

Relay Selection in Cooperative Power Line Communication: A Multi-Armed Bandit Approach

Babak Nikfar and A. J. Han Vinck

Abstract: Power line communication (PLC) exploits the existence of installed infrastructure of power delivery system, in order to transmit data over power lines. In PLC networks, different nodes of the network are interconnected via power delivery transmission lines, and the data signal is flowing between them. However, the attenuation and the harsh environment of the power line communication channels, makes it difficult to establish a reliable communication between two nodes of the network which are separated by a long distance. Relaying and cooperative communication has been used to overcome this problem. In this paper a two-hop cooperative PLC has been studied, where the data is communicated between a transmitter and a receiver node, through a single array node which has to be selected from a set of available arrays. The relay selection problem can be solved by having channel state information (CSI) at transmitter and selecting the relay which results in the best performance. However, acquiring the channel state information at transmitter increases the complexity of the communication system and introduces undesired overhead to the system. We propose a class of machine learning schemes, namely multi-armed bandit (MAB), to solve the relay selection problem without the knowledge of the channel at the transmitter. Furthermore, we develop a new MAB algorithm which exploits the periodicity of the synchronous impulsive noise of the PLC channel, in order to improve the relay selection algorithm.

Index Terms: Cooperative communication, multi-armed bandit, power line communication (PLC), relay selection.

I. INTRODUCTION

POWER line communication (PLC) is the technology of transferring data signals through existing power delivery infrastructures, with applications in smart grids, in-vehicle communication, etc. The data signal is generated as a differential voltage between two power delivery conductors, and propagates through the transmission line from source to destination. In power line communication networks, multiple networks nodes are interconnected via the transmission lines and the data signals flow between different nodes of the network. If the distance between source and destination in a communication scenario is long, the limitation of the transmission range of the PLC node prevents the establishment of a reliable communication. This limitation is due to the harsh environment of the PLC channel, for instance fading, noise, interference, and receiver sensitivity. To overcome this problem, cooperative communication is used to transmit the data signals from source to destination with help of one or more intermediate nodes. In this case, the source

node can communicate directly with nodes within its transmission range, and these nodes in turn can forward the message to the destination node. The intermediate nodes are called *relays* and the process of transmitting signals with the help of relays is referred to as multi-hop communication or relaying.

The application of relaying in cooperative wireless communication has been studied to a great extent. The use of relaying in cooperative power line communication has been mentioned and studied for certain communication scenarios as well. For example in [1] and [2], PLC relaying based on single frequency networking has been introduced and its performance has been discussed. Distributed space-time coding for multi-hop transmission and decode-and-forward relaying has been studied in [3] and [4], respectively. The concept of cooperative multi-hop communication for PLC has been first introduced and discussed in [5]–[7]. Existence of many intermediate nodes between source and destination, results in the existence of many optional paths or routes to follow. This situation gives rise to the problem of proper relay selection, where the challenge is to pick the optimal path that satisfies the needed performance requirements. In [7] the relay selection problem has been discussed and the link rate has been introduced as a figure of merit for different relay selection criteria.

In this paper, we consider the two-hop cooperative communication scenario, in which the transmitted signal from source travels to an intermediate relay before reaching the destination. We consider N available intermediate relay nodes, from which one of them as the relaying node is to be selected. However, the proper relay selection policy requires the availability of channel side information at the transmitter, which in turn requires an increased complexity of signal processing and introduces a lot of overhead in the system. In order to avoid this problem, we introduce a class of machine learning algorithms, namely *multi-armed bandit (MAB)*, which helps us find the best relay from available relays, based on the performance of the channel in both of the communications hops. Different algorithms of MAB has been developed in the field of machine learning, for example [8]–[10]. In wireless communication, MAB has been used in order to solve problems which are dealing with acquiring the best selection policy. For example, in [11] and [12], MAB has been used for relay selection problem in a stationary wireless channel. In PLC, MAB has been introduced in [13] to solve the channel selection problem in a multichannel PLC system. In this paper, we introduce the MAB tool to solve the relay selection problem in two-hop cooperative PLC without access to the channel side information. Furthermore, we propose two new algorithms based on MAB to further improve the performance of the relay selection based on the channel characteristics of the power line systems. The proposed algorithms are adapted

Manuscript received February 25, 2016; approved for publication by Caijun Zhong, Division I Editor, July 29, 2016.

The authors are with the Department of Digital Signal Processing, University of Duisburg-Essen, Germany, email: {babak.nikfar, han.vinck}@uni-due.de.

