Nomikos, Nikolaos; Charalambous, Themistoklis; Trakadas, Panagiotis; Wichman, Risto

Bandit-Based Learning-Aided Full-Duplex/Half-Duplex Mode Selection in 6G Cooperative Relay Networks

Published in:
IEEE Open Journal of the Communications Society

DOI:
10.1109/OJCOMS.2024.3370476

Published: 01/01/2024

Document Version
Publisher's PDF, also known as Version of record

Received 29 January 2024; accepted 23 February 2024. Date of publication 26 February 2024; date of current version 8 March 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.3370476

Bandit-Based Learning-Aided Full-Duplex/Half-Duplex Mode Selection in 6G Cooperative Relay Networks

NIKOLAOS NOMIKOS[1] (Senior Member, IEEE), THEMISTOKLIS CHARALAMBOUS[2,3,4] (Senior Member, IEEE), PANAGIOTIS TRAKADAS[1], AND RISTO WICHMAN[5] (Senior Member, IEEE)

[1]Department of Ports Management and Shipping, National and Kapodistrian University of Athens, 34400 Euboea, Greece

[2]Department of Electrical and Computer Engineering, School of Engineering, University of Cyprus, 1678 Nicosia, Cyprus

[3]Department of Electrical Engineering and Automation, School of Electrical Engineering, Aalto University, 02150 Espoo, Finland

[4]FinEst Centre for Smart Cities, 12616 Tallinn, Estonia
[5]Department of Signal Processing and Acoustics, School of Electrical Engineering, 02150 Espoo, Finland

Corresponding author: N. NOMIKOS (e-mail: nomikosn@pms.uoa.gr)

This work was supported in part by the HORSE Project, funded by the Smart Networks and Services Joint Undertaking (SNS JU) through the European Union's Horizon Europe Research and Innovation Programme under Grant 101096342 (www.horse-6g.eu), and in part by the Project MINERVA, funded by the European Research Council (ERC) through the European Union's Horizon 2022 Research and Innovation Programme under Grant 101044629.

ABSTRACT The high level of autonomy and intelligence that is envisioned in sixth generation (6G) networks necessitates the development of learning-aided solutions, especially in cases in which conventional Channel State Information (CSI)-based network processes introduce high signaling overheads. Moreover, in wireless topologies characterized by fast varying channels, timely and accurate CSI acquisition might not be possible and the transmitters (CSIT) only have statistical CSI available. This work focuses on the appropriate selection of relaying mode in a cooperative network, comprising a single information source, one buffer-aided (BA) relay with full-duplex (FD) capabilities, and a single destination. Here, prior to each transmission, the relay should select to operate either in FD mode with power control, or, resort to half-duplex (HD) relaying when excessive self-interference (SI) arises. Targeting the selection of the best relaying mode, we propose an FD/HD mode selection mechanism, namely multi-armed bandit-aided mode selection (MABAMS), relying on reinforcement learning and the processing of acknowledgements/negative-acknowledgements (ACK/NACK) packets for acquiring useful information on channel statistics. As a result, MABAMS does not require continuous CSI acquisition and exchange and nullifies the negative effect of outdated CSI. The proposed algorithm's average throughput performance is evaluated, highlighting a performance-complexity trade-off over alternative solutions, based on pilot-based channel estimation that result in spectral and energy costs while obtaining instantaneous CSI.

INDEX TERMS 6G, full-duplex, buffer-aided relays, multi-armed bandits (MAB), relay mode selection, reinforcement learning.

I. INTRODUCTION

TO SATISFY the targets of sixth generation (6G) networks, the adoption of novel paradigms for network coordination are required, relying on machine learning (ML) to achieve fully autonomous network operation. Moreover, towards improving wireless connectivity cooperative relays have been proposed as an efficient solution for increasing wireless transmissions' quality. By relying on different duplex modes, full-duplex (FD) relays are able to simultaneously receive/transmit while half-duplex (HD) relays avoid self-interference (SI), inherent to FD operation, at the cost of inefficient radio-resource utilisation. As FD relays transmit

and receive on the same temporal and spectral resources, resource-efficient network operation is enabled. Nonetheless, excessive SI might arise from the transmitting antenna to the receiving antenna of the relay that should be treated with appropriate interference mitigation measures [1].

Considering the massive amount of nodes that is expected in future 6G deployments, it is necessary to integrate ML-aided solutions that can efficiently acquire the optimal network operation mode, according to the desired performance target and with low-complexity and processing requirements. Towards this end, we adopt reinforcement learning (RL) and more specifically, multi-armed bandits (MAB), a subcategory of reward-based learning algorithms. In the past, the MAB framework has been employed in the context of wireless communications; see, for example, [2], [3] in which it was shown that MAB-based algorithms can support various 5G/6G use cases with reduced network coordination complexity. Still, to the best of our knowledge, this is the first work to use MAB for the problem of FD/HD mode selection in cooperative relay networks.

A. BACKGROUND

In wireless communications systems, conventional channel estimation relying on pilot signals allows receivers to estimate the wireless channel state at the cost of radio-resource and energy consumption due to overheads. Recently, Xu et al. [4] studied massive multiple-input multiple-output (MIMO) FD networks with simultaneous wireless information and power transfer (SWIPT), using the energy signals to harvest energy and conduct channel estimation. Simulation results highlight spectral efficiency gains by the proposed protocol on SE over conventional massive MIMO SWIPT protocol. However, the authors do not provide a solution to reduce the amount of channel state information (CSI) overheads. In addition, in a number of cases, wireless networks might be characterized by non-stationary channels, introducing challenges for the channel estimation process. In the study by Shi et al. [5] pilots and interpolation schemes were used to timely acquire CSI while Abdul Careem and Dutta [6] mitigated the effect of channel impairments by modifying the modulation vectors. In settings with non-stationary channels, both works employ ML for extracting channel statistics. Both schemes provided improved performance over pilot-based methods but the amount of coordination and computation overheads reduction has not been quantified. Furthermore, in industrial environments, Lu et al. [7] have calculated the non-stationary parameters of Rician channels, using iterative sub-component discrimination and the Gaussian mixture model, resulting in near-optimal channel estimation. Then, for vehicular topologies, Pan et al. [8] present data pilot-aided (DPA) deep learning channel estimation, leveraging as pilots, the de-mapped data symbols. Moreover, to extract time-frequency correlation, they combine DPA with a long short-term memory network and a multi-layer perceptron network. Results for fast time-varying channel, large packets and high modulation order reveal performance gains over conventional DPA solutions. Although several works have provided important contributions to improving the performance of channel estimation methods, online policies to facilitate the network to operate with increased autonomy, interacting and adapting to the wireless environment is missing.

In the context of 6G networks, comprising an increased number of users and machines, a high level of signaling and feedback messages is required to ensure efficient operation [9]. Still, such overheads threaten the network's performance, especially when centralized network coordination is adopted and resource- and energy constrained devices must participate in data routing processes [10], [11]. In addition, recently, the integration of ML in wireless communications has shown attractive performance while offering low-complexity coordination mechanisms (see, for example, [12], [13] and references therein). By carefully investigating the different ML families, significant radio-resource allocation autonomy can be provided by RL-based approaches [14]. More specifically, under by integrating RL in network operation procedures, intelligent agents, in the sense of network nodes, exploit feedback from previously taken actions, and adapt their behaviour, considering the state of the wireless environment, their radio-resources, energy availability and desired Quality-of-Service (QoS) of the application, among others. A popular reward-based category of learning algorithms relies on the MAB framework [15], [16]. In MAB, a player (user) picks an action from a given set of actions, targeting the maximization of her cumulative expected reward. Since MAB enables the system to learn unknown environments during network deployment, it can significantly assist distributed allocation of radio-resources, such as spectrum, time and power [17].

Achieving resource efficiency is an enabler for 6G networks, and in this context, the development of novel FD techniques is vital for maximizing the spectral efficiency of wireless communications. Power control represents a popular solution to reduce the level of SI in the network, facilitating the receiving antenna of the FD node to correctly decode the signal. Riihonen et al. [18] presented opportunistic FD/HD relay mode switching by exploiting instantaneous CSI availability at the relay to perform transmit power adaptation. This process provided instantaneous and average spectral efficiency maximization in the uplink and downlink. In topologies with multi-antenna FD relays, Suraweera et al. [19] focused on performance improvement through power allocation and transmit antenna selection, considering different cases of CSI availability. Their solutions managed to surpass the fixed transmit power's zero diversity effect by adopting a simple power allocation mechanism. Next, Tran et al. [20] provided an optimal power allocation algorithm to increase the diversity order of FD amplify-and-forward (AF) relay network. More specifically, the closed-form expression of the pairwise-error probability was derived and optimal power allocation was

conducted, based on the bisection method, under statistical source-relay ($\{S \rightarrow R\}$) CSI at the relay, instantaneous relay-destination ($\{R \rightarrow D\}$) CSI at the destination. Recently, ML-aided approaches for FD communications have been introduced, aiming to combat the impact of SI, mainly using ML data-driven algorithms digital SI cancellation to reduce the complexity of traditional methods in terms of CSI overheads [21]. Still, most of the existing ML-aided solutions rely on offline-trained ML algorithms for SI estimation over static SI channels. In practice, user mobility and/or environmental changes can affect the SI characteristics over time, and the ML algorithms should be retrained to adapt to time-varying SI channels.

Another viable option for SI mitigation is to allow for hybrid FD/HD operation, activating most appropriate duplexing method in each transmission period. When buffers are not available at the relays, hybrid FD/HD switching was proposed in [18], offering improved performance over standalone FD or HD relaying. For buffer-aided (BA) increased flexibility in network scheduling can be provided, resulting in improved performance; see, e.g., [22] and references therein. For HD cooperative networks, the max-link policy was given in [23], activating a BA relay for reception or transmission, providing a diversity order equal to times the number of relays, when large buffers are available. In *single-relay* topologies, Zlatanov et al. [24] showed that the throughput of BA relaying can improve the throughput of FD networks. The study in [25] studied a single-relay topology where the source is not saturated while statistical CSIT was obtained. Still, that scheme only relied on FD relaying to achieve end-to-end capacity maximization. In multi-relay networks with buffers, successive opportunistic relaying (SOR) is possible to recover radio-resource losses, related to HD relay operation. In [26], Nomikos et al. assume a saturated source, transmitting with fixed-rate, multiple FD relays with buffering capabilities, and availability of instantaneous CSI at the reception while the transmitter only have statistical CSIT knowledge. In this setting, a hybrid FD/SOR/HD policy was developed to maximize the throughput per energy unit of the transmission. Another study by Della Penda et al. presented joint relay mode selection and power allocation, assuming Rician fading channels [27]. The proposed algorithm, activated a set of wireless links for power consumption minimizing, and provided success probability guarantees. Unfortunately, at the moment, there exists a gap in the literature, as ML-aided approaches for FD/HD relay mode selection are missing.

