Received 31 January 2022; accepted 17 February 2022. Date of publication 28 February 2022; date of current version 8 March 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3154292

Bandit-Based Power Control in Full-Duplex Cooperative Relay Networks With Strict-Sense Stationary and Non-Stationary Wireless Communication Channels

NIKOLAOS NOMIKOS[1] (Senior Member, IEEE), MOHAMMAD SADEGH TALEBI[2], THEMISTOKLIS CHARALAMBOUS[3] (Senior Member, IEEE), AND RISTO WICHMAN[3] (Member, IEEE)

[1] IRIDA Research Centre for Communication Technologies, University of Cyprus, Nicosia 1678, Cyprus

[2] Department of Computer Science, University of Copenhagen, 1165 Copenhagen, Denmark

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

CORRESPONDING AUTHOR: T. Charalambous (e-mail: themistoklis.charalambous@aalto.fi)

Preliminary results of this work have been published in [1] [DOI: 10.1109/ICC42927.2021.9501010].

ABSTRACT Full-duplex relaying is an enabling technique of sixth generation (6G) mobile networks, promising tremendous rate and spectral efficiency gains. In order to improve the performance of full-duplex communications, power control is a viable way of avoiding excessive loop interference at the relay. Unfortunately, power control requires channel state information of source-relay, relay-destination and loop interference channels, thus resulting in increased overheads. Aiming to offer a low-complexity alternative for power control in such networks, we adopt reward-based learning in the sense of multi-armed bandits. More specifically, we present bandit-based power control, relying on acknowledgements/negative-acknowledgements observations by the relay. Our distributed algorithms avoid channel state information acquisition and exchange, and can alleviate the impact of outdated channel state information. Two cases are examined regarding the channel statistics of the wireless network, namely, strict-sense stationary and non-stationary channels. For the latter, a sliding window approach is adopted to further improve the performance. Performance evaluation highlights a performance-complexity trade-off, compared to optimal power control with full channel knowledge and significant gains over cases considering channel estimation and feedback overheads, outdated channel knowledge, no power control and random power level selection. Finally, it is shown that the sliding-window bandit-based algorithm provides improved performance in non-stationary settings by efficiently adapting to abrupt changes of the wireless channels.

INDEX TERMS Full-duplex relaying, power control, reinforcement learning, multi-armed bandits, non-stationary wireless channels, outdated CSI, sliding-window, upper confidence bound policies.

I. INTRODUCTION

A. BACKGROUND

SIXTH generation (6G) mobile networks are envisioned to support dense topologies of small cells where coexisting user devices and machines will compete for wireless resources [2], [3]. As a result, the available radio spectrum is expected to get even more crowded and, hence, efficient spectral usage is critical. Towards tackling this issue, novel approaches departing from orthogonal temporal and spectral resource allocation are currently being developed; for example, interference coordination/mitigation mechanisms with multiple-input, multiple-output (MIMO) multi-antenna transceiver technologies [4]. Furthermore, tremendous rate gains are expected through the use of full-duplex (FD) transceivers, offering simultaneous transmission and reception on the same spectral and temporal resources and the use of multi-antenna deployments with increased antenna numbers [5]. Various antenna solutions and digital loop interference (LI) cancellation algorithms have shown the feasibility of FD relay communication with low-cost deployments in mobile networks [6]–[9].

In wireless systems, pilot-aided channel estimation enables receivers to estimate the wireless environment and facilitates signal detection at the cost of consuming radio-resources. Recently, in massive MIMO FD networks with simultaneous wireless information and power transfer (SWIPT), Xu *et al.* [10] have used the energy signals for both energy harvesting and channel estimation. However, in several cases, wireless networks are characterized by non-stationary channels, thus posing difficulties to channel estimation. In sub-6 GHz networks, Shi *et al.* [11] use pilots and interpolation schemes to acquire channel state information (CSI) and Careem and Dutta [12] appropriately adjust the modulation vectors to diminish channel impairments. In non-stationary environments, both studies integrate machine learning to capture the statistical channel properties. In industrial settings, Lu *et al.* [13] calculate the non-stationary Rician channel parameters through a non-data aided method, based on the Gaussian mixture model and iterative sub-component discrimination, achieving near-optimal estimation accuracy. In vehicle-to-everything (V2X) networks, Pan *et al.* [14] propose data pilot-aided (DPA) deep learning-based channel estimation, exploiting de-mapped data symbols as pilots. DPA is integrated with a long short-term memory network and a multi-layer perceptron network to obtain time-frequency correlation. Performance evaluation reveals improved performance over conventional DPA schemes in scenarios with fast time-varying channels, high modulation order and large packets.

In such complex networks, a significant amount of signaling and feedback messages is necessary for efficient operation, thus threatening the network's performance when centralized solutions are adopted [15]. Meanwhile, in recent years, the coupling of machine learning techniques and wireless communications has shown promising results for providing low-complexity coordination mechanisms (see, for example, [16]–[20] and references therein). Among the different machine learning categories, increased autonomy in wireless resource allocation can be achieved through reinforcement learning [21]. Reinforcement learning algorithms operate by using feedback on previously taken actions, adapting their behavior to the wireless environment. A popular reward-based class of learning algorithms is based on the multi-armed bandit (MAB) framework [22], [23]. MAB enables a player (user) to pick an action from a given set of actions, aiming to maximize her cumulative expected reward. As MAB allows for learning unknown environments during network deployment, it can be of great importance for distributed resource allocation, such as spectrum and power [24].

A popular method of enabling FD relay operation is related to the design of efficient power control mechanisms towards avoiding excessive LI and mitigating its malicious effect on the end-to-end rate. Riihonen *et al.* [25] presented opportunistic relay mode selection, switching between half-duplex (HD) and FD relaying. By exploiting instantaneous CSI knowledge at the relay, transmit power adaptation maximized the instantaneous and average spectral efficiency in the uplink and downlink. In MIMO FD relay networks, Suraweera *et al.* [26] investigated the performance gains of power allocation and transmit antenna selection under various cases of CSI availability. Through a simple power allocation mechanism, the zero diversity effect of using fixed power was surpassed. Then, Tran *et al.* [27] investigated optimal power allocation for improving the diversity of amplify-and-forward (AF) FD relaying. The closed-form expression of the derivative of the pairwise-error probability was derived and bisection was used to find the optimal power allocation, assuming that the relay had statistical knowledge of the source-relay ($\{S \rightarrow R\}$) channel, while the destination had full knowledge of the relay-destination ($\{R \rightarrow D\}$) channel. In settings where FD relays were equipped with buffers, statistical and instantaneous CSI availability was exploited to conduct power adaptation at both the source and the selected relay, in order to maximize the chances of LI cancellation or avoidance and improve the average throughput of the network [28], [29]. Finally, power adaptation in FD relay-aided device-to-device networks can lead to LI mitigation and improved coverage without compromising the end-to-end rate due to additional multi-hop transmissions [30]. Penda *et al.* proposed a joint relaying-operation selection and power-allocation scheme under Rician fading, selecting a set of wireless links to minimize the power consumption and provide success probability guarantees [31]. By using the concept of coherent-measure-of-risk from the field of finance, the nonconvexity of the outage probability constraints was overcome, and improved energy efficiency was achieved, relying only on statistical CSI.

B. CONTRIBUTIONS

Inspired by the increased density of forthcoming 6G networks and the CSI overheads of conventional power control, we aim at developing a low-complexity power control mechanism for FD relay networks. Towards this end, we invoke reinforcement learning and more specifically, MAB, an important framework of reward-based learning algorithms. The MAB framework is not new in the context of wireless communications; see, for example, [32], [33] in which it is reported that the MAB framework was adopted in several 5G cases aiming to overcome the complexity of network coordination through learning. Nevertheless, to the best of the authors' knowledge, its use has not been investigated before for the problem of power control in FD relay networks.

More specifically, an online transmit power selection policy in each time-slot is developed and modeled as a MAB game. Thus, in each time-slot, the relay observes the acknowledgement/negative-acknowledgement (ACK/NACK) message from the destination for the previous transmissions, as well as whether or not the receptions from the source were successful. At each time slot, the online policy chooses a power level as a function of past decisions and observations. It is guaranteed that most of the time (i.e., except for a number of slots sublinearly growing with time), the policy selects the power level offering the maximum end-to-end throughput. Our contributions are the following.

