, which is the upper bound of conditional variances of random variables $\tilde{R}^{(k)}_{(i\to j),t-1}$ [39].

10: Let $\tilde{R}^{(k)}_{(i\to j),t-1}=\tilde{R}^{(k)}_{(i\to j),t-1}-\sqrt{1+\sqrt{2/N_k}}\sigma_{(i\to j),t-1}\times\sqrt{\ln(t)}$ [39].

11: Randomly select a perturbation vector $\underline{\mu}_t$ with $N_k(N_k-1)$ elements from a two-sided exponential distribution with width ϵ_t.

12: Consider a selection rule that selects the action $(i\to j)$ given by

$$\operatorname{argmax}\left\{\tilde{R}^{(k)}_{(i\to j),t-1}+\mu_{(i\to j),t}\right\} \tag{17}$$

for $(i\to j)\in\{1,\dots,N_k(N_k-1)\}$. Note that in our setting, $\tilde{R}_{(i\to j)}$ denotes the estimated regret of *not* playing action $(i\to j)$; hence, we find the action with the largest $\tilde{R}$.

13: By using (20), calculate the probability $\delta^{(k)}_{(i\to j),t}$ assigned to each pair $(i\to j)$.

14: Given $\delta^{(k)}_{(i\to j),t}$, solve the following fixed-point equation to find $\mathbf{P}^{(k)}_t$:

$$\mathbf{P}^{(k)}_t=\sum_{(i\to j):i\neq j}\mathbf{P}^{(k)}_{(i\to j),t}\delta^{(k)}_{(i\to j),t}. \tag{18}$$

15: The final probability distribution yields

$$\mathbf{P}^{(k)}_t=(1-\gamma_t)\mathbf{P}^{(k)}_t+\frac{\gamma_t}{N_k}. \tag{19}$$

16: Using the final $\mathbf{P}^{(k)}_t$, given by (19), select an action.

17: Play and observe the reward.

18: **end for**

Algorithm 2 requires knowledge of the probability assigned to each action by the *follow the perturbed leader* strategy at every trial. However, in contrast to NR-BEWAS, these probabilities are not explicitly assigned; therefore, we explain how to calculate these values. From (17), the selection probability of some virtual action $(i\to j)\in\{1,\dots,N_k(N_k-1)\}$ is the probability that $\tilde{R}_{(i\to j),t-1}$ plus perturbation $\mu_{(i\to j),t}$ is larger than those of other actions, i.e.,

$$\begin{aligned}\delta^{(k)}_{(i\to j),t}&=\Pr\Big[\tilde{R}^{(k)}_{(i\to j),t-1}+\mu_{(i\to j),t}\geq\tilde{R}^{(k)}_{(i'\to j'),t-1}\\&\qquad+\mu_{(i'\to j'),t}\quad\forall(i\to j)\neq(i'\to j')\Big]\\&=\int_{-\infty}^{\infty}\Pr\Big[\tilde{R}^{(k)}_{(i\to j),t-1}+\mu_{(i\to j),t}=m\wedge\tilde{R}^{(k)}_{(i'\to j'),t-1}\\&\qquad+\mu_{(i'\to j'),t}\leq m\quad\forall(i\to j)\neq(i'\to j')\Big]\,dm\\&=\int_{-\infty}^{\infty}\Pr\Big[\tilde{R}^{(k)}_{(i\to j),t-1}+\mu_{(i\to j),t}=m\Big]\\&\quad\times\prod_{(i'\to j')\neq(i\to j)}\Pr\Big[\tilde{R}^{(k)}_{(i'\to j'),t-1}+\mu_{(i'\to j'),t}\leq m\Big]\,dm.\end{aligned} \tag{20}$$

Since $\underline{\mu}_t$ is distributed according to a two-sided exponential distribution with width ϵ_n, the terms under the integral can be easily calculated [40].

Now, we are in a position to show some properties of NR-BFPLS (see Algorithm 2).

Proposition 5: Let $\epsilon_t=\epsilon=(\sqrt{\ln(n)}/3\sqrt{N_kn})$ and $\gamma_t=\gamma=\min(1,N_k\epsilon_t)$. Then, Algorithm 2 (NR-BFPL) yields vanishing internal regret with $R^{(k)}_{\mathrm{Int}}\in O((nN_k^2\ln(N_k))^{1/2})$.

Proof: By [39], we know that if the BFPLS is applied to N_k actions, then $R^{(k)}_{\mathrm{Ext}}\in O((nN_k\ln(N_k))^{1/2})$. Using this, the proof proceeds along similar lines as the proof of Proposition 3 and is therefore omitted. ■

Corollary 3: If the game horizon is known and all players play according to NR-BFPLS, then the minimum required number of trials to achieve ϵ-correlated equilibrium yields $\max_{k=1,\dots,K}\epsilon^{-2}O((N_kK)(N_k^2\ln(N_k)+K^2\ln(K)))$, which is proportional to ϵ^{-2} and polynomially increases in the number of actions and players.

Proof: The proof is a result of the bound of Proposition 5 and [32, Remark 7.6]. ■

VI. Bandit Experimental Regret-Testing Strategy (BERTS)

Experimental regret-testing belongs to the large family of exhaustive search algorithms and is comprehensively discussed in [31] and [32] for bandit games. Here, we briefly review this approach and investigate its performance numerically later in Section VII-A.

First, the time is divided into periods $m=1,2,\dots$ of length T so that for each m, we have $t\in[(m-1)T+1,mT]$. At the beginning of period m, any player k randomly selects a mixed strategy, denoted by $\mathbf{P}^{(k)}_m$. Moreover, some random variable $U^{(m)}_{k,t}\in\{1,\dots,N_k\}$ is defined as follows. For $t\in[(m-1)T+1,mT]$ and for each action $i\in\{1,\dots,N_k\}$, there are exactly s values of t such that $U^{(m)}_{k,t}=i$, and $U^{(m)}_{k,t}=0$ for the remaining $t=T-sN_k$ trials. At time t, the action $I^{(k)}_t$ is selected to be [37]

$$I^{(k)}_t:\begin{cases}\text{is distributed as }\mathbf{P}^{(k)}_m, & \text{if } U^{(m)}_{k,t}=0\\ \text{equals } i, & \text{if } U^{(m)}_{k,t}=i.\end{cases} \tag{21}$$

At the end of period m, player k calculates the experimental regret of playing each action i as [37]

$$\begin{aligned}\hat{r}^{(k)}_{i,m}=&\frac{1}{T-sN_k}\sum_{t=(m-1)T+1}^{mT}g^{(k)}_t(\mathbf{I}_t)\mathbf{1}_{\left\{U^{(m)}_{k,t}=0\right\}}\\&-\frac{1}{s}\sum_{t=(m-1)T+1}^{mT}g^{(k)}_t\left(i,\mathbf{I}^-_{k,t}\right)\mathbf{1}_{\left\{U^{(m)}_{k,t}=i\right\}}.\end{aligned} \tag{22}$$

If the regret is smaller than an acceptable threshold ρ, the player continues to play its current mixed strategy. Otherwise, another mixed strategy is selected. The procedure is summarized in Algorithm 3. It is known that if the parameters of BERTS (e.g., T and ρ) are appropriately chosen, then in the long run, the played mixed strategy profile is an approximate Nash

equilibrium almost all the time. Details can be found in [32] and, hence, are omitted.