Digital object identifier: 10.1109/JCN.2017.000003

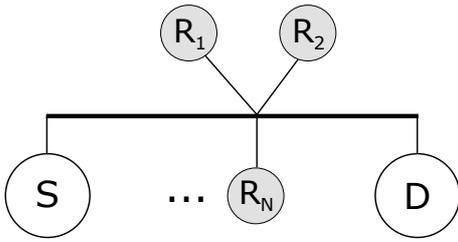


Fig. 1. PLC network with a transmitter, a receiver, and N relay nodes.

in such way to exploit the specific features of the PLC channel, namely the periodicity of the synchronous impulsive noise of the network. This adaptation to PLC of the proposed algorithms proves to be advantageous as illustrated in the numerical results.

This paper is organized as follows. Section II describes the system and channel model of the PLC network. The cooperative PLC and the relay selection problem are presented in Section III. Multi-armed bandit model is introduced in Section IV as well as the representation of the existing MAB algorithms. The proposed algorithms are described in Section V and simulation results are presented in Section VI. Finally, Section VII concludes the paper.

II. SYSTEM MODEL

Power line communication channels can be characterized as frequency-selective time-variant channels. We consider a PLC network, consisting of a transmitter node, a receiver node, and N intermediate nodes known as relays. In PLC networks, due to line attenuations, fading, noise, and signal interferences, transmission over long distances results in a degraded performance which may lie below the performance requirements of the system. Therefore, each node can communicate directly with those nodes which fall into its transmission range. These intermediate nodes transmit the message to the destination, constructing a two-hop communication. The structure of a two-hop communication can be seen in Fig. 1, where S and D represent source (transmitter) and destination (receiver), respectively, and R_i , $i \in \{1, \dots, N\}$ are N available relay nodes which can help the transmission of PLC signals from transmitter to receiver.

We use the bottom-up deterministic approach in PLC channel modeling, namely the transmission line theory model. Therefore, we can model each two nodes of the PLC network with voltages $V_i(f)$ and $V_j(f)$, respectively, and corresponding currents $I_i(f)$ and $I_j(f)$, with ABCD-parameters as follows [6], [7]. Note that both voltages and currents are frequency dependent.

$$\begin{bmatrix} V_i(f) \\ I_i(f) \end{bmatrix} = \begin{bmatrix} A(f) & B(f) \\ C(f) & D(f) \end{bmatrix} \begin{bmatrix} V_j(f) \\ I_j(f) \end{bmatrix}. \quad (1)$$

The corresponding transfer function between these two nodes can be derived as

$$H_{ij}(f) \triangleq \frac{V_j(f)}{V_i(f)} = \frac{Z_j(f)}{A(f)Z_j(f) + B(f)}, \quad (2)$$

where $Z_j(f) = V_j(f)/I_j(f)$ is the impedance of node j . The transfer function of the channel between two network nodes i

and j , with an intermediate node k in between, can therefore be calculated as

$$H_{ij}(f) = H_{ik}(f) \cdot H_{kj}(f). \quad (3)$$

ABCD-parameters or transmission line parameters depend on the power line characteristics as well as the length ℓ of the transmission line, and can be calculated as

$$\begin{bmatrix} A(f) & B(f) \\ C(f) & D(f) \end{bmatrix} = \begin{bmatrix} \cosh(\gamma\ell) & Z_0 \sinh(\gamma\ell) \\ \frac{1}{Z_0} \sinh(\gamma\ell) & \cosh(\gamma\ell) \end{bmatrix}, \quad (4)$$

where $Z_0 = \sqrt{(R + j2\pi fL)/(G + j2\pi fC)}$ is the characteristic impedance of the transmission line per unit length and $\gamma = \sqrt{(R + j2\pi fL)(G + j2\pi fC)}$ is the propagation constant, and they are related to the primary cable parameters R, G, L, and C, representing resistance, conductance, inductance, and capacitance of the line per unit length. The primary cable parameters depend on the physical cable characteristics, and are derived for a PLC channel in [14]. The propagation constant is a complex quantity and can be written as $\gamma = \alpha(f) + j\beta(f)$, where $\alpha(f)$ is the attenuation constant, and $\beta(f)$ is the phase constant. The channels in this model are considered to be independent and channel correlation is neglected. This assumption is due to independent loads in a PLC network and the geographically distributed routs of a relay network which makes the independent channel assumption reasonable.

In in-home PLC applications, dominant noise components consist of background noise, aperiodic impulsive noise, as well as synchronous and asynchronous impulsive noise. A prominent noise in PLC is the periodic impulsive noise which is synchronous to the alternating current (AC) of the mains [15]. Therefore, the noise is considered to be *cyclostationary* and the periodic instantaneous noise power is derived in [16] as

$$\sigma_N^2(t) = \sum_{l=0}^{L-1} A_l \left| \sin \left(\frac{2\pi t}{T_{AC}} + \theta_l \right) \right|^{n_l}, \quad (5)$$

where L represents the number of noise classes (for narrowband PLC, $L = 3$ [16]), A_l , θ_l , and n_l are different characteristic parameters of the l th noise class, and T_{AC} is the period of the mains voltage.