- A bandit-based power control (BB-PC) algorithm is proposed, relying on local observation by the relay of the received signal transmitted by the source and ACK/NACK feedback from the destination.
- Two cases for the wireless channel statistics are investigated; namely, strict-sense stationary and non-stationary channels. For the latter, the sliding window (SW) approach [34], [35] is adopted to extend BB-PC, in order to better adapt to abrupt changes of the wireless environment.
- BB-PC with various upper confidence bound (UCB) policies is evaluated in terms of outage probability, average throughput, and regret against optimal power control with CSI at the transmitter (CSIT) and the cases with outdated CSI, no power control and random power level selection.

From the performance comparisons, it is observed that BB-PC provides performance gains over random power selection and the case without power control. More importantly, in the majority of the considered scenarios, BB-PC outperforms optimal power control when channel estimation and feedback overheads are taken into consideration, while the impact of non-stationary wireless environments is efficiently mitigated from the sliding window approach. Meanwhile, complexity concerns and possible errors in the CSI acquisition and exchange process, as well as issues related to outdated CSI for power control are completely eliminated, as BB-PC relies on only 1-bit ACK/NACK packets.

C. STRUCTURE

The remainder of this paper is organized as follows. In Section II, we introduce the system model and the main assumptions. In Section III, we provide in detail, the MAB modeling of the power control process in FD relay networks. The proposed bandit-based power control for FD relaying is described in Section IV, while performance evaluation is provided in Section V. Finally, conclusions and future directions are given in Section VI.

II. SYSTEM MODEL AND PRELIMINARIES

A. SYSTEM MODEL

In this work, a two-hop cooperative network, comprising a source node S, a single destination D, and a single FD decode-and-forward (DF) relay R, is examined. The relay is equipped with two antennas and operates in the FD mode, resulting in simultaneous transmission and reception of signals. It is considered that direct transmissions from the source towards the destination are not possible, due to severe fading conditions and communication can only be established through the FD relay. Fig. 1 depicts an instance of the two-hop FD cooperative relay network. This simple setup is emblematic of a wide range of wireless communication applications.

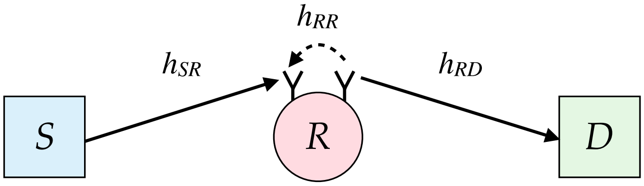

FIGURE 1. The two-hop relay-assisted topology where a source S communicates with a single destination D via a single FD relay R experiencing LI.

Time is assumed to be divided in time-slots, where source node S and relay R transmit using power levels P_S and P_R, respectively. In order to reduce the amount of LI from the output antenna of the relay to its input antenna, the relay is able to choose among different power levels. A saturated source is assumed, having always data for transmission, while the information rate is equal to r_0. Retransmissions rely on an ACK/NACK mechanism, where the receivers (either the activated relay or the destination) broadcast short-length packets, assumed error-free via a separate narrow-band link, informing the network on whether or not, the packet transmission was successful. Furthermore, it is assumed that the wireless channel quality is degraded by additive white Gaussian noise (AWGN) and frequency flat block fading. For simplicity, the power of the AWGN is assumed to be normalized with zero mean and unit variance. Also, since the relay operates in the FD mode, LI arises and h_{RR} denotes the instantaneous residual LI between the two antennas of relay R, following a complex Gaussian distribution mean 0 and variance $(0, \sigma^2_{RR})$.

Since the relay operates in the FD mode, the HD loss of conventional relays is surpassed and the destination receives one packet in each time-slot. Nonetheless, FD operation introduces LI and the power control algorithm must take into consideration the interference level arising by each transmit power level. In an arbitrary time-slot, a packet is successfully forwarded from relay R towards the destination D if the signal-to-noise Ratio (SNR), denoted by SNR_{RD}, is greater than or equal to a threshold γ_D, called the capture ratio, i.e.,

$$\frac{|h_{RD}|^2 P_R}{n_D} \geq \gamma_D, \tag{1}$$

where h_{RD} refers to the channel coefficient between the relay and the destination and n_D denotes thermal noise variance at the destination, which is considered to be AWGN.

A packet transmission from source S to relay R is successful, if the SINR at the relay, denoted by SINR_{SR} is greater than or equal to γ_R, i.e.,

$$\frac{|h_{SR}|^2 P_S}{|h_{RR}|^2 P_R + n_R} \geq \gamma_R. \tag{2}$$

where h_{SR}, h_{RR} refer to the channel coefficient between the source and the relay, and the relay's output and input antenna, respectively.

B. ESTIMATION AND FEEDBACK ERRORS

In general, the reliability of control channels is vital to wireless communication systems. For example, stringent quality of service requirements of ultra-reliable low latency communications (URLLC) service category in 5G new radio impose strict requirements for hybrid automatic repeat request (HARQ) processes. Decoding NACK erroneously as ACK introduces delay outage while the opposite error event causes redundant transmissions and waste of radio resources. The effect of the two error events is asymmetric and can be fine-tuned by false-alarm rate and the corresponding detection threshold to optimize the system performance.

3GPP TS 38.212, Multiplexing and channel coding, specifies several different ways to encode HARQ feedback with uplink control information using repetition coding, simplex coding, Polar coding, or Reed-Muller coding with variable coding rates and overheads [36]. Varying the encoding methods brings up several options to tune detection error and false alarm rates in fading channels according to the service requirements and radio propagation environments.

The effect of errors in a control channel on wireless systems is a multidimensional problem and warrants a study by itself. For simplicity, we assume that the errors in decoding ACK/NACK are negligible and can be ignored.

C. OUTDATED CSI

In practical systems, the CSIT used for the selection of a transmit power level is different from the one during the transmission in that link, because of the delays inherited by the feedback mechanism. In greater detail, outdated CSI might be available due to channel variations during the period from the end of channel estimation and the start of the transmission [37] or because it might not be constantly fed-back, towards avoiding excessive coordination overhead [38].

So, the case of outdated CSI is also considered and its effect on the relay's CSI-based power control is examined. In a system with CSI feedback delays, the actual channel response h_{ij} conditioned on the channel response $\hat{h}_{ij}$ that was estimated in the $\{i \rightarrow j\}$ link, during power control is given by [37]

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

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

D. CSI-BASED POWER ALLOCATION

If the power levels at the source and the relay can be jointly decided, in the case for which CSI is available, it is sufficient to find the minimum P_S and P_R such that inequalities (1) and (2) are satisfied with equality. In this case, the optimal power levels (P_S^*, P_R^*) are given by:

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

For allocating the optimal power levels, the source is required to know: the channel gain from the source to the relay, as well as that from the relay to itself, the optimal power of the relay, the thermal noise at the relay, and the decoding threshold at the relay. Furthermore, at the relay, only the channel gain from the relay to the destination is needed, apart from the thermal noise and the decoding threshold at the destination.

If the power level at the relay is fixed and only the power level at the source is optimized, then the source will need all the aforementioned information, except the power level of the relay, assuming that the fixed power level of the relay is known to the source. If, however, the power level at the source is fixed (as it is the case in the setup studied herein), then the minimum power P_R, denoted by $P_R^\dagger$, is given by:

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

provided

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

Eqs (5a)–(5b) signify that, for using the optimal power level (eq. (5a)) and for checking if the solution is feasible (eq. (5b)), the relay is required to know all three involved channels, the thermal noises at both the relay and the destination, the power level of the source, as well as the decoding thresholds γ_R and γ_D. Even if we assume that the thermal noises at both the relay and the destination, the power level of the source, and the decoding thresholds remain constant and are known, relay R needs to estimate the channels $|h_{SR}|^2$, $|h_{RR}|^2$, and $|h_{RD}|^2$. However, if the relay just chooses the optimal power $P_R^\dagger$ without checking whether it is feasible (in the case eq. (5b) does not hold, no relay power level P_R would be feasible), it only requires to know $|h_{RD}|^2$, γ_D, and n_D.