Algorithm 3 BERTS [32]

1: Set T (period length), ρ (acceptable regret threshold), $\xi \ll 1$ (exploration parameter), $m = 1$ (period index). Notice that for each period $m = 1, \ldots, M$, we have $t \in [(m-1)T + 1, mT]$.
2: Select a mixed strategy $\mathbf{P}_m^{(k)}$ according to the uniform distribution, from the probability simplex with N_k dimensions.
3: For each $i \in \{1, \ldots, N_k\}$, select s exploring trials at random. Exploration trials that are dedicated to different actions should not overlap.
4: **for** $t = (m-1)T + y$, where $1 \leq y < T$ **do**
5: **if** t is an exploring trial dedicated to action i **then**
6: play action i and observe the reward;
7: **else**
8: select an action using $\mathbf{P}_m^{(k)}$. Play and observe the reward.
9: **end if**
10: **end for**
11: Calculate the experimental regret of period m, $\hat{r}_{m,i}^{(k)}$, using (22).
12: **if** $\max_{i=1,\ldots,N_k} \hat{r}_{i,m}^{(k)} > \rho$, **then**
13: 1) set $m = m + 1$, 2) go to line 2;
14: **else**
15: • with probability ξ: 1) set $m = m + 1$, 2) go to line 2.
• with probability $1 - \xi$: 1) let $\mathbf{P}_{m+1}^{(k)} = \mathbf{P}_m^{(k)}$, 2) set $m = m + 1$, 3) go to line 3.
16: **end if**

VII. Numerical Analysis

Numerical analysis consists of two parts. In Section VII-A, we consider a simple network and clarify the work flow of the developed resource allocation schemes. In Section VII-B, we consider a larger network and study the performance of the proposed game model and algorithmic solutions in comparison with some other selection strategies.

A. Part One

1) Network Model: The network consists of two transmitter–receiver pairs, which are referred to as users. There exist two orthogonal channels, i.e., C_1 and C_2, and two power levels, i.e., P_1 and P_2. Hence, the action set of each user yields $\{a_1 : (C_1, P_1), a_2 : (C_1, P_2), a_3 : (C_2, P_1), a_4 : (C_2, P_2)\}$. The distribution of channel gains changes at each trial. We assume that the variance of mean values of these distributions is relatively small, which corresponds to low dynamicity.[7] Channel matrices are $\mathrm{H}_1 = \begin{bmatrix} [0.50, 0.80] & [0.15, 0.20] \\ [0.01, 0.05] & [0.01, 0.09] \end{bmatrix}$ and $\mathrm{H}_2 = \begin{bmatrix} [0.02, 0.05] & [0.02, 0.06] \\ [0.05, 0.15] & [0.75, 0.95] \end{bmatrix}$, where $\mathrm{H}_{l,(u,v)}$, $u, v, l \in \{1, 2\}$, corresponds to the interval from which $|h_{uv,l}|$ is selected at each trial. Moreover, we assume $P_1 = 1$, $P_2 = 5$, and $\alpha = 10^{-3}$. Except for their instantaneous rewards, no other information is revealed to the users. This information can be provided by the receiver feedback to the transmitter. With these settings, it is easy to see that $((C_1, P_2), (C_2, P_2))$ is the unique pure strategy Nash equilibrium of this game.

[7]Note that this assumption is made to simplify the implementation; as theoretically established, both proposed strategies converge to equilibrium for general time varying distributions.

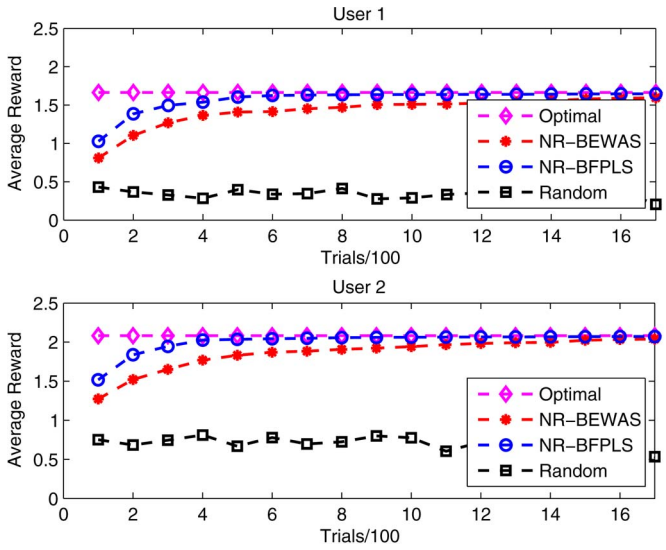

Fig. 1. Performance of four selection strategies. Both NR-BEWAS and NR-BFPLS exhibit vanishing regret, that is, their average rewards converge to that of optimal (centralized) selection.

2) Results and Discussion: We investigate the performance of selection strategies NR-BEWAS, NR-BFPLS, and BERTS. The following strategies are also considered as benchmark:

- optimal (centralized) action (channel and power level) assignment by using global statistical channel knowledge and through an exhaustive search, performed by a central controller;
- uniformly random selection.

Fig. 1 compares the average reward achieved by NR-BEWAS and NR-BFPLS by those of random and optimal selections. In the figure, despite being provided with no information, both NR-BFPLS and NR-BEWAS exhibit vanishing regret, in the sense that the achieved average reward converges to that of the centralized scenario.

Figs. 2 and 3 show the evolution of mixed strategies of the two users when NR-BEWAS is used. Figs. 4 and 5, on the other hand, show the same variable when actions are selected by using NR-BFPLS. For both cases, the first and second users respectively converge to (C_1, P_2) and (C_2, P_2).

The performance of BERTS, however, is not an explicit function of the game duration. As described before, the procedure continues to search the strategy space until a suitable (mixed or pure) strategy, which yields a regret less than the selected threshold, is captured. In [32, Th. 7.8], the minimum game duration to guarantee the convergence of BERTS

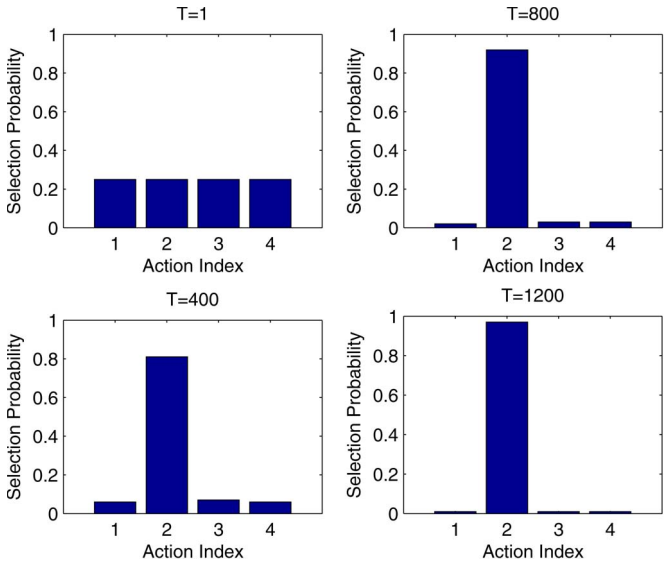

Fig. 2. Evolution of the mixed strategy of User 1, applying NR-BEWAS. The mixed strategy of User 1 converges to (0, 1, 0, 0).