In dense mobile network deployments relying on small cells, e.g., in Industry 4.0 settings, it is expected that mobile users and machines will coexist and strain wireless resources. In such complex wireless settings, a large amount of overheads for signaling and feedback messages is introduced to ensure robust network operation. However, these overheads may severely affect network performance and as a remedy, ML integration in wireless networks has introduced promising gains towards low-complexity network operation (see, for example, [12], [28]). Furthermore, reward-based learning, as it is the case of MAB, has already been adopted in various 5G/6G cases; see, e.g., [3], [29], [30], where power control in a single FD relay network was studied, and modelled as a MAB game.

B. CONTRIBUTIONS

This work presents a reinforcement-based online policy to choose the relay's operation mode and in the case of FD transmission/reception, to appropriately set the relay transmit power. Thus, FD/HD relay modes and power control are integrated in a MAB framework where during each time frame, the destination sends ACKs/NACKs that are exploited by the relay, resulting in autonomous network operation. Moreover, this work extends the study in [31] by providing further details on the operation of MABAMS and considers the practical cases of outdated CSI and non-stationary wireless channels, also presenting relevant performance evaluation results. Our contributions are the following.

1) Contrary to [29], [30], here, we aim to tackle a more complex and practical problem, by equipping the relay with a buffer to overcome cases when FD operation is infeasible due to excessive SI, thus efficiently switching to HD mode. Under this mode, the relay receives and stores packets in its buffer or extracts packets from its buffer and forwards them towards the destination.
2) We propose bandit-based mode selection (MABAMS) by the relay, using only local observations of source's signal, and ACKs/NACKs from the destination. Thus, feedback of CSI from the destination to the relay is avoided which overcomes the issue of outdated CSI and significantly reduces feedback overhead.
3) Different cases of channel stationarity are evaluated and MABAMS is compared against other CSI- and learning-based policies, in terms of average throughput, showing that our learning-based hybrid FD/HD provides significant performance gains.

The proposed online and bandit-based solution for FD/HD relay mode selection avoids pilot transmissions and processing and introduces the following gains: 1) The impact of outdated CSI is nullified, as network operation only relies on one-bit ACK/NACK feedback, enabling the relay to timely decide which duplexing mode should be adopted and which transmit power level must be used; and 2) coordination overheads are avoided, compared to CSI-based FD/HD mode selection, introducing energy and computations gains, being especially important in networks with resource-constrained devices, as it is the case in industrial Internet-of-Things settings. Overall, MABAMS learning-based approach supports the vision for fully autonomous network operation which is of utmost importance in the forthcoming 6G era.

TABLE 1. List of acronyms.

Acronym	Definition
6G	Sixth generation
ACK	Acknowledgment
AF	Amplify-and-forward
AWGN	Additive white Gaussian noise
BA	Buffer-aided
BSI	Buffer state information
CSIT	Channel state information at the transmitter
DF	Decode-and-forward
DPA	Data pilot-aided
HD	Half-duplex
FD	Full-duplex
HARQ	Hybrid automatic repeat request
MABAMS	Multi-armed bandit mode selection
MIMO	Multiple-input multiple-output
ML	Machine learning
NACK	Negative acknowledgment
QoS	Quality-of-Service
RL	Reinforcement learning
SI	Self-interference
SINR	Signal-to-interference-and-noise ratio
SNR	Signal-to-noise ratio
SOR	Successive opportunistic relaying
SSS	Strict-sense stationary
UCB	Upper confidence bound
URLLC	Ultra-reliable low-latency communications

TABLE 2. Summary of notation.

Symbol	Description
L	Maximum no. of packets that can be stored
Q	No. of packets in the relay's buffer in the network
h_{ij}	Fading coefficient of link ij
g_{ij}	Channel gain of link ij
σ_{ij}^2	Channel gain variance of link ij
Γ_{ij}	Signal-to-noise ratio at receiver j of link ij
r_0	Information rate
σ_i^2	Thermal noise variance at receiver i
P_i	Transmit power at transmitter i
$\hat{h}_{ij}$	Estimated channel response of $\{i \to j\}$ link
ρ_i	Correlation coefficient between h_{ij} and $\hat{h}_{ij}$
T	time horizon
q_{th}	Success transmission probability threshold
$F_W(w)$	Cumulative distribution function (cdf) of random variable W
Π	Set of all feasible policies
$R^\pi(T)$	Regret of a policy $\pi \in \Pi$
$U_{\ell,t}$	Instantaneous utility by selecting link ℓ at round t
$U_{I_t^\pi,t}$	Instantaneous utility by choosing link I_t^π under policy π at round t
$\mathcal{P}_R$	The relay's discrete power level set
$\lvert\mathcal{P}_R\rvert$	No. of power levels at the relay
$\boldsymbol{\mu}$	Vector of mean rewards of various arms
$r_{j,t}$	Reward of arm j at round t
I_t	Selected arm at round t
$n_{j,t}$	No. of plays of arm j up to round t
$\mathbb{1}_A$	Indicator function of the event A
$\hat{q}_{j,t}$	Empirical average reward of arm j up to time-slot t
Υ_T	No. of breakpoints before time T

C. STRUCTURE

This paper is structured as follows. Section II provides details on the system model while Section III formulates the problem that we aim to tackle. Next, Section IV includes the MAB modeling of FD/HD mode selection. Bandit-based learning-aided mode selection is given in Section V, while Section VI presents the performance evaluation. Finally, Section VII provides the conclusions of our work and various interesting future directions.

II. SYSTEM MODEL

This section presents the system model that is adopted in our study. In Table 2, the notation used in this paper is given. A cooperative network with a single source, S, a single destination, D, and one FD decode-and-forward (DF) relay node R is assumed, as shown in Fig. 1. As a consequence of severe fading, end-to-end communication is only possible in two-hops and through the relay. To enable further scheduling flexibility, the relay has a buffer with size L (in packets). By Q ($Q \in \{0, 1, \ldots, L\}$), we denote the number of stored packets in the relay's buffer.

Network operation is divided into "frames" that are equal to the duration of one packet and at any arbitrary frame K, the channel coefficient h_{ij} of link $\{i \to j\}$, is modeled as an independent complex normal random variable, having zero mean, and variance σ_{ij}^2, i.e., $h_{ij} \sim \mathcal{CN}(0, \sigma_{ij}^2)$. The channel

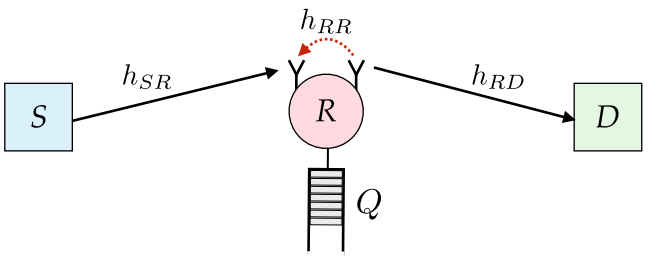

FIGURE 1. The buffer-aided cooperative relay network with hybrid half-duplex/full-duplex capabilities.

coefficient envelope follows the Rayleigh distribution, i.e., $|h_{ij}| \sim \text{Rayleigh}(\sigma_{ij})$. Therefore, the channel gains $g_{ij} \triangleq |h_{ij}|^2$ are exponentially distributed, i.e., $g_{ij} \sim \text{Exp}(\sigma_{ij}^{-2}/2)$. In addition, the wireless channels' distribution is considered to be either strict-sense stationary (SSS) or non-stationary during network operation.

It is assumed that the source is saturated and always has packets scheduled for transmissions. Also, the required

information rate, r_0, to successfully receive the packets at the receivers is fixed and application-dependent. As a result, the transmission of transmitter i towards receiver j will be successful, when SNR Γ_{ij} at the reception will be greater than or equal to the *capture ratio* γ_j, i.e., the *capture model* is adopted in this work. Regarding the variance of thermal noise at the relay node and the destination, it is denoted by σ_R^2 and σ_D^2, respectively, being modelled as additive white Gaussian noise (AWGN). As hybrid relay operation is considered in our network, at each frame, two duplexing modes are possible, i.e., *FD* and *HD max-link* [23] for enhanced reliability.

In FD, simultaneous transmissions by the source and the relay are performed, adopting transmit power levels P_S and P_R, respectively. In this case, SI arises and by h_{RR}, we denote the instantaneous residual SI from the output to the input antenna of the relay. Residual SI follows a complex Gaussian distribution, with values within $(0, \sigma_{RR}^2)$. The $\{S \rightarrow R\}$ transmission will be successful when the signal-to-interference-and-noise ratio (SINR) satisfies

$$\Gamma_R(P_S) = \frac{g_{SR}P_S}{g_{RR}P_R + \sigma_R^2} \geq \gamma_R. \tag{1a}$$

Next, the $\{R \rightarrow D\}$ transmission will be successful if the signal-to-noise ratio (SNR) at the destination is such that

$$\Gamma_D(P_R) = \frac{g_{RD}P_R}{\sigma_D^2} \geq \gamma_D. \tag{1b}$$

In our network, we consider a fixed source power of (P_S) while on the contrary, the relay power (P_R) is adjustable for mitigating the impact of SI and enhancing the end-to-end throughput performance.

In max-link, in frame, the source or the relay is activated for packet transmission. This operation leads the network to schedule only one node each time, and SI is avoided. Thus, towards increasing the $\{R \rightarrow D\}$ SNR, the relay will adopt its maximum power level, i.e., $P_R = P_{R,\max}$. Furthermore, in the $\{S \rightarrow R\}$ link SI does not exist and hence, (1a) will be equal to

$$\Gamma_R(P_S) = \frac{g_{SR}P_S}{\sigma_R^2} \geq \gamma_R. \tag{1c}$$

Packet retransmission relies on ACK/NACK packet by the receives that broadcast error-free and short-length packets via a separate narrow-band link. A link is assumed to be *feasible* when it does not experience an outage and also, the queue conditions are satisfied, i.e., non-full buffers for $\{S \rightarrow R\}$ transmissions and non-empty buffers for $\{R \rightarrow D\}$ transmissions.