A discussion on the complexity of estimating the channel quality of the communication links is provided in Section V (Performance Evaluation).

III. MAB MODELING

A. THE MAB PROBLEM

MAB refers to a class of sequential decision problems of resource allocation among several competing entities in unknown environments with an exploration-exploitation trade-off, i.e., searching for a balance between exploring all possible decisions to learn their reward distributions while selecting the best decision more often to acquire more reward. For a detailed discussion on the topic, see, for example, [23], [39]. In the classical stochastic MAB problem, introduced by Robbins [40], a player has access to a finite set of arms, and to each arm j, a probability distribution with an initially unknown mean q_j is associated. At each round t, the player chooses an arm j and receives a random reward $r_{j,t}$ drawn from the distribution associated to arm j (whose mean is q_j). In our setup, the relay transmits with a power from a set of discrete power levels, $\mathcal{P}_R$. The number of power levels, $|\mathcal{P}_R|$, and their values depend on the radio configuration. Therefore, in the MAB framework, each arm corresponds to one of the $|\mathcal{P}_R|$ power levels.

The goal of the learner is to maximize the expected accumulated reward in the course of her interaction. If the reward distributions were known, this goal would have been achieved by always selecting the arm with the highest mean reward. To identify the optimal arm, the learner has to play various arms so as to learn their reward distributions (exploration) while ensuring that the gathered knowledge on reward distributions is exploited so that arms with higher expected rewards are preferred (exploitation). The performance of the learner in implementing such an *exploration-exploitation trade-off* is measured through the notion of *regret*, which compares the cumulative reward of the learner to that achieved by always selecting the optimal arm. It is defined as the difference between the reward achieved when the optimal arm is pulled and the player's choice. For our setup, the objective is to identify a policy over a finite time horizon T that maximizes the expected number of successfully transmitted packets or simply, what we call the throughput. Equivalently, we target the design of a sequential power control algorithm that minimizes the *regret*. The regret of a policy $\pi \in \Pi$ (Π being the set of all feasible policies) is defined as the performance loss and it is found by comparing the performance achieved under policy π to that of the best static policy, i.e.,

$$R^{\pi}(T) = \max_{\ell \in \mathcal{P}_R} \mathbb{E}\left\{\sum_{t=1}^{T} r_{\ell,t}\right\} - \mathbb{E}\left\{\sum_{t=1}^{T} r_{I_t^{\pi},t}\right\}, \tag{6}$$

where I_t^{π} denotes the chosen power level under policy π at time-slot t.

In their seminal paper, Lai and Robbins [22] characterize a problem-dependent lower bound on the regret of any adaptive policy (or algorithm), indicating that the lower bound grows logarithmically with time horizon T. More precisely, they show that for any *uniformly good* adaptive learning algorithm π,[1]

$$\liminf_{T\to\infty} \frac{R^{\pi}(T)}{\log(T)} \geq c(\boldsymbol{q}), \tag{7}$$

where $\boldsymbol{q}$ denotes the vector of mean rewards of various arms, and $c : [0,1]^{|\mathcal{P}_R|} \to \mathbb{R}_+$ is a deterministic and explicit function presented in [22].

B. UPPER CONFIDENCE BOUND POLICIES

A big class of policies for MAB problems, whose regret grows logarithmically over the time horizon, are based on the *optimism in the face of uncertainty* principle (or for short, the *optimistic* principle) proposed by Lai and Robbins [22]. The idea behind an *optimistic algorithm* is to replace the unknown mean rewards of each arm with a high-probability *Upper Confidence Bound (UCB)* on it. To further specify the generic form of an optimistic algorithm, let us first introduce some notations. In what follows, when the choice of the algorithm is clear from the context, we let I_t denote the arm selected at time t. Furthermore, we let $n_{j,t}$ denote the number of plays of arm j up to round t, i.e., $n_{j,t} := \sum_{s=1}^{t-1} \mathbb{1}_{\{I_s=j\}}$, where $\mathbb{1}_A$ denotes the indicator function of event A. We let $\hat{q}_{j,t}$ represent the empirical average reward of arm j built using the observations from j up to t:

$$\hat{q}_{j,t} = \frac{1}{n_{j,t}} \sum_{s=1}^{t-1} r_{j,s} \mathbb{1}_{\{I_s=j\}}, \tag{8}$$

where $r_{j,s}$ is the reward of arm j at round s.

An optimistic algorithm π maintains an index function $\bar{q}_{j,t}$ for each arm j, which depends only on the past observations of j, and satisfies: $\bar{q}_{j,t} \geq q_j$ with high probability for all $t \geq 1$. Then, π simply consists in selecting the arm with the largest index $\bar{q}_{j,t}$ at each round t:

$$I_t = \arg\max_{j \in \mathcal{P}_R} \bar{q}_{j,t}. \tag{9}$$

In the sequel, we briefly introduce some popular index policies for stochastic MABs. In the rest of this section, we assume that the reward realizations of arm j belong to the interval $[0, 1]$ almost surely.

1) UCB1 [41]

UCB1 is an index policy designed based on Hoeffding's concentration inequality for bounded random variables. The UCB1 index function (or for short, UCB) is defined as follows:

$$\bar{q}_{j,t}^{\mathrm{UCB}} = \widehat{q}_{j,t} + \sqrt{\frac{3\log(t)}{2n_{j,t}}}. \tag{10}$$

2) KL-UCB [42]

KL-UCB is an index policy designed based on a novel concentration inequality for bounded random variables, and relies on the following index function:

$$\bar{q}_{j,t}^{\mathrm{KL\text{-}UCB}} = \sup\left\{\lambda \in \left[\widehat{q}_{j,t}, 1\right] : \mathrm{kl}\left(\widehat{q}_{j,t}, \lambda\right) \leq \frac{\log(t) + 3\log(\log(t))}{n_{j,t}}\right\}, \tag{11}$$

where $\mathrm{kl}(x, y)$ is the Kullback-Leibler divergence between two Bernoulli distributions with means x and $y : \mathrm{kl}(x,y) := x\log(\frac{x}{y}) + (1-x)\log(\frac{1-x}{1-y})$. When the reward distribution

[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, $n_{i,t}^{\pi}$, satisfies: $\mathbb{E}[n_{i,t}^{\pi}] = o(t^{\alpha})$, for all $\alpha > 0$.

of arms are Bernoulli distributions, KL-UCB achieves the problem-dependent lower bound (7), and is hence said to be *asymptotically optimal*.[2] We remark that computing $\overline{q}_{j,t}^{\texttt{KL-UCB}}$ corresponds to finding the roots of a strictly convex and increasing function.[3] Therefore, $\overline{q}_{j,t}^{\texttt{KL-UCB}}$ can be computed using simple line search methods, such as bisection.

3) KL-UCB^{++} [43]

KL-UCB^{++} is a variant of KL-UCB, which enjoys both asymptotic and minimax optimality in stochastic MABs simultaneously. It relies on the following index function:

$$\overline{q}_{j,t}^{\texttt{KL-UCB}^{++}} = \sup\left\{\lambda \in \left[\widehat{q}_{j,t}, 1\right] : \mathrm{kl}\left(\widehat{q}_{j,t}, \lambda\right) \le g(n_{j,t})/n_{j,t}\right\}, \tag{12}$$

where

$$g\left(n_{j,t}\right) = \log_{+}\left(\frac{t}{|\mathcal{P}_R| n_{j,t}}\left(\log_{+}^{2}\left(\frac{t}{|\mathcal{P}_R| n_{j,t}}\right) + 1\right)\right),$$

with $\log_{+}(x) = \max(\log(x), 0)$.

