Energy-Efficient Power Control for Multiple-Relay Cooperative Networks Using Q-Learning

Farshad Shams, Giacomo Bacci, *Member, IEEE*, and Marco Luise, *Fellow, IEEE*

***Abstract*—In this paper, we investigate the power control problem in a cooperative network with multiple wireless transmitters, multiple amplify-and-forward relays, and one destination. The relay communication can be either full- or half-duplex, and all source nodes interferer with each other at every intermediate relay node, and all active nodes (transmitters and relay nodes) interferer with each other at the base station. A game-theory-based power control algorithm is devised to allocate the powers among all active nodes: the source nodes aim at maximize their energy efficiency (in bits per Joule per Hertz), whereas the relays aim at maximizing the network sum-rate. We show that the proposed game admits multiple pure/mixed-strategy Nash equilibrium points. A Q-learning-based algorithm is then formulated to let the active players converge to the best Nash equilibrium point that combines good performance in terms of both energy efficiency and overall data rate. Numerical results show that the full-duplex scheme outperforms half-duplex configuration, Nash bargaining solution, the max-min fairness, and the max-rate optimization schemes in terms of energy efficiency, and outperforms the half-duplex mode, Nash bargaining system and the max-min fairness scheme in terms of network sum-rate.**

***Index Terms*—Energy efficiency, reinforcement learning algorithms, relay-assisted communications, full-duplex communications, mixed-strategy Nash equilibria, power control.**

I. Introduction

A. Motivation

EVERY new generation of wireless communication networks aims at using more and more intelligent wireless devices, that should be able to enhance the quality of service (QoS), increase the achievable throughput, and increase the battery life, while not overusing the common scarce wireless network resources, such as spectrum and battery energy. In recent years, the idea of *cooperative diversity* [1], [2] has been considered as a promising technique to significantly enhance the transmission of information.

The basic idea is that transmit data can be aided by several relays (intermediate nodes) in a communal manner with the goal of either decode-and-forward (DF), compress-and-forward (CF), or amplify-and-forward (AF) [3] the transmitted data by the sender, and then retransmit such data to the receiver. DF cooperative transmission requires the relay to successfully decode and then retransmit the (encoded) received information to the destination. The CF strategy exploits the side information available at the decoder. Applying CF, the relay estimates the sent data and then retransmit an encoded version of the received information. The AF cooperative communication technique forwards a strengthened copy of the received signal to the destination. Since AF is the easiest strategy to implement for relaying, in this work we will focus on AF relays, for the sake of mathematical tractability. A similar way of reasoning can be applied to the other two relay techniques, thus extending the present contribution to DF and CF relays as well.

Cooperative relaying communication can be accomplished in either half-duplex or full-duplex mode [3], [4]. In half-duplex mode, source, relay, and destination nodes are equipped with single transmit and receive antennas, so that a first time slot is used by the relay to receive the transmitted signal, and a second time slot is used to forward a delayed/processed version of the signal to the destination node while the source node is silent. In full-duplex mode, during the second time slot the transmitter sends a new symbol, while the relays is forwarding the transmitter's first symbol.

B. Related works

In the literature, there exist many attempts to properly allocate the resources of a relay-aided network. Just to mention a few relevant applications in this field, in [5], optimal power allocation for a narrow-band end-to-end connection assisted by a single DF relay is investigated. The goal of the proposed algorithm is achieving the minimum bit error rate (BER) under an overall power constraint. In [4], an optimal power control algorithm is introduced to maximize the channel capacity of a network with one source/destination pair and several linear AF relays. The power allocation problem for a two hop communication assisted by one relay in an ad hoc network is studied in [6], wherein different strategies (AF, DF, and CF)

Manuscript received March 16, 2014; revised August 26, 2014, and October 23, 2014; accepted November 4, 2014. The associate editor coordinating the review of this paper and approving it for publication was Prof. Jianwei Huang. This work was supported by the People Programme (Marie Curie Actions) of the European Union's FP7 under REA Grant agreement no. PIOF-GA-2011-302520 GRAND-CRU Game-theoretic Resource Allocation for wireless Networks based on Distributed and Cooperative Relaying Units, and by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM# (Grant agreement no. 318306).

F. Shams is with the Dept. Computer Science and Engineering, IMT Institute for Advanced Studies, Piazza San Ponziano, 6, 55100, Lucca, Italy (email: f.shams@imtlucca.it).

G. Bacci was with the Dipartimento di Ingegneria dell'Informazione, University of Pisa, Pisa, Italy, and he is now with Mediterranean Broadband Infrastructure (MBI) srl, Via F. Squartini, 7, 56121 Pisa, Italy (email: gbacci@mbigroup.it).

M. Luise is with the Dipartimento di Ingegneria dell'Informazione, University of Pisa, Via G. Caruso, 16, 56122 Pisa, Italy (email: marco.luise@unipi.it).

Part of this work has been presented to the *Intl. Conf. Cognitive Radio Oriented Wireless Networks (CROWNCOM) 2014*, Oulu, Finland (June 2014), and to the *XXXI URSI General Assembly and Scientific Symposium*, Beijing, China (Aug. 2014).

Digital Object Identifier XX.XXX/TWC.2015.YYYYYYY

at relay nodes are considered. The stochastic power control algorithm proposed in [7] enables nodes to dynamically learn about adjusting power in a multi source-destination pairs communication relayed by multiple intermediate AF nodes. Reference [8] introduces an algorithm to adjust the magnitude and the phase of the received signals at relay nodes in end-to-end communications assisted by multiple relays under individual power constraints. Reference [9] devises a max-min algorithm for the same scenario, but without direct link between two end-nodes, with the goal of maximizing the mutual information between two end-users.

It is reasonable to assume that transmitters do not call for any relaying if their demanded QoS is achievable through the direct link to destination. In a large reliable network, the questions of "to relay or not to relay" and "the best relay node(s) election policy" is based on the quality of the channel parameters. Shan *et al.* in [10] devise a cross-layer protocol to choose the relays which are able to increase the network throughput as much as possible. Reference [11] proposes a method to choose the best relay depending on geographic positions, based on the geographic random forwarding protocol proposed in [12].

A number of game-theoretic schemes (e.g., [13]–[26]) are proposed to allocate power resources to different transmitters and relay nodes, and to choose the best relay node. However, such approaches show a relatively high computational complexity, which could seriously undermine their applicability. Reference [13] investigates the problem of relay selection in cluster-based geographical routing for wireless sensor networks. The authors devise an auction mechanism which deals with contention avoidance at medium access control (MAC) layer and contention resolution in routing between source nodes and relays, in a cross-layer approach. References [14], in a single relay cooperative network, and [15], [16], in multi-AF-relay communications, seek an answer to the same questions as above applying game-theoretic tools. In these works, the powers of the source nodes are known, and the optimization variables are the power levels the selected relays must spend in a multiple end-to-end reliable network. The authors of [14], [15] propose an auction based resource allocation, and [16] employs a distributed Stackelberg game [27], whereas [17] studies the problem of power allocation in a relay-aided network with half-duplex relays as a noncooperative game, using the framework of quasi-variational inequality [28] to derive a practical iterative algorithm. Reference [18] introduces a relay selection algorithm for end-to-end communications relayed by multiple DF intermediate nodes with the aim of minimizing the outage probability. References [19], [20] devise repeated noncooperative games for the best relay assignment in a dynamic ad-hoc environment. The derived utility function in [19] achieves a max-min fairness data rate vector, whereas that of [20] aims at packet delivery reliability at the expense of delivery time and power consumption, and a dynamic learning based algorithm leads players to a correlated equilibrium [27].

In [21], a two-user network where each user can also work as a relay for the other is studied. By employing a two-user Nash bargaining game, the best power allocation is found. Reference [22] introduces a pricing game criterion which motivates cooperation via repayments to the relays in the network. The coalitional-based algorithm proposed by [23] determines the best relay nodes for each coalition of source nodes, each aiming at transmitting towards a base station (BS) at a minimum power expenditure. Reference [23] also presents a noncooperative power control game between coalitions. The work in [24] studies power control and pricing problems in a multi-user single-relay network, where the relay adjusts the energy price factor to maximize its revenue, and each user adjusts its transmit power to maximize its own utility. A distributed iterative scheme is proposed to achieve the unique Nash equilibrium point of the game. Reference [25] investigates the problem of peer-to-peer connection assisted by multiple relays and propose a noncooperative game to select the best relay and power control at the source node. Reference [26] investigates the problem of relay power control in a multi-user single-relay communication, and proposes a bargaining game solution to achieve a fair allocation of the relay power.

C. *Summary of the contributions*

In this contribution, we elaborate on consolidated tools, such as noncooperative game theory paired with reinforcement learning methods, to investigate the problem of power allocation at both source nodes and relay nodes, in a multiple-access multiple-relayed cooperative network scenario, and to eventually derive a low-complexity distributed algorithm. The intermediate relay nodes are assumed to be parallel, i.e., there is no link between the relay nodes, and each relay applies the AF strategy, and all source nodes interfere with each other at all relays and at the destination. To study the resource allocation problem, we consider both the full-duplex and the half-duplex modes from an information-theory viewpoint, and we then model the power control problem as a distributed noncooperative game, in which all source nodes and relays act as players that adjust their transmit powers in order to approach an energy efficient data rate vector with maximum spectral efficiency. Each source node has a minimum data rate demand, and individually tries to adjust its own transmitting power in order to achieve the highest Shannon channel capacity, while minimizing the transmit power consumption. Each relay node individually adapts its own transmitting power to maximize the network sum-rate while guaranteeing each user's minimum rate demand. To the best of our knowledge, this contribution is the first work that considers multiple relays in designing power allocation schemes at *both* source nodes and relay nodes in a *full-duplex* communication mode. Based on this formulation, we show that this game admits mixed-strategy Nash equilibria, and we propose a Q-learning-based algorithm [29] to achieve one of these equilibria, using a theoretical analysis, also supported by simulation results. The equilibrium point is shown to represent a tradeoff between the interests of the source nodes and those of the relays, i.e., it is the point at which the achieved data rates by source nodes is as much as possible close to max-rate solution, and the energy efficiency (EE) (measured in bits per Joule per Hertz) of each individual source node is as high as possible. Furthermore, we show that full-duplex relaying configuration, which is considered as one of the key technologies for next-generation wireless communications, enhances

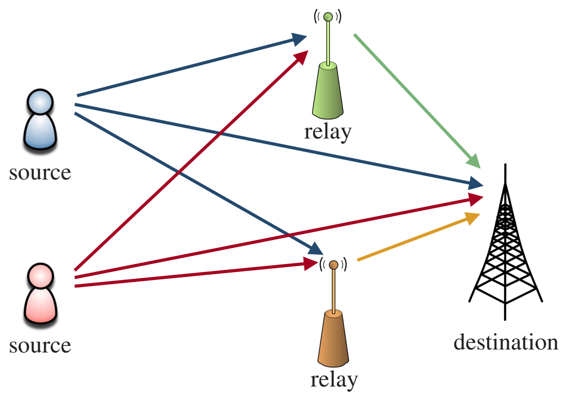


Fig. 1: Multi-relay assisted cooperative system model in an uplink direction.

the performance, in terms of both spectral efficiency and EE of the system, at the cost of a more elaborate resource management of the network, and of a more expensive and sensitive hardware architecture [30].

D. Structure of the paper

The remainder of the paper is structured as follows. Sect. II contains the model of the network and the formulation of the resource allocation problem as a noncooperative game, whose solution is computed in Sect. III using a reinforcement-learning method. Sect. IV compares the simulation results of the proposed algorithm with that of other methods available in the literature, and Sect. V concludes the paper.

II. Statement of the problem

A. Network model

The basic idea of cooperative relay-aided uplink of a multi-access network is depicted in Fig. 1, wherein multiple sources reach their common destination (the BS) through multiple parallel relays. In particular, we will study a network populated by multiple sources, multiple AF relay nodes, and one destination. The DF and CF cooperation strategies can be analyzed in a similar fashion. The relaying process is done in parallel, and both half- and full-duplex modes are investigated. Every relay transmits a new version of the observed signal to the base-station. In the full-duplex mode, unlike the half-duplex one, all links are always busy, and data reception and forwarding are performed concurrently. At the same time, all relays are synchronized with each other to amplify the respective received data and simultaneously forward them to the BS.