III. COOPERATIVE PLC AND PROBLEM FORMULATION

The idea of cooperative communication has been well investigated in wireless communication, e.g., in [17] and [18]. The principle of this idea is to realize spatial diversity without the use of multiple antennas. In this case, cooperative users generate a virtual antenna array to achieve the desired cooperative diversity. This concept has been extended to relay networks with multi-hop transmission between source and destination, e.g., in [19]. The concept of cooperative diversity has been introduced in PLC for the first time in [5], and has been further discussed in [6] and [7]. It has been shown in [7] that the PLC relay channel consists of two keyhole channels, and thus a diversity gain as observed for wireless relaying cannot be achieved for PLC. However, despite the lack of the cooperative diversity advantage, cooperative multi-hop transmission can provide significant

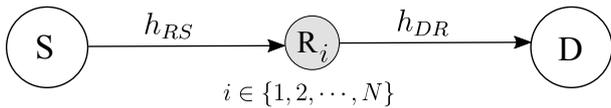


Fig. 2. A two-hop transmission in cooperative PLC.

power gains. In this paper, we assume a two-hop transmission, that is, a source node transmits the message to a destination node through a relay node between them as depicted in Fig. 2. A generalization to a multi-hop transmission is straightforward.

In a two-hop cooperative PLC, as depicted in Fig. 2, three nodes are available, namely source node S , selected relay node R_i , which is selected from a sequence of all the available relays $i \in \{1, 2, \dots, N\}$, and destination node D . We assign a number n to each node so that $\{S, R_i, D\} \sim \{1, 2, 3\}$. The cooperative transmission is assumed to follow a time division protocol, which means at each time instant, only one node (either the source or the relay) can transmit data with a fixed transmit power. As a figure of merit, we consider the end-to-end achievable rate, also known as the *end-to-end capacity*. The end-to-end capacity in a two-hop transmission for a link between node n and node $n + 1$ is expressed as $C_{n,n+1}$, which is a strictly increasing function of the transmitted signal-to-noise ratio (SNR) [15]. The SNR performance of the transmission links directly exhibits the overall performance of transmission, hence forming a reliable figure of merit.

We use the conventional strategy of *fixed-rate two-hop* transmission [20]. The reason to choose the fixed-rate strategy is due to the fact that in other two-hop strategies such as rate-adaptive scheme, the link capacities are required before transmission so that the transmitter can adjust the rate to follow the link capacity. In a fixed-rate cooperative transmission scheme, the relay node between the source and destination nodes, retransmits the received message from source using the same transmission scheme and thus link rate. Hence, the maximum rate that can be achieved over this route is determined by the minimum of the rates achievable on the individual links. Let us denote the data rate over each hop as $R_1 = R_2 = R$, for some fixed value of R . In order to ensure reliable communication, $R \leq C_{n,n+1}$ must be satisfied over all hops. It is shown that maximum end-to-end capacity can be achieved by choosing $R = \min_{n=1,2} C_{n,n+1}$ [20]. The end-to-end capacity for a fixed-rate two-hop transmission can be expressed as

$$C_{total} = \frac{1}{2} \min_{n=1,2} C_{n,n+1}. \quad (6)$$

Due to the time-variant nature of the PLC systems, in order to follow the changes of the channel and selecting the best relay, channel state information (CSI) should be available at the transmitter at all times. However, acquiring CSI requires constant transmission of pilot signals throughout the transmission which in turn enormously increases the overhead and decreases the throughput. Moreover, channels formed by all of the available relays must be evaluated in order to perform an optimal relay selection, which with a high number of available relays is not an easy task. We wish to have a relay selection strategy which results in a high end-to-end capacity C_{total} . Formally, we

aim to solve the following maximization problem

$$\underset{R_i \in \{R_1, R_2, \dots, R_N\}}{\text{maximize}} \quad C_{total}. \quad (7)$$

We consider three relay selection strategies to maximize the total end-to-end capacity without any information at transmitter.

1. Fixed selection: In this strategy, a fixed relay node is assigned for cooperative transmission between the source and destination, regardless of the instantaneous channel conditions. This strategy neglects the variations in the PLC channel and therefore is not an optimal method of relay selection.
2. Random selection: In this strategy, at each transmission time interval a random relay is selected from the sequence of N relays. This method neglects the variations of the channel over time as well; however, the randomness of the relay selection may decrease the probability of selecting a bad relay compared to the fixed selection method.
3. The proposed learning algorithms: We propose two new learning algorithms based on the MAB model in machine learning. In our approach we consider the variations of the channel over time which occurs in a cyclostationary manner as described before, and try to adapt the relay selection with these variations without the knowledge of CSI at the transmitter. The concept of multi-armed bandit and our algorithmic solutions are discussed in the following sections. Furthermore, we show through numerical results that our proposed algorithms result in a better performance compared to other relay selection approaches.