IV. BANDIT-BASED POWER CONTROL

A. ONLINE LEARNING MODEL

We now turn to model the power control problem as a MAB. Each power level corresponds to an arm, and pulling an arm leads to a packet transmission using the selected power level. More formally, if power level j is selected in time-slot t, a reward $r_{j,t}$ is obtained, where

$$r_{j,t} = \begin{cases} 1, & \text{if packet received successfully,} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The power level selection yields a random reward from an unknown joint probability distribution, which corresponds to the links (i.e., links $\{S \to R_j\}$, $\{R \to R\}$, and $\{R_j \to D\}$). In other words, pulling arm j at round t results in an end-to-end packet transmission via relay R. If the packet is successfully received by D, a reward $r_{j,t} = 1$ is obtained. If an outage occurs, no reward is obtained.

Hence, the sequence $(r_{j,t})_{t \ge 1}$ of rewards of power level j follows a Bernoulli distribution, whose mean corresponds to the probability of successful transmission using j.

We consider two scenarios depending on whether the probabilities of successful transmission evolve over time or not:

1) CASE 1: STRICT-SENSE STATIONARY CHANNELS (HENCE, FIXED) SUCCESS PROBABILITIES

In this case, success probabilities of the *SR* and *RD* channels are assumed to be fixed but unknown. Hence, for each j, $(r_{j,t})_{t \ge 1}$ 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 set of power levels chosen by the algorithm before round t, and their realized rewards.

Algorithm 1: Bandit-Based Power Control (BB-PC) Mechanism

```
Input: Set of power levels $\mathcal{P}_R$, capture ratios $\gamma_R$ and $\gamma_D$.
for $t = 1, 2, \dots$ do
    compute $\hat{q}_{j,t}$ (8) and then $\bar{q}_{j,t}$ according to UCB used
    select power level $j$ for transmission at time-slot $t$ using (9)
    receive packet from $S$ and transmit it to $D$
    $n_{j,t+1} \leftarrow n_{j,t} + \mathbb{1}_{\{I_t = j\}}$ for all $j$
    if reception and transmission are successful then
        $r_{I_t,t} = 1$
    end
end
```

2) CASE 2: NON-STATIONARY CHANNELS (HENCE, TIME-VARYING) SUCCESS PROBABILITIES

This case corresponds to a system, where channel statistics may change over time. Here, we consider an *abruptly-changing* environment, in which the success probabilities may undergo *abrupt changes* over time. Specifically, $(r_{j,t})_{t \ge 1}$ is a sequence of independent Bernoulli random variables with $\mathbb{E}[r_{j,t}|\mathcal{F}_{t-1}] = q_{j,t}$ for all t. It is worth remarking that this implies that the optimal arm may also change over time, so in the definition of regret in (6), the maximizer in the first term changes over time. In other words, the optimal arm (power level) is not fixed and could change over time. Following the terminology used in the literature on non-stationary MABs, we refer to time instants at which such abrupt changes occur as *breakpoints* [34]. It is also assumed that breakpoints occur independently of the channel selection strategy or of the sequence of rewards. We denote the number of breakpoints before time T by Υ_T. We also assume that Υ_T grows sublinear with T: $\Upsilon_T = o(T)$; otherwise, there is no hope that one could learn the changing optimal power level, and hence, achieve a sublinear regret.

B. ONLINE LEARNING ALGORITHMS

We are now ready to describe our bandit-based power control algorithms for strict-sense stationary and non-stationary wireless channels. We first present an algorithm for strict-sense stationary channels.

1) STRICT-SENSE STATIONARY CHANNELS

For strict-sense stationary wireless environments, we present Bandit-Based Power Control (BB-PC), whose pseudo-code is presented in Algorithm 1.

BB-PC follows the optimistic principle and relies on a generic index function. In other words, for any choice of index function presented in Section III-B, we have a variant of BB-PC. For a given choice of index function, in each round t, the algorithm first computes the empirical estimate of each j, $\widehat{q}_{j,t}$, using (8). It then computes the UCB for each power level j, and selects the one with the largest UCB (ties are broken arbitrarily), denoted by I_t:

$$I_t \in \arg\max_{j \in \mathcal{P}_R} \bar{q}_{j,t}.$$

[2] Indeed KL-UCB is shown to be asymptotically optimal for a wider class of MABs whose reward distributions are taken within one-parameter exponential families, provided that one replaces the Kullback-Leibler divergence of Bernoulli distributions with an appropriate divergence.

[3] Note that $v \mapsto \mathrm{kl}(u, v)$ is strictly convex and increasing for $v \ge u$.

Then, a packet transmission from S to D will occur using power level I_t. Upon a successful transmission and reception, a reward of 1 is collected, $r_{I_t,t} = 1$. Otherwise, $r_{I_t,t} = 0$.

2) NON-STATIONARY CHANNELS

We now consider the case of non-stationary wireless environments with abrupt changes of channel statistics. Algorithms for strict-sense stationary environments, which assume fixed success probabilities, can incur a linear regret in such environments. To achieve a sublinear regret, one should use an algorithm tailored to the non-stationary nature of the environment.

In non-stationary environments, using the empirical estimate $\widehat{q}_{j,t}$ in (8) would lead to a biased and inaccurate estimation of $q_{j,t}$. One remedy to this issue is to use an estimator tailored to the time-varying nature of the environment. One prominent estimator widely used in the literature on non-stationary MAB is the one constructed using a *sliding-window (SW)* approach. Such an SW estimator uses only the rewards collected within a sliding-window of observations. This is done, e.g., in the SW-UCB algorithm of [34] and in the SW-UCB# algorithm of [35]. Precisely speaking, we introduce

$$\mathcal{T}_{t,\theta} = \{t-\theta, t-\theta+1, \ldots, t-1\},$$

as a SW of width θ at time t. (The choice of θ will be discussed later.) We then build an estimate of $q_{j,t}$ using ACKs/NACKs received within $\mathcal{T}_{t,\theta}$ as follows:

$$\hat{q}_{j,t,\mathcal{T}_{t,\theta}} = \frac{1}{|\mathcal{T}_{t,\theta}|} \sum_{s \in \mathcal{T}_{t,\theta}} r_{j,s} \mathbb{1}_{\{I_s = j\}}. \tag{14}$$

Contrasting $\hat{q}_{j,t,\mathcal{T}_{t,\theta}}$ to $\hat{q}_{j,t}$ in (8), one can observe that the former discards observations collected prior to the SW, as it is hypothesized that they likely come from a distribution with a different mean. The ideal situation happens when a given SW $\mathcal{T}_{t,\theta}$ contains no breakpoint, in which case all the observations collected during $\mathcal{T}_{t,\theta}$ come from the same distribution and thus, $\hat{q}_{j,t,\mathcal{T}_{t,\theta}}$ accurately estimates $q_{j,t}$. However, as breakpoints are not known *a priori*, $\hat{q}_{j,t,\mathcal{T}_{t,\theta}}$ may have some bias.

If the width of $\mathcal{T}_{t,\theta}$, θ, is too small, there is a high chance that $\mathcal{T}_{t,\theta}$ contains no breakpoint, and hence, the observations come from the same distribution. But $\hat{q}_{j,t,\mathcal{T}_{t,\theta}}$ becomes sample-inefficient due to ignoring many samples, and thus, inaccurate. On the other hand, if θ is chosen too large, then there is a high chance that $\mathcal{T}_{t,\theta}$ contains some breakpoint(s), and therefore, $\hat{q}_{j,t,\mathcal{T}_{t,\theta}}$ may inaccurately estimate $q_{j,t}$. Therefore, there is a trade-off to choose θ.

In this paper, similarly to [35], we set θ as

$$\theta := \theta(\alpha, \mu, t) := \min\{\lceil \mu t^{\alpha} \rceil, t\},$$

where $\alpha \in (0, 1]$ and $\mu \geq 0$ are input parameters that control θ. In order to guarantee a sublinear regret, one must choose α in accordance to the frequency of breakpoints. It is shown that when the number of breakpoints grows as $\Upsilon_T = O(T^{\upsilon})$, for some $\upsilon \in [0, 1)$ known in advance, then the best choice is $\alpha = \frac{1-\upsilon}{2}$. Now, using the estimator for $q_{j,t}$ based on the SW approach, the SW-UCB# index is defined as:

$$\bar{q}^{\texttt{SW-UCB\#}}_{j,t,\mathcal{T}_{t,\theta}} = \widehat{q}_{j,t,\mathcal{T}_{t,\theta}} + \sqrt{\frac{(1+\alpha)\log(t)}{n_{j,t}(\theta)}}, \tag{15}$$

where $n_{j,t}(\theta)$ denotes the number of times j is selected in the SW $\mathcal{T}_{t,\theta}$, i.e., $n_{j,t}(\theta) = \sum_{s \in \mathcal{T}_{t,\theta}} \mathbb{1}_{\{I_s = j\}}$.