In the network, each source node $s \in \mathcal{S} = \{1, \ldots, S\}$ has a minimum end-to-end data-rate demand towards the base-station d, possibly exploiting (all) relay nodes $\mathcal{N} = \{1, \ldots, n, \ldots, N\}$. Every relay node connects all transmitters to the base-station d. Before starting to formulate our resource allocation problem, we review the formulation of the quasi-static additive white Gaussian noise (AWGN) capacity in cooperative communications using AF relaying (i.e., there is no regeneration at relay nodes). We assume that we have $S+1$ available Gaussian code spaces: $\mathcal{X}_s$ at each source node s

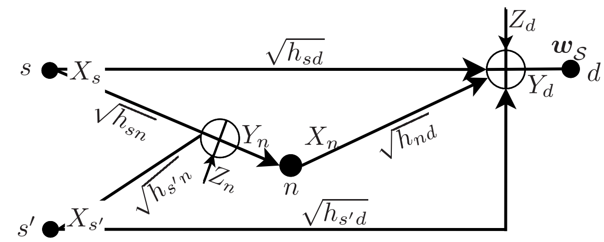


Fig. 2: Multiple sources, multiple parallel relays, one destination communication network scenario.

encoder, and $\mathcal{Y}_d$ at the destination's decoder. Each source s wishes to transmit a message w_s partitioned into a sequence of B sub-messages $w_s^1, ..., w_s^b, ..., w_s^B$ which are going to be reliably conveyed to the unique destination d. Each sub-message $w_s^b \in \mathcal{W}_s = \{0, 1, ..., 2^{nR_s} - 1\}$ is independently and fully encoded by its own transmitter. Each source node's encoder is a function $\mathcal{W}_s \to \mathcal{X}_s$ which maps a Gaussian random codeword $X_s[b]\left(w_s^b\right)$ to each sub-message w_s^b. Each symbol sequence $X_s[1]\left(w_s^1\right), \ldots, X_s[b]\left(w_s^b\right), \ldots, X_s[B]\left(w_s^B\right)$ is bounded by an individual average power of $\mathbf{E}\{X_s^2\} \leq \overline{p}_s$, with $\mathbf{E}\{X_s\} = 0$. Note that the $\mathcal{X}_s$'s are statistically uncorrelated, and there is no cooperation across source nodes. For simplicity we focus on a real-valued AWGN channel. Every relay is shared by all source nodes, and thus all transmitters' outputs interfere at every relay node. Then, in each block index b, each relay n generates a new signal $X_n[b]$, which is the scaled version of the analog observed signal $Y_n[b-1]$ (see Fig. 2):

$$Y_n[b] = \sum_{s\in\mathcal{S}} \sqrt{h_{sn}}\, X_s[b]\left(w_s^b\right) + Z_n \tag{1}$$

where h_{sn} denotes the real-valued power channel gain between source s and relay n, and $Z_n \sim \mathcal{CN}\left(0, \sigma_w^2\right)$ is the AWGN at the relay n. The received signals at the relays and destination are statistically correlated. Each relay node has its own power constraint $\mathbf{E}\{X_n^2\} \leq \overline{p}_n$, with $\mathbf{E}\{X_n\} = 0$. In each block index b, the received signal at each relay n is related to $\boldsymbol{w}_{\mathcal{S}}^{b-1} = (w_1^{b-1}, \ldots, w_S^{b-1})$.

Each AF relay node n amplifies the signal received in the previous block index, and generates the information $X_n[b]\left(\boldsymbol{w}_{\mathcal{S}}^{b-1}\right)$ and then re-sends it into the $n \to d$ link. In each block index b, each relay n scales the amplitude of the analog observed signal $Y_n[b-1]$ as:

$$\begin{aligned} X_n[b]\left(\boldsymbol{w}_{\mathcal{S}}^{b-1}\right) &= \alpha_n[b] \cdot Y_n[b-1] \\ &= \alpha_n[b] \cdot \left(\sum_{s\in\mathcal{S}} \sqrt{h_{sn}}\, X_s[b-1]\left(w_s^{b-1}\right) + Z_n\right) \end{aligned} \tag{2}$$

wherein the scaling factor α_n is chosen to satisfy each relay's power constraint $\mathbf{E}\{X_n^2\} \leq \overline{p}_n$. Hence, in AWGN mode, this bound translates into

$$|\alpha_n[b]| \leq \sqrt{\frac{\overline{p}_n}{\sigma_w^2 + \sum_{s\in\mathcal{S}} h_{sn} p_s}} \tag{3}$$

$$Y_d[b] = \underbrace{\sqrt{h_{sd}}X_s + \sum_{n\in\mathcal{N}} |\alpha_n|\sqrt{h_{sn}h_{nd}}X_s}_{(a)} + \underbrace{\sum_{s\neq s'\in\mathcal{S}} \sqrt{h_{s'd}}X_{s'} + \sum_{n\in\mathcal{N}}\sum_{s\neq s'\in\mathcal{S}} |\alpha_n|\sqrt{h_{s'n}h_{nd}}X_{s'}}_{(b)} + \underbrace{\sum_{n\in\mathcal{N}} |\alpha_n|\sqrt{h_{nd}}Z_n + Z_d}_{(c)} \tag{7}$$

$$\gamma_s = \frac{\left(\sqrt{h_{sd}} + \sum_{n\in\mathcal{N}} |\alpha_n|\sqrt{h_{sn}h_{nd}}\right)^2 \cdot p_s}{\sum_{s'\in\mathcal{S}, s'\neq s}\left(\sqrt{h_{s'd}} + \sum_{n\in\mathcal{N}} |\alpha_n|\sqrt{h_{s'n}h_{nd}}\right)^2 \cdot p_{s'} + \left(1 + \sum_{n\in\mathcal{N}} |\alpha_n|^2 h_{nd}\right)\cdot \sigma_w^2} \tag{8}$$

that does not depend on b based on the assumption of quasi-stationarity of the channel. As can be seen, if $\sigma_w^2 + \sum_{s\in\mathcal{S}} h_{sn}p_s \gg \overline{p}_n$, the relay is basically useless. The destination's decoder tries to reconstruct the whole sequences of B sub-messages $\boldsymbol{w}_{\mathcal{S}}^b = \left(w_1^b, \ldots, w_S^b\right)$. The decoding process is a (Gaussian) joint de-mapping function $\mathcal{Y}_d \to \mathcal{W}_1 \times \cdots \times \mathcal{W}_S$. The decoder at the destination jointly estimates the messages $\hat{\boldsymbol{w}}_{\mathcal{S}} = (\hat{w}_1, \ldots, \hat{w}_S)$, block-by-block or after having collected the whole sequence of $Y_d[1], \ldots, Y_d[B+1]$.

The probability of error at destination's decoder is formulated as:

$$P_e^n = \prod_{s\in\mathcal{S}} 2^{-nR_s} \cdot \sum_{\boldsymbol{w}_{\mathcal{S}}=(w_1,\ldots,w_S)} \Pr\left\{\hat{\boldsymbol{w}}_{\mathcal{S}} \neq \boldsymbol{w}_{\mathcal{S}} | \boldsymbol{w}_{\mathcal{S}} \text{ is sent}\right\} \tag{4}$$

which is based on the assumption that the messages are independent, and uniformly distributed over their respective alphabet ranges.

Let us try to model the performance at the receiver as a function of the signal parameters in full- and half-duplex modes, respectively.

1) Full-duplex mode: At the destination, in every block index b, the BS receives a combination of the signals $X_s[b]\left(w_s^b\right)$ $\forall\, s \in \mathcal{S}$ and $X_n[b]\left(\boldsymbol{w}_{\mathcal{S}}^{b-1}\right)$ $\forall\, n \in \mathcal{N}$ over a multiple-access channel (see Fig. 2):

$$Y_d[b] = \sum_{s\in\mathcal{S}} \sqrt{h_{sd}}X_s[b]\left(w_s^b\right) + \sum_{n\in\mathcal{N}} \sqrt{h_{nd}}X_n[b]\left(\boldsymbol{w}_{\mathcal{S}}^{b-1}\right) + Z_d \tag{5}$$

where $Z_d \sim \mathcal{CN}\left(0, \sigma_w^2\right)$. In fact, each received $Y_d[b] \in \mathcal{Y}_d$ is a linear combination of the sent signals based on $w_1^b, \ldots, w_S^b$ over the direct channels, and N different relayed signals based on $\boldsymbol{w}_{\mathcal{S}}^{b-1}$. Combining (2) and (5) yields the following observed signal at the destination:

$$\begin{aligned} Y_d[b] &= \sum_{s\in\mathcal{S}} \sqrt{h_{sd}}\, X_s[b]\left(w_s^b\right) \\ &+ \sum_{s\in\mathcal{S}}\sum_{n\in\mathcal{N}} |\alpha_n| \sqrt{h_{sn}\, h_{nd}}\, X_s[b-1]\left(w_s^{b-1}\right) \\ &+ \sum_{n\in\mathcal{N}} |\alpha_n|\sqrt{h_{nd}}\, Z_n + Z_d. \end{aligned} \tag{6}$$

Since every relay node amplifies whatever it receives, including noise and interference, and noting that all source nodes interferer at every relays, the AF technique is mainly useful in sufficiently high signal-to-noise ratio (SNR) environments. Increasing the amplification factor α_n has the effect of also increasing interference at the destination. If the network is able to adjust the power of the relay, the relay should thus transmit with a properly fine-tuned power. This is the reason that motivates us to include the relays into our resource allocation problem. When the relays' transmitting powers are adjusted, full-duplex relaying also proves to be better than half-duplex relaying, whose exploitation of capacity is limited.

We decompose now the received signal $Y_d[b]$ as in (7), shown at the top of this page, where (a) is the useful signal term, (b) is the multiple access interference (MAI), and (c) is the Gaussian thermal noise. Consequently, for each source node in the AWGN mode, we can express the received signal-to-interference-plus-noise ratio (SINR) at the destination as in (8), where

$$|\alpha_n| = \sqrt{\frac{p_n}{\sigma_w^2 + \sum_{s\in\mathcal{S}} h_{sn}p_s}}. \tag{9}$$

Using (8), we can approximate the (normalized) Shannon capacity achievable by the single point-to-point link source node s as:

$$R_s = \log_2\left(1 + \gamma_s\right) \qquad [\mathrm{b/s/Hz}]. \tag{10}$$

Hence, when inspecting (8), it is straightforward to note the coupling among the powers of all nodes (sources and relays) on the rate achievable by each source node.