A. MAX-LINK: ADAPTIVE LINK AND RELAY SELECTION

For convenience, in this subsection, details on the HD max-link policy are provided. Initially, BA opportunistic relaying policies, relied on two-slots protocols and fixed scheduling, where odd time-slot were reserved for the source's transmission and even time-slot for the relay's transmission. This inefficient approach was surpassed in [23], where each slot was flexibly allocated to the source or the relay, according to the instantaneous CSI and the relays' buffer status. In multi-relay setups, max-link exploited the buffering flexibility and compared the channel gains of the feasible links to activate the strongest one, as follows

1. When an $\{S \rightarrow R\}$ link is the strongest one prevails, the source will transmit to the respective relay. An $\{S \rightarrow R\}$ link is available for selection, as long as the relay's buffer has space to store another packet from the source.
2. On the contrary, if an $\{R \rightarrow D\}$ link prevails during the selection process, the respective relay transmit towards the destination. An $\{R \rightarrow D\}$ link is available when the relay has stored at least one packet in its buffer.

Max-link operation for selecting the best relay to receive/transmit is expressed as

$$R^* = \arg\max_{R_k \in \mathcal{C}} \left\{ \bigcup_{R_k \in \mathcal{C}: \Psi(Q_k) \neq L} \{g_{SR_k}\}, \bigcup_{R_k \in \mathcal{C}: \Psi(Q_k) \neq 0} \{g_{R_kD}\} \right\}, \tag{2}$$

where R^* is the activated relay while the function $0 \leq \Psi(Q_k) \leq L$ denotes the number of stored packets in k-th relay's buffer Q_k.

B. ESTIMATION AND FEEDBACK ERRORS

In wireless communication systems, having reliable control channels is vital. More specifically, stringent quality of service demands imposed by the ultra-reliable low-latency communications (URLLC) service type, setting strict requirements on hybrid automatic repeat request (HARQ) procedures. When a NACK is incorrectly decoded as an ACK, it can lead to delays, while the opposite error event results in unnecessary retransmissions and waste of radio-resources. It should be noted that the impact of these two error types can be adjusted by fine-tuning the rate of false alarms and the respective detection threshold value towards optimizing the network's performance.

The 3GPP TS 38.212 specification on multiplexing and channel coding offers various methods for encoding HARQ feedback with uplink control information. These methods include repetition coding, Polar coding, simplex coding, or Reed-Muller coding, each with varying coding rates and overheads [32]. This diversity in encoding methods provides different options to adjust false alarm rates or detection error in fading channels, satisfying service requirements and adapting to different radio propagation conditions.

The negative impact of control channel errors on wireless system performance is a complex problem that is out of scope for the current study. So, for simplicity, in this work, we assume that errors related to decoding ACKs/NACKs can be considered negligible and are ignored in the results.

C. OUTDATED CSI

In practice, the obtained CSIT for transmit power level selection does not match perfectly to the wireless link,

as a consequence of the feedback mechanism's delays. More specifically, outdated CSI might be attributed to channel variations, occurring between the beginning of the channel estimation process and the actual start of the transmission [33]. Another case, involves low-complexity approaches, avoiding continuous feedback to reduce network coordination overheads [34].

To enhance the practical aspects of our study, outdated CSI is considered and its impact on the power control process is investigate. In settings characterized by CSI feedback delays, the actual channel response h_{ij} conditioned on the estimated channel response $\hat{h}_{ij}$ of $\{i \to j\}$ link, prior to power control is expressed by [33]

$$h_{ij}|\hat{h}_{ij} \sim \mathcal{CN}\left(\rho_i \hat{h}_{ij}, 1-\rho_i^2\right), \tag{3}$$

where $\rho_i \in [0, 1)$ is the h_{ij} and $\hat{h}_{ij}$ correlation coefficient.

D. CSI-BASED POWER CONTROL

In case the source and relay power levels can be jointly set, as it is the case with full CSI availability, it suffices to calculate the minimum P_S and P_R values, so that inequalities (1b) and (1c) can be satisfied with equality. Here, the optimal transmit power levels (P_S^*, P_R^*) are as follows

$$\left(P_S^*, P_R^*\right) = \left(\frac{\gamma_R\left(|h_{RR}|^2 P_R^* + \sigma_R^2\right)}{|h_{SR}|^2}, \frac{\gamma_D \sigma_D^2}{|h_{RD}|^2}\right). \tag{4}$$

Towards obtaining the optimal transmit power levels, the source must acquire the $\{S \to R\}$ and $\{R \to R\}$ channel gains, the optimal transmit power of the relay, the relay's thermal noise value, and finally, the relay's decoding threshold. Meanwhile, relay only needs to know the $\{R \to D\}$ channel gain, as well as the destination's thermal noise and decoding threshold.

If fixed relay transmit power is considered and only the source transmit power level is optimized, then the source should have obtained all necessary information, apart from the relay transmit power level, that is known to the source. If, still, the fixed source transmit power is considered, as in our setup, the minimum transmit power level P_R, denoted by $P_R^\dagger$, is expressed as

$$P_R^\dagger = \frac{\gamma_D \sigma_D^2}{|h_{RD}|^2}, \tag{5a}$$

provided

$$P_R^\dagger \leq \frac{|h_{SR}|^2 P_S - \gamma_R \sigma_R^2}{\gamma_R |h_{RR}|^2}. \tag{5b}$$

From eqs. (5a)–(5b) it can be deduced that to use the optimal power level (eq. (5a)) and ensuring that the solution is feasible (eq. (5b)), the relay should obtain information of all three involved channels, the thermal noise values at the relay and destination, the transmit power level of the source, and γ_R and γ_D values. Considering that the thermal noise values, the power level of the source, and the decoding thresholds are known, still, the relay must estimate $|h_{SR}|^2$, $|h_{RR}|^2$, and $|h_{RD}|^2$. However, in case the relay chooses the optimal power $P_R^\dagger$ without examining if it is feasible (when eq. (5b) does not hold, then, there no relay transmit power level P_R exists that can support the desired transmission rate), it only has to acquire the values of $|h_{RD}|^2$, γ_D, and σ_D^2.

III. PROBLEM STATEMENT

It might be the case that, at each frame, only statistical CSIT is available, and the wireless network will aim to achieve a success transmission probability over a link that will be greater than or equal to an application- or condition-dependent threshold q_{th}, i.e., $\mathbb{P}\{\Gamma_i(P_j) \geq \gamma_i\} \geq q_{\text{th}}$. If this condition cannot be simultaneously fulfilled by both $\{S \to R\}$ and $\{R \to D\}$ links during FD operation, some works develop policies that switch to adaptive link selection; see, e.g., [26]. Here, as we assume that channel distribution is SSS, the statistical CSIT is not available prior to transmission. Thus, the network does not know *a priori* which relay mode, among FD and HD max-link, should be chosen and how to set relay's power level. As a remedy, we deploy a reinforcement learning algorithm that enables the relay, being the decision maker in this network, to decide which operation mode is the best one and, in case FD is activated, to adjust its power level for maximizing the end-to-end throughput.

In order to highlight the problem's complexity, below, we provide the success probabilities for each link and the various transmission options at each time frame. First, inequality (1a) can be written as

$$g_{SR} P_S - g_{RR} \gamma_R P_R \geq \gamma_R \sigma_R^2. \tag{6}$$

In (6) the exponentially distributed variables $g_{SR} \sim \text{Exp}(\mu)$ and $g_{RR} \sim \text{Exp}(\lambda)$, $\lambda, \mu > 0$, are linearly combined and as a result, the formed distribution becomes [26]

$$f_X(x) = \frac{\lambda\mu}{\lambda P_S + \mu \gamma_R P_R} \begin{cases} \exp\left(-\frac{\mu}{P_S} x\right), & \text{if } x \geq 0, \\ \exp\left(\frac{\lambda}{\gamma_R P_R} x\right), & \text{if } x < 0. \end{cases}$$

The probability that inequality (1b) holds, $\mathbb{P}((1b))$, since g_{RD} is exponentially distributed (i.e., $g_{RD} \sim \text{Exp}(\nu)$, $\nu > 0$), is expressed as

$$\mathbb{P}\big((1b)\big) = 1 - F_{g_{RD}}\left(\frac{\gamma_D \sigma_D^2}{P_R}\right) = \exp\left(-\nu \frac{\gamma_D \sigma_D^2}{P_R}\right), \tag{7}$$

where $F_W(w)$ is the cumulative distribution function (cdf) of a random variable W; specifically, for the exponential distribution, the cdf is given by $F_W(w) = 1 - \exp(-\lambda_W w)$. Likewise, the probability for inequality (1a) to hold, $\mathbb{P}((1a))$, is given by

$$\begin{aligned} \mathbb{P}\big((1a)\big) &= 1 - F_{g_{SR}P_S - \gamma_R g_{RR} P_R}\left(\gamma_R \sigma_R^2\right) \\ &= \frac{P_S \lambda}{P_S \lambda + \gamma_R P_R \mu} \exp\left(-\mu \frac{\gamma_R \sigma_R^2}{P_S}\right). \end{aligned} \tag{8}$$

When the $\{S \to R\}$ link is not affected by interference, the distribution reverts to an exponential distribution, i.e.,

$$\mathbb{P}\big((1c)\big) = \exp\left(-\mu \frac{\gamma_R \sigma_R^2}{P_S}\right). \tag{9}$$

At each time frame k, the relay must select among the following three options:

- O_1 Activate FD operation, enabling concurrent transmissions by the source and the relay. The source transmit with fixed power P_S, while the relay chooses a power level $P_R[k]$. $P_R[k]$ is set with the target of end-to-end throughput maximization (as it will be explained in Section V). Here, at least one packet must be stored in the relay's buffer ($Q > 0$) and less than L packets ($Q < L$), as a packet will stay in its queue, in case the transmission is not successful).
- O_2 Employ a fixed power level (P_S) transmission by the on the $\{S \to R\}$ link. For this case to occur, the relay should have maximum $L-1$ packets in its buffer ($Q < L$).
- O_2 Prompt a fixed power level relay transmission with $P_R = P_{R,\max}$, where $P_{R,\max}$ is the maximum power level of the relay, on the $\{R \to D\}$ link. Here, at least one packet must be stored in the relay's buffer ($Q > 0$).

IV. THE MAB MODEL

A. THE SETUP

A player has a set of possible actions to choose from, usually referred to as arms, over T rounds. In each round, the player chooses an arm and collects a reward for this arm. Note that the player observes only the reward for the selected action and not the rewards for other actions that could have been selected; this is called bandit feedback. For each action taken, the reward is sampled independently by the associated reward distribution. The reward distributions are initially unknown to the player. The player's goal is to maximize its expected accumulated reward over the T rounds.