Algorithm 2: Sliding-Time Window Bandit-Based Power Control (SW-BB-PC) Mechanism

```
Input: Set of power levels $\mathcal{P}_R$, capture ratios $\gamma_R$ and $\gamma_D$,
       sliding-time window parameters $\alpha$ and $\mu$.
for $t = 1, 2, \ldots$ do
    compute the sliding-time window width
     $\theta = \min\{\lceil \mu t^{\alpha} \rceil, t\}$
    compute $\hat{q}_{j,t,\mathcal{T}_{t,\theta}}$ (14) and then $\bar{q}^{\texttt{SW-UCB\#}}_{j,t,\mathcal{T}_{t,\theta}}$ (15)
    select power level $j$ for transmission at time-slot $t$ using
     (9)
    receive packet from $S$ and transmit it to $D$
    $n_{j,t+1} \leftarrow n_{j,t} + \mathbb{1}_{\{I_t = j\}}$ for all $j$
    if reception and transmission are successful then
        $r_{j,t} = 1$
    end
end
```

The SW-based BB-PC in non-stationary wireless environments, namely SW-BB-PC, is given in Algorithm 2. The algorithm proceeds quite similarly to BB-PC, with two exceptions: (i) it uses a SW-based estimate of success probabilities; and (ii) it requires additional input parameters α and μ to determine the SW width θ. This entails some prior knowledge on the frequency of breakpoints.

Remark 1: It is often the case that we are not able to know *a priori* whether the channels are going to be strict-sense stationary or not. So, it is better to adopt the non-stationary approach in order to avoid severe performance degradation in case the channel is non-stationary. The performance degradation obtained by erroneously assuming that the channels are non-stationary is negligible (if any), as shown in Fig. 10 in Section V, for the period of time that the channel is strict-sense stationary. When there is an abrupt change in the channel quality, then the performance superiority of SW-BB-PC over BB-PC is exemplified.

V. PERFORMANCE EVALUATION

In this section, BB-PC is evaluated, in terms of outage probability, average throughput and total accumulated regret over time. Two different BB-PC versions, based on UCB1 [41] and kl-UCB^{++} [43] are compared against CSI-based optimal power control (opt), optimal power control with channel estimation and feedback overheads, accounting for 10% or 20% of a time-slot's duration, power control with outdated CSI, characterized by $\rho = 0.8$, no power control (no-PC) and random power level selection (rnd). In each link, the transmit SNR ranges from 0 dB to 40 dB and it represents the ratio

TABLE 1. Simulation parameters.

Parameter	Value
No. of transmissions per SNR value	10^4
Transmission rate r_0	3 bps/Hz
CSIT overheads per transmission $1 - A$	{10, 20} %
Outdated CSI coefficient ρ	0.8
Average LI channel SNR $\bar{\gamma}_{LI}$	{-30, -10} dB
Transmit SNR $P_{\max}/n_D$ range	{0, 40} dB
No. of relay power levels	6
Wireless channel types	Rayleigh, Rician
Rician factor K_{Rice}	10 dB
Stationarity breakpoints	1 (at $t = 5000$)
Sliding window parameter μ	15
Sliding window parameter α	0.2

of the maximum available transmit power at each transmitter, assuming $P_{S,\max} = P_{R,\max} = P_{\max}$ over the noise power. Furthermore, in the outage and throughput results, the x-axis corresponds to the transmit SNR in the $\{R \to D\}$ link, being equal to $P_{\max}/n_D$. For each transmit SNR value, 10^4 transmissions are performed over which, the results are averaged. Moreover, a fixed transmission rate $r_0 = 3$ bps/Hz is considered in a topology with a single relay node being able to select among six different transmit power levels, i.e., $P_1 = P_{\max}$, $P_2 = 0.50P_{\max}$, $P_3 = 0.30P_{\max}$, $P_4 = 0.20P_{\max}$, $P_5 = 0.05P_{\max}$, $P_6 = 0.01P_{\max}$ [44]. Outages occur when the relay or the destination fail to perform a reception with the pre-determined rate r_0.

Regarding the wireless setting, two cases are examined. In the first case, strict-sense stationary wireless channels are considered whose statistics remain the same for the whole transmission duration. Furthermore, i.i.d. $\{S \to R\}$ and $\{R \to D\}$ channels with average channel SNR $\bar{\gamma}_{\{S\to R\}} = \bar{\gamma}_{\{R\to D\}} = 0$ dB are assumed. Also, two LI cases are considered with average channel SNR $\bar{\gamma}_{LI}$ taking values from the set $\{-30, -10\}$ dB.

The second case represents a non-stationary wireless environment where the $\{R \to D\}$ channel statistics abruptly change at one breakpoint ($t = 5000$). Initially, the $\{R \to D\}$ link enjoys line-of-sight (LoS) conditions with Rician fading, characterized by a Rician factor $K_{Rice} = 10$ dB. Meanwhile, the $\{S \to R\}$ link experiences non-LoS conditions with Rayleigh fading, characterized by $\bar{\gamma}_{\{S\to R\}} = 0$ dB. After the breakpoint, the $\{R \to D\}$ link reverts to Rayleigh fading with $\bar{\gamma}_{\{R\to D\}} = -10$ dB while the SR fading conditions do not change, i.e., $\bar{\gamma}_{\{S\to R\}} = 0$ dB. Here, the LI channel is characterized by an average SNR $\bar{\gamma}_{LI}$ taking values from the set $\{-30, -10\}$ dB. Finally, SW-based BB-PC, namely SW-BB-PC is evaluated with parameters $\mu = 15$ and $\alpha = 0.2$.

Table 1 lists the simulation parameters that are considered in the performance comparisons.

A. STRICT-SENSE STATIONARY CASE

Fig. 2 depicts the outage probability performance for different power control algorithms under a weak LI channel, characterized by $\bar{\gamma}_{LI} = -30$ dB. Here, it is clear that when full CSI is available, the optimal power control algorithm has the best outage performance. Also, after 28 dB, kl-UCB^{++} has an advantage over UCB1, while both BB-PC versions avoid the outage floor of the case without power control, as the LI impact is mitigated. Meanwhile, BB-PC significantly outperforms the case where transmit power is randomly selected.

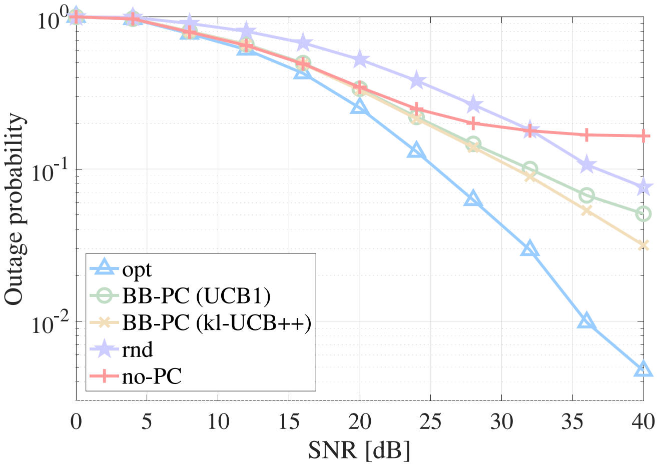

FIGURE 2. Outage probability comparisons for different power control algorithms and $\bar{\gamma}_{LI} = -30$ dB (strict-sense stationary case).