2) Half-duplex mode: In half-duplex mode, the communication is done in two stages (times) and sources and relays share the time equally, and sources remains silent while relays transmit in the second half of the time. In the first transmitting stage (time), sources broadcast symbols to relays and the destination, while relays are silent. So, the received signal at each relay is the same as (1), and the received signal at the destination is described as:

$$Y_d[b] = \sum_{s\in\mathcal{S}} \sqrt{h_{sd}}\, X_s[b]\left(w_s^b\right) + Z_d. \tag{11}$$

Thus, in the first stage, the achieved SINR by each source s can be expressed by

$$\dot{\gamma}_s = \frac{h_{sd}p_s}{\sum\limits_{s'\in\mathcal{S}, s'\neq s} h_{s'd}p_{s'} + \sigma_w^2}. \tag{12}$$

In the second stage, sources are silent and each relay transmits a scaled version of the signal received in the first stage to the destination. The received signal at the destination is described by

$$Y_d[b] = \sum_{n\in\mathcal{N}} \sqrt{h_{nd}}\, X_n[b]\left(\boldsymbol{w}_{\mathcal{S}}^{b-1}\right) + Z_d \tag{13}$$

$$\ddot{\gamma}_s = \frac{\left(\sum_{n\in\mathcal{N}} |\alpha_n| \sqrt{h_{sn}\, h_{nd}}\right)^2 \cdot p_s}{\sum_{s'\in\mathcal{S}, s'\neq s}\left(\sum_{n\in\mathcal{N}} |\alpha_n| \sqrt{h_{s'n} h_{nd}}\right)^2 \cdot p_{s'} + \left(1+\sum_{n\in\mathcal{N}} |\alpha_n| \sqrt{h_{nd}}\right)^2 \cdot \sigma_w^2} \tag{15}$$

By substituting $X_n[b]$ from (2) in the above equation, the received signal at destination is described as:

$$Y_d[b] = \sum_{s\in\mathcal{S}} \sum_{n\in\mathcal{N}} |\alpha_n| \sqrt{h_{sn}\, h_{nd}}\, X_s[b-1] \left(w_s^{b-1}\right) + \sum_{n\in\mathcal{N}} |\alpha_n| \sqrt{h_{nd}}\, Z_n + Z_d \tag{14}$$

where $|\alpha_n|$ is equal to (9). By decomposing $Y_d[b]$ as in the case of the full-duplex mode, the achieved SINR by each source s in the second stage can be formulated by (15), shown at the top of this page.

Using (12) and (15), we can approximate the Shannon capacity achievable by the single point-to-point link source s as [31]:

$$R_s = \frac{1}{2} \cdot \log_2 \left(1 + \dot{\gamma}_s + \ddot{\gamma}_s\right) \qquad [\,\text{b/s/Hz}] \tag{16}$$

where the coefficient $1/2$ accounts for the fact that the throughput is accomplished in two consecutive stages for each block index b.

Hence, in both the full- and half-duplex schemes, when inspecting (10) and (16), it is straightforward to note the coupling among the powers of all nodes (sources and relays) on the rate achievable by each source node. It is important to note that a change in the transmit power of each active node, be either a source or a relay, significantly impacts the achieved SINRs by all source nodes, and this is the reason for studying this kind of interaction as a game [27]. A joint power control to be implemented by all nodes is thus highly desirable to increase the performance of the network, as described in the following subsection.

B. Problem formulation

Our power control strategy consists in finding a vector of transmit powers $[p_1, \ldots, p_S]$ and $[p_1, \ldots, p_N]$, wherein $p_s \leq \overline{p}_s$ and $p_n \leq \overline{p}_n$ represent the powers allocated by wireless terminal s and relay node n, respectively, bounded by their respective maximum levels, while the QoS constraint on the minimum rates $\underline{R}_s$ is fulfilled at each source node s.

To investigate the solution to this problem, we will use the analytical tools of game theory, by modeling the interaction among different nodes as the following noncooperative game:

$$\mathcal{G} = \left\{\mathcal{M},\ \{\mathcal{P}_m\}_{m\in\mathcal{M}},\ \{u_m\left(p_m; \mathbf{p}_{-m}\right)\}_{m\in\mathcal{M}}\right\} \tag{17}$$

where $\mathcal{M} = \mathcal{S} \cup \mathcal{N}$, with $|\mathcal{M}| = M = S + N$, is the set of all active nodes (both source and relay nodes), that represent the players of the game; $u_m\left(p_m; \mathbf{p}_{-m}\right)$ is the utility function of each user $m \in \mathcal{M}$, detailed in the following; and $\mathcal{P}_m$ is the *discrete* set of user s's transmit power, defined as $\mathcal{P}_m = \{0, \Delta p_m, 2\Delta p_m, \cdots, K_m \cdot \Delta p_m\} = \{k\Delta p_m\}_{k=0}^{K_m}$, where $(K_m + 1)$, with $1 \leq K_m < \infty$, denotes the number of power levels (including zero power), and $\Delta p_m = \overline{p}_m / K_m$ is the power step, with $\overline{p}_m$ denoting user m's maximum transmit power. Without loss of generality, for simplicity we assume $K_m = K\ \forall\, m \in \mathcal{M}$. Note that, as the number of players M is finite, and the number $K+1$ of actions available to each player is also finite, $\mathcal{G}$ is called a *finite* game [27].

To account for the different needs demanded by the two classes of users, we will define two different utility functions, one for the source nodes, and another one for the relay nodes. The goal of each source node is to trade off its achieved channel capacity with its minimum power consumption. On the other hand, the goal of each relay node is to adjust its transmit power in order to increase the spectral efficiency of the source nodes. To this end, we define the utility function for source nodes as follows:

$$u_s\left(p_s\,;\, \mathbf{p}_{-s}\right) = \frac{R_s(\mathbf{p})}{p_s + p_c} \qquad \text{s.t.} \quad R_s \geq \underline{R}_s \ \text{ and } \ p_s > 0 \tag{18}$$

where $\mathbf{p}_{-s} = \mathbf{p} \setminus p_s$ is the power vector of all nodes (including both sources and relays) excluding source s's power p_s, with $\mathbf{p}$ denoting the $M \times 1$ power vector collecting the transmit powers by all M nodes in the network; $R_s(\mathbf{p}) = \log\left(1 + \gamma_s(\mathbf{p})\right)$ is Shannon capacity achievable by transmitter s, with the SINR $\gamma_s(\mathbf{p}) = \gamma_s(p_s\,;\, \mathbf{p}_{-s})$ defined as in (8); and $p_c > 0$ is the background power consumption of the terminal, independent of the transmission rate and modeled as in [32], [33]. To explicitly account for the constraints in (18), the payoff of a source node is equal to zero when either $R_s < \underline{R}_s$ or $p_s = 0$.

On the other hand, as the relays are just ancillary nodes that aim at increasing the network performance while not showing significant power-saving constraints, their main purpose is to increase the spectral efficiency of the system. Hence, we define the utility function for each relay node as follows:

$$u_n\left(p_n\,;\, \mathbf{p}_{-n}\right) = \sum_{s\in\mathcal{S}} \frac{R_s(\mathbf{p})}{p_c} \qquad \text{s.t.} \quad R_s \geq \underline{R}_s\ \forall\, s \in \mathcal{S} \tag{19}$$

where the parameter p_c is used just to let $u_n\left(p_n\,;\, \mathbf{p}_{-n}\right)$ have the same unit of measure of $u_s\left(p_s\,;\, \mathbf{p}_{-s}\right)$; $R_s(\mathbf{p}) = R_s(p_n\,;\, \mathbf{p}_{-n}) = \log\left(1 + \gamma_s(p_n\,;\, \mathbf{p}_{-n})\right)$ is again the Shannon capacity achievable by transmitter s, in which we explicitly emphasize the impact of each relay node n's transmit power p_n on the SINR γ_s as outlined in (8), since each relay n has control only on its own p_n. Similarly to (18), we assign $u_n\left(p_n\,;\, \mathbf{p}_{-n}\right) = 0$ when $R_s(p_n) < \underline{R}_s$ for at least one source node s. Note that, for a given power vector $\mathbf{p} = \{p_m\}_{m=1}^{M}$, all relay nodes $n \in \mathcal{N}$ earn the same payoff $u_n\left(\mathbf{p}\right)$. This is because we model the relays as altruistic players which aim at maximizing the overall data rate. The transmit power selected by each relay impacts on the data rate of all source nodes, and hence each relay individually tries to increase the performance of the network, which measures its own utility. A close inspection of the utilities (18) and (19) reveals that including the

QoS constraints on the minimum rate demands $\underline{R}_s$ introduces a coupling between the power sets that provide positive utilities for all players $m \in \mathcal{M}$. While there is a competition between players $m \in \mathcal{S}$ to individually increase their own EE, the players $m \in \mathcal{N}$ interact with the source nodes to increase the sum-rate. In the remainder of the paper, we assume that the network setup (that includes channel realizations and set of minimum rate demands) is such that all sources are able to meet their requirements $\underline{R}_s$, otherwise the problem is declared to be unfeasible, as better detailed in Sect. III.[1]

Given the game formulation detailed above, it is apparent that there exists a tradeoff between achieving a high network sum-rate (in terms of achievable Shannon capacity) and maximizing the EE of each individual source node (without considering the power expenditure at the relay nodes). Each source node tries to pick its optimal power transmission level in order to to maximize the surplus achieved Shannon channel capacity above the proper demand, while each relay node tries to pick its best transmit power such that the achieved data rate vector of source nodes coincides with the maximum sum rate. Note that the source nodes' power allocations are coupled in a conflicting way, as increasing each sources' power level increases its own SINR while generating a higher interference level to the others at both the relays and the destination. Similarly, the relay nodes' power allocation affects the performance of all sources in a conflicting way according to the coupling among source-to-relay and relay-to-destination links. To solve the maximization problem[2]

$$p_m^* = \arg \max_{p_m \in \mathcal{P}_m} u_m \left(p_m\,;\, \mathbf{p}_{-m}\right) \tag{20}$$

in a scalable and distributed way, and thus keeping its complexity low, we can make use of the analytical tools of noncooperative game theory [27], whose aim is to help us predict the behavior of rational agents with conflicting interests competing for some common resources. One such prediction is the *pure-strategy Nash equilibrium*: each player chooses an action $p_m \in \mathcal{P}_m$ that is its "best response" (in the sense of utility maximization) to the other players' choices. Unfortunately, not all games have pure-strategy Nash equilibria.

A generalization of this concept is represented by *mixed-strategy Nash equilibria*, which are probabilistic distributions on the set of actions available to each player. In general, a mixed-strategy vector $\boldsymbol{\sigma}_m = \{\sigma_m(p_m)\}_{p_m \in \mathcal{P}_m}$ is a probability distribution over the strategy set $\mathcal{P}_m$, where $\sigma_m : \mathcal{P}_m \to [0,1]$ is the probability function associated to action p_m, with $\sum_{p_m \in \mathcal{P}_m} \sigma_m(p_m) = 1$. A mixed-strategy Nash equilibrium is a joint probabilistic distribution on the set of actions of each player, with the property that each player's distribution is a best response to the others' distributions, i.e., it is the maximizer of each player's *expected payoff* to the joint probabilistic distribution of all others:

[1]The theoretical analysis of the feasibility of the problem is out of scope of the present contribution, and is left as a future work. The feasibility of the problem is assessed numerically, as detailed in Sect. III.

[2]Note that, although all relay nodes earn the same payoff for a given power allocation $\mathbf{p}$, the optimal transmit powers p_n^* for $n \in \mathcal{N}$ are in general not equal, due to the different channel link conditions.

Definition 1: A mixed-strategy Nash equilibrium for a game $\mathcal{G}$ is a M-tuple of vectors $[\boldsymbol{\sigma}_1^*, \ldots, \boldsymbol{\sigma}_M^*]$, with $\boldsymbol{\sigma}_m^* \in [0,1]^{K+1}$, such that, for all $m \in \mathcal{M}$, and for all $\boldsymbol{\sigma}_m \in [0,1]^{K+1}$,

$$\begin{aligned}\sum_{p_m \in \mathcal{P}_m} \sum_{\mathbf{p}_{-m} \in \boldsymbol{\mathcal{P}}_{-m}} \boldsymbol{\sigma}_{-m}^*(\mathbf{p}_{-m}) \sigma_m^*(p_m) u_m\left(p_m\,;\,\mathbf{p}_{-m}\right) \geq \\ \geq \sum_{p_m \in \mathcal{P}_m} \sum_{\mathbf{p}_{-m} \in \boldsymbol{\mathcal{P}}_{-m}} \boldsymbol{\sigma}_{-m}^*(\mathbf{p}_{-m}) \sigma_m(p_m) u_m\left(p_m\,;\,\mathbf{p}_{-m}\right)\end{aligned} \tag{21}$$

where $p_m \in \mathcal{P}_m$ is a pure strategy, $\boldsymbol{\mathcal{P}}_{-m}$ is the cartesian product of all strategy sets other than m, i.e.

$$\boldsymbol{\mathcal{P}}_{-m} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_{m-1} \times \mathcal{P}_{m+1} \times \cdots \times \mathcal{P}_M \tag{22}$$

and, likewise, $\boldsymbol{\sigma}_{-m}^*(\mathbf{p}_{-m})$ is the product of probability of the opponents' joint strategy $\mathbf{p}_{-m}$, given by

$$\begin{aligned}&\boldsymbol{\sigma}_{-m}^*(\mathbf{p}_{-m}) = \\ &\quad = \sigma_1^*(p_1) \cdots \sigma_{m-1}^*(p_{m-1}) \cdot \sigma_{m+1}^*(p_{m+1}) \cdots \sigma_M^*(p_M) \\ &\quad = \textstyle\prod_{i \in \mathcal{M}, i \neq m} \sigma_i^*(p_i)\end{aligned} \tag{23}$$

where the product stems from the independence of each player's action with respect to the other ones. ■

The interest in considering mixed-strategy Nash equilibria for the game $\mathcal{G}$ introduced above comes from the following theorem, which constitutes a seminal result in the framework of noncooperative game theory.

