Nomikos, Nikolaos; Charalambous, Themistoklis; Wichman, Risto

MABAMS: Multi-Armed Bandit-Aided Mode Selection in Cooperative Buffer-Aided Relay Networks

Published in:
2022 IEEE Globecom Workshops (GC Wkshps)

DOI:
10.1109/GCWkshps56602.2022.10008754

Published: 12/01/2023

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Nomikos, N., Charalambous, T., & Wichman, R. (2023). MABAMS: Multi-Armed Bandit-Aided Mode Selection in Cooperative Buffer-Aided Relay Networks. In *2022 IEEE Globecom Workshops (GC Wkshps)* (pp. 1230-1235). Article 10008754 IEEE. https://doi.org/10.1109/GCWkshps56602.2022.10008754

MABAMS: Multi-Armed Bandit Aided Mode Selection in Cooperative Buffer-Aided Relay Networks

Nikolaos Nomikos
IRIDA Research Centre
University of Cyprus
Nicosia, Cyprus
nomikos.nikolaos@ucy.ac.cy

Themistoklis Charalambous
School of Electrical Engineering
Aalto University
Espoo, Finland
themistoklis.charalambous@aalto.fi

Risto Wichman
School of Electrical Engineering
Aalto University
Espoo, Finland
risto.wichman@aalto.fi

***Abstract*—In networks with fast changing environments, the continuous acquisition of Channel State Information (CSI) is often infeasible. In such cases, only statistical CSI is available at the transmitters (CSIT). In this paper, we study the mode selection problem for a cooperative network, consisting of a source, a buffer-aided (BA) Full-Duplex (FD) relay, and a destination. In this setting, at every time frame, the network can operate in either FD mode (in different power levels), or, switch to half-duplex (HD) transmission/reception when FD operation is not feasible. Aiming to choose the best mode of operation, a mode selection mechanism is proposed, named MABAMS, which is based on a multi-armed bandit learning approach that uses the acknowledgements/negative-acknowledgements (ACK/NACK) observations in order to extract useful information about the statistics of the channels. As a consequence, MABAMS avoids the need for channel state information acquisition and exchange. We assess the performance in terms of outage probability, average throughput, and accumulated regret over time, in order to demonstrate an interesting performance-complexity trade-off when compared to the case in which the channel statistics are known. In addition, we demonstrate significant performance improvements over the cases without any power adaptation.**

***Index Terms*—Full-duplex relaying, buffer-aided relaying, multi-armed bandits, mode selection, statistical CSIT.**

I. Introduction

Towards meeting the increasingly demanding objectives of sixth generation (6G) wireless communications, cooperative relaying has evolved as an important technique to improve the quality of wireless transmissions. Based on the ability of the relay for simultaneous transmission and reception, Full-Duplex (FD) and Half-Duplex (HD) relaying provide a trade-off between spectral efficiency and interference avoidance. FD relays transmit and receive at the same time and on the same channel, thus achieving higher spectral efficiency. However, Loop Interference (LI) from the relay's transmitting antenna to its receiving antenna leads to performance degradation [1].

For relays without buffers, a hybrid FD/HD scheme was proposed in [2], outperforming schemes based on either FD or HD relaying. When relays are equipped with buffers, Buffer-Aided (BA) relaying offers increased flexibility in algorithmic design and improved performance; see, e.g., [3] and references therein. In a HD network, $\max-\text{link}$ selection was proposed in [4] where in each time frame, a BA relay is selected to receive or transmit, thus achieving a diversity gain twice the number of relays for large buffer size. For *single-relay* networks, as is the case in this paper, Zlatanov *et al.* [5] showed that BA relaying improves the throughput of FD systems, due to improved robustness. The authors in [6] consider a single-relay network with a non-saturated source and FD/HD relaying and where statistical CSIT is exploited, in order to maximize the packet arrival rate at the source under a power constraint. However, their scheme concerned only FD relaying and the main goal was to maximize the end-to-end capacity. When multiple BA relays are available, Successive Opportunistic Relaying (SOR) can also be deployed for recovering the spectral loss of HD relaying. For example, Nomikos *et al.* [7] consider a network with a saturated source, fixed-rate transmissions, multiple BA FD relays, statistical CSIT and instantaneous CSI at the receivers and they proposed a hybrid FD/SOR/HD link selection algorithm that aimed at maximizing the throughput per energy unit of the network.