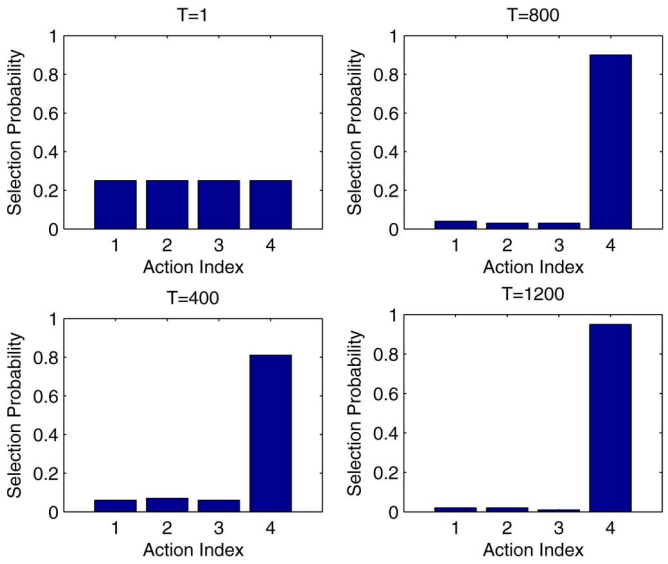

Fig. 3. Evolution of the mixed strategy of User 2, applying NR-BEWAS. The mixed strategy of User 2 converges to (0, 0, 0, 1).

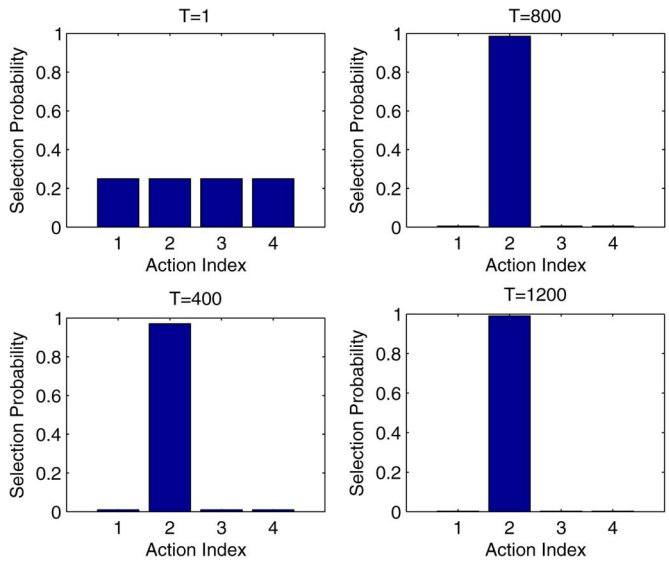

Fig. 4. Evolution of the mixed strategy of User 1, applying NR-BFPLS. The mixed strategy of User 1 converges to (0, 1, 0, 0).

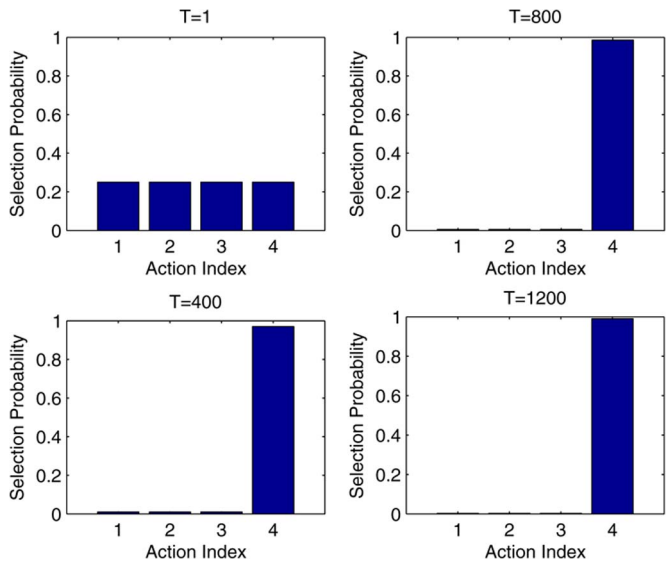

Fig. 5. Evolution of the mixed strategy of User 2, applying NR-BFPLS. The mixed strategy of User 2 converges to (0, 0, 0, 1).

is specified, which is relatively long, even for small number of users and actions. Nevertheless, similar to other search-based algorithms, there is also the possibility of finding some acceptable strategy at the early stages of the game. As a result, for relatively short games, the performance of BERTS is rather unpredictable. The other issue is the effect of regret threshold. On the one hand, a larger threshold reduces the search time, since the set of acceptable strategies is large. On the other hand, a large regret threshold could possibly lead to performance loss, since the user might get locked at some suboptimal strategy at the early stages, thereby incurring large accumulated regret. It is worth noting that due to its simplicity and despite unpredictable performance, BERTS is an appealing approach in cases where computational effort should be minimized. Fig. 6 summarizes the results of few exemplary performances of BERTS. The parameters are selected as $T = 80$, $M = 1500$, and $\rho = 0.16$ (see Section VI). Simulation is performed for six *independent* rounds. The curve on the left side in Fig. 6 shows the period $(1 \leq m \leq 1500)$ at which the algorithm finds an acceptable strategy. As expected, the results exhibit no specific pattern. The four subfigures on the right show the mixed strategies selected by BERTS at rounds 1 and 2, together with average rewards. In this figure, at round 2, acceptable strategies are found earlier than round 1 by both users, leading to better average performance. It is also worth noting that for User 2, the strategy of round 1 is in essence better than that of round 2; nevertheless, it is found too late. As a result, the average performance of round 2 is superior to that of round 1.

B. Part Two

Here, we consider a wireless network consisting of five users (transmitter–receiver pairs), which compete for access to three orthogonal channels at two possible power levels (hence, six

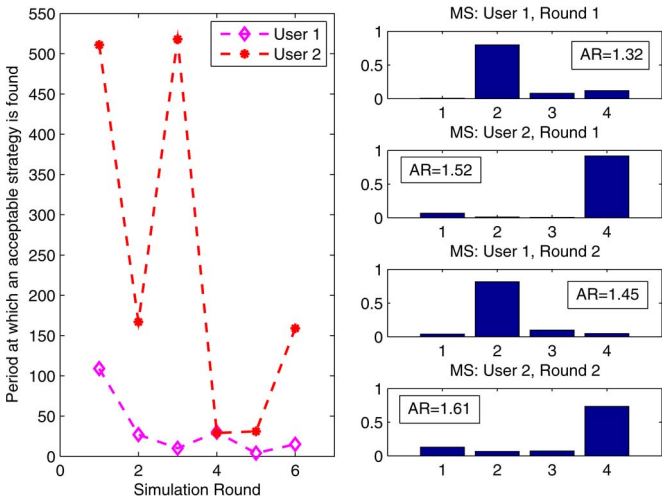

Fig. 6. Performance of BERTS. On the left, the vertical and horizontal axes show the periods and round number, respectively. The two curves depict the period at which a suitable mixed strategy (MS) is found at each of the six rounds. On the right, these mixed strategies are shown for both users at rounds 1 and 2, together with average rewards (AR). The horizontal and vertical axes, respectively, depict actions' indexes and their selection probabilities.