Theorem 1 (Nash [34]): In every finite static game $\mathcal{G}$ there exists at least one mixed-strategy Nash equilibrium.
The proof makes use of the Brouwer-Kakutani fixed-point theorem and can be found in [27]. ■

Once the existence of such equilibrium point is assessed, the next question is how to find it in a practical system.

III. Computing mixed-strategy Nash equilibria via Q-learning

We aim now at finding the mixed-strategy Nash equilibrium points of the proposed game. In general, there does not exist a specific algebraic method to solve mixed-strategy best response equations, and solving such problems is typically NP-hard, in particular when the number of strategies and players is huge. The question we seek to answer here is: *"How to find a mixed-strategy Nash equilibrium when there are a large number of strategies and players?"*

The computational complexity of solving (21) is discussed in [35], which shows that the problem is NP-hard, and the convergence to such equilibrium point(s) might take prohibitively long times. For a two-player game, in the literature there are some efforts for explicitly computing the Nash equilibrium points. The main algorithm for doing so is the Lemke-Howson algorithm [36]. More effective algorithms in terms of computational complexity are introduced in [37], [38]. The algorithm GAMBIT [39], and the techniques introduced in [40] and [41] "approximate" the Nash equilibrium with a certain error. Furthermore, three methods for detecting the Nash equilibrium based on computational intelligence methods are described in [42]. Unfortunately, all these algorithms focus either on two-player or three-player games with a (very) small number of strategies.

To address the question that opened this section, we can resort to *learning methods* [43], that are able to let the players interact so that they can learn about the game and gather information about each other in the course of playing, to finally end up with computing the mixed-strategy Nash equilibrium point. Learning in strategic environments presents in fact some phenomena not found in individual decision-making, because the environment in which each individual gains experience includes the others players, whose behavior changes as they gain experience as well. Roth and Erev in [44] show how players converge to equilibrium points by applying a simple model of learning, also highlighting the impact of different parameters of the learning model on the convergence speed of the algorithm.

Another variant of the learning techniques is the *reinforcement learning* [45], well-suited for multi-agent systems, where each agent knows little about the other agents. Reinforcement learning deals with a learning agent that interacts with a unknown and possibly stochastic environment, in order to learn optimal control policies. During the play, each agent adapts its behavior, i.e., evolves by learning, based on its own experience gathered by the environment and adapted to the other agents' behaviors. To cache the results of experience, most reinforcement learning methods use Q-*value* functions [29], that lead agents to the optimal strategy (i.e., the mixed-strategy Nash equilibrium). Q-learning has been deeply investigated, and possesses a firm foundation in the theory of Markov decision processes. In addition, it is also quite easy to use, and has been widely employed in many fields of application, such as communications and networking [46].

For the reader's convenience, let us introduce some typical notation used for Q-learning-based techniques, specifically adapted to the game $\mathcal{G}$ introduced in Sect. II-B. Q-learning defines a value Q_m for each agent m. At the beginning of the algorithm, each agent starts with an arbitrary initial value $Q_m^{t=0}(\mathbf{p})$ for all pure joint actions $\mathbf{p} = (p_m;\, \mathbf{p}_{-m}) \in \mathcal{P}_1 \times \ldots \times \mathcal{P}_M$. At each time step t, according to the seminal paper of Watkins and Dayan [29], each agent m:

1) computes its optimal probability associated to each action $p_m \in \mathcal{P}_m$, given joint strategies $(p_m;\, \mathbf{p}_{-m})$, for all $\mathbf{p}_{-m} \in \boldsymbol{\mathcal{P}}_{-m}$, based on the received payoff $u_m\,(p_m;\, \mathbf{p}_{-m})$;[3]
2) according to some individual (unilateral) criteria, chooses its best joint strategy $\tilde{\mathbf{p}}_m \in \mathcal{P}_m \times \boldsymbol{\mathcal{P}}_{-m}$; and
3) adjusts the proper Q-value according to:

$$\begin{aligned} Q_m^{t+1}(\tilde{\mathbf{p}}_m) \longleftarrow &\left(1 - f_m^{t+1}\right) \cdot Q_m^t(\tilde{\mathbf{p}}_m) \\ &+ f_m^{t+1} \cdot \left(u_m\,(\tilde{\mathbf{p}}_m) + \delta_m \cdot Q_m^t(\tilde{\mathbf{p}}_m)\right), \end{aligned} \tag{24}$$

where $\delta_m \in [0, 1]$ is a "discount factor"; $u_m\,(p_m;\, \mathbf{p}_{-m})$ is the utility defined as in (18)-(19); and f_m^t is the *learning rate*, which is a function of t. In [29], it is shown that the Q-learning algorithm converges provided that, for all $m \in \mathcal{M}$, the following conditions hold:

- the payoff $u_m\,(p_m;\, \mathbf{p}_{-m})$ is bounded;
- the learning rate f_m is such that $0 \le f_m^t < 1$;
- $\sum_{t=1}^{\infty} f_m^t = \infty$; and
- $\sum_{t=1}^{\infty} \left(f_m^t\right)^2 < \infty$.

[3]Note that, if the network setup is such that the problem is not feasible, as discussed in Sect. II-B, then the algorithm exits, as better detailed at the end of this section.

Even-Dar *et al.* in [47] show that, with a polynomial learning rate defined by $f_m^t = t^{-\alpha}$, such that $\alpha \in (0.5, 1)$, the convergence rate is the minimum one. In the simulation results, we will apply $f_m^t = t^{-0.8}$ as the learning rate. The Q-learning algorithm and other reinforcement learning techniques, such as the Boltzmann-Gibbs learning and the fuzzy learning, only guarantee the convergence to an operating point. This implies that they do not *per se* guarantee the convergence to the optimal joint strategy representing a Nash equilibrium [48]. There are some learning algorithms [39]–[41] that estimate the Nash equilibrium with a certain error. On the contrary, in this work, we will introduce a QNash learning algorithm which converges to the Nash equilibrium point with probability one.

One of the major challenges of Q-learning is the strategy of choosing the best joint action by each individual agent. When the number of strategies and players are large, the number of time steps to achieve an optimal joint action exponentially increases. It is fairly clear that the best manner is to start with "exploration" of different strategies; and then focus on "exploitation" of the strategies with the best value of Q. Kaelbling *et al.* in [49] make use of the Boltzmann function [45] as an efficient strategy selection to strike a balance between exploration and exploitation. At each time step $t+1$, every player will individually select the joint strategy $\mathbf{p} = (p_m;\, \mathbf{p}_{-m})$ with a probability $\pi_m^t(p_m;\, \mathbf{p}_{-m})$ defined as

$$\pi_m^t(p_m;\, \mathbf{p}_{-m}) = \frac{\left[\exp\left\{u_m\,(p_m;\, \mathbf{p}_{-m})\right\}\right]^{\Omega_m^t}}{\sum_{\mathbf{p} \in \boldsymbol{\mathcal{P}}} \left[\exp\left\{u_m\,(\mathbf{p})\right\}\right]^{\Omega_m^t}} \tag{25}$$

wherein $\boldsymbol{\mathcal{P}} = \mathcal{P}_m \times \boldsymbol{\mathcal{P}}_{-m}$, and $\Omega_m^t = \sum_{\ell=0}^{t} (\delta_m)^{\ell} / T_m^t$ properly discounts the reward for taking joint action $(p_m;\, \mathbf{p}_{-m})$ by the user m in time step t, with the parameter T_m^t being a function which provides a randomness component to control exploration and exploitation of the actions. In practice, the "temperature function" T_m^t is a decreasing function over time to decrease exploration and increase exploitation. High values of T_m^t yield a small $\pi_m^t(p_m;\, \mathbf{p}_{-m})$ value, and this encourages exploration, whereas a low T_m^t makes $Q_m(p_m;\, \mathbf{p}_{-m})$ more important, and this encourages exploitation. In the simulation results, we will show that the temperature function significantly impacts the convergence to the Nash equilibrium point.

A great deal of attention in multi-agent learning schemes has been paid to coordination in order to converge to a desired equilibrium point. Hu and Wellman in [50] design a multi-agent stochastic Q-learning method, called *QNash algorithm*, which converges to one of the (mixed or pure) Nash equilibria under specific conditions. They show that, when the Nash equilibrium is unique, the proposed algorithm converges to it with probability one, and when there are multiple Nash equilibria, the agents surely converge to one of them. When there exist multiple Nash equilibria in the game, we can properly tune the learning parameters and the game parameters (notably, the temperature function) to let the QNash algorithm converge to the optimal strategies, as described in Sect. IV.

$$\sigma_m^*(p_m) = \sum_{p_1\in\mathcal{P}_1} \cdots \sum_{p_{m-1}\in\mathcal{P}_{m-1}} \sum_{p_{m+1}\in\mathcal{P}_{m+1}} \cdots \sum_{p_M\in\mathcal{P}_M} \pi_m^t(p_1,\dots,p_{m-1},p_m,p_{m+1},\cdots,p_M) \tag{28}$$

Initialization:
  `for` every $m \in \mathcal{M}$ `do`
    `for` every $\mathbf{p} \in \boldsymbol{\mathcal{P}} = \mathcal{P}_m \times \boldsymbol{\mathcal{P}}_{-m}$ `do`
      set $Q_m^0(\mathbf{p}) = u_m(\mathbf{p})$;
    `end for`
    set $\tilde{\mathbf{p}}_m$ using (26);
  `end for`
Feasibility check:
  `if` $\left(\exists\, m \in \mathcal{M} \quad \text{s.t.} \quad Q_m^0(\mathbf{p}) = 0\ \forall\, \mathbf{p} \in \boldsymbol{\mathcal{P}}\right)$
  `then` **exit;**
  `else` **set** $t = 0$ and a tolerance $\varepsilon \ll 1$;
Loop:
  `repeat`
    {*updating probabilities*}
    `for` every $m \in \mathcal{M}$ `do`
      `for` every $\mathbf{p} \in \boldsymbol{\mathcal{P}} = \mathcal{P}_m \times \boldsymbol{\mathcal{P}}_{-m}$ `do`
        update $\pi_m^t(\mathbf{p})$ using (25);
      `end for`
    `end for`
    {*updating Q-values*}
    `for` every $m \in \mathcal{M}$ `do`
      update $Q_m^{t+1}(\tilde{\mathbf{p}}_m)$ using (27);
    `end for`
    update $t = t + 1$;
  `until` $\max_{m\in\mathcal{M}} \left|Q_m^t(\tilde{\mathbf{p}}_m) - Q_m^{t-1}(\tilde{\mathbf{p}}_m)\right| \le \varepsilon$
Output:
  compute the mixed-strategy Nash equilibrium $\sigma_m^*(p_m)$ using (28).

Table I: The QNash algorithm [50].