In 6G mobile networks, dense deployments of small cells are envisioned where machines and mobile users will be competing for wireless access. In such complex scenarios with dense deployments, overhead for signaling and feedback messages is necessitated for efficient operation of the network. However, this overhead may become so significant that it can affect the performance of the network. To alleviate this overhead, machine learning techniques have been deployed in wireless communications, showing promising results for facilitating low-complexity coordination mechanisms (see, e.g., [8]–[10]). We aim at developing a low-complexity mode selection mechanism for a simple BA FD relay network. Towards this end, we adopt a reward-based machine learning framework, called multi-armed bandits (MAB), which has been already deployed in several 5G cases aiming to overcome the complexity of network coordination through learning; see, e.g., [11]. The problem that is closest to ours is [12], where the problem of power control in a network with a single FD relay is investigated, modeling the power level selection as a MAB game.

In this paper, an online policy for selecting the mode of operation of the cooperative network and assigning the power level in the case of FD is developed. The different modes of operation, as well as the power levels in the FD mode, are combined in a MAB game. Thus, in each time frame, the

relay observes the ACK/NACK messages from the destination for the previous transmissions, as well as whether or not the receptions from the source were successful. Our contributions are the following.

- A bandit-based mode selection mechanism (MABAMS) is proposed, relying on local observation by the relay of the outcome of the source's signal reception and ACK/NACK feedback from the destination, such that the relay does not need to perform channel estimation.
- Different versions of MABAMS, based on various upper confidence bound (UCB) policies are evaluated and results are given in terms of outage probability, average throughput and accumulated regret against: *i)* FD with optimal power allocation (known CSIT), *ii)* FD without power allocation, and *iii)* HD (max − link).

II. System model

A two-hop network consisting of a source, S, a destination, D, and a BA FD decode-and-forward (DF) relay R is considered, as depicted in Fig. 1. Due to severe fading, communication is only established via the relay, which is equipped with a buffer of size L, where L denotes the maximum number of packets that can be stored. The number of packets in R's buffer is denoted by Q ($Q \in \{0, 1, \ldots, L\}$).

Time is divided into "frames" of one packet duration (e.g., fixed-size packets). At any arbitrary time-frame K, for the link $\{i \to j\}$, the channel coefficient h_{ij} is modeled as statistically independent complex normal random variables with zero mean, and variance σ_{ij}^2, i.e., $h_{ij} \sim \mathcal{CN}(0, \sigma_{ij}^2)$. The envelope of the channel coefficients is Rayleigh distributed, i.e., $|h_{ij}| \sim \text{Rayleigh}(\sigma_{ij})$. The channel gains $g_{ij} \triangleq |h_{ij}|^2$ are, therefore, exponentially distributed, i.e., $g_{ij} \sim \text{Exp}(\sigma_{ij}^{-2}/2)$. We assume that the distribution of the channels is strict-sense stationary (SSS) during the operation of the system.

The source is assumed to be saturated and the information rate, r_0, for successful reception at the destination is fixed (and depends on the application). So, a transmission from a transmitter i to a receiver j is successful, if the SNR Γ_{ij} at the receiver is greater than or equal to the *capture ratio* γ_j. This framework is equivalent to the *capture model*. The thermal noise variance at the relay and the destination are denoted by σ_R^2 and σ_D^2, respectively, and they are assumed to be additive white Gaussian noise (AWGN). At each time-slot, there are two modes of operation; namely, *full-duplex* and *max-link* [4].