IV. MULTI-ARMED BANDIT PROBLEM MODELING

MAB is a class of decision making problems introduced in machine learning field, where an agent sequentially selects an arm (action, interchangeably) from a set of predefined arms (actions), and receives a reward drawn from some a priori unknown distribution. Only the reward of played arm is observable. As a result of lack of information, at each trial, the player may choose an inferior arm in terms of average reward, yielding a *regret* that is quantified by the difference between the reward that would have been achieved if the agent would have selected the best arm with the highest reward and the actual achieved reward.

We model the PLC relay selection problem as a MAB. Let us assume a set \mathcal{K} of available relay nodes, here and thereafter known as *actions* or *arms*. Each frame of data is transmitted by selecting a relay node $k \in \mathcal{K}$, $k = \{1, 2, \dots, N\}$, resulting in a particular total end-to-end capacity, here and thereafter known as *reward*. At each time slot t (corresponding to one frame of data), an action a_t is selected, yielding the instantaneous reward $X_t(a_t)$. The rewards $\{X_t(k)\}_{t \geq 1}$ for each arm $k \in \{1, 2, \dots, N\}$ is calculated according to the received signal and this information is fed back to the transmitter via a robust mode of transmission and the transmitter chooses an arm at each trial according to a policy π . Let us denote the expectation of the reward $X_t(k)$ by $\mu_{k,t}$. Let k_t^* denote the optimal arm at time t , with expected reward $\mu_{k_t^*,t}$, where by definition $\mu_{k^*,t} = \mu_t^* = \max_{k \in \mathcal{K}} \mu_{k,t}$. We define the *instantaneous regret* at time t as the difference between mean rewards of the selected

arm and the optimal arm. The expected regret of a decision making policy π after T trials, therefore, can be expressed as

$$R_{\pi,T} = E_{\pi} \left[\sum_{t=1}^T (\mu_t^* - \mu_{a_t,t}) \right], \quad (8)$$

where $E[\cdot]$ represents the mathematical expectation. The goal of a good policy is to select the optimal arm at each trial, which results in a minimum expected regret over all trials. Therefore, the goal of the MAB problem is to minimize the expected regret with a certain decision making policy π , or equivalently

$$\underset{\pi}{\text{minimize}} R_{\pi,T}. \quad (9)$$

A. UCB and Modified UCB Algorithms

Upper-confidence bound (UCB) algorithms are deterministic MAB policies which have been introduced and analyzed by [8]. In the seminal UCB algorithm, an upper-bound for the expected reward $\mu_{k,t}$ is evaluated which its calculation is based on the previous rewards of that particular arm and some uncertainty factor. During the t th round, the user selects an arm which maximizes the upper-bound of the confidence interval for expected reward $\mu_{k,t}$. The algorithm starts the selection process by the initialization phase, in which during the first N rounds, action a_k , $k \in \{1, 2, \dots, N\}$ is selected successively and after transmission at each arm, the corresponding UCB index $I_{k,t}$ of that arm is calculated as

$$I_{k,t} = \bar{X}_t(k) + c_t(k), \quad (10)$$

where

$$\bar{X}_t(k) = \frac{1}{N_t(k)} \sum_{s=1}^t X_s(k) \mathbb{1}_{\{a_s=k\}} \quad (11)$$

is the empirical mean of the previous rewards up to the time t , and $N_t(k) = \sum_{s=1}^t \mathbb{1}_{\{a_s=k\}}$ is the number of times arm k has been selected up to time t , and $\mathbb{1}_{\{a_s=k\}}$ returns one if $a_s = k$ and zero otherwise. The second term in (10) is referred to as a padding function. It describes the uncertainty factor of the corresponding arm, which has high values for less selected arms and vice versa. The purpose of the padding function is to ensure an exploration-exploitation policy in which the less we have played an arm, the more uncertain we are about the calculated empirical mean and thus, a bigger padding function. This results in selecting the less selected arms due to their bigger padding functions, hence explore the other arms. A standard choice for the padding function is

$$c_t(k) = B \sqrt{\frac{\xi \log(t)}{N_t(k)}}, \quad (12)$$

where B is an upper-bound on the rewards and is selected as the maximum value of the observed reward through many trials of the algorithm. $\xi > 0$ is a parameter which is used to tune the algorithm to obtain the best results and is selected based on empirical applications of the algorithm [10]. Finally, for the next rounds, the arm yielding the maximum UCB index is selected and after the transmission through the selected arm, the corresponding index is updated. The selection policy for UCB, therefore, can be expressed as

$$a_t = \arg \max_{1 \leq k \leq N} \bar{X}_t(k) + c_t(k). \quad (13)$$

Theorem 1: Upper-confidence bound algorithm is optimal in the sense that its expected regret matches the lower bound regret of all policies for stationary bandit problems [8].