The algorithm proposed in this paper, which adapts the one derived in [50], starts with an initialization of the Q-values to $Q_m^{t=0}(p_m;\, \mathbf{p}_{-m}) = u_m(p_m;\, \mathbf{p}_{-m})$ for all $m \in \mathcal{M}$ and for all $\mathbf{p} \in \boldsymbol{\mathcal{P}}$, also selecting the best joint strategy $\tilde{\mathbf{p}}_m$, based on

$$\tilde{\mathbf{p}}_m = \arg\max_{\mathbf{p}\in\boldsymbol{\mathcal{P}}} u_m(\mathbf{p})\,. \tag{26}$$

Then, at the beginning of each time step t, each agent m individually updates the table $\boldsymbol{\pi}_m^t = [\pi_m^t(\mathbf{p})]$ for all $\mathbf{p} \in \mathcal{P}_m \times \boldsymbol{\mathcal{P}}_{-m}$ using (25), and updates its own Q-value according to the following recursion:

$$Q_m^{t+1}(\tilde{\mathbf{p}}_m) \longleftarrow \left(1 - f_m^{t+1}\right) \cdot Q_m^t(\tilde{\mathbf{p}}_m) + f_m^{t+1} \cdot \left(u_m(\tilde{\mathbf{p}}_m) + \delta_m \cdot Q_m^t(\tilde{\mathbf{p}}_m) \cdot \prod_{i=1}^{M} \pi_i^t(\tilde{\mathbf{p}}_m)\right). \tag{27}$$

In order to calculate this scalar product, agent m would need to know the information about others agents at beginning of each time step. Hu and Wellman in [50] proves that the Q-values updates by recursion (27) for all agents $m \in \mathcal{M}$ converge (in the long-run) with probability one to a Nash equilibrium point. To show the convergence of Q-values updated by (27), and to prove that the convergence point of the QNash learning algorithm is actually the Nash equilibrium point, Hu and Wellman in [50] resort to fixed-point theorem [27]. The computational complexity of the proposed algorithm is unknown, however it is obvious that learning rate, discount factor, and probability distribution impact the convergence speed, as discussed in Sect. IV. When all Q-values converge, the profile $\boldsymbol{\sigma}_m^* = [\sigma_m^*(p_m)]$, with elements (28) for all $p_m \in \mathcal{P}_m$, coincides with a mixed-strategy Nash equilibrium of the game $\mathcal{G}$. For the reader's convenience, the QNash learning algorithm is summarized in Table I. Note that, if there exists some agent $m \in \mathcal{M}$, such that its initial Q-values $Q_m^{t=0}(p_m;\, \mathbf{p}_{-m}) = u_m(p_m;\, \mathbf{p}_{-m}) = 0$ for all $\mathbf{p} \in \boldsymbol{\mathcal{P}}$, then its strategy set becomes empty. In this case, the problem is declared to be unfeasible, as the network resources are not enough to accommodate all users given their QoS constraints and the channel realizations. Since we are focusing on an infrastructure network, this task can be naturally accomplished by the destination node (the BS). As is customarily assumed in all infrastructure networks, the source nodes undergo a phase of network association with the BS: during this stage, the BS can check whether the network possesses enough resources to accommodate the incoming nodes with their own QoS requirements, and thus perform some form of admission control.

It is worth emphasizing that, as is apparent in Table I, most complexity of the algorithm is devoted to *i*) initializing the utilities $\{u_m(\mathbf{p})\}_{\mathbf{p}\in\boldsymbol{\mathcal{P}},m\in\mathcal{M}}$, that requires a matrix with $M \times (K+1)^M$ entries; and *ii*) updating the probabilities $\{\pi_m^t(\mathbf{p})\}_{\mathbf{p}\in\boldsymbol{\mathcal{P}},m\in\mathcal{M}}$ at each step of the algorithm t. As better detailed in Sect. IV, the algorithm achieves good results even with very low power levels K, and hence the size of the matrix can be reasonably kept low for a large range of M. However, in order to guarantee the scalability of the system (i.e., very large M's), we need to equip the BS with very high computational power (i.e., by exploiting some parallel computing) to initialize the utility matrix. Interestingly, computing $\pi_m^t(\mathbf{p})$ as in (25) only requires the exponentiation of $\exp\{u_m(\mathbf{p})\}$ (that can be easily accesses by looking up the table built during the initialization step) with the coefficient Ω_m^t, that does *not* depend on $\mathbf{p}$. Hence, the complexity of (25) can be significantly reduced in practice by an efficient use of look-up tables.

IV. Numerical results

In this section, we compare the performance of the proposed algorithm with that of well known power allocation schemes, namely: *i*) Nash bargaining solution (NBS) [51]; *ii*) max-min fairness solution [52]; and *iii*) max rate solution [53]. Then, we show that the proposed power control at both source nodes and relay nodes outperforms a power control algorithm where only the source nodes update their power levels.

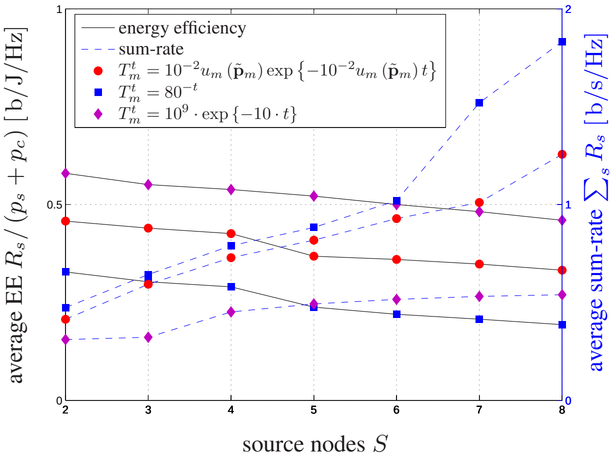


Fig. 3: Performance of the QNash algorithm as a function of S for different exploration policies.

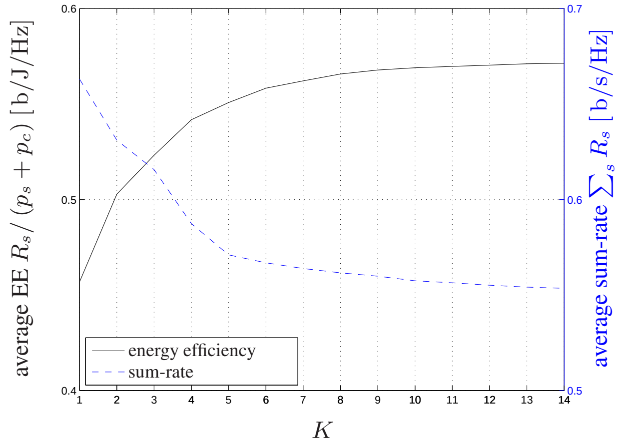


Fig. 5: Performance of the QNash algorithm as a function of K ($S = N = 3$).

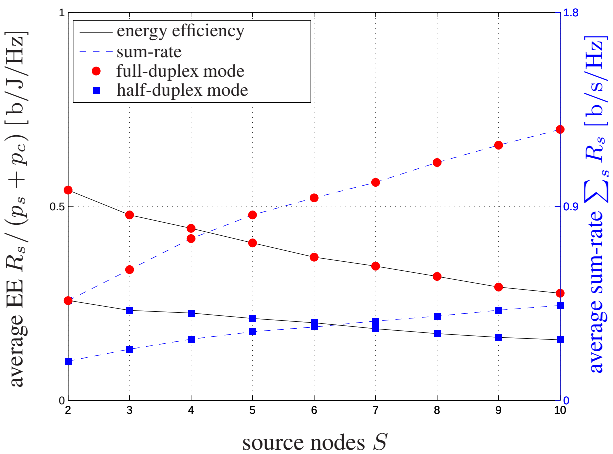


Fig. 4: Full-duplex performance vs. half-duplex performance as a function of S with $N = 3$ and $K = 2$.

Throughout the simulations, unless otherwise specified, we make use of the following system parameters: $\sigma_w^2 = 10\,\text{nW}$, $p_c = 100\,\text{mW}$ [33], and $\overline{p}_m = 1\,\text{W}$ for all $m \in \mathcal{M}$. For simplicity, we set $\underline{R}_s = 0.1$ b/s/Hz for all source nodes $s \in \mathcal{S}$. The distances of relays and source nodes from the base stations are assumed to be uniformly distributed between 10 and 100 m, whereas the path gains Γ_{ij} between transmitter i and receiver j is modeled as

$$\Gamma_{ij} = G_{\text{tx}} \cdot G_{\text{rx}} \cdot \left(\frac{\lambda_0}{4\pi}\right)^2 \cdot d_{ij}^{-\varsigma} \cdot \left(\frac{d_R}{10}\right)^{\varsigma-2} \tag{29}$$

wherein $G_{\text{tx}} = G_{\text{rx}} = 1$ are the gains of the transmit receive antennas, respectively, assumed to be omnidirectional; the parameter $\lambda_0 \approxeq 0.12\,\text{m}$ represents the carrier wavelength; d_{ij} denotes the distance between any two nodes i and j; $d_R = 100\,\text{m}$ denotes the radius of the cell; and the path loss exponent is set to $\varsigma = 4$. To reproduce the effects of shadowing and scattering, we use a 24-tap channel model [54].

As mentioned in Sect. III, one of the most important parameters of the Q-learning algorithm is the temperature function, which significantly impacts on the system performance. Unfortunately, it is not possible to mathematically derive the optimal temperature function for the problem at hand. To this aim, we use a heuristic approach to find the best exploration policy, by applying the QNash algorithm of Table I with a number of tentative temperature functions, to a single-cell network, consisting of $N = 2$ relay nodes and a variable number of source nodes S, all using $K = 2$ power steps. We also set the learning rate $f_m^t = t^{-0.8}$ and the discount factor $\delta_m = 0.85$ in Q-value updating steps (25) and (27). Fig. 3 reports the performance of the QNash as a function of S: solid black lines, corresponding to the left axis, represent the average EE of the source nodes, whereas dashed blue lines, corresponding to the right axis, represent the average sum-rate achieved by the source nodes. Three different temperature functions (all decreasing with t) are used: $T_m^t = 80^{-t}$ (squares), $T_m^t = 10^9 \cdot \exp\{-10 \cdot t\}$ (diamonds), and $T_m^t = 10^{-2} \cdot u_m(\tilde{\mathbf{p}}_m) \cdot \exp\left\{-10^{-2} \cdot u_m(\tilde{\mathbf{p}}_m) \cdot t\right\}$ (circles), where $u_m(\tilde{\mathbf{p}}_m) = \max_{\mathbf{p} \in \boldsymbol{\mathcal{P}}} u_m(\mathbf{p})$, as follows from (26). As can be seen, a good tradeoff between achieving a good EE while ensuring high data rates is provided by $T_m^t = 10^{-2} \cdot u_m(\tilde{\mathbf{p}}_m) \cdot \exp\left\{-10^{-2} \cdot u_m(\tilde{\mathbf{p}}_m) \cdot t\right\}$, which is then used in all subsequent simulations for the application of the proposed algorithm, together with $f_m^t = t^{-0.8}$ and $\delta_m = 0.85$.

Fig. 4 exhibits a comparison between the performance of full-duplex mode and half-duplex mode as a function of the number of source nodes in a network using $N = 3$ relays, applying QNash algorithm with $K = 2$. In this simulation, we assume $\underline{R}_s = 0.04\,\text{b/s/Hz}$, to allow feasible scenarios for the half-duplex mode. As can be seen, full-duplex mode shows much better performance in both terms of EE and overall rate. In terms of sum-rate, the increasing rate of full-duplex mode is higher than that of half-duplex. In terms of EE, the decreasing rate of the half-duplex mode is lower than that of the full-duplex mode. This is because, accomplishing the data transfer in two consecutive phases in half-duplex mode reduces the MAI more efficiently than in the case of the full-duplex mode.