In the full-duplex mode, the source and the relay transmit a packet simultaneously, using power levels P_S and P_R, respectively. As FD relaying is supported, self-interference (SI) exists and h_{RR} denotes the instantaneous residual SI between the two antennas of relay R, following a complex Gaussian distribution and taking values in the range $(0, \sigma_{RR}^2)$. For the packet on the $\{S \to R\}$ link to be successfully received, we require that the signal-to-interference-and-noise ratio (SINR) is such that

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

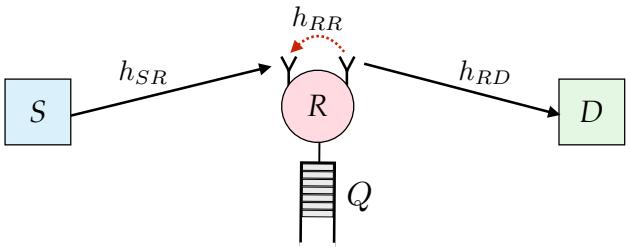

Fig. 1. A buffer-aided full-duplex relay network.

For the packet on the $\{R \to D\}$ link to be successfully received, we require that the signal-to-noise ratio (SNR) is such that

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

In this setup, it is assumed that the power of the source (P_S) is fixed, whereas the power of the relay (P_R) can be adjusted in order to maximize the end-to-end throughput.

In the max-link mode, either the source or the relay transmits a packet. As a result, no interference component is included. In this case, in order to maximize the SNR on the $\{R \to D\}$ link, the maximum power at the relay is deployed, i.e., $P_R = P_{R,\max}$. On the $\{S \to R\}$ link there is no interference and hence (1a) becomes

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

Retransmissions are based on an Acknowledgements-/Negative-Acknowledgements (ACKs/NACKs), where receivers broadcast short-length error-free packets over a separate narrow-band channel. A link is *feasible* if it is not in outage and fulfills the queue conditions (i.e., for non-full buffers in $\{S \to R\}$ links and for non-empty buffers in $\{R \to D\}$ links).

III. Problem Statement

Usually, it is assumed that statistical CSIT is available at each frame, and the during operation the target is to ensure that the success transmission probability over a link is greater than or equal to a pre-defined threshold q_{th} (that can be application- or condition-dependent), i.e., $\mathbb{P}\{\Gamma_i(P_j) \geq \gamma_i\} \geq q_{\text{th}}$. When this condition cannot be satisfied by both $\{S \to R\}$ and $\{R \to D\}$ links during full duplex operation, some papers propose switching to a single-link selection instead; see, e.g., [7]. In this paper, while we assume that the distribution of the channels is SSS throughout the operation, the statistical CSIT of the links is not initially available. As a result, it is not possible to know *a priori* which mode of operation (full-duplex or max-link) is the best and which power level the relay should choose. Towards resolving this issue, we aim at deploying a learning algorithm that will allow the relay (who is the main decision maker in this network setup) to decide which mode of operation is better and if FD is selected finds the power level that maximizes the end-to-end throughput.

In order to understand the complexity of the problem, in what follows we provide the probabilities for successful transmission for each link and the different options for transmission at each time frame.

Inequality (1a) can be written as

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

This linear combination of exponentially distributed variables $g_{SR} \sim \text{Exp}(\mu)$ and $g_{RR} \sim \text{Exp}(\lambda)$, $\lambda, \mu > 0$, can be shown to have the following distribution [7]:

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

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

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

where $F_W(w)$ denotes the cumulative distribution function (cdf) of a random variable W; for the exponential distribution, this is given by $F_W(w) = 1 - \exp(-\lambda_W w)$. Similarly, the probability that inequality (1a) holds, $\mathbb{P}\big((1a)\big)$, is given by

$$\begin{aligned} \mathbb{P}\big((1a)\big) &= 1 - F_{g_{SR}P_S - \gamma_R g_{RR} P_R}(\gamma_R\sigma_R^2) \\ &= \frac{P_S\lambda}{P_S\lambda + \gamma_R P_R \mu} \exp\left(-\mu\frac{\gamma_R\sigma_R^2}{P_S}\right) \,. \end{aligned} \quad (3)$$

When there is no interference on the $\{S{\to}R\}$ link, the distribution reverts to an exponential distribution, i.e.,

$$\mathbb{P}\big((1c)\big) = \exp\left(-\mu\frac{\gamma_R\sigma_R^2}{P_S}\right). \quad (4)$$

At each time frame k, the relay (decision maker) has to select between three options:

- O_1 Deploy the FD mode and have both the source and the destination transmit a packet. The source transmits with fixed power P_S, whereas the relay transmits with power level $P_R[k]$. The power level $P_R[k]$ is selected with the aim of maximizing the end-to-end throughput of the transmission; details will be explained in Section V. In this case, the relay needs to have at least one packet in the buffer ($Q > 0$) and no more than $L-1$ packets in the buffer ($Q < L$) (since a packet remains in the queue if it is not successfully transmitted).
- O_2 Have the source transmit only on the $\{S{\to}R\}$ link with fixed power P_S. In this case, the relay needs to have at most $L-1$ packets in the buffer ($Q < L$).
- O_2 Have the relay transmit only on the $\{R{\to}D\}$ link with fixed power $P_R = P_{R,\max}$, where $P_{R,\max}$ is the maximum power with which the relay can transmit. In this case, the relay needs to have at least one packet in the buffer ($Q > 0$).

IV. MAB Modeling

A. The MAB Problem

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. The regret is defined as the difference between the reward achieved when the best 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 packets successfully transmitted or simply what we call the throughput. Equivalently, we aim at designing a link selection and power level selection policy that minimize the *regret*. The regret of a policy $\pi \in \Pi$ (Π being the set of all feasible policies) is defined by the performance loss and it is found by comparing the performance achieved under policy π to that of the best static policy, i.e.,

$$R^\pi(T) = \max_{\ell\in\mathcal{L}} \mathbb{E}\left\{\sum_{t=1}^{T} U_{\ell,t}\right\} - \mathbb{E}\left\{\sum_{t=1}^{T} U_{I_t^\pi,t}\right\}, \quad (5)$$

where $U_{\ell,t}$ denotes the instantaneous utility obtained from choosing link ℓ at time-slot t under feasible configuration $\ell \in \mathcal{L}$. Moreover, $U_{I_t^\pi,t}$ denotes the instantaneous utility obtained from the link I_t^π chosen under policy π at time-slot t. In our setup, the relay either receives, transmits with $P_{R,\max}$, or is in FD mode and 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 $O_{1,i}$ with i one of the $|\mathcal{P}_R|$ power levels, O_2 and O_3 (so, in total there are $|\mathcal{P}_R| + 2$ arms).

In their seminal paper, Lai and Robbins [13] characterize a problem-dependent lower bound on the regret of any adaptive policy, 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],

$$\liminf_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \geq c(\boldsymbol{\mu}) \,, \quad (6)$$

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

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

B. Upper Confidence Bound (UCB) Policies

UCB policies are based on the *optimism in the face of uncertainty* principle (or for short, the *optimistic* principle) proposed by Lai and Robbins [13], and for MAB problems their regret grows logarithmically with the time horizon. The underlying idea of an *optimistic algorithm* is to replace the unknown mean rewards of each arm with a high-probability 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., $n_{j,t} \coloneqq \sum_{s=1}^{t} \mathbb{1}_{\{I_s=j\}}$, where $\mathbb{1}_A$ denotes the indicator function of the 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:

$$\hat{q}_{j,t} = \frac{1}{n_{j,t}} \sum_{s=1}^{t} r_{j,s} \mathbb{1}_{\{I_s=j\}} \,, \tag{7}$$