For known reward distributions, this goal can be achieved through the selection of the arm with the highest average reward. In order to identify the optimal arm, various arms must be played by the player, for learning their reward distributions (exploration) while ensuring that the obtained knowledge on reward distributions is leveraged in the sense of preferring arms with higher expected rewards (exploitation). To measure the performance of the player in conducting such an *exploration-exploitation trade-off* the notion of *regret* is adopted, in which, the learner's cumulative reward is compared against that achieved by always choosing the optimal arm. The regret is defined as the difference between the acquired reward by pulling the best arm and the learner's choice. In our network, the target is the identification of a policy over a finite time horizon T maximizing the expected number of packets that are successfully transmitted (throughput). Thus, we are targeting the design of a relay mode selection and power control policy that will minimize the *regret*. The regret for a policy $\pi \in \Pi$ (Π being the set of all feasible policies) is measured by the performance loss and it can be calculated through the comparison of the performance provided by policy π to that of the best static policy, i.e.,

$$R^{\pi}(T) = \max_{\ell \in \mathcal{L}} \mathbb{E}\left\{\sum_{t=1}^{T} U_{\ell,t}\right\} - \mathbb{E}\left\{\sum_{t=1}^{T} U_{I_t^{\pi},t}\right\}, \tag{10}$$

where $U_{\ell,t}$ represents the instantaneous utility obtained from selecting link ℓ at time-slot t, assuming a feasible configuration $\ell \in \mathcal{L}$. Additionally, $U_{I_t^{\pi},t}$ is the instantaneous utility acquired from link I_t^{π}, selected through policy π at time-slot t. In this cooperative network, the relay node will either receive, transmit with $P_{R,\max}$, or operate in FD mode and transmit, using a power level from a finite set of discrete power levels, $\mathcal{P}_R$. This set depends on the radio specifications and configuration. As a result, in MAB problems, each arm corresponds to $O_{1,i}$ with i, denoting one of the $|\mathcal{P}_R|$ power levels in FD mode, or to HD modes O_2 and O_3 (hence, there will be $|\mathcal{P}_R| + 2$ arms in total).

In the seminal paper by Lai and Robbins [15] the characterization of a problem-dependent lower bound on the regret of any adaptive policy is presented, showing that the lower bound grows logarithmically over the time horizon T. More specifically, they prove that for any *uniformly good* adaptive learning algorithm π,[1]

$$\liminf_{T \to \infty} \frac{R^{\pi}(T)}{\log(T)} \geq c(\boldsymbol{\mu}), \tag{11}$$

where $\boldsymbol{\mu}$ represents the vector of mean rewards of different arms, and $c : [0,1]^{|\mathcal{L}|} \to \mathbb{R}$ denotes a deterministic and explicit function.

B. UPPER CONFIDENCE BOUND (UCB) POLICIES

UCB policies is an approach for adaptive exploration, which relies on the principle of *optimism under uncertainty* (*optimistic* principle) proposed by Lai and Robbins [15]. The intuition is that by assuming that each arm is as good as it can possibly be given the observations so far, it is the best option to choose the best arm based on these optimistic estimates.

To describe the generic form of such policies, we introduce some notation. We let I_t denote the arm selected at time t. Also, we let $n_{j,t}$ represent the number of plays of arm j until round t, i.e., $n_{j,t} := \sum_{s=1}^{t} \mathbb{1}_{\{I_s = j\}}$, where $\mathbb{1}_A$ is the indicator function of the event A. We let $\hat{q}_{j,t}$ denote the empirical average reward of arm j accumulated though the observations from j up to t:

$$\hat{q}_{j,t} = \frac{1}{n_{j,t}} \sum_{s=1}^{t} r_{j,s} \mathbb{1}_{\{I_s = j\}}, \tag{12}$$

where $r_{j,t}$ represents the reward of arm j at round t.

[1] An algorithm π is uniformly good if for any sub-optimal arm i, the number of times arm i is selected up to round t, $n_i(t)$, fulfils : $\mathbb{E}[n_i(t)] = o(t^{\alpha})$, for all $\alpha > 0$.

A UCB policy π maintains an index function $\bar{q}_j$ for each action (arm) j, depending on previous observations of j only (e.g., $\widehat{q}_{j,t}$, $n_{j,t}$, etc.), and that $\bar{q}_{j,t} \geq q_j$ with high probability for all $t \geq 1$. Then, π will simply consist in choosing the arm with the largest index $\bar{q}_{j,t}$ at each round t:

$$I_t = \arg\max_{j \in \mathcal{L}} \bar{q}_{j,t}. \tag{13}$$

Below, we outline UCB1 [35], a simple policy designed based on Hoeffding's inequality for bounded random variables. The UCB1 index (or for short, UCB) is defined as

$$\bar{q}_{j,t}^{\text{UCB}} = \widehat{q}_{j,t} + \sqrt{\frac{3\log(t)}{2n_{j,t}}}. \tag{14}$$

From (14), it can be deduce that an arm which has not been explored as often as other arms will have a bigger UCB1 index, thus enabling that arm to be picked.

V. MAB-AIDED AND FULL-DUPLEX/MAX-LINK RELAY MODE SELECTION

A. PRELIMINARIES

The relay aims at minimizing the long term regret (10) of the overall system; this is equivalent to maximizing the end-to-end throughput of the system. Towards that direction, at each time frame k, the relay has to decide *i)* in which mode to operate: HD (either O_2 or O_3) or FD ($O_{1,i}$) and *ii)* its power level P_R. In standard MAB communication problems the actions are decoupled. However, in this setup, the options are coupled in several ways. For example, option O_1 includes the $\{R \to D\}$ link of option O_3.

The optimization problem of maximizing the end-to-end throughput can be expressed as follows:

$$P1: \quad \max_{P_R} \left\{ \min\left\{ \mathbb{P}\big((1b)\big), \mathbb{P}\big((1a)\big) \right\} \right\} \tag{15a}$$

$$\text{s.t } 0 \leq P_R \leq P_{R,\max}. \tag{15b}$$

The epigraph form of optimization problem $P1$ is given below:

$$P2: \quad \max_{P_R} e \tag{16a}$$

$$\text{s.t } \mathbb{P}\big((1b)\big) \geq e, \tag{16b}$$

$$\mathbb{P}\big((1a)\big) \geq e, \tag{16c}$$

$$0 \leq P_R \leq P_{R,\max}, \tag{16d}$$

$$0 < e \leq 1. \tag{16e}$$

where e serves as the epigraph of the function

$$g(P_R) \triangleq \min\left\{ \mathbb{P}\big((1b)\big), \mathbb{P}\big((1a)\big) \right\}.$$

Proposition 1: Suppose that the parameters of the distribution of the channels, μ, ν and λ, are already known. Also, parameters, P_S, γ_R, γ_D, σ_R^2, and σ_D^2 are assumed to be known. Then, the optimal relay power, P_R^*, such that the end-to-end throughput in the FD mode is maximized is given by:

$$P_R^* = \frac{\nu\gamma_D\sigma_D^2}{\ln(1/e^*)} = \frac{P_S\lambda(\beta(\mu) - e^*)}{\gamma_R\mu}, \tag{17}$$

where e^* is the solution to the optimization problem $P2$ and $\beta(\mu) = \exp\left(-\mu\frac{\gamma_R\sigma_R^2}{P_S}\right)$.

Proof: Since $\mathbb{P}((1b))$ and $\mathbb{P}((1a))$ are monotonically increasing and decreasing functions of P_R, respectively, the optimal solution to this problem is achieved with equality for both (16b) and (16c). Therefore, after algebraic manipulations, we obtain (17). Note that z^* can be obtained by solving the equality in (17) (using simple line search methods, such as bisection), since all the parameters, P_S, γ_R, γ_D, σ_R^2, and σ_D^2 are assumed to be known. Note that whether the boundary conditions of P_R are satisfied can be justified separately. ■

Hence, with known channel distribution parameters, one could compute P_R^* directly, thus, finding which mode of operation is the optimal with respect to the overall throughput of the system.

B. ONLINE LEARNING MODEL

In case the wireless channels' distributions are SSS, the success probabilities for each mode of operation will be fixed but unknown. For computing these success probabilities and hence the optimal power levels for each mode of operation, in what follows, we propose an algorithm, herein called MABAMS, with which the channel distribution parameters are implicitly estimated and the best available option is obtained.

Using the MAB framework, we assign each discrete power level to an arm. Hence, pulling an arm is equivalent to a packet transmission using the selected power level. As illustrated in Fig. 2, depending on the relay mode of operation, i.e., HD or FD, a different reward is obtained. More specifically, in the considered two-hop cooperative relay network, if during time frame k FD transmission is selected with power level j, a reward $r_{j,k}^{(FD)}$ is obtained, where

$$r_{j,k}^{(FD)} = \begin{cases} 2, & \text{if } \{S \to R\} \text{ and } \{R \to D\} \text{ trans. successful,} \\ 1, & \text{if one of the trans. successful,} \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

If, instead, HD transmission is selected (either $\{S \to R\}$ or $\{R \to D\}$ link), then a reward $r_{j,k}^{(HD)}$ is obtained, where

$$r_{j,k}^{(HD)} = \begin{cases} 1, & \text{if trans. successful,} \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

Note that the arm selection yields a random reward, which reveals information about the link/links (i.e., links $\{S \to R\}$, $\{R \to R\}$, and $\{R \to D\}$) under consideration. Despite the fact that there exist correlations between the (herein assumed independent) arms and one can infer information from one outcome about the other, in this paper we do not exploit this available side information.

We take into consideration the following two scenarios, depending on whether the probabilities of successful transmission evolve over time or not:

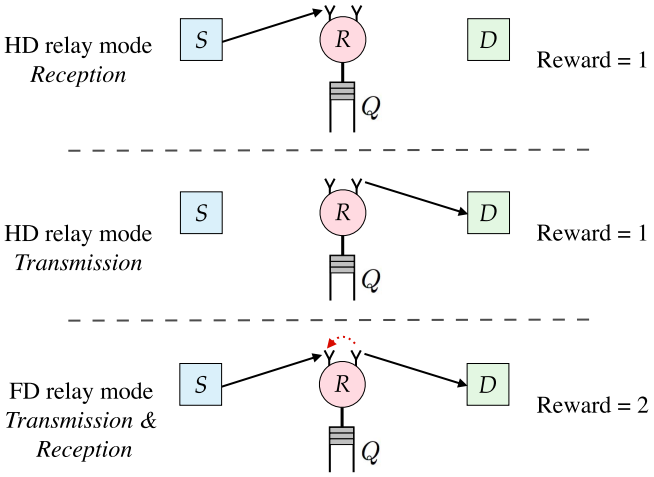

FIGURE 2. The obtained reward for the different relay modes.