actions). We compare NR-BFPLS and NR-BEWAS with the following selection approaches.[8]

- optimal (centralized) action assignment as described in Section VII-A2;
- centralized no-collision action selection, where no reward is assigned to users that access the same channel; hence, users are encouraged to avoid collision. This curve can be considered as an upper bound for the performance of learning algorithms that select actions based on collision avoidance, such as [20];
- ϵ-greedy algorithm, where at each trial, with probability ϵ (exploration parameter), an action is uniformly selected at random, whereas with probability $1-\epsilon$, the best action so far is played. The player's estimation of the average reward of the selected action is improved after each play [41]. For stationary environments, ϵ is usually time varying and converges to zero in the limit, whereas in adversarial cases, ϵ is preferred to remain fixed. Here, we choose $\epsilon = 0.1$;
- greedy approach, where at the beginning of the game, some trials are reserved for exploration, in which actions are selected at random (exploration period). The length of this period is a predefined fraction of the entire game duration. Based on the rewards of exploration period, the best possible action is selected, and it is played for the rest of the game (exploitation period) [32]. This approach is simple to implement; however, to our best knowledge, there is no analysis on the optimal length of the exploration period;
- uniformly random selection.

[8] As mentioned before, observing the joint action profile and/or information exchange is not required for implementing NR-BEWAS, NR-BFPLS, and BERTS. Therefore, they cannot be compared with strategies that include mutual observation and/or communication. A good example of such algorithms is the widely used *best response dynamics*, where the strategy of each player is to play with the best response to either the historical [10] or the predicted [5] joint action profile of opponents. Another example is the strategy suggested in [20], which is a combination of learning and auction algorithms where users communicate with each other.

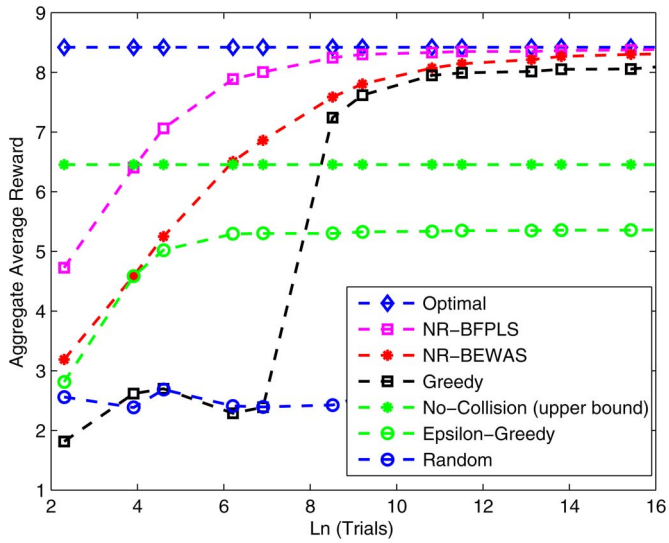

Fig. 7. Aggregate average reward of NR-BFPLS and NR-BEWAS compared with some other selection strategies.

The results are shown in Fig. 7. In this figure, we can conclude the following.

- The performance of interference-avoidance strategies is strongly influenced by channel matrices and tends to be poor specifically when the number of channels is less than that of users. The reason is that the sum reward of multiple interfering users with limited transmission power might be larger than the maximum achievable reward of any single user.
- The performance of both NR-BFPLS and NR-BEWAS converges to that of the centralized approach. As expected, NR-BFPLS converges faster than NR-BEWAS. We also point out that the convergence speed of both algorithms would be dramatically enhanced if some side information were available to players, e.g., if users observed the actions of each other, or if information exchange were allowed. It is also worth noting that although NR-BFPLS converges faster than NR-BEWAS, the computation of integral (20) might be involved, particularly for a large number of actions [40].
- In general, ϵ-greedy and greedy approaches can be easily implemented with low computational cost; nevertheless, it can be seen that the greedy approaches are inferior to NR-BEWAS and NR-BFPLS in terms of asymptotic performance. Basically, these approaches are more suitable for static environments.

VIII. Conclusion and Remarks

We have considered a resource allocation problem in multiuser infrastructureless wireless networks. The problem of utility maximization has been formulated using the multiplayer multiarmed adversarial bandit framework. Given no side information, the users aim at minimizing some regret expressed

in terms of the loss of reward by selecting appropriate actions on a given space of transmit power levels and orthogonal frequency channels. Based on some recent mathematical results, we developed two selection strategies, which not only provide vanishing regret for every player but guarantee that the game asymptotically converges to the set of correlated equilibria as well. We have also studied an experimental regret-testing strategy that asymptotically converges to the set of Nash equilibria. Numerical results confirmed the applicability of the proposed game model and allocation strategies to wireless channel selection and power control.

APPENDIX A
SOME AUXILIARY RESULTS

Here, we state some auxiliary results and materials from game theory and adversarial MABs that are necessary for the proofs.

1) Game Theory: We consider a game $\mathfrak{G}$ consisting of a set of K players where the strategy set of each player $k \in \{1, \ldots, K\}$ is denoted by $\mathcal{I}^{(k)}$ with a generic element $I^{(k)} = (I_1^{(k)}, \ldots, I_M^{(k)})$. Similarly, the set of joint strategy profiles of players is denoted by $\mathcal{I}$ with a generic element $\mathbf{I} = (I^{(1)}, \ldots, I^{(K)})$ and $\mathbf{I}_k^-$ stands for the joint action profile of all players except for player k. Moreover, $f^{(k)}(\mathbf{I})$ stands for the (bounded) average reward function of some player k.

Definition 4 (Smooth Game): A game $\mathfrak{G}$ is smooth if, for each $k \in \{1, \ldots, K\}$, $f^{(k)}(\mathbf{I})$ has continuous partial derivatives with respect to the components of $I^{(k)}$.

Definition 5 (Strictly Monotone Payoff Gradient): Let $\nabla f^{(k)} = (\partial f^{(k)}/\partial I_1^{(k)}, \ldots, \partial f^{(k)}/\partial I_M^{(k)})$, and call $(\nabla f^{(k)})_{k\in\{1,\ldots,K\}}$ the payoff gradient of a smooth game $\mathfrak{G}$. We say that the payoff gradient is strictly monotone if

$$\sum_{k=1}^{K} \left(\nabla f^{(k)}(\mathbf{I}) - \nabla f^{(k)}(\mathbf{J})\right)^T \left(I^{(k)} - J^{(k)}\right) < 0 \qquad (23)$$

holds for all $\mathbf{I}, \mathbf{J} \in \mathcal{I}$ with $\mathbf{I} \neq \mathbf{J}$.

Theorem 2 [42]: Consider a smooth game $\mathfrak{G}$ with compact strategy sets. If the payoff gradient of $\mathfrak{G}$ is strictly monotone, then it has a unique correlated equilibrium, which places probability one on a unique pure strategy Nash equilibrium.