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

An optimistic algorithm π maintains an index function $\bar{q}_j$ for each arm j, which depends only on the past 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 simply consists in selecting the arm with the largest index $\bar{q}_{j,t}$ at each round t:

$$I_t = \arg\max_{j \in \mathcal{L}} \bar{q}_{j,t}. \tag{8}$$

In the sequel, we briefly introduce UCB1 [14], a popular index policy for stochastic MABs. UCB1 is an index policy designed based on Hoeffding's concentration inequality for bounded random variables. The UCB1 index (or for short, UCB) is defined as follows:

$$\bar{q}_{j,t}^{\text{UCB}} = \widehat{q}_{j,t} + \sqrt{\frac{3\log(t)}{2n_{j,t}}} \,. \tag{9}$$

Note that, other policies, such as KL-UCB [15] and KL-UCB^{++} [16] can be used for improved performance.

V. MABAMS: MAB-Aided Full-Duplex/Max-Link Mode Selection

A. Preliminaries

At each time frame k, the relay has to decide in which mode to operate: HD (either O_2 or O_3) or FD (O_1) along with its power level P_R in order to minimize the long term regret of the overall system. Unlike standard MAB communication problems (described in Section IV), in this case, the choices are coupled in many ways. For example, option O_1 includes the $\{R \to D\}$ link of option O_3, but while it is difficult to extract information about the $\{S \to R\}$ link in O_1 due to the complicated distribution (see, equation (3)), we can extract information about the distribution of the $\{R \to R\}$ link, that can help us better understand the effect of SI.

If we first assume that the distribution of the channels are already known (which is not the case in this work), we will find the optimal relay power, P_R^*, such that the end-to-end throughput in the FD mode is maximized. The optimization problem can be expressed as follows:

$$\textbf{P1:} \max_{P_R} \left\{ \min \left\{ \mathbb{P}\big((1\text{b})\big), \mathbb{P}\big((1\text{a})\big) \right\} \right\} \tag{10a}$$

$$\text{s.t} \quad 0 \leq P_R \leq P_{R,\max}. \tag{10b}$$

This problem is equivalently written as follows:

$$\textbf{P2:} \max_{P_R} z \tag{11a}$$

$$\text{s.t} \quad \mathbb{P}\big((1\text{b})\big) \geq z, \tag{11b}$$

$$\mathbb{P}\big((1\text{a})\big) \geq z, \tag{11c}$$

$$0 \leq P_R \leq P_{R,\max}, \tag{11d}$$

$$0 < z \leq 1. \tag{11e}$$

where z serves as the epigraph of the function

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

Since $\mathbb{P}\big((1\text{b})\big)$ is a monotonically increasing function of P_R and $\mathbb{P}\big((1\text{a})\big)$ is a monotonically decreasing function of P_R, the optimal solution to this problem is achieved with equality for both (11b) and (11c). Therefore, after algebraic manipulations, we obtain the following:

$$P_R^* = \frac{\nu \gamma_D \sigma_D^2}{\ln\left(1/z^*\right)} = \frac{P_S \lambda \left(\alpha(\mu) - z^*\right)}{\gamma_R \mu} \,, \tag{12}$$

where $\alpha(\mu) = \exp\left(-\mu \frac{\gamma_R \sigma_R^2}{P_S}\right)$ and z^* is the solution to the optimization problem **P2**, which is obtained by solving the equality in (12), since all the distribution parameters, P_S, γ_R, γ_D, σ_R^2, and σ_D^2 are assumed to be known (can be computed using simple line search methods, such as bisection). Note that the boundaries of P_R can be justified separately.

From this preliminary discussion, we observe that if the channel distribution parameters λ, μ, and ν were known, one could compute directly P_R^*, thus concluding directly which mode of operation is better, based on the overall throughput of the system for each mode of operation. However, in this case, the distributions are initially unknown. For solving this problem, 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.