1) Case 1: SSS channels (hence, fixed) success probabilities: Here, success probabilities of the *SR* and *RD* channels are fixed but unknown. Hence, for each j, $(r_{j,t})_{t\geq 1}$ is a sequence of i.i.d. Bernoulli random variables with $\mathbb{E}[r_{j,t}|\mathcal{F}_{t-1}] = q_j$ for all t, where $\mathcal{F}_{t-1}$ denotes the history of power levels chosen by the proposed algorithm up to round $t-1$, and their corresponding rewards.

2) Case 2: non-stationary channels (hence, time-varying) success probabilities: This case corresponds to a system, with varying channel statistics may change over time. More specifically, the environment is time-varying, and as a consequence, the success probabilities change over time. Specifically, $(r_{j,t})_{t\geq 1}$ is a sequence of independent Bernoulli random variables with $\mathbb{E}[r_{j,t}|\mathcal{F}_{t-1}] = q_{j,t}$ for all t. Note that this behaviour implies that the optimal arm (and hence, power level) may also change over time, so in the definition of regret in (10), the maximizer in the first term changes over time. According to the terminology used for non-stationary MABs, the time instants at which such (abrupt) changes occur are called *breakpoints* [36]. In this work, we assume that breakpoints occur independently of the channel selection strategy or of the sequence of rewards. Let Υ_T denote the number of breakpoints before time T. For being able to learn the optimal changing power level, we additionally assume that Υ_T grows sublinearly with T, i.e., $\Upsilon_T = o(T)$. This assumption is necessary for being able to achieve a sublinear regret.

C. THE MABAMS ALGORITHM

In what follows, we describe MABAMS, developed for cooperative relay networks. It is expected that after the initial exploration phase, and under the assumptions aforementioned, MABAMS will reach the best mode of relaying operation (FD or HD) along with the power level (if the best mode is FD) yielding the maximum reward, in terms of end-to-end throughput.

Algorithm 1 MABAMS at Each Time Frame

input Power levels $\mathcal{P}_R$, $P_{R,\max}$ set, Q, SNR thresholds γ_R and γ_D, thermal noise variance at R (σ_R^2) and D (σ_D^2)
set $P_R[0] = P_{R,\max}$
for $k = 0, 1, 2, \ldots$ **do**
  if $Q = 0$ **then**
    Select O_2 and calculate $\hat{q}_{\mu,k}$ (12) and then $\bar{q}_{\mu,k}$
  else if $Q = L$ **then**
    Select O_3 and calculate $\hat{q}_{v,k}$ and $\hat{q}_{\lambda,k}$ (12) and then $\bar{q}_{v,k}$ and, $\bar{q}_{\lambda,k}$, respectively
  else if $0 < Q < L$ **then**
    Calculate $\hat{q}_{j,k}$ (12) and then $\bar{q}_{j,k}$,, where $j \in \{O_{1,i}, O_2, O_3\}$, $i = \{1, 2, \ldots, M\}$
    Choose mode j (with power level $i \in \{1, 2, \ldots, M\}$ if option O_1 is selected) for transmission at time-slot k, employing (13)
    $n_{j,k+1} \leftarrow n_{j,k} + \mathbb{1}_{\{I_k=j\}}$ for all j
    if $j = O_2$ OR $j = O_3$ **then**
      if transmission is successful **then**
        $r_{j,k}^{(HD)} = 1$
      end if
    end if
    if $j = O_{1,i}$ (i.e., with power level i) **then**
      if transmission is successful on both links **then**
        $r_{j,k}^{(FD)} = 2$
      else if transmission is successful on one link **then**
        $r_{j,k}^{(FD)} = 1$
      end if
    end if
  end if
end for

The procedure followed in each time frame is summarized in Algorithm 1. In what follows, the steps taken by the relay in the algorithm are described.

1) First, it should be noted that the relay is always aware of its queue size Q, capture ratios γ_R and γ_D, thermal noises σ_R^2 and σ_D^2.
2) It sets its initial power to the maximum allowable power level, i.e., $P_R[0] = P_{R,\max}$
3) At each time step k, the relay has to select among three options, depending on the size of its buffer:
4) If the buffer is empty, then the relay has nothing to transmit and in the current time slot it can only receive a packet from the source. Hence, the source is requested to transmit a packet to the relay. Then, based on the outcome of the transmission, the relay computes the empirical average reward of the $\{S \rightarrow R\}$ link built using the observations of the link from up to step k.
5) If, however, the buffer is full of packets (and hence cannot receive any), it is forced to transmit a packet (with maximum power) to the destination while the source remains silent. Similarly, based on the outcome of the transmission (that the relay is informed about via ACK/NACK feedback from the destination), the relay computes the empirical average reward of the $\{R \rightarrow D\}$ link.

TABLE 3. Required overheads of CSI-based FD/HD mode selection and MABAMS at each time-slot.

	Pilot transmissions			CSI estimations		
	S	R	D	S	R	D
CSI-based FD/HD selection	1	1	1	0	2	1
MABAMS	0	0	0	0	0	0

6) If the buffer is neither full nor empty, the relay may either deploy the FD mode (in which both the source and the destination transmit simultaneously) or choose a HD mode aforementioned. If the FD mode is deployed, the power level of the source is fixed, but the relay can transmit with a power from a set of discrete power levels. The power that maximizes the index in (13).

Remark 1: In conventional CSI-based power control, as described in Section II-D, at the start of each time-slot, the source, the relay and the destination must transmit pilot sequences in order to obtain the instantaneous $\{S \rightarrow R\}$, $\{R \rightarrow R\}$, and $\{R \rightarrow D\}$ links, respectively. Having acquired this information, a pre-specified node, e.g., the source or the destination can evaluate which power value should be adopted by each relay and decide which relay node should be adopted in each transmission period. Optimal power control and power allocation between pilot and data symbols in full-duplex relays require numerical optimization and transmission of pilot symbols within the channel coherence time [37]. On the contrary, the online and learning-based framework of MABAMS enables the network to adaptively and recursively learn the statistics, compared to collecting enough data for parameter estimation. In this way, the robustness of the network against the inherent characteristics of wireless channels is increased by: 1) avoiding issues related to phase noise that degrade channel estimation quality [38] and 2) overcoming the impact of outdated CSI in fast-changing wireless environments [33].

Towards better highlighting the importance of online learning for FD/HD relay mode selection, Table 3 includes a comparison in terms of CSI overheads among the conventional CSI-based approach and ML-aided MABAMS. From this comparison, it is evident that MABAMS can also guarantee improved scalability when multi-relay deployments are employed, as the amount of CSI and BSI overheads is proportional to the number of relays in the network. Moreover, in fast-changing environments, where only outdated CSI might be acquired,the performance gain (in terms of throughput) of MABAMS is illustrated in Section VI (Performance Evaluation). Note that this gain becomes even more noticeable when CSI estimation overheads are also taken into consideration when assessing the throughput performance. Finally, one should consider that by completely avoiding pilot transmissions and processing, ensures energy gains and reduced computation requirements in relay networks, thus facilitating the operation of resource-constrained devices in such settings.

VI. PERFORMANCE EVALUATION

Here, performance evaluation, in terms of average throughput for MABAMS and other CSI- and learning-based policies are presented. In greater detail, MABAMS is compared to FD relaying with power control, using outdated CSI (out-CSI PC) with $\rho = 0.5$, FD relaying without power control (no-PC), FD relaying with random power level selection (rnd) and learning-aided BB-PC where only the FD mode is employed [29], [30]. As a performance upper-bound for the FD relay operation, a CSI-based FD relaying policy is included (opt-PC) where power control at the relay sets the transmit power according to eq. (17).

We assume that for each wireless link, the transmit SNR varies between 0 dB to 30 dB. In this work, we consider that the transmit SNR corresponds to the ratio of the maximum available transmit power at the transmitter, being equal for all network nodes, i.e., $P_S = P_{R,\max} = P_{\max}$ over the noise power. Furthermore, in the comparison figures, the x-axis is the $\{R \rightarrow D\}$ link's transmit SNR, i.e., $P_{\max}/\sigma_D^2$. Also, for each transmit SNR value, we conduct 10^4 packet transmissions over which, we obtain the average throughput. Moreover, a transmission rate equal to $r_0 = 3$ bps/Hz is considered in this single-relay cooperative network where power control chooses among six different levels, i.e., $P_1 = P_{\max}$, $P_2 = 0.50P_{\max}$, $P_3 = 0.30P_{\max}$, $P_4 = 0.20P_{\max}$, $P_5 = 0.05P_{\max}$, $P_6 = 0.01P_{\max}$ [39]. When the FD relaying mode is activated, the SI channel at the relay is characterized by average channel SNR $\bar{\gamma}_{SI} \in \{-10, 0, 10\}$ dB.

Two cases are examined for the wireless setting. First, strict-sense stationary channels are assumed with fixed statistics during the whole transmission process. In addition, $\{S \rightarrow R\}$ and $\{R \rightarrow D\}$ links are assumed to be i.i.d. and characterized by average SNR $\bar{\gamma}_{\{S\rightarrow R\}} = \bar{\gamma}_{\{R\rightarrow D\}} = 0$ dB. The second case corresponds to a non-stationary wireless environment with abruptly changing $\{R \rightarrow D\}$ statistics at one breakpoint ($t = 5000$). When transmission begins, the $\{R \rightarrow D\}$ link is a line-of-sight (LoS) one, with Rician factor $K_{Rice} = 10$ dB. At the same time, $\{S \rightarrow R\}$ communication is characterized by non-LoS conditions, corresponding to Rayleigh fading conditions with $\bar{\gamma}_{\{S\rightarrow R\}} = 0$ dB. Following the breakpoint, the $\{R \rightarrow D\}$ channel is characterized by Rayleigh fading with $\bar{\gamma}_{\{R\rightarrow D\}} = 0$ dB. At the same time, the fading conditions do not change for the SR link, i.e., $\bar{\gamma}_{\{S\rightarrow R\}} = 0$ dB. Table 4 includes the adopted performance evaluation parameters.

A. STRICT-SENSE STATIONARY WIRELESS CHANNELS

The first comparison is illustrated in Fig. 3, assuming that SI has a severe impact on the FD operation of the relay. The hybrid nature of MABAMS shows its advantage for low to medium SNR values, as FD operation cannot be performed due to the very strong SI channel. As a consequence, network operation relies on HD relay, avoiding outages at

TABLE 4. Simulation parameters.