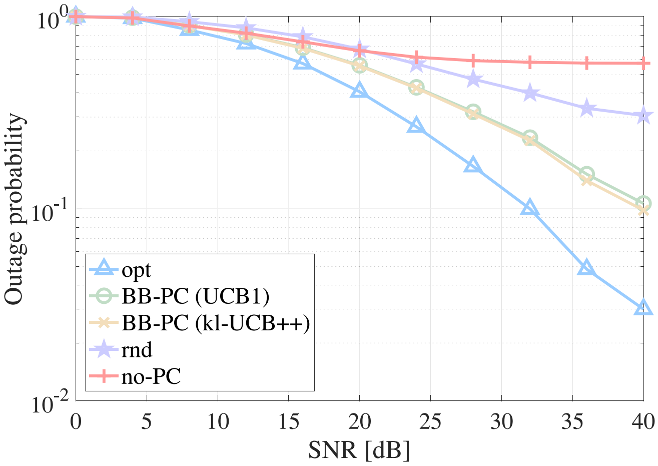

FIGURE 3. Outage probability comparisons for different power control algorithms and $\bar{\gamma}_{LI} = -10$ dB (strict-sense stationary case).

The second outage comparison is presented in Fig. 3 when $\bar{\gamma}_{LI} = -10$ dB. Here, the necessity for power control is clearly shown, as it is revealed by the performance of the case without power control at the relay, experiencing an outage floor after 20 dB. Meanwhile, random power selection cannot provide a satisfactory outage performance throughout the SNR range. On the contrary, both BB-PC algorithms offer improved outage performance and avoid an outage floor, whereas optimal power control with full CSI exhibits the best outage performance at the cost of increased overheads.

Average throughput comparisons under a weak LI channel, characterized by $\bar{\gamma}_{LI} = -30$ dB are shown in Fig. 4. It can be seen that CSI-based optimal power control provides the throughput upper bound, being closely followed by the

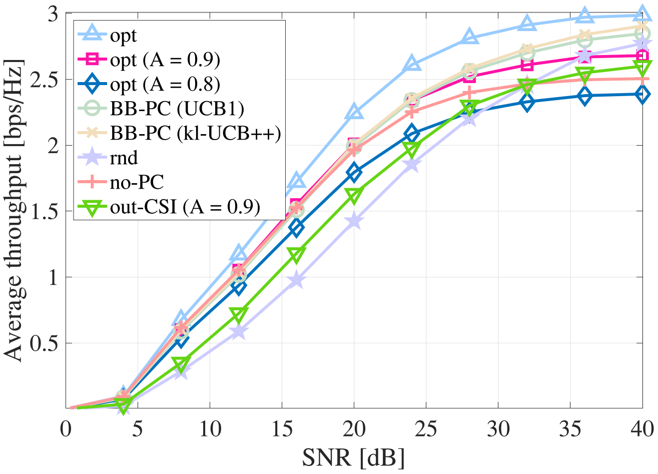

FIGURE 4. Average throughput comparisons for different power control algorithms and $\bar{\gamma}_{LI} = -30$ dB (strict-sense stationary case).

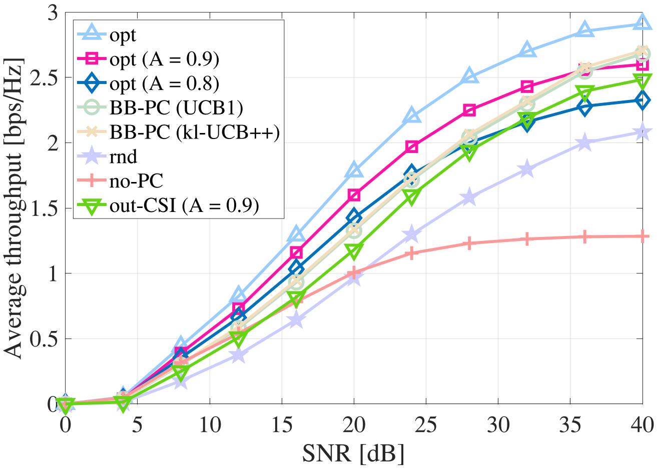

FIGURE 5. Average throughput comparisons for different power control algorithms and $\bar{\gamma}_{LI} = -10$ dB (strict-sense stationary case).

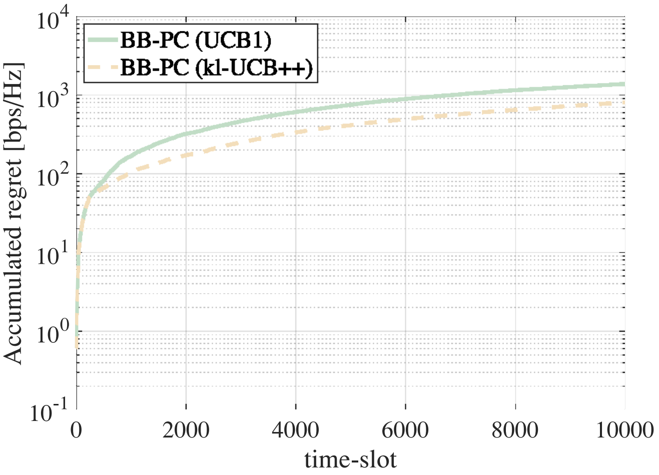

FIGURE 6. Total accumulated regret over time for the two BB-PC versions, $\bar{\gamma}_{LI} = -30$ dB for a transmit SNR in the $\{R \rightarrow D\}$ link equal to 40 dB (strict-sense stationary case).

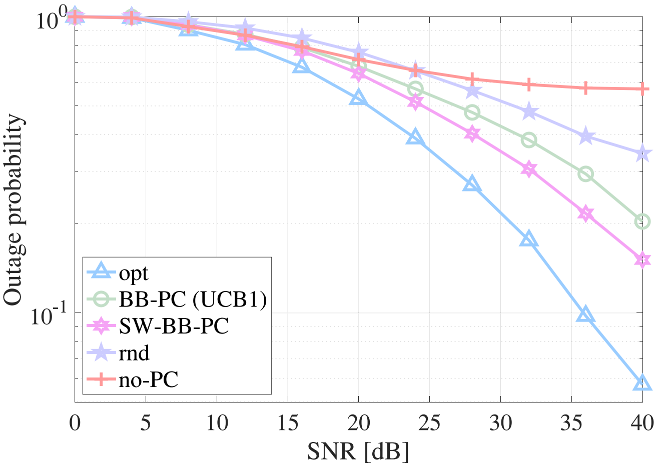

FIGURE 7. Outage probability comparisons for different power control algorithms and $\bar{\gamma}_{LI} = -10$ dB (the non-stationary case).

BB-PC algorithms. However, BB-PC outperforms optimal power control with CSI overheads when $A = 0.8$ and they even manage to surpass CSI-based power control when $A = 0.9$ for very high SNR values. In addition, CSI-based power control with outdated CSI and $A = 0.9$ (out-CSI) exhibits a large performance gap against BB-PC. Among the two BB-PC algorithms, kl-UCB^{++} offers a noticeable performance gain after 28 dB. Furthermore, when power control is not employed, significantly worse average throughput is obtained, while random transmit power selection has the worst performance, until 32 dB.

Then, average throughput results are illustrated in Fig. 5 when $\bar{\gamma}_{LI} = -10$ dB. Here, the LI channel power is not negligible and power control is vital to maintain adequate throughput performance. When CSI overheads are not considered, CSI-based power control reaches the throughput upper-bound for high SNR values. Nonetheless, when practical considerations are made, both BB-PC versions outperform CSI-based power control after 24 dB when $A = 0.8$ and 36 dB when $A = 0.9$. Additionally, the outdated CSI case with $A = 0.9$ falls behind BB-PC throughout the SNR range. Finally, the case with fixed transmit power (no-PC) has by far, the worst performance while better results are observed through random power selection after 20 dB.

An important performance metric for all bandit-based algorithms is related to the accumulated regret over time, consisting of 10000 time-slots. Fig. 6 presents the accumulated regret, in terms of throughput for the two BB-PC algorithms. It is clear that kl-UCB^{++} experiences less accumulated regret compared to UCB1, thus revealing that it converges faster to the optimal transmit power level.