B. Online Learning Model

Since we consider a scenario in which the distributions of the channels are SSS during the operation of the system, the success probabilities for the different modes and power levels are assumed to be fixed but unknown. Each power level corresponds to an arm, and pulling an arm corresponds to a packet transmission using the selected power level.

More formally, if in time frame k FD transmission is selected with power level j, a reward $r_{j,k}^{(FD)}$ is obtained, where

$$r_{j,k}^{(FD)} = \begin{cases} 2, & \text{if } \{S \to R\} \text{ and } \{R \to D\} \text{ trans. successful,} \\ 1, & \text{if one of the trans. successful,} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

If, however, in time frame k HD transmission is selected (either $\{S\rightarrow R\}$ or $\{R\rightarrow D\}$ link), then a reward $r_{j,k}^{(HD)}$ is obtained, where

$$r_{j,k}^{(HD)} = \begin{cases} 1, & \text{if trans. successful,} \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Even though there are correlations between the arms and one can infer information from one outcome about the other, in this work we do not exploit this available information.

C. MABAMS *Algorithm*

We are now ready to describe MABAMS. We expect that after the initial exploration phase, MABAMS will reveal the best mode of operation (FD or HD) and if FD is the best mode of operation, also provide the power level yielding the maximum reward, in terms of end-to-end throughput in order to minimize the regret. Note that the arm selection yields a random reward, which corresponds to the link/links (i.e., links $\{S\rightarrow R\}$, $\{R\rightarrow R\}$, and $\{R\rightarrow D\}$) involved. The procedure followed in each time frame is given in Algorithm 1.

Algorithm 1 MABAMS at each time frame.

input Set of power levels $\mathcal{P}_R$, $P_{R,\max}$, Q, capture ratios γ_R and γ_D, thermal noises σ_R^2 and σ_D^2
set $P_R[0] = P_{R,\max}$
for $k = 0, 1, 2, \ldots$ **do**
  if $Q = 0$ **then**
    Choose O_2 and compute $\hat{q}_{\mu,k}$ (7) and then $\bar{q}_{\mu,k}$
  else if $Q = L$ **then**
    Choose O_3 and compute $\hat{q}_{\nu,k}$ and $\hat{q}_{\lambda,k}$ (7) and then $\bar{q}_{\nu,k}$ and, $\bar{q}_{\lambda,k}$, respectively
  else if $0 < Q < L$ **then**
    Compute $\hat{q}_{j,k}$ (7) and then $\bar{q}_{j,k}$, , where $j \in \{O_{1,i}, O_2, O_3\}$, $i = \{1, 2, \ldots, M\}$
    Select mode j (with power level $i \in \{1, 2, \ldots, M\}$ if option O_1 is chosen) for transmission at time-slot k using (8)
    $n_{j,k+1} \leftarrow n_{j,k} + \mathbb{1}_{\{I_k=j\}}$ 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

VI. Performance Evaluation

Here, the average throughput performance of MABAMS is evaluated. More specifically, MABAMS is compared against FD relaying with CSI-based optimal power control (opt-PC FD), no power control (no-PC), random power level selection (rnd) and BB-PC [12]. In each link, the transmit SNR ranges from 0 dB to 40 dB, corresponding to the ratio of the maximum available transmit power at each transmitter, considering $P_S = P_{R,\max} = P_{\max}$ over the noise power. Also, in the figures, the x-axis corresponds to the transmit SNR in the $\{R\rightarrow D\}$ link, being equal to $P_{\max}/n_D$. Then, for each transmit SNR value, we perform 10^4 transmissions over which, the results are averaged. Furthermore, a fixed transmission rate $r_0 = 3$ bps/Hz is assumed in a single-relay topology where transmit power is selected from six different levels, i.e., $P_1 = P_{\max}$, $P_2 = 0.50P_{\max}$, $P_3 = 0.30P_{\max}$, $P_4 = 0.20P_{\max}$, $P_5 = 0.05P_{\max}$, $P_6 = 0.01P_{\max}$ [17]. When FD relaying takes place, the SI channel is characterized by average channel SNR $\bar{\gamma}_{SI} = 0$ dB or $\bar{\gamma}_{SI} = -10$. Table I lists the simulation parameters that are considered in the performance comparisons.