Parameter	Value
No. of transmissions per SNR value	10^4
Buffer size L [23]	10 packets
Transmission rate r_0	3 bps/Hz
Outdated CSI coefficient ρ [34], [35]	0.5
Average SI channel SNR $\bar{\gamma}_{SI}$	{-10, 0, 10} dB
Transmit SNR $P_{\max}/\sigma_D^2$ range	{0, 30} dB
No. of relay power levels [40]	6
Wireless channel types	i.i.d. / i.n.i.d. Rayleigh / Rice
Rician factor K_{Rice}	10 dB
Stationarity breakpoints [37]	1 (at $t = 5000$)

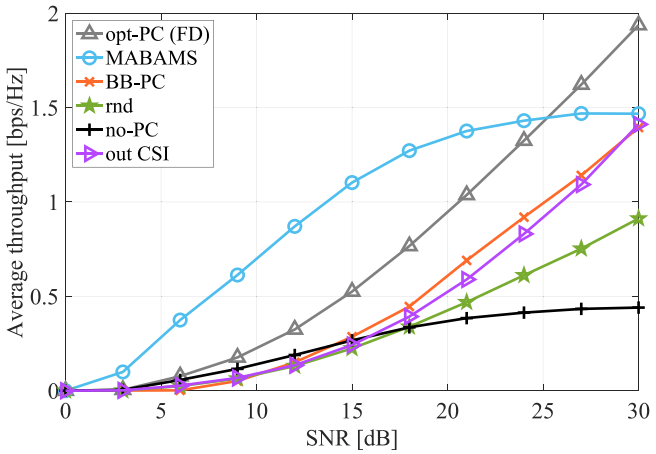

FIGURE 3. The average throughput performance comparisons for $\bar{\gamma}_{SI} = 10$ dB and various power control algorithms (strict-sense stationary case).

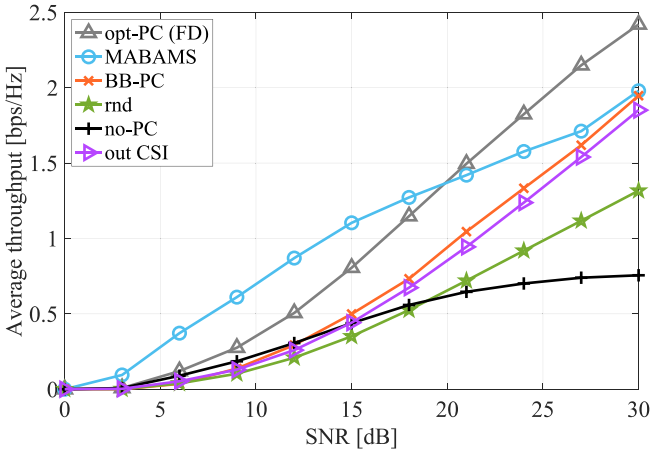

FIGURE 4. The average throughput performance comparisons for $\bar{\gamma}_{SI} = 0$ dB and various power control algorithms (strict-sense stationary case).

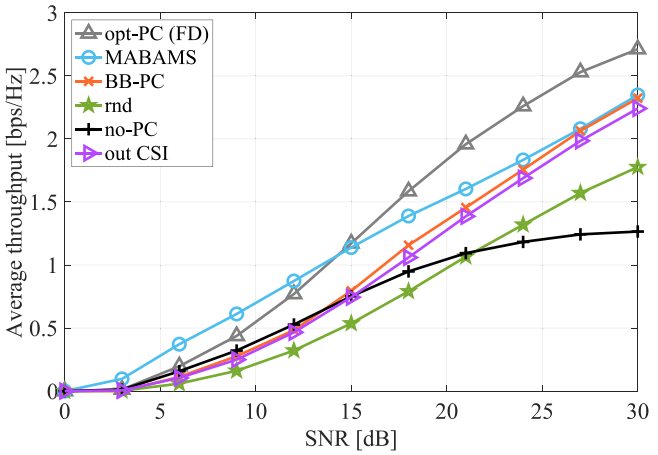

FIGURE 5. The average throughput performance comparisons for $\bar{\gamma}_{SI} = -10$ dB and various power control algorithms (strict-sense stationary case).

the cost of lower throughput. The other algorithms suffer significant throughput losses and only opt-PC manages to outperform MABAMS after 25 dB where power control is capable of appropriately setting the relay's transmit power to satisfy the required rate and mitigate the impact of SI. Also, the learning-aided BB-PC which only relies on FD operation offers higher throughput than the CSI-based policy with outdated channel estimation, while random PC and especially, no-PC cannot combat the very strong SI conditions.

Next, the average throughput performance under an SI channel with $\bar{\gamma}_{SI} = 0$ dB is depicted in Fig. 4. This comparison shows that MABAMS has the best performance up to 20 dB. As reinforcement learning-aided mode switching is enabled, MABAMS is able to overcome cases where FD relaying is infeasible, leveraging the HD and BA relaying capabilities. After 20 dB, opt-PC provides the best performance due to optimal power control but entails higher complexity as accurate and timely channel estimation must be ensured to acquire instantaneous CSI from the $\{S \to R\}$, $\{R \to R\}$ and $\{R \to D\}$ channels. It should be underlined that in practice, a fraction of each time-slot is dedicated for channel estimation and thus, the actual performance of CSI-based PC policies, i.e., opt-PC and out-CSI will be worse. For high SNR, MABAMS more often activates the FD mode with power control and surpasses the performance of the rest of the policies. Most notably, MABAMS provides higher throughput compared to the policy with outdated CSI, highlighting the benefits of learning-aided relay mode switching. For higher transmit SNR values, the FD-only BB-PC provides similar throughput with MABAMS while, random power level selection and no-PC fail to combat the strong SI conditions.

Then, average throughput results for $\bar{\gamma}_{SI} = -10$ dB are provided in Fig. 5. As optimal CSI-based power control is able to adopt relay transmit power levels that can more efficiently mitigate the weak SI, after 15 dB, opt-PC exhibits the best throughput performance. In addition, it can be seen that power control with outdated CSI falls behind the two learning-based policies, i.e., MABAMS and BB-PC throughout the transmit SNR range. Moreover, when the

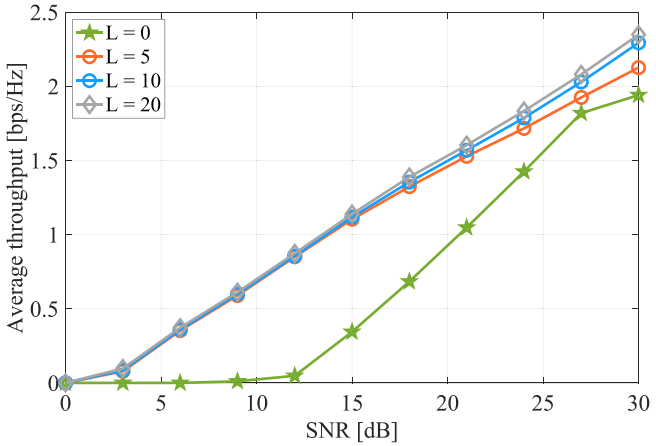

FIGURE 6. The average throughput performance of MABAMS for $\bar{\gamma}_{SI} = -10$ dB and varying buffer size L (strict-sense stationary case).

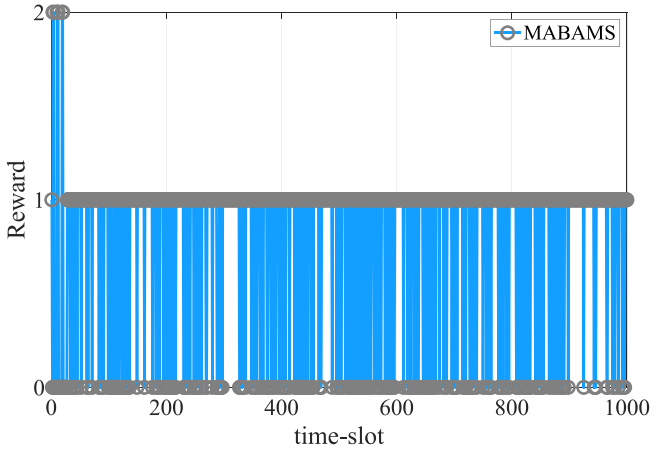

FIGURE 7. MABAMS convergence over time for $\bar{\gamma}_{SI} = -10$ dB and a transmit SNR equal to 15 dB (strict-sense stationary case).

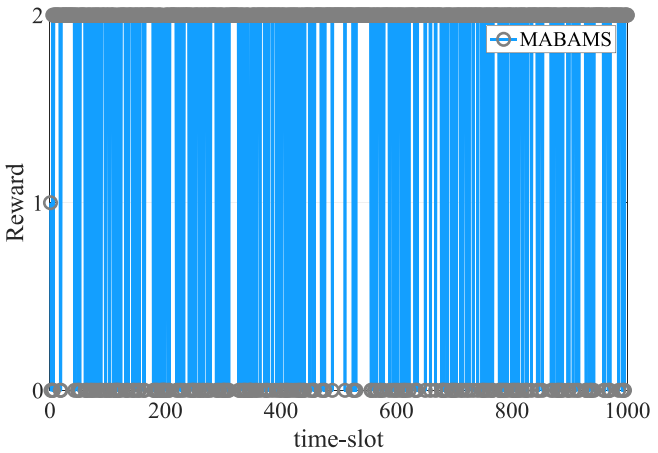

FIGURE 8. MABAMS convergence over time for $\bar{\gamma}_{SI} = -10$ dB and a transmit SNR equal to 30 dB (strict-sense stationary case).

network experiences low transmit SNR, the best performance is provided by MABAMS, as the relay reverts to HD transmissions, leveraging its buffer. In addition, until 15 dB MABAMS has the best performance, without introducing CSI estimation overheads and the associated energy costs which are critical, especially in resource-constrained networks, relying on sensors and other battery-dependent devices. In this context, it must be emphasized that the opt-PC and out-CSI PC policies experience additional performance losses due to channel estimation overheads that are not considered in the comparisons. Finally, the least throughput is offered from the algorithm without PC, as SI cannot be combated, and even the random transmit power selection algorithm offers higher throughput after 20 dB.