Theorem 1 denotes the optimality of UCB policy for stationary bandits. However, for non-stationary bandits UCB cannot be considered as an optimal policy. For non-stationary bandits the rewards are assumed to be non-stationary, therefore the optimal policy must have the ability of adaptation to the changes of the statistical characteristics of the rewards. For piece-wise stationary bandits, *discounted upper-confidence bound* (D-UCB) algorithm has been introduced in [9], where a discount factor $\gamma \in (0, 1)$ has been introduced to the UCB algorithm to mark the effects of the time in which the selected arm has been played. This means the more past actions do not have equal weights in calculating the empirical mean of the rewards, whereas the more recent selected actions weigh more in the calculation of the confidence bound index. Details of the D-UCB algorithm is described in Algorithm 1.

Algorithm 1 Discounted Upper-Confidence Bound

- 1: **for** $t = 1, \dots, N$ **do**
- 2: Let $a_t = t$, play and observe the reward.
- 3: Calculate the discounted empirical mean as

$$\bar{X}_t(\gamma, k) = \frac{1}{N_t(\gamma, k)} \sum_{s=1}^t \gamma^{t-s} X_s(k) \mathbb{1}_{\{a_s=k\}},$$

$$N_t(\gamma, k) = \sum_{s=1}^t \gamma^{t-s} \mathbb{1}_{\{a_s=k\}}.$$

- 4: Calculated the discounted padding function as

$$c_t(\gamma, k) = 2B \sqrt{\frac{\xi \log(n_t(\gamma))}{N_t(\gamma, k)}},$$

$$n_t(\gamma) = \sum_{k=1}^N N_t(\gamma, k).$$

- 5: Update the discounted UCB index $I_{k,t,\gamma} = \bar{X}_t(\gamma, k) + c_t(\gamma, k)$ of the arm.
 - 6: **end for**
 - 7: **for** $t = K + 1, \dots, T$ **do**
 - 8: Let $a_t = \arg \max_{1 \leq k \leq N} \bar{X}_t(\gamma, k) + c_t(\gamma, k)$, play and observe the reward.
 - 9: Update the discounted UCB index of the selected arm.
 - 10: **end for**
-

Theorem 2: Discounted upper-confidence bound algorithm is almost optimal in the sense that its expected regret matches the lower bound regret of all policies for piece-wise stationary bandit problems [9], [10].

Theorem 2 denotes that the D-UCB algorithm is proved to be almost optimal for piece-wise stationary bandits. However, in PLC network relay selection problem, the reward of each arm is denoted by the end-to-end capacity of that arm, which in turn depends inversely on the noise power. As mentioned in Section II, the noise in PLC networks is a cyclostationary process. We propose in this paper, two more appropriate variants of UCB algorithm, designed specifically for cyclostationary behavior of the PLC channel.

V. THE PROPOSED ALGORITHMIC SOLUTIONS

In a cyclostationary process, the statistical characteristics of the process repeat periodically. Therefore, in the calculation of the empirical mean and the padding function, considering all the previous actions with the same weight, as in UCB, is not an optimal policy. Furthermore, the mere consideration of the recent past actions as major contributors to the calculation of the confidence index, as in D-UCB, may result in a sub-optimal policy as well, since the far past actions in the same cycle and hence with the same statistical characteristics are neglected due to their low discount weight. To overcome this problem, we propose two novel algorithms, namely *cyclo-discounted upper-confidence bound* algorithm and *sinusoidal upper-confidence bound* algorithm. Furthermore, through simulation results, we show that for a cyclostationary system like PLC, these algorithms result in a better selection policy, and therefore a better performance.