B. NON-STATIONARY CASE

The outage probability for the non-stationary case when $\bar{\gamma}_{LI} = -10$ dB is depicted in Fig. 7. Here, the increased LI channel power degrades reliability of the transmission. Still, as transmit SNR increases, BB-PC efficiently determines which power level should be employed. Furthermore, SW-BB-PC outperforms UCB1 by better adapting to the

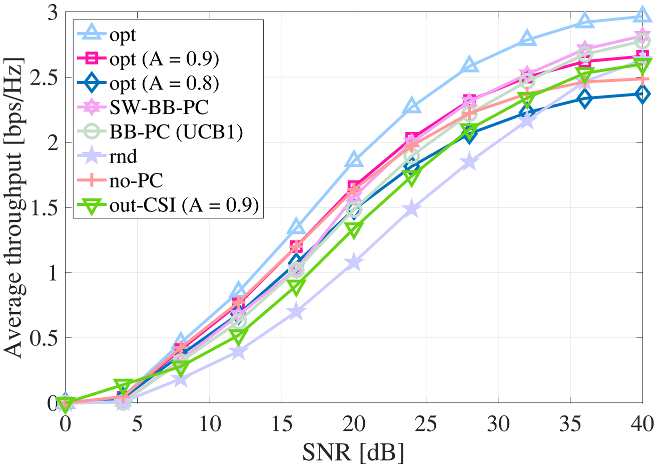

FIGURE 8. Average throughput comparisons for different power control algorithms and $\bar{\gamma}_{LI} = -30$ dB (the non-stationary case).

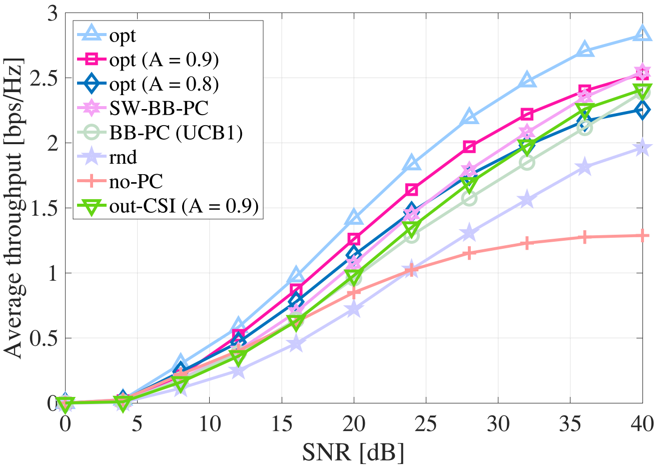

FIGURE 9. Average throughput comparisons for different power control algorithms and $\bar{\gamma}_{LI} = -10$ dB (the non-stationary case).

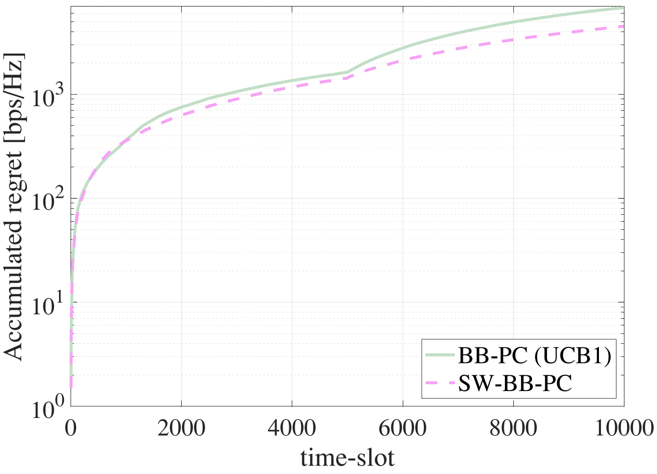

FIGURE 10. Total accumulated regret over time for the two BB-PC versions, $\bar{\gamma}_{LI} = -10$ dB for a transmit SNR in the $\{R \rightarrow D\}$ link equal to 30 dB (non-stationary case).

non-stationary wireless environment when the $\{R \rightarrow D\}$ fading conditions change from LoS to non-LoS after 5000 time-slots. At the same time, the random power selection and fixed transmit power cases have a large performance gap, outlining the importance of power control when high residual LI remains and fading conditions abruptly change.

Then, Fig. 8 shows average throughput comparisons under weak LI, i.e., $\bar{\gamma}_{LI} = -30$ dB and non-stationary wireless channels. In this comparison, it is observed that BB-PC surpass the performance of optimal power control for transmit SNR values above 16 dB, when $A = 0.8$ and after 32 dB when $A = 0.9$. Also, both BB-PC algorithms offer high performance gains over CSI-based power control with outdated CSI and $A = 0.9$. Moreover, as low LI levels are assumed, the case without power control is not severely affected and maintains satisfactory performance until 24 dB. Finally, SW-BB-PC exhibits improved throughput performance over UCB1 after 16 dB by adapting to the abruptly changing wireless conditions and determining the relay's transmit power with ACKs/NACKs from smaller time periods.

Another throughput comparison for the non-stationary case is included in Fig. 9 when $\bar{\gamma}_{LI} = -10$ dB. In this case, SW-BB-PC provides improved throughput over CSI-based power control with channel knowledge acquisition and exchange overheads when $A = 0.8$ after 28 dB and only matches the performance of the case with $A = 0.9$ at 40 dB. Thus, a trade-off arises between BB-PC and CSI-based solutions when higher LI levels exist in the network. Meanwhile, SW-BB-PC provides superior performance over CSI-based power control with outdated CSI and $A = 0.9$, showing that it can efficiently alleviate the impact of CSI feedback delays. Among the two bandit-based versions, SW-BB-PC has the edge over the non-SW alternative, especially after 20 dB. Finally, the bandit-based algorithms offer higher throughput than the random power selection and fixed transmit power cases.

In order to better illustrate the performance improvement provided by SW-BB-PC when compared to the regular UCB1 BB-PC, Fig. 10 shows the accumulated regret over time, compared to CSI-based optimal power control, when $\bar{\gamma}_{LI} = -10$ dB and the transmit SNR in the $\{R \rightarrow D\}$ link is equal to 30 dB. It can be seen that before the breakpoint, the two bandit-based algorithms exhibit almost the same performance in terms of regret. Then, after $t = 5000$ time-slots, SW-BB-PC maintains lower regret compared to UCB1 and their gap increases. Thus, the efficiency of adopting a dynamic SW is revealed, as SW-BB-PC is able to select a more appropriate power level compared to BB-PC without SW when $\{R \rightarrow D\}$ fading conditions abruptly change from LoS to non-LoS.

VI. CONCLUSION

A. CONCLUSION

Power control is an important technique to guarantee the performance of full-duplex cooperative relaying. However, the selection of an appropriate power level entails significant coordination overheads as full channel state information must be acquired. Targeting to reduce the complexity of this process and provide autonomous network operation, we have adopted the MAB framework on a stochastic wireless setting and developed relevant power control algorithms. The learning process was based only on ACK/NACK observations, adjusting the relay's transmit power in such a way, so as to reduce the impact of loop interference and ensure increased throughput. Furthermore, aiming to highlight the efficacy of bandit-based power control in different wireless settings, both strict-sense stationary and non-stationary channels were considered and for the latter, a sliding-window approach was adopted, enabling improved power control in abruptly changing wireless environments. Performance evaluation showed that the proposed algorithms closely followed the optimal power control with full channel state information for different cases of loop interference severity, while providing significant gains over CSI-based power control when channel acquisition and exchange overheads were considered. More importantly, when compared against practical CSI-based power control, considering the effect of overheads and outdated CSI, our bandit-based solutions exhibited high performance gains.

B. FUTURE DIRECTIONS

In this work, we assumed that the channel conditions for different power levels are independent, i.e., no inference for other power levels is made based on the outcome of the trial for a specific power level. In fact they are not, and our method, by assuming this independence, provides the worst case scenario, and converging to the optimal power is slower than what it could be. Part of ongoing work focuses on exploiting the correlation between the outcomes of the trials for different power levels.

Also, the proposed MAB-based framework can be applied in a variety of wireless communication areas. The investigation of a multi-relay setup can provide further performance gains to full-duplex transmissions while avoiding excessive coordination overheads through distributed timer-based coordination [45]. In addition, multi-antenna topologies can provide further performance gains by increasing the diversity of the transmission, as long as CSI overheads are efficiently tackled. Here, bandit-based solutions can facilitate the operation of multi-antenna networks by selecting not only the optimal transmit power level but also, designing appropriate beamforming vectors towards improved self-interference cancellation in the spatial domain.