To reduce the computational burden of the proposed algorithm, which is exponentially increasing with the number of power steps $K + 1$, we aim at identifying a suitable tradeoff between the system performance and a low K. Fig. 5 reports the performance in terms of EE (black line, left axis) and sum-

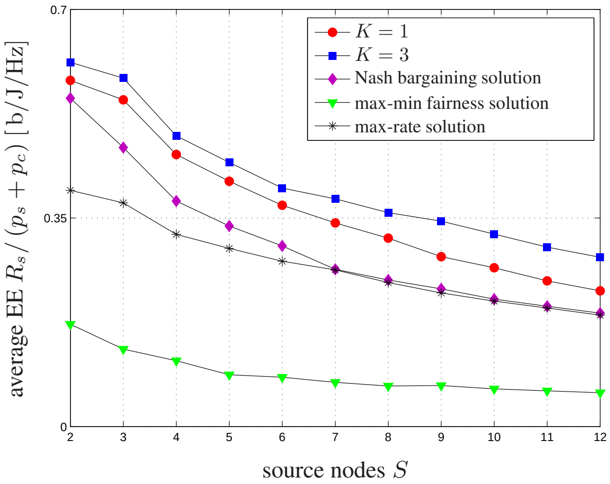


Fig. 6: Average source node's EE as a function of S ($N = 4$).

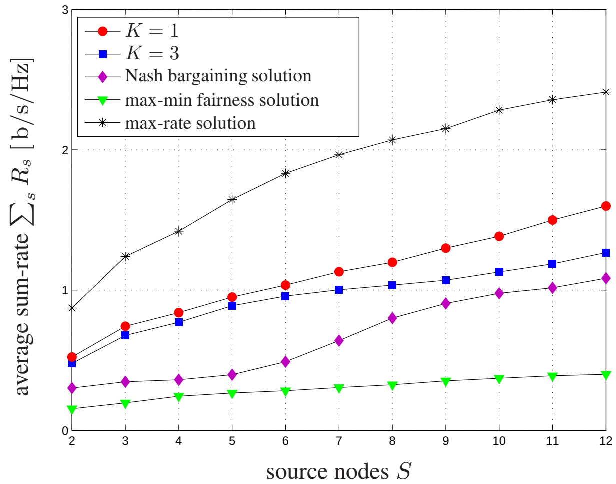


Fig. 8: Average sum-rate as a function of S ($N = 4$).

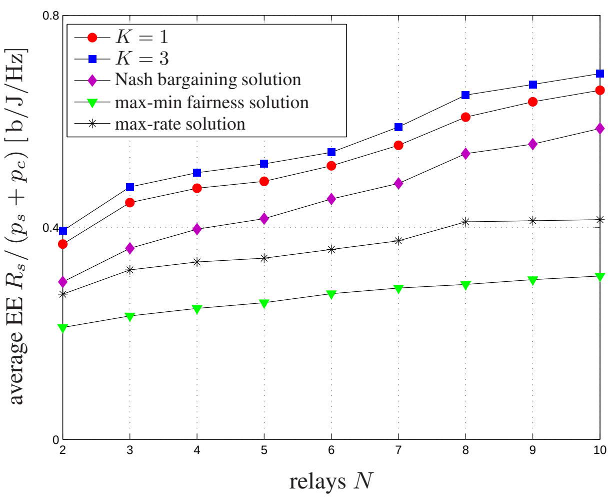


Fig. 7: Average source node's EE as a function of N ($S = 4$).

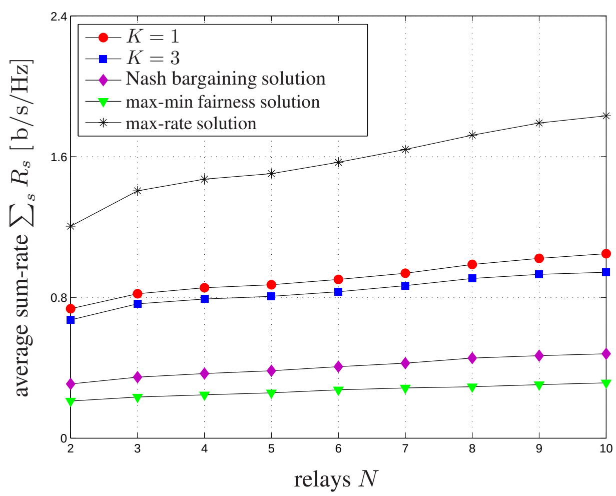


Fig. 9: Average sum-rate as a function of N ($S = 4$).

rate (blue line, right axis) as a function of K, by averaging over $2,000$ random realizations of a network with $S = N = 3$ nodes. As can be seen, increasing the number of power steps $K+1$ increases the EE while decreasing the spectral efficiency (in terms of sum-rate). In the following simulations, we will thus select the cases $K = 1$, corresponding to the situation in which the sum-rate is maximized and the computational load is the minimum one, at the cost of a reduced EE, and $K = 3$, which provides an interesting tradeoff between spectral efficiency, EE, and computational complexity of the algorithm.

In the following set of simulations, we will compare the performance of our proposed algorithm, using $K = 1$ (circles) and $K = 3$ (squares), with the following optimization techniques, formalized below, and depicted by diamonds, lower triangles, and asterisks, respectively:[4]

$$\text{NBS fairness:} \quad \max_{p_m \in [0, \overline{p}_m]} \prod_{s \in \mathcal{S}} \frac{R_s - \underline{R}_s}{p_s + p_c} \tag{30a}$$

$$\text{max-min fairness:} \quad \max_{p_m \in [0, \overline{p}_m]} \min_{s \in \mathcal{S}} \frac{R_s}{p_s + p_c} \tag{30b}$$

$$\text{max rate:} \quad \max_{p_m \in [0, \overline{p}_m]} \sum_{s \in \mathcal{S}} R_s \tag{30c}$$

Figs. 6 and 7 report the average EE as functions of the number of source nodes S, and the number of relays N, respectively. As expected, the EE is a decreasing function of S and an increasing function of N. When N is fixed, increasing S increases the MAI, thus reducing the EE. On the contrary, increasing N while S is constant increases the overall data rate in the long run, without additional power expenditure required at the source side. Furthermore, in accordance with Fig. 5, the case $K = 3$ outperforms the case $K = 1$. However, note that, even in the extreme case $K = 1$ (i.e., each node selects either zero power or its maximum one), the QNash algorithm outperforms the well-known solutions (30).

Similar conclusions can be drawn for the average spectral efficiencies, reported in Figs. 8 and 9, where the average sum-rate is reported as a function of S and N, respectively. Obviously, by definition the maximum sum-rate is given (30c), and confirmed by our numerical results. However, the performance gap paid

[4]Throughout the simulations, and unlike our proposed algorithm, the transmit power set is assumed to be continuous in $[0, \overline{p}_m]$ for all three methods.

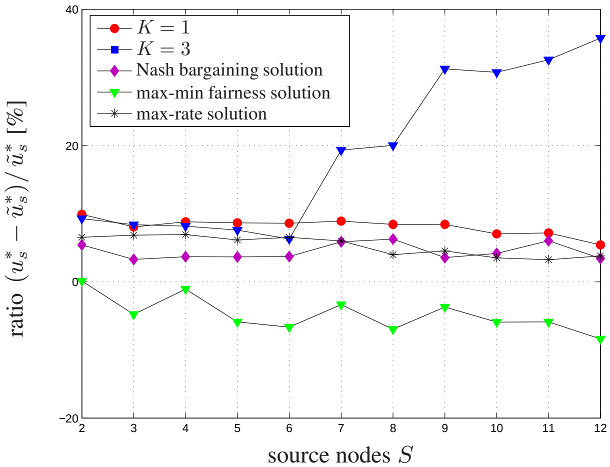


Fig. 10: Percentage of EE improvement as a function of S compared to the case $p_n = 1$ W for all $n \in \mathcal{N}$ ($N = 4$).

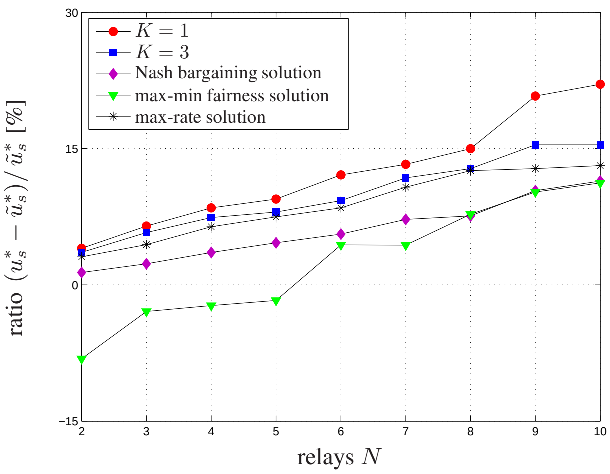


Fig. 11: Percentage of EE improvement as a function of N compared to the case $p_n = 1$ W for all $n \in \mathcal{N}$ ($S = 4$).

by the proposed algorithm is still acceptable, especially when compared with NBS and max-min approaches.

To measure the improvement achieved by including the relay powers into the resource allocation problem, which, to the best of our knowledge, is not present in the available literature, we now compare the performance of the proposed algorithm, where $\mathcal{M} = \mathcal{S} \cup \mathcal{N}$, with that achieved by the same algorithm, but with $\mathcal{M} = \mathcal{S}$, while each relay nodes $n \in \mathcal{N}$ adopts the same power level $p_n = \overline{p}$ at all steps of the algorithm. For the sake of completeness, we also report the comparison between the source-plus-relay versus source-only power control schemes for the NBS, max-min fairness, and max-rate solutions. Figs. 10 and 11 report the percentage of improvement in terms of average EE at the mixed-strategy Nash equilibrium of the game, expressed as the ratio of the average utility u_s^* achieved by the source nodes using the proposed algorithm, to the average utility $\tilde{u}_s^*$, achieved by regulating the power for source nodes only, using $\overline{p} = 1$ W for the relays, as a function of S, using $N = 4$, and of N, using $S = 4$, respectively. Figs. 12 and 13 report analogous results when comparing the

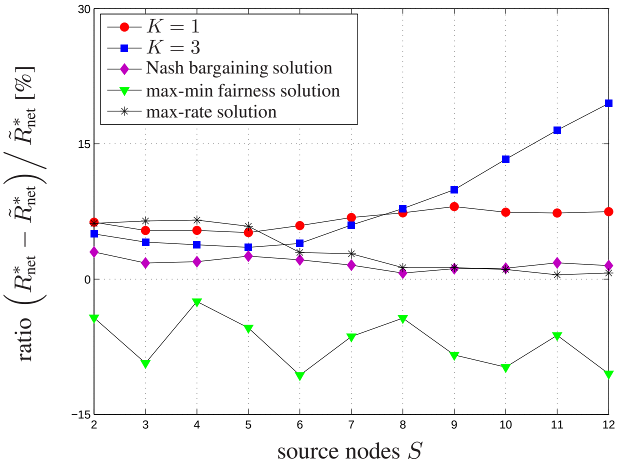


Fig. 12: Percentage of sum-rate improvement as a function of S compared to the case $p_n = 1$ W for all $n \in \mathcal{N}$ ($N = 4$).

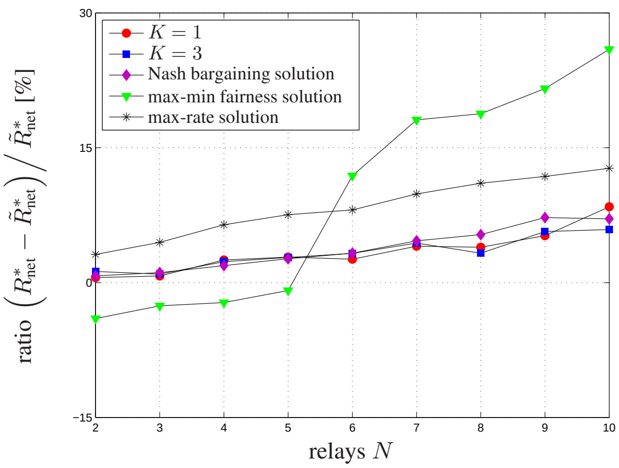


Fig. 13: Percentage of sum-rate improvement as a function of N compared to the case $p_n = 1$ W for all $n \in \mathcal{N}$ ($S = 4$).