A. The Proposed Cyclo-Discounted UCB Algorithm

Let us assume the period of the AC waveform of the power lines as T_{AC} with the noise power and hence the rewards of each arm as a cyclostationary process with T_{AC} duration of each cycle. Up to the time index $t = T$, the total number of complete cycles, denoted by P , can be calculated as

$$P = \left\lfloor \frac{T}{T_{AC}} \right\rfloor. \quad (14)$$

Furthermore, the empirical mean value of each arm at time t , as well as the padding function, are to be calculated in a way that all the last cycles are included in the calculation. The weighing factor should be chosen in a way that it involves the cyclostationary behavior of the reward. For this purpose, let us consider a single period of time with duration T_{AC} . In this period, the first samples will have lower weights and the last samples will have higher weights. In other words, we use a D-UCB discount factor for each period. For the last P complete periods, we calculate the empirical means as

$$\begin{aligned} \Xi_2 &= \Xi_2(t, T_{AC}, \gamma, k) \\ &= \sum_{p=1}^P \sum_{s=t-pT_{AC}}^{t-(p-1)T_{AC}} \gamma^{t-(p-1)T_{AC}-s} X_s(k) \mathbb{1}_{\{a_s=k\}}, \end{aligned} \quad (15)$$

where the term $\gamma^{t-(p-1)T_{AC}-s}$ applies a discounted factor for each period separately. For the incomplete period at the beginning of the time index (as depicted in Fig. 3), we calculate the corresponding portion of the empirical mean as

$$\begin{aligned} \Xi_1 &= \Xi_1(t, T_{AC}, \gamma, k) \\ &= \sum_{s=1}^{t-PT_{AC}} \gamma^{t-PT_{AC}-s} X_s(k) \mathbb{1}_{\{a_s=k\}}. \end{aligned} \quad (16)$$

Therefore, for $1 \leq t \leq T$, the empirical mean can be calculated as

$$\bar{X}_t(T_{AC}, \gamma, k) = \frac{\Xi_1 + \varkappa \Xi_2}{N_t(T_{AC}, \gamma, k)}, \quad (17)$$

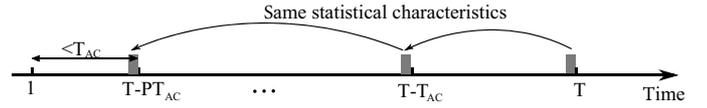


Fig. 3. The cyclostationary behavior of the channels with mains period T_{AC} .

where $\varkappa = \text{sign}(P)$, and $N_t(T_{AC}, \gamma, k)$ is described as

$$\begin{aligned} N_t(T_{AC}, \gamma, k) &= \sum_{s=1}^{t-PT_{AC}} \gamma^{t-PT_{AC}-s} \mathbb{1}_{\{a_s=k\}} \\ &+ \varkappa \sum_{p=1}^P \sum_{s=t-pT_{AC}}^{t-(p-1)T_{AC}} \gamma^{t-(p-1)T_{AC}-s} \mathbb{1}_{\{a_s=k\}}. \end{aligned} \quad (18)$$

The equations (15)–(18) are used to calculate the empirical mean so that the periodic elements at each cycle have more contribution to the final value. This results in a more accurate empirical mean for a cyclostationary process. Furthermore, the padding function is calculated in the same way as before, but with the new value of $N_t(T_{AC}, \gamma, k)$ which contains the cyclostationary weighing method as well. Finally, at time T , the arm with the highest UCB index is selected as the next route for transmission. This algorithm is summarized in Algorithm 2.

Algorithm 2 Proposed Cyclo-Discounted Upper-Confidence Bound

- 1: **for** $t = 1, \dots, N$ **do**
- 2: Let $a_t = t$, play and observe the reward.
- 3: Calculate the discounted empirical mean $\bar{X}_t(T_{AC}, \gamma, k)$ as described in (15)–(18).
- 4: Calculated the discounted padding function as

$$c_t(T_{AC}, \gamma, k) = 2B \sqrt{\frac{\xi \log(n_t(T_{AC}, \gamma))}{N_t(T_{AC}, \gamma, k)}},$$

$$n_t(T_{AC}, \gamma) = \sum_{k=1}^N N_t(T_{AC}, \gamma, k)$$

- 5: Update the cyclo-discounted UCB index $I_{k,t,\gamma,T_{AC}} = \bar{X}_t(T_{AC}, \gamma, k) + c_t(T_{AC}, \gamma, k)$ of the arm.
 - 6: **end for**
 - 7: **for** $t = K + 1, \dots, T$ **do**
 - 8: Let $a_t = \arg \max_{1 \leq k \leq N} \bar{X}_t(T_{AC}, \gamma, k) + c_t(T_{AC}, \gamma, k)$, play and observe the reward.
 - 9: Update the cyclo-discounted UCB index of the selected arm.
 - 10: **end for**
-

B. The Proposed Cyclic-Window UCB Algorithm

Similar to the approach in cyclo-discounted UCB algorithm, we propose another algorithm which its weight factor is periodic, hence adapted to the cyclostationary behavior of the channel. In cyclo-discounted UCB algorithm, the statistical characteristics of the reward function at time $t = T$ and $t = T - T_{AC}$ are the same, and therefore the weight factor at these times is at its maximum. However, at time $t = T - T_{AC} + 1$, the reward function enters the next cycle and results in a low value of weight factor. Although, the statistical characteristics of the reward function at time $t = T - T_{AC}$ and $t = T - T_{AC} + 1$