Definition 6 (Potential Game): A game $\mathfrak{G}$ is potential if there exists a potential function $v : \mathcal{I} \to \mathbb{R}$ such that

$$f^{(k)}\left(I, \mathbf{I}_k^-\right) - f^{(k)}\left(J, \mathbf{I}_k^-\right) = v\left(I, \mathbf{I}_k^-\right) - v\left(J, \mathbf{I}_k^-\right) \qquad (24)$$

for all $I, J \in \mathcal{I}^{(k)}$, and $k \in \{1, \ldots, K\}$.

Theorem 3 [43]: Let $\mathfrak{G}$ be a smooth potential game with a strictly concave potential function. Then, a strategy profile is the unique pure strategy Nash equilibrium if and only if it is the potential maximizer.

Lemma 1 [42]: Let $\mathfrak{G}$ be a smooth potential game. A potential of $\mathfrak{G}$ is strictly concave if and only if the payoff gradient of $\mathfrak{G}$ is strictly monotone.

2) Bandit Theory:

Lemma 2: Let R_n and R_{Ext} be given by (1) and (2), respectively. Then, for any $\delta \in (0, 1/2]$, we have[9]

$$\Pr\left(|R_n - R_{\mathrm{Ext}}| \leq \sqrt{\frac{n}{2}\ln\left(\frac{1}{\delta}\right)}\right) \geq 1 - 2\delta \qquad (25)$$

from which it follows that if $R_n \in o(n)$, then we have $R_{\mathrm{Ext}} \in o(n)$, with arbitrarily high probability.[10]

Proof: By comparing (1) and (2), it suffices to show that

$$\Pr\left(\left|\sum_{t=1}^{n} g_t(I_t) - \sum_{t=1}^{n} \bar{g}_t(\mathbf{P}_t)\right| \leq \sqrt{\frac{n}{2}\ln\left(\frac{1}{\delta}\right)}\right) \geq 1 - 2\delta. \qquad (26)$$

To this end, define $S := \sum_{t=1}^{n} g_t(I_t)$, where $g_t(I_t) \in [0, 1]$, $1 \leq t \leq n$, are independent random variables (see also Section II-A). Further note that $\bar{S} = \mathrm{E}[S] = \sum_{t=1}^{n} \bar{g}_t(\mathbf{P}_t)$, where $\mathrm{E}[\cdot]$ denotes the expectation. Therefore, by Hoeffding's inequality [32]

$$\begin{aligned}
&\Pr\left(|R_n - R_{\mathrm{Ext}}| \geq \sqrt{\frac{n}{2}\ln\left(\frac{1}{\delta}\right)}\right) \\
&\quad = \Pr\left(|S - \bar{S}| \geq \sqrt{\frac{n}{2}\ln\left(\frac{1}{\delta}\right)}\right) \\
&\quad \leq 2\exp\left(-\frac{2\frac{n}{2}\ln\left(\frac{1}{\delta}\right)}{n}\right) = 2\delta. \qquad (27)
\end{aligned}$$

Therefore, by using $\Pr(|R_n - R_{\mathrm{Ext}}| \leq \sqrt{(n/2)\ln(1/\delta)}) = 1 - \Pr(|R_n - R_{\mathrm{Ext}}| \geq \sqrt{(n/2)\ln(1/\delta)})$, the lemma follows. ■

Lemma 3: Let R_{Ext} be given by (2). Moreover, define $\tilde{R}_n = \max_{i=1,\ldots,N} \sum_{t=1}^{n} g_t(i) - \sum_{t=1}^{n} \tilde{g}_t(\mathbf{P}_t)$, where $\tilde{g}_t(\mathbf{P}_t) = \sum_{i=1}^{N} p_{i,t}\tilde{g}_t(i)$ and $\tilde{g}_t(i)$ is given by (12). Then, we have

$$\Pr\left(|\tilde{R}_n - R_{\mathrm{Ext}}| \leq \sqrt{\frac{n}{2}\ln\left(\frac{1}{\delta}\right)}\right) \geq 1 - 2\delta. \qquad (28)$$

Hence, for sufficiently small $\delta > 0$, $R_{\mathrm{Ext}} \in o(n)$ implies that $\tilde{R}_n \in o(n)$, with arbitrarily high probability.

Proof: Similar to the proof of Lemma 2, it follows from (2) and the definition of $\tilde{R}_n$ that it is sufficient to show that for $\delta \in (0, 1/2]$:

$$\Pr\left(\left|\sum_{t=1}^{n} \tilde{g}_t(\mathbf{P}_t) - \sum_{t=1}^{n} \bar{g}_t(\mathbf{P}_t)\right| \leq \sqrt{\frac{n}{2}\ln\left(\frac{1}{\delta}\right)}\right) \geq 1 - 2\delta. \qquad (29)$$

To this end, note that $\tilde{g}_t(\mathbf{P}_t) \in [0, 1]$, $1 \leq t \leq n$ are independent random variables. Moreover, since $\tilde{g}_t(i)$ is an unbiased estimate of $g_t(i)$, we have $\mathrm{E}[\sum_{t=1}^{n} \tilde{g}_t(\mathbf{P}_t)] = \sum_{t=1}^{n} \bar{g}_t(\mathbf{P}_t)$. Hence, defining $S = \bar{g}_t(\mathbf{P}_t) - \tilde{g}_t(\mathbf{P}_t)$ and proceeding as in the proof of Lemma 2 with Hoeffding's inequality in hand proves the lemma. ■

[9]Throughout this section and to simplify the notation, the player index (k) is omitted, unless ambiguity arises.

[10]Here and hereafter, the statement "$X(n) \in o(n)$ with arbitrarily high probability" for some nonnegative random sequence $X(n) \in \mathbb{R}$ means that the probability of $X(n) \notin o(n)$ can be made arbitrarily small, provided that some parameter is chosen sufficiently small.

Proposition 6: Let R_n be given by (1) and $\tilde{R}_n$ be defined as in Lemma 3. Then, $R_n \in o(n)$ implies that $\tilde{R}_n \in o(n)$.

Proof: Lemma 2 implies that $R_n \in o(n) \Rightarrow R_{\text{Ext}} \in o(n)$ with arbitrarily high probability, whereas by Lemma 3, we have $R_{\text{Ext}} \in o(n) \Rightarrow \tilde{R} \in o(n)$. Therefore, if $R_n \in o(n)$, then $\tilde{R} \in o(n)$ with arbitrarily high probability. ■

Theorem 4 [32]: Let $\Phi(\mathbf{U}) = \psi(\sum_{i=1}^{N} \phi(u_i))$, where $\mathbf{U} = (u_1, \ldots, u_N)$. Consider a selection strategy, which at time t selects action I_t according to distribution $\mathbf{P}_t$, whose elements $p_{i,t}$ are defined as

$$p_{i,t} = (1-\gamma_t)\frac{\phi'(R_{i,t-1})}{\sum_{i=1}^{N}\phi'(R_{i,t-1})} + \frac{\gamma_t}{N} \tag{30}$$

where $R_{i,t-1} = \sum_{s=1}^{t-1}(g_s(i) - g_s(I_s))$. Assume the following:

A1) $\sum_{t=1}^{n}(1/\gamma_t^2) = o(n^2/\ln(n))$.