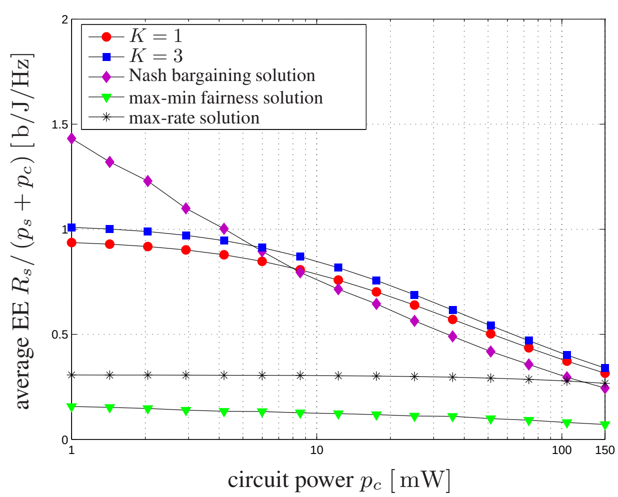


Fig. 14: Average source node's EE as a function of p_c ($S = 6$, $N = 4$).

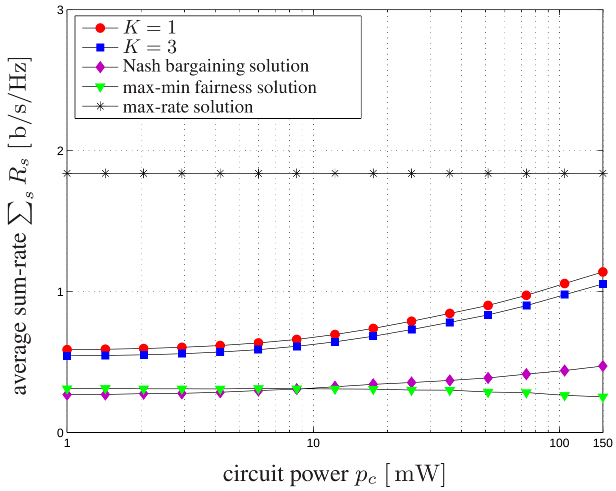


Fig. 15: Average sum-rate as a function of p_c ($S = 6$, $N = 4$).

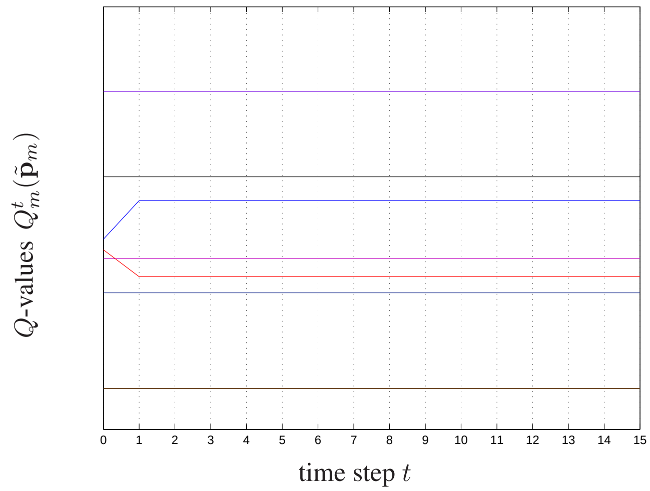


Fig. 16: Q-values as functions of the time step t ($K = 1$, $S = 6$, $N = 4$).

improvement in terms of spectral efficiency, measured as the ratio of achieved sum-rates at the equilibrium, where R^*_{net} and $\tilde{R}^*_{\text{net}}$ are the average sum-rates at the equilibrium in the above-mentioned cases. Similar results, not reported for brevity, can be obtained with different relay powers. As can be seen, apart from the max-min fairness case,[5] allowing the relays to regulate their transmit powers jointly with the source nodes is always beneficial in terms of both average EE and average sum-rate. This is particularly apparent for the case $K = 3$ when S increases, but the same behavior can also be observed for all other curves. More importantly, the proposed source-plus-relay approach outperforms not only the source-only case, as the improvement is always positive, but also both versions (i.e., source-plus-relay and source-only) at all other schemes in terms of energy efficiencies, and NBS and max-min fairness in terms of sum-rates.

Let us now investigate the impact of the circuit power p_c on the performance of the resource allocation techniques introduced above. As mentioned in Sect. II-B, p_c depends upon characteristics of electrical architecture of the nodes. Using some results available in the literature (e.g., [61]–[63]), it is known that the EE increases as p_c decreases. To verify this behavior in our simulation results, we measure the performance as a function of circuit power for a network consisting of $S = 6$ source nodes and $N = 4$ relay nodes. Figs. 14 and 15 depict the EE and the overall data rate of the network, respectively. As can be seen, p_c has a weak impact on max-min fairness and max-rate solutions, whose performance in terms of EE is always lower than that achieved by the technique illustrated in Sect. III. On the contrary, the NBS outperforms the proposed algorithm for low values of p_c (in this particular scenario, when $p_c < 7\,\text{mW}$). However, note that such small values are not yet available in practical systems (e.g., see [64]), and typical values of p_c are in the order of $100\,\text{mW}$ [33], [64], wherein the algorithm derived in this paper shows significantly higher performance.

Finally, Fig. 16 shows an example of the convergence of the Q-values for the network investigated above ($S = 6$ and $N = 4$),[6] using $K = 1$. As can be seen, in this scenario with $M = S + N = 10$ active players and $(K+1)^M = 2^{10}$ entries in the payoff matrix of the game, Q-values converge in the first time step. Our experiments with different network scenarios show that the convergence speed always happens before the third time step, thus confirming the hypothesis of quasi-static channels. The major reason for a cheap time complexity of the algorithm is an appropriate choice of the learning rate ($f_m^t = t^{-0.8}$) to update the Q-values [47].

V. Conclusion

In this work, we considered a reliable wireless communication network consisting of multiple source nodes, multiple parallel amplify-and-forward relays, and one destination. First, we modeled the interactions between active nodes, source nodes, and intermediate relay nodes, using an information-theoretic viewpoint to derive the relevant system parameters, namely, signal-to-interference-plus-noise ratios and maximum achievable rates, in both the full- and the half-duplex modes. Then, we modeled each active wireless node as a player in a noncooperative finite game, in which the strategy of each player is to adjust its transmit power using a discrete set of radiative powers. Under the assumption that each node has its own power constraint, we formulated each node's interest as follows: each source node is a self-interested player which aims at maximizing its own energy efficiency in terms of the ratio between its achievable rate and its total power (including the nonradiative one); and each relay is an altruistic player which aims at maximizing the network sum-rate with constraints on the minimum rates of all sources. The mixed-strategy Nash equilibrium points of the game are then computed using a Q-

[5]Here, we fall into a Braess-type paradox [55], i.e., a situation in which adding extra strategies to a network when the entities operate selfishly can in some cases reduce the overall performance, already observed in other different contexts (such as routing in [56], [57], and wireless communications in [58]–[60]).

[6]Since all $N = 4$ relays earn the same payoffs, Fig. 16 reports $S + 1 = 7$ different curves only.

learning-based algorithm, which leads the players to compute a probability distribution among all available pure strategies.

Numerical results show that our proposed algorithm in the full-duplex mode outperforms the half-duplex counterpart, the Nash bargaining solution and the max-min fairness approach in terms of both energy efficiency and network sum-rate, and significantly outperforms the max-rate solution in terms of energy efficiency, which is our major performance metric, while paying a tolerable performance gap in terms of network sum-rate. This performance improvement is obtained with a low-complexity algorithm, that uses a low number of power levels and converges after a few iteration steps, thus providing a practical, scalable, and adaptive solution. Simulations also show that our method increases the energy efficiency of the system when either the number of relays increases, or the number of sources decreases, or the nonradiative powers of the active nodes decreases, in accordance with recent studies in this field.

Further work is needed: *i*) to assess the feasibility of the problem given a particular network realization; *ii*) to assess the complexity of the proposed algorithm as a function of the system parameters; and *iii*) to extend the formulation of the problem to a multicarrier system, so that the additional degrees of freedom may improve the both the energy and the spectral efficiency of the network.

REFERENCES

[1] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity. Part I. System description," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.

[2] ——, "User cooperation diversity. Part II. Implementation aspects and performance analysis," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1939–1948, Nov. 2003.

[3] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. Information Theory*, vol. 51, no. 9, pp. 3037–3063, Sept. 2005.

[4] I. Hammerström, M. Kuhn, and A. Wittneben, "Impact of relay gain allocation on the performance of cooperative diversity networks," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, vol. 3, Los Angeles, CA, Sept. 2004, pp. 1815–1819.

[5] Z. Sahinoglu and P. Orlik, "Regenerator versus simple-relay with optimum transmit power control for error propagation," *IEEE Commun. Letters*, vol. 7, no. 9, pp. 416–418, Sept. 2003.

[6] S. Serbetli and A. Yener, "Optimal power allocation for relay assisted F/TDMA ad hoc networks," in *Proc. Intl. Conf. Wireless Networks, Commun. Mobile Comp.*, vol. 2, Maui, HI, June 2005, pp. 1319–1324.

[7] D. Wang, X. Wang, and X. Cai, "Optimal power control for multi-user relay networks over fading channels," *IEEE Trans. Wireless Commun.*, vol. 10, no. 1, pp. 199–207, 2011.

[8] J. Paredes and A. Gershman, "Relay network beamforming and power control using maximization of mutual information," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4356–4365, Dec. 2011.

[9] Y. Jing and S. Shahbazpanahi, "Max-min optimal joint power control and distributed beamforming for two-way relay networks under per-node power constraints," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6576–6589, 2012.

[10] H. Shan, W. Zhuang, and H. T. Cheng, "Cross-layer protocol design for distributed wireless networks with novel relay selection," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Miami, FL, Dec. 2010, pp. 1–6.

[11] B. Zhao and M. Valenti, "Practical relay networks: A generalization of hybrid-ARQ," *IEEE J. Select. Areas Commun.*, vol. 23, no. 1, pp. 7–18, Jan. 2005.

[12] M. Zorzi and R. Rao, "Geographic random forwarding (GeRaF) for ad hoc and sensor networks: multihop performance," *IEEE Trans. Mobile Computing*, vol. 2, no. 4, pp. 337–348, Oct.-Dec. 2003.

[13] C. Lima and G. de Abreu, "Game-theoretical relay selection strategy for geographic routing in multi-hop WSNs," in *Proc. Workshop on Positioning, Navigation and Communication (WPNC)*, Hannover, Germany, March 2008, pp. 277–283.

[14] J. Huang, Z. Han, M. Chiang, and H. V. Poor, "Auction-based distributed resource allocation for cooperation transmission in wireless networks," in *Proc. IEEE Global Telecommunications Conf. (GLOBECOM)*, Washington DC, Nov. 2007, pp. 4807–4812.

[15] ——, "Auction-based resource allocation for multi-relay asynchronous cooperative networks," in *Proc. IEEE Intl. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Nev, Apr. 2008, pp. 5356–5359.

[16] B. Wang, Z. Han, and K. Liu, "Distributed relay selection and power control for multiuser cooperative communication networks using Stackelberg game," *IEEE Trans. Mobile Computing*, vol. 8, no. 7, pp. 975–990, July 2009.

[17] I. Stupia, L. Vandendorpe, L. Sanguinetti, and G. Bacci, "Distributed energy-efficient power optimization for relay-aided heterogeneous networks," in *Proc. Intl. Workshop Wireless Networks: Communication, Cooperation and Competition*, Hammamet, Tunisia, May 2014.

[18] Y. Zhang, Y. Xu, and Y. Cai, "Relay selection utilizing power control for decode-and-forward wireless relay networks," in *Proc. Intl. Conf. Signal Processing and Communication Systems (ICSPCS)*, Gold Coast, Australia, Dec. 2008, pp. 1–5.