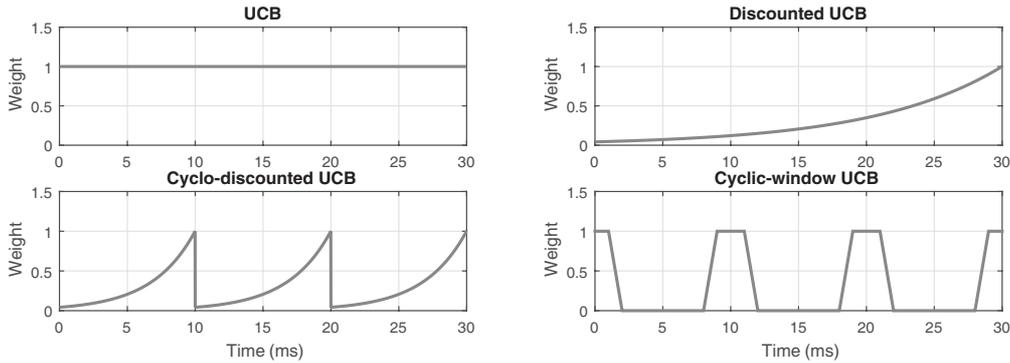


Fig. 4. Weight factors in confidence index calculation for UCB, discounted UCB, proposed cyclo-discounted UCB, and proposed cyclic-window UCB algorithms.

are not that different. This problem can be addressed with the proposed *cyclic-window UCB* algorithm.

In cyclic-window UCB algorithm, the empirical mean and the padding function are not discounted as in cyclo-discounted UCB, but are windowed periodically. The windowing period is chosen to be matched with the periodic behavior of the reward and hence matched with T_{AC} . The window size, W , is selected to maximize the effect of windowing. Formally, we can calculate the empirical mean as

$$\bar{X}_t(W, T_{AC}, k) = \frac{1}{N_t(W, T_{AC}, k)} \sum_{p=0}^P \sum_{s=1}^t w(s-t+pT_{AC}) X_s(k) \mathbb{1}_{\{a_s=k\}}, \quad (19)$$

where P is defined in (14), and $w(s)$ is the window function and is defined as

$$w(s) = \begin{cases} 1, & |s| < \frac{W}{2}, \\ 0, & \text{otherwise.} \end{cases} \quad (20)$$

The term $w(s-t+pT_{AC})$ in (19) denotes that the windowing is performed at the current time in addition to the multiple times of the mains period before the current time (see Fig. 4). The term $N_t(W, T_{AC}, k)$ can be described as

$$N_t(W, T_{AC}, k) = \sum_{p=0}^P \sum_{s=1}^t w(s-t+pT_{AC}) \mathbb{1}_{\{a_s=k\}}. \quad (21)$$

The weight factors in confidence index calculation of the proposed algorithms as well as that of the basic UCB and discounted UCB algorithms are depicted in Fig. 4. The proposed cyclic-window UCB algorithm is summarized in Algorithm 3.

The proposed algorithms consist of an initialization phase which its duration is proportional to the number of relays. Then at each frame length the end-to-end capacity has to be obtained and instantaneous rewards are calculated at the transmitter. The calculation of reward for the selected arm consists of a linear calculation of the empirical mean in addition to the calculation of the padding function which consists of a square root and a logarithm function. These calculations are only done in one arm at each frame time. The algorithms are bounded in the sense that after a limited amount of time, the arm with the best rewards can be detected and selected for most of the consecutive

Algorithm 3 Proposed Cyclic-Window Upper-Confidence Bound

- 1: **for** $t = 1, \dots, N$ **do**
- 2: Let $a_t = t$, play and observe the reward.
- 3: Calculate the discounted empirical mean $\bar{X}_t(W, T_{AC}, k)$ as described in (19)–(21).
- 4: Calculate the discounted padding function as

$$c_t(W, T_{AC}, k) = 2B \sqrt{\frac{\xi \log(n_t(W, T_{AC}))}{N_t(W, T_{AC}, k)}},$$

$$n_t(W, T_{AC}) = \sum_{k=1}^N N_t(W, T_{AC}, k)$$

- 5: Update the cyclic-window UCB index $I_{k,t,W,T_{AC}} = \bar{X}_t(W, T_{AC}, k) + c_t(W, T_{AC}, k)$ of the arm.
 - 6: **end for**
 - 7: **for** $t = K + 1, \dots, T$ **do**
 - 8: Let $a_t = \arg \max_{1 \leq k \leq N} \bar{X}_t(W, T_{AC}, k) + c_t(W, T_{AC}, k)$, play and observe the reward.
 - 9: Update the cyclic-window UCB index of the selected arm.
 - 10: **end for**
-