A2) For all vectors $\mathbf{V}_t = (v_{1,t}, \ldots, v_{n,t})$ with $|v_{i,t}| \leq (N/\gamma_t)$, we have

$$\lim_{n\to\infty} \frac{1}{\psi(\phi(n))} \sum_{t=1}^{n} C(\mathbf{V}_t) = 0 \tag{31}$$

where $C(\mathbf{V}_t) = \sup_{\mathbf{U}\in\mathbb{R}^N} \psi'(\sum_{i=1}^{N}\phi(u_i)) \sum_{i=1}^{N}\phi''(u_i)v_{i,t}^2$.

A3) For all vectors $\mathbf{U}_t = (u_{1,t}, \ldots, u_{n,t})$, with $u_{i,t} \leq t$

$$\lim_{n\to\infty} \frac{1}{\psi(\phi(n))} \sum_{t=1}^{n} \gamma_t \sum_{i=1}^{N} \nabla_i \Phi(\mathbf{U}_t) = 0. \tag{32}$$

A4) For all vectors $\mathbf{U}_t = (u_{1,t}, \ldots, u_{n,t})$, with $u_{i,t} \leq t$,

$$\lim_{n\to\infty} \frac{\ln(n)}{\psi(\phi(n))} \sqrt{\sum_{t=1}^{n} \frac{1}{\gamma_t^2} \left(\sum_{i=1}^{N} \nabla_i \Phi(\mathbf{U}_t)\right)^2} = 0. \tag{33}$$

Then, the selection strategy satisfies

$$\lim_{n\to\infty} \frac{1}{n} \left(\max_{i=1,\ldots,N} \sum_{t=1}^{n} g_t(i) - \sum_{t=1}^{n} g_t(I_t) \right) = 0 \tag{34}$$

or equivalently, $R_n \in o(n)$, where R_n is given by (1).

Appendix B
Proof of Proposition 1

To prove Proposition 1, we use Theorem 2. As the strategy set is compact, to use this theorem, we show that 1) the game is smooth, and 2) the payoff gradient is strictly monotone.

According to our system model, by changing the channel index, i.e., $I'^{(k)}$, the channel gain and interference changes. Therefore, we define $I_1^{(k)} := |h_{kk',I'^{(k)}}|^2/(\sum_{q\in\mathcal{Q}^{(k)}} I''^{(q)} \times |h_{qk',I'^{(k)}}|^2 + N_0)$ and $I_2^{(k)} := I''^{(k)}$, from which we have $f^{(k)}(\mathbf{I}) = \log(I_1^{(k)} I_2^{(k)}) - \alpha I_2^{(k)}$. This results in[11]

$$\frac{\partial f^{(k)}}{\partial I_1^{(k)}} = \frac{1}{I_1^{(k)}} \tag{35}$$

$$\frac{\partial f^{(k)}}{\partial I_2^{(k)}} = \frac{1}{I_2^{(k)}} - \alpha. \tag{36}$$

[11]The constant factor $(\ln(2))$ is omitted from derivations as it does not have any impact on the final result.

Hence, by Definition 4, the game is smooth. On the other hand, given $f^{(k)}$, we have

$$\begin{aligned}
&\left(\nabla f^{(k)}(\mathbf{I}) - \nabla f^{(k)}(\mathbf{J})\right)^T \left(I^{(k)} - J^{(k)}\right) \\
&= \left[\frac{1}{I_1^{(k)}} - \frac{1}{J_1^{(k)}} \;\; \frac{1}{I_2^{(k)}} - \frac{1}{J_2^{(k)}}\right] \begin{bmatrix} I_1^{(k)} - J_1^{(k)} \\ I_2^{(k)} - J_2^{(k)} \end{bmatrix} \\
&= \left(\frac{1}{I_1^{(k)}} - \frac{1}{J_1^{(k)}}\right)\left(I_1^{(k)} - J_1^{(k)}\right) \\
&\quad + \left(\frac{1}{I_2^{(k)}} - \frac{1}{J_2^{(k)}}\right)\left(I_2^{(k)} - J_2^{(k)}\right)
\end{aligned} \tag{37}$$

which is always negative as for any $x, y > 0$ and $x \neq y$, $x - y > 0$ yields $(1/x) - (1/y) < 0$, and *vice versa*. Thus

$$\sum_{k=1}^{K} \nabla f^{(k)} < 0 \tag{38}$$

i.e., the payoff gradient is strictly monotone by Definition 5. As a result, by Theorem 2, the game has a unique correlated equilibrium that places probability one on the unique Nash equilibrium.

Appendix C
Proof of Proposition 2

First, we point out that the game is a potential game with a potential function being $v(\cdot) = \sum_{k=1}^{K} f^{(k)}(\cdot)$, by simply inserting $f^{(k)}$ and v into condition (24). Moreover, similar to Proposition 1, it can be easily shown that the game is smooth and the payoff gradient is strictly monotone (define $I_1^{(k)} := |h_{kk',I'^{(k)}}|^2/N_0$ and $I_2^{(k)} = I''^{(k)}$).[12] Therefore, by Lemma 1, the game is a smooth potential game with a strictly concave potential function. As a result, by Theorem 2, it has a unique pure strategy Nash equilibrium, which is the potential maximizer. On the other hand, by Lemma 2, the game has a unique correlated equilibrium that places probability one on the unique pure strategy Nash equilibrium.

Appendix D
Proof of Proposition 3

We first notice that $R_{\text{Ext}} \in O((nN)^{2/3}(\ln(N))^{1/3})$, as stated by the following lemma.[13]

Lemma 4: Consider a BEWAS that uses $\mathbf{P}_t = (p_{1,t}, \ldots, p_{N,t})$ to select an action among N possible choices, where $p_{i,t}$ is calculated as

$$p_{i,t} = (1-\gamma)\frac{\exp(\eta\tilde{R}_{i,t-1})}{\sum_{j=1,..,N}\exp(\eta\tilde{R}_{j,t-1})} + \frac{\gamma}{N} \tag{39}$$

and $\tilde{R}_{i,t-1}$ denotes the estimated accumulated regret of not playing action i.[14] Then, selecting γ and η, as given by Proposition 3, yields $R_{\text{Ext}} \in O((nN)^{2/3}(\ln(N))^{1/3})$.

[12]Details are similar to Proposition 1 and are thus omitted.

[13]Throughout this section and to simplify the notation, the player index (k) is omitted, unless ambiguity arises.

[14]This definition should not be mistaken for the general regret defined in Section II-A.

Proof: The proof is a direct corollary of [32, Th. 6.6]. ■

Given Lemma 4, we follow the approach in [33] for the rest of the proof.

Recall that by Section II-C and Algorithm 1, the mixed strategy of each player is defined by

$$\mathbf{P}_t = \sum_{(i\to j):i\neq j} \mathbf{P}_{(i\to j),t}\delta_{(i\to j),t}. \tag{40}$$

Hence

$$\bar{g}_t(\mathbf{P}_t) = \sum_{(i\to j):i\neq j} \bar{g}_t\left(\mathbf{P}_{(i\to j),t}\right)\delta_{(i\to j),t}. \tag{41}$$

Lemma 4 specifies the growth rate of external regret. On the other hand, as described in Section II-C, the convergence approach applies the BEWAS algorithm for $N(N-1)\leq N^2$ actions. Therefore, (41) together with Lemma 4 yields

$$\max\sum_{t=1}^{n}\bar{g}_t\left(\mathbf{P}_{(i\to j),t}\right) - \sum_{t=1}^{n}\bar{g}_t(\mathbf{P}_t) \in O\left((N^2n)^{2/3}\left(2\ln(N)\right)^{1/3}\right) \tag{42}$$

and the definition of internal regret ensures that $\max_{i\neq j} R_{(i\to j),n}\in O((N^2n)^{2/3}(\ln(N))^{1/3})$, which concludes the proof. Details can be found in [33] and, hence, are omitted.