In Fig. 6, the impact of varying buffer size L on the average throughput performance of MABAMS is depicted. In greater detail, it can be observed that when buffering is not possible at the relay, network reliability is threatened as relay cannot store any packets and only FD relaying is possible. Then, when buffer-aided relaying is performed, the cases of $L = 5, 10, 20$ offer similar performance until 20 dB, as mainly HD relaying is activated and with increased diversity. After 20 dB, the case of $L = 5$ exhibits a small performance gap, as instances of empty buffers in the network slightly affect the average throughput performance. The other two cases, ensure higher packet availability in the network, and thus, both FD and HD relaying can be activate with enhanced reliability. It should be noted that when SNR conditions allow the FD mode to be selected, the value of L does not play a major role as usually, packets are forwarded right after their reception at the relay.

Next, the convergence of our algorithm is studied for different transmit SNR cases. It should be noted that the first 1,000 time-slots are shown in order to ensure visibility, as MABAMS does not change its action afterwards. In Fig. 7, a transmit SNR equal to 15 dB is assumed and MABAMS is shown to activate the HD duplexing mode, apart from a brief period where FD relaying is activated. For the rest of the time-slots, HD relaying is activated and that action is preferred throughout the duration of the experiment.

The second evaluation of convergence is included in Fig. 8 when the transmit SNR is equal to 30 dB. As in this case, the increased transmit SNR allows FD relaying to be activated, MABAMS easily identifies the correct action and remains constant during the whole transmission period. As a result, for very high transmit SNR, the operation of MABAMS almost exclusively relies on FD relaying.

B. NON-STATIONARY WIRELESS CHANNELS

The next set of comparisons assumes a non-stationary wireless environment where the $\{R \rightarrow D\}$ conditions abruptly change at $t = 5000$ time-slots. In Fig. 9, a very strong SI channel with $\bar{\gamma}_{SI} = 10$ dB is assumed. Such a setting emphasizes on the importance of FD/HD relay switching when power control is not able to mitigate the impact of SI at the relay. As a result, MABAMS clearly achieves to support network operation while all the FD policies experience outages. The opt-PC requires very high SNR in order to adopt transmit power levels that overcome the

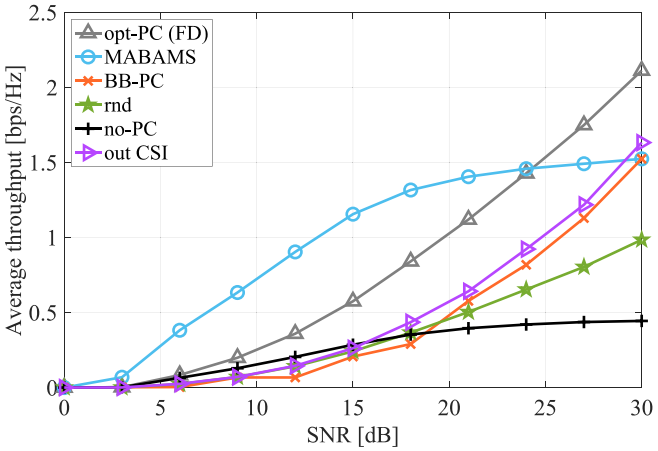

FIGURE 9. The average throughput performance comparisons for $\bar{\gamma}_{SI} = 10$ dB and various power control algorithms (non-stationary case).

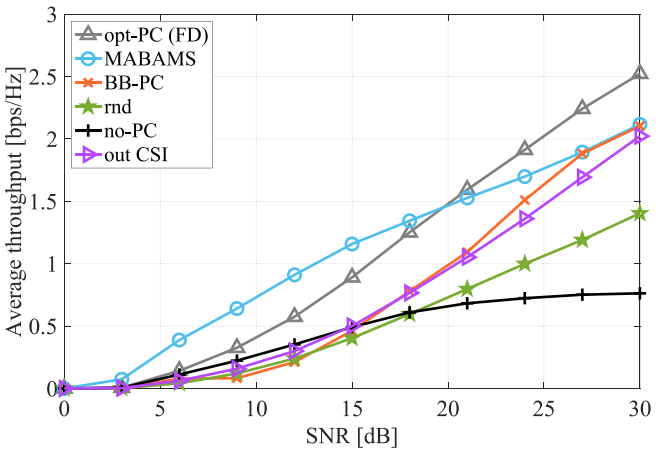

FIGURE 10. The average throughput performance comparisons for $\bar{\gamma}_{SI} = 0$ dB and various power control algorithms (non-stationary case).

SI's severity while out-CSI only manages to outperform MABAMS for an SNR value of 30 dB but without including the throughput losses due to CSI estimation that do not exist for the learning-aided policies, i.e., MABAMS and BB-PC.

Fig. 10 illustrates average throughput results under strong SI conditions with $\bar{\gamma}_{SI} = 0$ dB. The advantage of CSI-based opt-PC is evident for high SNR while MABAMS exploits the BA HD relaying mode to avoid outages for low to medium SNR values. For the first 5,000 time-slots, the $\{R \rightarrow D\}$ link experiences Rician fading and thus, after 20 dB, efficient learning-based power control enables the relay to switch to FD operation with low transmit power to achieve improved SI mitigation. Still, even when the $\{R \rightarrow D\}$ channel switches to non-LoS conditions after $t = 5000$ time-slots, MABAMS maintains its advantage over the rest of the policies, excluding opt-PC. However, as discussed for the previous comparisons, CSI-based policies incur spectral and energy costs while estimating CSI, rendering them less attractive for resource-constrained settings.

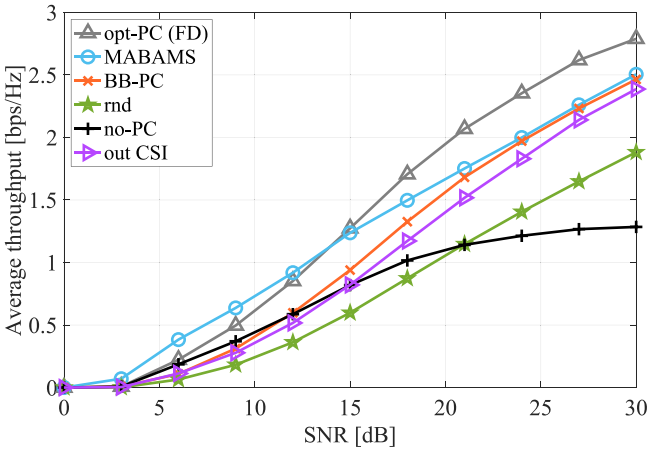

FIGURE 11. The average throughput performance comparisons for $\bar{\gamma}_{SI} = -10$ dB and various power control algorithms (non-stationary case).

Finally, throughput results for the case of weak SI characterized by $\bar{\gamma}_{SI} = -10$ dB and a non-stationary setting are presented in Fig. 11. Here, all the considered policies offer improved throughput even for SNR values below 20 dB, compared to the previous comparison. In greater detail, opt-PC manages to exceed the throughput performance of MABAMS after 15 dB. Then, when MABAMS is employed, increased throughput is provided compared to the rest of the policies throughout the SNR range. Again, it is clear that for fast-changing environments where outdated CSI might be obtained, relying on MABAMS results in significant performance gains that will be more pronounced when CSI estimation overheads are taken into consideration in the throughput performance.

From the performance evaluation, it has been shown that MABAMS is able to achieve autonomous network operation by avoiding channel estimation, and efficiently converging to the best duplexing mode, depending on wireless conditions. In addition, the performance gap that is observed in the case where optimal power control is performed due to full CSI availability will diminish when the impact of pilot transmissions, processing, and reporting are taken into consideration. For example, assuming that this process takes place at the start of each time-slot, it will lead to throughput losses proportional to the fraction of the time-slot that is allocated to CSI estimation. In most comparisons, allocating around 15-20% of each time-slot for CSI estimation will result in MABAMS outperforming the throughput performance of opt-PC in most SNR regions.

VII. CONCLUSION AND FUTURE DIRECTIONS

A. CONCLUSION

This paper presented a bandit-based algorithm, namely MABAMS, for conducting half-duplex/full-duplex relay mode selection without the need for pilot transmissions and processing, thus avoiding spectral and energy costs that are associated with conventional channel estimation procedures. In greater detail, MABAMS relies on reinforcement learning

and one-bit ACKs/NACKs to decide the duplexing mode and the transmit power level of the relay in case, full-duplex relaying is selected. It is shown that improved performance is achieved over schemes strictly relying on full-duplex relaying with channel state information-based power control. More importantly, apart from nullifying the amount of network coordination overheads, MABAMS enables autonomous network operation and its distributed operation guarantees that timely decisions are made. In this manner, MABAMS overcomes the issue of outdated channel state information, inherent to fast-changing wireless environments and networks with resource-constrained devices that cannot perform accurate channel estimation, e.g., industrial settings comprising a plethora of sensors.

B. FUTURE DIRECTIONS

Current research, including the proposed algorithm, is dedicated to developing efficient methods for acquiring the network's parameters while simultaneously maximizing network throughput, being the primary performance objective. This work reveals a lot of opportunities to further enhance performance in full-duplex relaying operating in fast-changing wireless conditions but also highlights some limitations of the proposed method. Specifically,

1) MAB-based approaches, such as MABAMS, assume discrete-valued variables. Hence, it is not possible to apply such an approach to problems in which the decision variables are continuous. In such a case, one has to deploy other methods, such as Bayesian-based methods.
2) An additional possible direction is to examine the case in which, the source power level is modeled as a (discrete) decision variable, although the number of variables will increase considerably, making the solution computationally expensive, and in need of a lot more additional trials.
3) Additionally, there can be a more efficient utilization of the information extracted from transmissions on the $\{S \rightarrow R\}$ link in different modes, an aspect that has not been studied in this work, due to the fact that there does not exist a straightforward way to extended the proposed algorithm to account for such coupling.
4) Finally, other performance metrics can be considered, such as latency, in which the main objective is to minimize latency subject to some fidelity criteria on the performance; such metrics are desirable in several industrial applications, such as flexible automation and autonomous cars. Our approach, however, does not provide a systematic way for choosing the reward function.

REFERENCES

[1] Z. Zhang, K. Long, A. V. Vasilakos, and L. Hanzo, "Full-duplex wireless communications: Challenges, solutions, and future research directions," *Proc. IEEE*, vol. 104, no. 7, pp. 1369–1409, Jul. 2016.

[2] S. Maghsudi and S. Stańczak, "Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1309–1322, Mar. 2015.

[3] S. Maghsudi and E. Hossain, "Multi-armed bandits with application to 5G small cells," *IEEE Wireless Commun.*, vol. 23, no. 3, pp. 64–73, Jun. 2016.