At the same time, the consideration of non-stationary channels represents a more practical setting for mobile relay networks and devising efficient practical learning algorithms should be prioritized [46]. More specifically, the consideration of transmitting using the millimeter waveband represents an important case of 6G communication and sliding-window-based learning, as well as other MAB-based approaches promise improved performance [47], [48].

REFERENCES

[1] N. Nomikos, T. Charalambous, and R. Wichman, "Bandit-based power control in full-duplex cooperative relay networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Montreal, QC, Canada, 2021, pp. 1–6.

[2] Y. Jiang *et al.*, "Toward URLLC: A full duplex relay system with self-interference utilization or cancellation," *IEEE Wireless Commun.*, vol. 28, no. 1, pp. 74–81, Feb. 2021.

[3] T. O. Olwal, K. Djouani, and A. M. Kurien, "A survey of resource management toward 5G radio access networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1656–1686, 3rd Quart., 2016.

[4] Z. Ding, I. Krikidis, B. Rong, J. S. Thompson, C. Wang, and S. Yang, "On combating the half-duplex constraint in modern cooperative networks: Protocols and techniques," *IEEE Wireless Commun.*, vol. 19, no. 6, pp. 20–27, Dec. 2012.

[5] X. Xia, K. Xu, Y. Wang, and Y. Xu, "A 5G-enabling technology: Benefits, feasibility, and limitations of in-band full-duplex mMIMO," *IEEE Veh. Technol. Mag.*, vol. 13, no. 3, pp. 81–90, Sep. 2018.

[6] M. Heino *et al.*, "Recent advances in antenna design and interference cancellation algorithms for in-band full duplex relays," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 91–101, May 2015.

[7] Z. Zhang, X. Chai, K. Long, A. V. Vasilakos, and L. Hanzo, "Full duplex techniques for 5G networks: Self-interference cancellation, protocol design, and relay selection," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 128–137, May 2015.

[8] 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.

[9] P. C. Sofotasios, M. K. Fikadu, S. Muhaidat, S. Freear, G. K. Karagiannidis, and M. Valkama, "Relay selection based full-duplex cooperative systems under adaptive transmission," *IEEE Wireless Commun. Lett.*, vol. 6, no. 5, pp. 602–605, Oct. 2017.

[10] 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.

[11] 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.

[12] M. A. A. 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.

[13] 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.

[14] 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.

[15] 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.

[16] M. Lelarge, A. Proutiere, and M. S. Talebi, "Spectrum bandit optimization," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2013, pp. 1–5.

[17] M. S. T. M. Shahi, "Minimizing regret in combinatorial bandits and reinforcement learning," Ph.D. dissertation, Dept. Autom. Control, KTH Roy. Inst. Technol., Stockholm, Sweden, 2017.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

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

[24] 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.

[25] 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, Jul. 2011.

[26] 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, Feb. 2014.

[27] N. H. Tran, L. J. 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.

[28] N. Nomikos, T. Charalambous, I. Krikidis, D. Vouyioukas, and M. Johansson, "Hybrid cooperation through full-duplex opportunistic relaying and max-link relay selection with transmit power adaptation," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Australia, 2014, pp. 5706–5711.

[29] 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.

[30] L. Wang, F. Tian, T. Svensson, D. Feng, M. Song, and S. Li, "Exploiting full duplex for device-to-device communications in heterogeneous networks," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 146–152, May 2015.

[31] D. D. 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 Wkshps)*, Singapore, 2017, pp. 1–6.

[32] 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.

[33] 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.

[34] 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.

[35] L. Wei and V. Srivatsva, "On abruptly-changing and slowly-varying multiarmed bandit problems," in *Proc. Annu. Amer. Control Conf. (ACC)*, Milwaukee, WI, USA, 2018, pp. 6291–6296.

[36] *NR; Multiplexing and Channel Coding*, 3GPP Standard TS38.212, 2021.

[37] 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.

[38] 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.

[39] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, 2012.

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

[41] 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.

[42] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz, "Kullback–Leibler upper confidence bounds for optimal sequential allocation," *Ann. Stat.*, vol. 41, no. 3, pp. 1516–1541, 2013.

[43] P. Mènard and A. Garivier, "A minimax and asymptotically optimal algorithm for stochastic bandits," in *Proc. 28th Int. Conf. Algorithmic Learn. Theory*, Sep. 2017, pp. 715–720.

[44] "Radio Transmit Power." Cisco. 2008. [Online]. Available: https://www.cisco.com/c/en/us/td/docs/routers/access/wireless/software/guide/RadioTransmitPower.html (accessed Oct. 10, 2020).

[45] N. Nomikos, M. S. Talebi, R. Wichman, and T. Charalambous, "Bandit-based relay selection in cooperative networks over unknown stationary channels," in *Proc. IEEE 30th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Espoo, Finland, 2020, pp. 1–5.

[46] E. T. Michailidis, N. Nomikos, P. Trakadas, and A. G. Kanatas, "Three-dimensional modeling of mmWave doubly massive MIMO aerial fading channels," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1190–1202, Feb. 2020.

[47] R. Gupta, K. Lakshmanan, and A. K. Sah, "Beam alignment for mmWave using non-stationary bandits," *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2619–2622, Nov. 2020.

[48] I. Chafaa, E. V. Belmega, and M. Debbah, "One-bit feedback exponential learning for beam alignment in mobile mmWave," *IEEE Access*, vol. 8, pp. 194575–194589, 2020.

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, where he worked as a Postdoctoral Researcher with the Information and Communication Systems Engineering Department from November 2014 to October 2019. From September 2018 to September 2019, he was an Adjunct Lecturer with the Open University of Cyprus. From January 2019 to December 2019, he worked as a Postdoctoral Researcher with the General Department, National and Kapodistrian University of Athens. He is currently a Research Associate with the IRIDA Research Centre for Communication Technologies, Department of Electrical and Computer Engineering, University of Cyprus. His research interests include cooperative communications, nonorthogonal multiple access, full-duplex communications, and machine learning for wireless networks optimization. He is a member of the IEEE Communications Society and the Technical Chamber of Greece.

MOHAMMAD SADEGH TALEBI received the B.S. degree in electrical engineering from the Iran University of Science and Technology, Tehran, Iran, in 2004, the M.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, in 2006, and the Ph.D. degree in electrical engineering from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2017. Prior to his Ph.D., he worked as a Research Engineer with the School of Computer Science, Institute for Research in Fundamental Science, Tehran. From June 2018 to January 2020, he was a Postdoctoral Researcher with the SequeL (currently, Scool) Team, Inria Lille—Nord Europe, Lille, France. Since February 2020, he has been a Tenure-Track Assistant Professor with the Department of Computer Science, University of Copenhagen, Copenhagen, Denmark. His primary research interests include multi-armed bandits, theoretical reinforcement learning, and adaptive control under uncertainty.

THEMISTOKLIS CHARALAMBOUS (Senior Member, IEEE) received the B.A. degree in electrical and information sciences from Trinity College Dublin, the M.Eng. degree in electrical and information sciences from the University of Cambridge in 2005, and the Ph.D. degree from the Control Laboratory, Engineering Department, University of Cambridge in 2010. He joined the Human Robotics Group, Imperial College London as a Research Associate 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 to January 2015, he worked with the Department of Automatic Control, School of Electrical Engineering, Royal Institute of Technology as a Postdoctoral Researcher. From April 2015 to December 2016, he worked as a Postdoctoral Researcher 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. Since September 2018, he has been a Research Fellow of the Academy of Finland. From July 2020 to August 2021, he was a Tenured Associate Professor and since then he has been a Visiting Professor. Since September 2021, he has been a Tenure-Track Assistant Professor with the University of Cyprus. His primary research interests include the design and analysis of (wireless) networked control systems that are stable, scalable, and energy efficient.

RISTO WICHMAN (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.