[19] D. Wu, Y. Cai, L. Zhou, Z. Zheng, and B. Zheng, "Cooperative strategies for energy-aware ad hoc networks: A correlated-equilibrium game-theoretical approach," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 2303–2314, 2013.

[20] X. Liang, M. Chen, and V. Leung, "A game-theoretic approach for relay assignment over distributed wireless networks," in *Proc. Intl. Wireless Commun. Mobile Comp. Conf. (IWCMC)*, Istanbul, Turkey, July 2011.

[21] G. Zhang, H. Zhang, L. Zhao, W. Wang, and L. Cong, "Fair resource sharing for cooperative relay networks using Nash bargaining solutions," *IEEE Communications Letters*, vol. 13, no. 6, pp. 381–383, 2009.

[22] C. Y. Ng, T. M. Lok, and T. Wong, "Pricing games for distributed cooperative transmission," *IEEE Trans. Veh. Technol.*, vol. 59, no. 7, pp. 3393–3406, 2010.

[23] H. Khayatian, R. Saadat, and J. Abouei, "Coalition-based approaches for joint power control and relay selection in cooperative networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 835–842, 2013.

[24] S. Ren and M. van der Schaar, "Pricing and distributed power control in wireless relay networks," *IEEE Trans. Signal Processing*, vol. 59, no. 6, pp. 2913–2926, 2011.

[25] W. A. Prasetyo, H. Lu, and H. Nikookar, "Optimal relay selection and power allocation using game theory for cooperative wireless networks with interference," in *Proc. European Microwave Conf. (EuMC)*, Manchester, UK, Oct. 2011, pp. 37–40.

[26] Q. Cao, H. Zhao, and Y. Jing, "Power allocation and pricing in multiuser relay networks using Stackelberg and bargaining games," *IEEE Trans. Veh. Technol.*, vol. 61, no. 7, pp. 3177–3190, 2012.

[27] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. Cambridge, MA: MIT Press, 1994.

[28] J.-S. Pang and F. Facchinei, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, ser. Springer Series in Operations Research and Financial Engineering. New York: Springer, 2003.

[29] C. J. Watkins and P. Dayan, "Q-Learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, 1992.

[30] D. Bharadia, E. McMilin, and S. Katti, "Full duplex radios," in *Proc. ACM SIGCOMM*, Hong Kong, China, Aug. 2013.

[31] M. Uysal, *Cooperative Communications for Improved Wireless Network Transmission: Framework for Virtual Antenna Array Applications*. Hershey, PA: IGI Global Publishing, 2009.

[32] S. Cui, A. Goldsmith, and A. Bahai, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," *IEEE J. Select. Areas Commun.*, vol. 22, no. 6, pp. 1089–1098, 2004.

[33] G. Miao, N. Himayat, and G. Li, "Energy-efficient link adaptation in frequency-selective channels," *IEEE Trans. Commun.*, vol. 58, no. 2, pp. 545–554, 2010.

[34] J. F. Nash, "Equilibrium points in n-person games," *Proc. National Academy of Sciences of the United States of America*, vol. 36, no. 1, pp. 48–49, Jan. 1950.

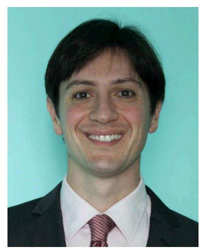
[35] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou, "The complexity of computing a Nash equilibrium," in *Proc. ACM Symp. Theory of Computing*, Seattle, WA, 2006, pp. 71–78.

[36] C. E. Lemke and J. T. Howson, "Equilibrium points of bimatrix games," *J. Society for Industrial and Applied Mathematics*, vol. 12, no. 2, pp. 413–423, 1964.

[37] R. Porter, E. Nudelman, and Y. Shoham, "Simple search methods for finding a Nash equilibrium," in *Proc. Nat. Conf. Artificial Intelligence (AAAI)*, San Jose, CA, 2004, pp. 664–669.

[38] T. Sandholm, A. Gilpin, and V. Conitzer, "Mixed-integer programming methods for finding Nash equilibria," in *Proc. Nat. Conf. Artificial Intelligence (AAAI)*, Pittsburgh, Pennsylvania, 2005, pp. 495–501.
[39] R. McKelvey, A. McLennan, and T. Turocy. (2010) Gambit: Software tools for game theory. [Online]. Available: http://www.gambit-project.org/
[40] H. Tsaknakis and P. G. Spirakis, "An optimization approach for approximate Nash equilibria," in *Proc. Intl. Conf. Internet and Network Economics (WINE)*, San Diego, CA, 2007, pp. 42–56.
[41] U. Boryczka and P. Juszczuk, "A new evolutionary approach for computing Nash equilibria in bimatrix games with known support," *Central European Journal of Computer Science*, vol. 2, no. 2, pp. 128–142, 2012.
[42] N. G. Pavlidis, K. E. Parsopoulos, and M. N. Vrahatis, "Computing Nash equilibria through computational intelligence methods," *J. Comput. Appl. Math.*, vol. 175, no. 1, pp. 113–136, Mar. 2005.
[43] R. Dilts and T. A. Epstein, *Dynamic Learning*. Capitola, CA: MeTa Publications, 1995.
[44] A. E. Roth and I. Erev, "Learning in extensive form games: experimental data and simple dynamic models in the intermediate term," *Games and Economic Behavior*, vol. 8, pp. 164–212, 1995.
[45] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.
[46] S. Singh and D. Bertsekas, "Reinforcement learning for dynamic channel allocation in cellular telephone systems," in *Proc. Conf. Advances in Neural Information Processing Systems*. MIT Press, 1997, pp. 974–980.
[47] E. Even-Dar and Y. Mansour, "Learning rates for Q-learning," *J. Machine Learning Research*, vol. 5, pp. 1–25, Dec. 2004.
[48] S. Lasaulce and H. Tembine, *Game Theory and Learning for Wireless Networks: Fundamentals and Applications*. Academic Press, 2011.
[49] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artificial Intelligence Research*, vol. 4, pp. 237–285, May 1996.
[50] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *J. Machine Learning Research*, vol. 4, pp. 1039–1069, 2003.
[51] T. Sanguanpuak and R. M. A. P. Rajatheva, "Power bargaining for amplify and forward relay channel," in *International Conference on Communications and Networking in China (ChinaCOM)*, Aug. 2009, pp. 1–5.
[52] K. Phan, T. Le-Ngoc, S. Vorobyov, and C. Tellambura, "Power allocation in wireless multi-user relay networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 5, pp. 2535–2545, May 2009.
[53] J. Zhang, F. Roemer, M. Haardt, A. Khabbazibasmenj, and S. Vorobyov, "Sum rate maximization for multi-pair two-way relaying with single-antenna amplify and forward relays," in *Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 2477–2480.
[54] ITU Radiocommunication Sector (ITU-R), "Guidelines for evaluation of radio transmission technology for IMT-2000," *Recommendation ITU-R M.1225*, 1997.
[55] S. M. Perlaza, "Game theoretic approaches to spectrum sharing in decentralized self-conguring networks," Ph.D. dissertation, École Nationale Supérieure des Télécommunications, Paris, France, 2011.
[56] T. Roughgarden, *Selfish Routing and the Price of Anarchy*. Cambridge, MA: MIT Press, 2005.
[57] E. Altman, T. Boulogne, R. El Azouzi, T. Jimenez, and L. Wynter, "A survey on networking games in telecommunications," *Computers and Operations Research*, vol. 33, no. 2, pp. 286–311, Feb. 2006.
[58] E. Altman, V. Kamble, and H. Kameda, "A Braess type paradox in power control over interference channels," in *Proc. Int. Symp. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, Berlin, Germany, Apr. 2008.
[59] L. Rose, S. M. Perlaza, and M. Debbah, "On the Nash equilibria in decentralized parallel interference channels," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kyoto, Japan, Jun. 2011.
[60] G. Bacci, L. Sanguinetti, M. Luise, and H. V. Poor, "Energy-efficient power control for contention-based synchronization in OFDMA systems with discrete powers and limited feedback," *EURASIP J. Wireless Commun. Networking (JWCN)*, vol. 2013, no. 1, p. 2013:192, Jul. 2013.
[61] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Information Theory*, vol. 36, no. 5, pp. 1019–1030, Sep. 1990.
[62] C. Isheden, Z. Chong, E. Jorswieck, and G. Fettweis, "Framework for link-level energy efficiency optimization with informed transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2946–2957, Aug. 2012.
[63] G. Bacci, E. V. Belmega, P. Mertikopoulos, and L. Sanguinetti, "Energy-aware competitive link adaptation in small-cell networks," in *Proc. Intl. Workshop Resource Allocation in Wireless Networks*, Hammamet, Tunisia, May 2014.
[64] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design for energy-efficient multi-user MIMO systems: Is massive MIMO the answer?" *IEEE Trans. Wireless Commun.*, 2015, to appear.

Farshad Shams received the M.E. degree in computer engineering from the University of Ferrara, Ferrara, Italy, in 2006, and the Ph.D. degree in computer science from the IMT Institute for Advanced Studies, Lucca, Italy, in 2012. Since 2007, he has been with IMT Institute for Advanced Studies, where is currently a post-doctoral research fellow. His research interests span several subjects in the broad area of computer science and information engineering, including: resource allocation for wireless communications, cooperative game theory and dynamic learning for wireless communications, information theory for relay-assisted systems, Java technologies for wireless devices, and mini-hardware programming. Dr. Shams has over eight years of experience as a network programmer, database analyzer and programmer, mini-hardware designer and programmer, and network engineer in different R&D centers in Italy and Iran.



Giacomo Bacci (S'07–M'09) received the B.E. and the M.E. degrees in telecommunications engineering and the Ph.D. degree in information engineering from the University of Pisa, Pisa, Italy, in 2002, 2004, and 2008, respectively. In 2006-07, he was a visiting student research collaborator at Princeton University, Princeton, NJ, USA. In 2008-14, he was a post-doctoral research fellow at the University of Pisa. In 2008-2012, he also joined Wiser srl, Livorno, Italy, as a software engineer, and in 2012-14 he was also enrolled as a visiting postdoctoral research associate at Princeton University. Since 2015, he has joined MBI srl, Pisa, Italy, as a project engineer.

Dr. Bacci is the recipient of the FP7 Marie Curie International Outgoing Fellowships for career development (IOF) 2011 GRAND-CRU, and received the Best Paper Award from the *IEEE Wireless Commun. and Networking Conference (WCNC) 2013*, the Best Student Paper Award from the *Intl. Waveform Diversity and Design Conf. (WDD) 2007*, and the Best Session Paper at the *ESA Workshop on EGNOS Performance and Applications 2005*. He is also the recipient of the 2014 URSI Young Scientist Award, and he has been name an Exemplary Reviewer 2012 for *IEEE Wireless Communications Letters*. His research interests are in the areas of digital communications, signal processing, and estimation theory. His current research topics focus on resource allocation for 5G wireless technologies.

Marco Luise (M'85–SM'94–F'12) is a Full Professor of Telecommunications at the University of Pisa, Italy. After receiving his M.E. and Ph.D. degrees in electronic engineering from the University of Pisa, he was a Research Fellow of the European Space Agency (ESA) at ESTEC Noordwijk, the Netherlands, a Researcher of the Italian National Research Council (CNR), at the Centro Studio Metodi Dispositivi Radiotrasmissioni (CSMDR), Pisa, and an Associate Professor at the Department of Information Engineering of the University of Pisa.

Prof. Luise served as an Editor of the *IEEE Transactions on Communications* and of the *European Transactions on Telecommunications*. He was the founder and first (co)-Editor-in-Chief of the *International Journal of Navigation and Observation*, and is now an Associate Editor of the *Journal of Communications and Networks*. Recently, Prof. Luise was the general chairman of EUSIPCO 2006, the general co-chair of European Wireless 2010, and the general co-chair of the IEEE Intl. Conf. Acoustics, Speech, and Signal Process. (ICASSP 2014). His main research interests lie in the broad area of communication theory, with particular emphasis on wireless communications, mobile and satellite communications, positioning systems, and software-defined radios.