Appendix E
Proof of Proposition 4

We first show that the algorithm has vanishing external regret, i.e., $R_{\text{Ext}}\in o(n)$, as formalized in the following.

Lemma 5: Assume that BEWAS uses $\mathbf{P}_t=(p_{1,t},\ldots,p_{N,t})$ to select an action among N possible choices, where $p_{i,t}$ is calculated as

$$p_{i,t} = (1-\gamma_t)\frac{\exp(\eta_t\tilde{R}_{i,t-1})}{\sum_{j=1,..,N}\exp(\eta_t\tilde{R}_{j,t-1})} + \frac{\gamma_t}{N} \tag{43}$$

and $\tilde{R}_{i,t-1}$ denotes the estimated accumulated regret of not playing action i. Then, for γ_t and η_t as given by Proposition 4, this strategy yields vanishing external regret, i.e., $R_{\text{Ext}}\in o(n)$.

Proof: By Proposition 6, if (34) is satisfied for a selection strategy (that is, if $R_n\in o(n)$), then the growth rate of the external regret caused by the bandit version of that strategy (which uses estimated rewards instead of true rewards) sublinearly grows in n, i.e., $\tilde{R}_n\in o(n)$. Therefore, to prove the proposition, we show that our selected parameters $\gamma_t=t^{-(1/3)}$ and $\eta_t=\gamma_t^3/N^2$ satisfy axioms A1–A4 of Theorem 4. In doing so, we omit for the lack of space simple calculus steps. Moreover, the reader should note that in our strategy, we have $\Phi(\mathbf{U})=(1/\eta_t)\ln(\sum_{i=1}^{N}\exp(\eta_t u_i))$.

A1) For $\gamma_t=t^{-(1/3)}$, we have

$$\sum_{t=1}^{n}\frac{1}{\gamma_t^2}=\sum_{t=1}^{n}t^{\frac{2}{3}}=\text{Harmonic Number}\left[n,-\frac{2}{3}\right]:=\text{H}_n\left[\frac{-2}{3}\right]. \tag{44}$$

Then

$$\lim_{n\to\infty}\frac{\ln(n)}{n^2}\sum_{t=1}^{n}\gamma_t^2=\lim_{n\to\infty}\frac{\ln(n)}{n^2}\text{H}_n\left[\frac{-2}{3}\right]=0. \tag{45}$$

A2) For $\psi(x)=(1/\eta_t)\ln(x)$ and $\phi(x)=\exp(\eta_t x)$, we obtain

$$C(\mathbf{V}_t)=\sup\left(\eta_t\sum_{i=1}^{N}v_{i,t}^2\right)=\frac{\eta_t N^3}{\gamma_t^2}. \tag{46}$$

Hence

$$\lim_{n\to\infty}\frac{1}{\psi\left(\phi(n)\right)}\sum_{t=1}^{n}C(\mathbf{V}_t)=\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}t^{\frac{-1}{3}} =\lim_{n\to\infty}\frac{1}{n}\text{H}_n\left[\frac{1}{3}\right]=0. \tag{47}$$

A3) For $\Phi(\mathbf{U})=(1/\eta_t)\ln(\sum_{i=1}^{N}\exp(\eta_t u_i))$, we have

$$\nabla_i\Phi(\mathbf{U}_t)=\frac{\exp(\eta_t u_i)}{\sum_{i=1}^{N}\exp(\eta_t u_i)}. \tag{48}$$

Therefore

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{\psi\left(\phi(n)\right)}\sum_{t=1}^{n}\gamma_t\sum_{i=1}^{N}\nabla_i\Phi(\mathbf{U}_t)&=\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}t^{\frac{-1}{3}}\sum_{i=1}^{N}\frac{\exp(\eta_t u_i)}{\sum_{i=1}^{N}\exp(\eta_t u_i)}\\&=\lim_{n\to\infty}\frac{1}{n}\text{H}_n\left[\frac{1}{3}\right]=0.\end{aligned} \tag{49}$$

A4) This simply follows by substituting (48) into (33).

Hence, all axioms A1–A4 are satisfied, and therefore, (34) holds, which, together with Proposition 6, completes the proof. ■

By Lemma 5, the external regret of BEWAS sublinearly grows in n. Therefore, similar to the proof of Proposition 3, (41) yields

$$\max\sum_{t=1}^{n}\bar{g}_t\left(\mathbf{P}_{(i\to j),t}\right)-\sum_{t=1}^{n}\bar{g}_t(\mathbf{P}_t)\in o(n) \tag{50}$$

and the definition of internal regret ensures that $\max_{i\neq j}R_{(i\to j),n}\in o(n)$, which concludes the proof.

Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

[1] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527–535, 1952.

[2] K. Liu, Q. Zhao, and B. Krishnamachari, "Distributed learning under imperfect sensing in cognitive radio networks," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2010, pp. 671–675.

[3] K. Liu and Q. Zhao, "Cooperative game in dynamic spectrum access with unknown model and imperfect sensing," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1596–1604, Apr. 2012.

[4] M. Di Felice, K. R. Chowdhury, and L. Bononi, "Learning with the bandit: A cooperative spectrum selection scheme for cognitive radio networks," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2011, pp. 1–6.

[5] S. Maghsudi and S. Stanczak, "Relay selection problem with no side information: An adversarial bandit approach," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2013, pp. 715–720.

[6] V. Krishnamurthy and D. V. Djonin, "Structured threshold policies for dynamic sensor scheduling—A partially observed Markov decision process approach," *IEEE Trans. Signal Process.*, vol. 55, no. 10, pp. 4938–4957, Oct. 2007.