TABLE I
Simulation Parameters

Parameter	Value
No. of transmissions per SNR value	10^4
Buffer size L	10 packets
Transmission rate r_0	3 bps/Hz
Average SI channel SNR $\bar{\gamma}_{SI}$	{-10, 0} dB
Transmit SNR $P_{\max}/n_D$ range	{0, 40} dB
No. of relay power levels	6

When a strong SI with $\bar{\gamma}_{SI} = 0$ dB degrades the FD performance, the average throughput results are depicted in Fig. 2. The most important observation here is the superior performance of MABAMS for transmit SNR values between 0 and 20 dB. As MAB-aided mode switching is enabled, MABAMS addresses cases where FD relaying can not be performed, exploiting the buffering capabilities of the relay. After 20 dB, the optimal CSI-based power control provides the best performance as the relay chooses the optimal power level. It should be noted that the throughput results for the optimal power control do not consider losses from CSI overheads, i.e. CSI acquisition and exchange, and thus, in practice, MABAMS outperforms the CSI-based approach when CSI overheads account for 10% or more of a time-slot's duration. For higher transmit SNR values, MABAMS switches to FD operation more often and performance is closer to BB-PC. Finally, random power level selection and no-PC can not cope with the strong SI conditions, offering the worst performance.

Next, an SI channel characterized by $\bar{\gamma}_{SI} = -10$ dB is assumed and the results are shown in Fig. 3. It can be observed that CSI-based optimal power control provides the throughput upper bound for transmit SNR values above 12 dB. Again, for low transmit SNR, MABAMS provides the best performance, as the network switches to the HD modes, exploiting buffer-aided operation. In addition, MABAMS outperforms BB-PC

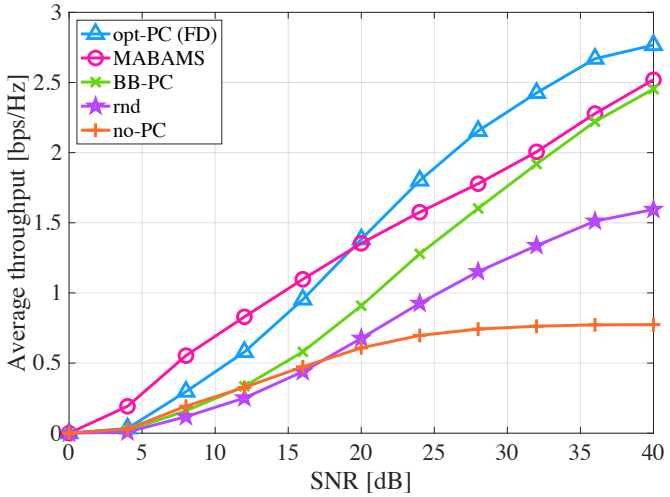

Fig. 2. Average throughput comparisons for different power control algorithms with $\bar{\gamma}_{SI} = 0$ dB.