[4] K. Xu, Z. Shen, Y. Wang, X. Xia, and D. Zhang, "Hybrid time-switching and power splitting SWIPT for full-duplex massive MIMO systems: A beam-domain approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7257–7274, Aug. 2018.

[5] Q. Shi, Y. Liu, S. Zhang, S. Xu, and V. K. N. Lau, "A unified channel estimation framework for stationary and non-stationary fading environments," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4937–4952, Jul. 2021.

[6] M. A. Abdul Careem and A. Dutta, "Real-time prediction of non-stationary wireless channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7836–7850, Dec. 2020.

[7] G. Lu, X. Dai, W. Zhang, Y. Yang, and F. Qin, "Nondata-aided Rician parameters estimation with redundant GMM for adaptive modulation in industrial fading channel," *IEEE Trans. Ind. Informat.*, vol. 18, no. 4, pp. 2603–2613, Apr. 2022.

[8] J. Pan, H. Shan, R. Li, Y. Wu, W. Wu, and T. Q. S. Quek, "Channel estimation based on deep learning in vehicle-to-everything environments," *IEEE Commun. Lett.*, vol. 25, no. 6, pp. 1891–1895, Jun. 2021.

[9] A. A. Tegos, S. A. Tegos, D. Tyrovolas, P. D. Diamantoulakis, P. Sarigiannidis, and G. K. Karagiannidis, "Breaking orthogonality in uplink with randomly deployed sources," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 566–582, 2024.

[10] Y. Teng, M. Liu, F. R. Yu, V. C. M. Leung, M. Song, and Y. Zhang, "Resource allocation for ultra-dense networks: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2134–2168, 3rd Quart., 2019.

[11] A. Slalmi, H. Chaibi, A. Chehri, R. Saadane, and G. Jeon, "Enabling massive IoT services in the future horizontal 6G network: From use cases to a flexible system architecture," *IEEE Internet Things Mag.*, vol. 6, no. 4, pp. 62–67, Dec. 2023.

[12] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, "Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks," *IEEE Access*, vol. 6, pp. 32328–32338, 2018.

[13] R. Shafin, L. Liu, V. Chandrasekhar, H. Chen, J. Reed, and J. C. Zhang, "Artificial intelligence-enabled cellular networks: A critical path to beyond-5G and 6G," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 212–217, Apr. 2020.

[14] H. Zhang, M. Feng, K. Long, G. K. Karagiannidis, and A. Nallanathan, "Artificial intelligence-based resource allocation in Ultradense networks: Applying event-triggered Q-learning algorithms," *IEEE Veh. Technol. Mag.*, vol. 14, no. 4, pp. 56–63, Dec. 2019.

[15] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[16] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[17] F. Li, D. Yu, H. Yang, J. Yu, H. Karl, and X. Cheng, "Multi-armed-bandit-based spectrum scheduling algorithms in wireless networks: A survey," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 24–30, Feb. 2020.

[18] T. Riihonen, S. Werner, and R. Wichman, "Hybrid full-duplex/half-duplex relaying with transmit power adaptation," *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, pp. 3074–3085, Sep. 2011.

[19] H. A. Suraweera, I. Krikidis, G. Zheng, C. Yuen, and P. J. Smith, "Low-complexity end-to-end performance optimization in MIMO full-duplex relay systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 913–927, Jan. 2014.

[20] N. H. Tran, L. Jiménez Rodríguez, and T. Le-Ngoc, "Optimal power control and error performance for full-duplex dual-hop AF relaying under residual self-interference," *IEEE Commun. Lett.*, vol. 19, no. 2, pp. 291–294, Feb. 2015.

[21] M. Elsayed, A. A. A. El-Banna, O. A. Dobre, W. Y. Shiu, and P. Wang, "Machine learning-based self-interference cancellation for full-duplex radio: Approaches, open challenges, and future research directions," *IEEE Open J. Veh. Technol.*, vol. 5, pp. 21–47, Nov. 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10314438

[22] N. Nomikos et al., "A survey on buffer-aided relay selection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1073–1097, 2nd Quart., 2016.
[23] I. Krikidis, T. Charalambous, and J. Thompson, "Buffer-aided relay selection for cooperative diversity systems without delay constraints," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1957–1967, May 2012.
[24] N. Zlatanov, D. Hranilovic, and J. S. Evans, "Buffer-aided relaying improves throughput of full-duplex relay networks with fixed-rate transmissions," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2446–2449, Dec. 2016.
[25] K. T. Phan and T. Le-Ngoc, "Power allocation for buffer-aided full-duplex relaying with imperfect self-interference cancellation and statistical delay constraint," *IEEE Access*, vol. 4, pp. 3961–3974, 2016.
[26] N. Nomikos, T. Charalambous, D. Vouyioukas, R. Wichman, and G. K. Karagiannidis, "Power adaptation in buffer-aided full-duplex relay networks with statistical CSI," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7846–7850, Aug. 2018.
[27] D. Della Penda, N. Nomikos, T. Charalambous, and M. Johansson, "Minimum power scheduling under Rician fading in full-duplex relay-assisted D2D communication," in *Proc. IEEE Globecom Workshops (GC)*, 2017, pp. 1–6.
[28] F. D. Calabrese, L. Wang, E. Ghadimi, G. Peters, L. Hanzo, and P. Soldati, "Learning radio resource management in RANs: Framework, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 138–145, Sep. 2018.
[29] N. Nomikos, T. Charalambous, and R. Wichman, "Bandit-based power control in full-duplex cooperative relay networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2021, pp. 1–6.
[30] N. Nomikos, M. S. Talebi, T. Charalambous, and R. Wichman, "Bandit-based power control in full-duplex cooperative relay networks with strict-sense stationary and non-stationary wireless communication channels," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 366–378, 2022.
[31] N. Nomikos, T. Charalambous, and R. Wichman, "MABAMS: Multi-armed bandit-aided mode selection in cooperative buffer-aided relay networks," in *Proc. IEEE Globecom Workshops (GC)*, 2022, pp. 1230–1235.
[32] "NR; Multiplexing and channel coding; (Release 15), Version 15.2.0," 3GPP, Sophia Antipolis, France, Rep. TS 38.212, 2021.
[33] J. L. Vicario, A. Bel, J. A. Lopez-salcedo, and G. Seco, "Opportunistic relay selection with outdated CSI: Outage probability and diversity analysis," *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 2872–2876, Jun. 2009.
[34] T. Islam, D. S. Michalopoulos, R. Schober, and V. K. Bhargava, "Buffer-aided relaying with outdated CSI," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1979–1997, Mar. 2016.
[35] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2, pp. 235–256, May 2002.
[36] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2011, pp. 174–188.
[37] M. Vehkaperä, T. Riihonen, R. Wichman, and B. Xu, "Power allocation for balancing the effects of channel estimation error and pilot overhead in full-duplex decode-and-forward relaying," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2016, pp. 1–5.
[38] R. Wang, H. Mehrpouyan, M. Tao, and Y. Hua, "Channel estimation, carrier recovery, and data detection in the presence of phase noise in OFDM relay systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1186–1205, Feb. 2016.
[39] "Radio transmit power." Cisco. 2008. Accessed: May 22, 2022. [Online]. Available: https://www.cisco.com/c/en/us/td/docs/routers/access/wireless/software/guide/RadioTransmitPower.html.

NIKOLAOS NOMIKOS (Senior Member, IEEE) received the Diploma degree in electrical engineering and computer technology from the University of Patras, Greece, in 2009, and the M.Sc. and Ph.D. degrees from the Information and Communication Systems Engineering Department, University of the Aegean, Samos, Greece, in 2011 and 2014, respectively. He is currently a Senior Researcher with the Department of Ports Management and Shipping, National and Kapodistrian University of Athens. His research interest is focused on 6G communications, NOMA, and machine learning-aided wireless networks. He is an Associate Editor of *Frontiers in Communications and Networks* and a member of the IEEE Communications Society and the Technical Chamber of Greece.

THEMISTOKLIS CHARALAMBOUS (Senior Member, IEEE) received the Ph.D. degree from the Control Laboratory, Engineering Department, Cambridge University in 2009. Following his Ph.D., he joined the Human Robotics Group as a Research Associate with Imperial College London for an academic year from September 2009 to September 2010. From September 2010 to December 2011, he worked as a Visiting Lecturer with the Department of Electrical and Computer Engineering, University of Cyprus. From January 2012 and January 2015, he worked with the Division of Decision and Control, Department of Intelligent Systems, Royal Institute of Technology (KTH) as a Postdoctoral Researcher. From April 2015 to December 2016, he worked as a Postdoctoral Researcher in the Unit of Communication Systems with the Department of Electrical Engineering, Chalmers University of Technology. In January 2017, he joined the Department of Electrical Engineering and Automation, School of Electrical Engineering, Aalto University as a Tenure-Track Assistant Professor. In September 2018, he was awarded the Academy of Finland Research Fellowship and in July 2020, he was appointed as a Tenured Associate Professor. In September 2021, he joined the Department of Electrical and Computer Engineering, University of Cyprus as a Tenure-Track Assistant Professor and he remains associated with Aalto University as a Visiting Professor. Since April 2023, he has been also a Visiting Professor with the FinEst Centre for Smart Cities.

PANAGIOTIS TRAKADAS received the Dipl.-Ing. degree in electrical and computer engineering and the Ph.D. degree from the National Technical University of Athens. In the past, he was worked with Hellenic Aerospace Industry as a Senior Engineer, on the design of military wireless telecommunications systems, and the Hellenic Authority for Communications Security and Privacy, where he was holding the position of the Director of the Division for the Assurance of Infrastructures and Telecommunications Services Privacy. He is currently an Associate Professor with the National and Kapodistrian University of Athens. He has been actively involved in many EU FP7 and H2020 Research Projects. He has published more than 130 papers in magazines, journals, and conference proceedings. His research interests include the fields of wireless and mobile communications, wireless sensor networking, network function virtualization, and cloud computing. He is a Reviewer in several journals, including IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON ELECTROMAGNETIC COMPATIBILITY.

RISTO WICHMAN (Senior Member, IEEE) received the M.Sc. and D.Sc. (Tech.) degrees in digital signal processing from the Tampere University of Technology, Finland, in 1990 and 1995, respectively. From 1995 to 2001, he worked with Nokia Research Center as a Senior Research Engineer. In 2002, he joined the Department of Signal Processing and Acoustics, School of Electrical Engineering, Aalto University, Finland, where he has been a Full Professor since 2008. His research interest includes signal processing techniques for wireless communication systems.