[7] J. Nino-Moraand and S. S. Villar, "Sensor scheduling for hunting elusive hiding targets via Whittle's restless bandit index policy," in *Proc. Int. Conf. Netw. Games, Control Optim.*, Oct. 2011, pp. 1–8.
[8] P. Arora, C. Szepesvari, and R. Zheng, "Sequential learning for optimal monitoring of multi-channel wireless networks," in *Proc. IEEE Int. Conf. Comput. Commun.*, Apr. 2011, pp. 1152–1160.
[9] R. Zheng, T. Le, and Z. Han, "Approximate online learning for passive monitoring of multi-channel wireless networks," in *Proc. IEEE Int. Conf. Comput. Commun.*, Apr. 2013, pp. 3111–3119.
[10] T. Chen, L. Zhu, F. Wu, and S. Zhong, "Stimulating cooperation in vehicular ad hoc networks: A coalitional game theoretic approach," *IEEE Trans. Veh. Technol.*, vol. 60, no. 2, pp. 566–579, Feb. 2011.
[11] W. Saad, Z. Han, T. Basar, M. Debbah, and A. Hjorungnes, "Hedonic coalition formation for distributed task allocation among wireless agents," *IEEE Trans. Mobile Comput.*, vol. 10, no. 9, pp. 1327–1344, Sep. 2011.
[12] A. Mukherjee and H. M. Kwon, "General auction-theoretic strategies for distributed partner selection in cooperative wireless networks," *IEEE Trans. Commun.*, vol. 58, no. 10, pp. 2903–2915, Oct. 2010.
[13] J. Sun, E. Modiano, and L. Zheng, "Wireless channel allocation using an auction algorithm," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 5, pp. 1085–1096, May 2006.
[14] O. Ileri, M. Siun-Chuon, and N. B. Mandayam, "Pricing for enabling forwarding in self-configuring ad hoc networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 151–162, Jan. 2005.
[15] S. Maghsudi and S. Stanczak, "A hybrid centralized-decentralized resource allocation scheme for two-hop transmission," in *Proc. Int. Symp. Wireless Commun. Syst.*, Nov. 2011, pp. 96–100.
[16] M. Guo, Y. Liu, and J. Malec, "A new Q-learning algorithm based on the metropolis criterion," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 34, no. 5, pp. 2140–2143, Oct. 2004.
[17] X.Fang, D.Yang, and G. Xue, "Taming wheel of fortune in the air: An algorithmic framework for channel selection strategy in cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 783–796, Feb. 2013.
[18] Y. Song, Y. Fang, and Y. Zhang, "Stochastic channel selection in cognitive radio networks," in *Proc. IEEE Global Telecommun. Conf.*, Nov. 2007, pp. 4878–4882.
[19] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1466–1478, Oct. 2012.
[20] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multi-armed bandits," in *Proc. IEEE Annu. Conf. Decis. Control*, Dec. 2012, pp. 3960–3965.
[21] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y. D. Yao, "Opportunistic spectrum access in unknown dynamic environment: A game-theoretic stochastic learning solution," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1380–1391, Apr. 2012.
[22] Y. Xu, Q. Wu, L. Shen, J. Wang, and A. Anpalagan, "Opportunistic spectrum access with spatial reuse: Graphical game and uncoupled learning solutions," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 4814–4826, Oct. 2013.
[23] Y. Xu, Q. Wu, J. Wang, L. Shen, and A. Anpalagan, "Opportunistic spectrum access using partially overlapping channels: Graphical game and uncoupled learning," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3906–3918, Sep. 2013.
[24] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple users: Learning under competition," in *Proc. IEEE Int. Conf. Comput. Commun.*, Mar. 2010, pp. 1–9.
[25] W. Xu *et al.*, "Performance enhanced transmission in device-to-device communications: Beamforming or interference cancellation?" in *Proc. IEEE Global Commun. Conf.*, Dec. 2012, pp. 4296–4301.
[26] K. Liu, Q. Zhao, and B. Krishnamachari, "Decentralized multi-armed bandit with imperfect observations," in *Proc. Annu. Allerton Conf. Commun., Control, Comput.*, Sep. 2010, pp. 1669–1674.
[27] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1902–1916, Mar. 2013.
[28] A. Blum and Y. Mansour, "From external to internal regret," *J. Mach. Learn. Res.*, vol. 8, pp. 1307–1324, Dec. 2007.
[29] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multi-armed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, Jan. 2003.
[30] J. Kujala and T. Elomaa, "On following the perturbed leader in the bandit setting," in *Proc. Algorithmic Learn. Theory*, Oct. 2005, pp. 371–385.
[31] F. Germano and G. Lugosi, "Global Nash convergence of Foster and Young's regret testing," *Games Econom. Behavior*, vol. 60, no. 1, pp. 135–154, Jul. 2007.
[32] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[33] G. Stoltz and G. Lugosi, "Internal regret in on-line portfolio selection," *J. Mach. Learn.*, vol. 59, no. 1, pp. 125–159, May 2005.
[34] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, 2012.
[35] N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, *Algorithmic Game Theory*. Cambridge, U.K.: Cambridge University Press, 2007.
[36] S. Hart and A. Mas-colell, "A general class of adaptive strategies," *J. Econom. Theory*, vol. 98, pp. 26–54, May 2001.
[37] N. Cesa-Bianchi and G. Lugosi, "Potential-based algorithms in on-line prediction and game theory," *J. Mach. Learn.*, vol. 51, no. 3, pp. 239–261, Jun. 2003.
[38] Y. D. Yao and A. U. H. Sheikh, "Approximation to Bayes risk in repeated play," *Contrib. Theory Games*, vol. 3, no. 39, pp. 97–139, 1957.
[39] J. Kujala and T. Elomaa, "Following the perturbed leader to gamble at multi-armed bandits," in *Proc. Algorithmic Learn. Theory*, 2007, vol. 4754, pp. 166–180.
[40] M. Hutter and J. Poland, "Adaptive online prediction by following the perturbed leader," *J. Mach. Learn. Res.*, vol. 6, pp. 639–660, Dec. 2005.
[41] J. Nie and S. Haykin, "A Q-learning-based dynamic channel assignment technique for mobile communication systems," *IEEE Trans. Veh. Technol.*, vol. 48, no. 5, pp. 1676–1687, Sep. 1999.
[42] T. Ui, "Correlated equilibrium and concave games," *Int. J. Game Theory*, vol. 37, no. 1, pp. 1–13, Apr. 2008.
[43] A. Nayman, "Correlated equilibrium and potential games," *Int. J. Game Theory*, vol. 26, no. 2, pp. 223–227, 1997.

Setareh Maghsudi received the B.Sc. degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran, in 2007 and the M.Sc. degree in digital communications from the University of Kiel, Kiel, Germany, in 2010, respectively. She is currently working toward the Ph.D. degree with the Communications and Information Theory Group, Technical University of Berlin, Berlin, Germany.

Sławomir Stańczak (M'04–SM'11) received the degree in control systems engineering from Wroclaw University of Technology, Wroclaw, Poland, and the Technical University of Berlin (TU Berlin), Berlin, Germany, where he also received the Dipl.-Ing. and the Dr.-Ing. degrees with distinction (summa cum laude) in electrical engineering in 1998 and 2003, respectively.

Since 2006, he has been a Habilitation degree (venialegendi) holder and an Associate Professor (privatdozent) with TU Berlin. Since 1997, he has been involved in research and development activities in wireless communications. Since 2003, he has been leading a research group with the Fraunhofer Heinrich Hertz Institute, and since 2010, he has been the acting Director of the Heinrich-Hertz-Lehrstuhl at TU Berlin. In 2004 and 2007, he was a Visiting Professor with RWTH Aachen, Aachen, Germany, and, in 2008, a Visiting Scientist with Stanford University, Stanford, CA, USA. He is a coauthor of two books and more than 150 peer-reviewed journal articles and conference papers in the area of information theory, wireless communications, and networking. He is a coauthor of the book *Fundamentals of Resource Allocation in Wireless Networks*.

Dr. Stańczak received research fellowships from the German Research Foundation and was a winner of the Best Paper Award from the German Communication Engineering Society in 2014. He was a Cochair of the 14th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2013). Between 2009 and 2011, he was an Associate Editor for the *European Transactions for Telecommunications* (information theory). Since 2012, he has been an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.