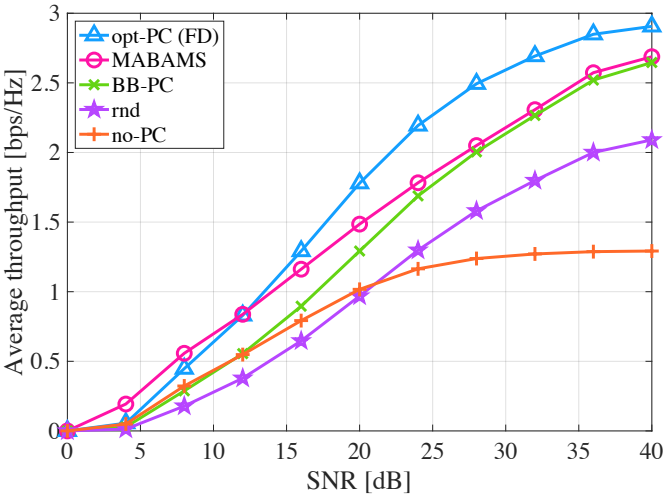

Fig. 3. Average throughput comparisons for different power control algorithms with $\bar{\gamma}_{SI} = -10$ dB.

throughout the transmit SNR range, while their performance gap is reduced in the high SNR regime since both algorithms adopt FD relaying with BB-aided power control. Moreover, the worst performance is provided from the scheme without PC as SI is not mitigated, being surpassed by random transmit power selection after 20 dB.

VII. Conclusions and Future Directions

In this paper, we propose MABAMS, a heuristic algorithm with which the mode of operation is chosen. While this algorithm seems to provide good efficiency and exploit the hybrid nature of the scheme, the optimal learning approach to solving this problem remains an open problem.

Ongoing research is focusing on the following issues: While we provide a method for reaching the best option eventually, it would be very interesting to consider a strategy tailored to the problem structure in order to find the *optimal* policy for learning the parameters of the network, while at the same time maximizing the network throughput (our main performance objective). Moreover, the transmissions on the $\{S \to R\}$ link should be efficiently used to extract information about the involved channels; this information has not been exploited in this work. Finally, it would be very interesting to consider the case in which the power level of the source is a (discrete) decision variable as well.

References

[1] Z. Zhang, K. Long, A. V. Vasilakos, and L. Hanzo, "Full-duplex wireless communications: Challenges, solutions, and future research directions," *Proceedings of the IEEE*.

[2] T. Riihonen, S. Werner, and R. Wichman, "Hybrid full-duplex/half-duplex relaying with transmit power adaptation," *IEEE Transactions on Wireless Communications*.

[3] N. Nomikos, T. Charalambous, I. Krikidis, D. N. Skoutas, D. Vouyioukas, M. Johansson, and C. Skianis, "A survey on buffer-aided relay selection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1073–1097, Secondquarter 2016.

[4] I. Krikidis, T. Charalambous, and J. Thompson, "Buffer-aided relay selection for cooperative diversity systems without delay constraints," *IEEE Transactions on Wireless Communications*, vol. 11, no. 5, pp. 1957–1967, May 2012.

[5] N. Zlatanov, D. Hranilovic, and J. S. Evans, "Buffer-aided relaying improves throughput of full-duplex relay networks with fixed-rate transmissions," *IEEE Communications Letters*, vol. 20, no. 12, pp. 2446–2449, 2016.

[6] 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.

[7] 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 Transactions on Vehicular Technology*, vol. 67, no. 8, pp. 7846–7850, 2018.

[8] 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. 32 328–32 338, 2018.

[9] 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, Sept. 2018.

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

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

[12] N. Nomikos, T. Charalambous, and R. Wichman, "Bandit-based power control in full-duplex cooperative relay networks," in *IEEE International Conference on Communications (ICC)*, 2021, pp. 1–6.

[13] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[14] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learn.*, vol. 47, no. 2, pp. 235–256, May 2002.

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

[16] P. Mènard and A. Garivier, "A minimax and asymptotically optimal algorithm for stochastic bandits," *Algorithmic Learning Theory*, pp. 715–720, Sept. 2017.

[17] Cisco, "Radio transmit power," 2008 (accessed May 22, 2022). [Online]. Available: https://www.cisco.com/c/en/us/td/docs/routers/access/wireless/software/guide/RadioTransmitPower.html.