selections. However, the amount of time needed for this convergence, denoted by $\tau_{alg}(N)$, is directly proportional to the number of relays N (see Fig. 10). Let us denote the time in which the relay channels remain in a particular state as τ_{ch} . In order to have a working algorithm, change of the channel in time must happen slower than the convergence time of the algorithm and $\tau_{alg}(N) < \tau_{ch}$ must hold. The exhaustive search method in the simulations, assumes the perfect CSI at transmitter and expectedly the returned reward is higher than learning algorithms. However, acquiring CSI is much more complicated than feeding back the observed reward. The reason for that is that the PLC channel is time-variant and frequency-selective. Therefore, pilot signals should be transmitted in all the subcarriers at pre-defined time intervals throughout the transmission. Moreover, the amount of overhead which this brings increases linearly by the number of available relays, since in order to react to the changes in the environment all the possible routs should be evaluated. On the other hand, feeding back the reward data can take place on the free bits of the ACK (acknowledgment) packet which is already being fed back to the transmitter and is not dependent on the number of available relays.

Parameters	Value
Number of subcarriers (samples)	128
Number of used subcarriers (samples)	102
Number of cyclic prefix (samples)	30
OFDM interval (μs)	640
Modulation	QPSK
Baseband sampling frequency (MHz)	0.6
Inter-carrier spacing (kHz)	4.6875

Table 1. System parameters.

VI. NUMERICAL RESULTS

We consider a two-hop cooperative communication with a PLC channel as described in Section II. The OFDM parameters of the PLC system is listed in Table 1. The number of available relays is considered to be 6 nodes, from which a single node is selected for transmission at each transmission instant. The only difference between the available relay nodes is the corresponding link rate, which in turn is dependent on the channel response as well as the noise power spectral density of the PLC channel. Given the frequency response $H_{n,n+1}(f)$ for a link from node n to node $n+1$, the link rates are computed as [7]

$$C_{n,n+1} = \int_{f_1}^{f_2} \log_2 \left(1 + \frac{S_T |H_{n,n+1}(f)|^2}{N_0 \Gamma} \right) df, \quad (22)$$

where S_T is the transmitter-side power spectral density, N_0 is the receiver-side noise power spectral density, and $\Gamma = 10$ is the margin taking into account the gap between information-theoretic capacity and achievable rate using practical coding and modulation schemes.

Different relay selection policies are applied to the PLC system to acquire a quantitative comparison between the existing policies in the literature and the selection policies described in the proposed algorithms. In the first method, an exhaustive search is performed to find the best available relay node for transmission with the assumption of perfect CSI. This returns the highest reward at each time instance, or equivalently the maximum achievable reward for other algorithms which is a function of the channel conditions. In the second method, a fixed relay node is selected at the beginning of the transmission in a random manner and the selected node is used for the entire transmission time, regardless of the changes in the channel. In the third method the relay node is selected randomly for each instant of transmission. Therefore the all the available nodes have the chance of being selected, but the changes in channel does not play a role in the selection procedure. The fourth and the fifth methods of relay selection are the basic UCB and Discounted-UCB algorithms described in Section IV. Finally, the sixth and seventh selection policies are the proposed algorithms described in Section V.

Figs. 5 and 6 show the average reward and the accumulated regret of the above mentioned relay selection policies, respectively. It can be seen that the best performance belongs to the exhaustive search method, as expected. The worst performance is from the random selection of the relay node and the fixed relay node, due to the fact that the changes in channel is disregarded in

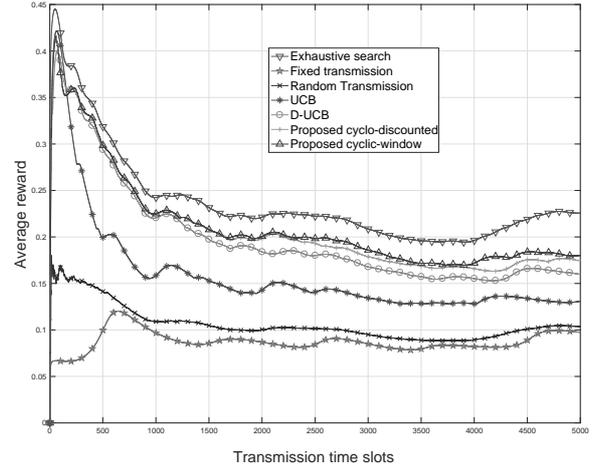


Fig. 5. Average reward comparison between different relay selection policies.

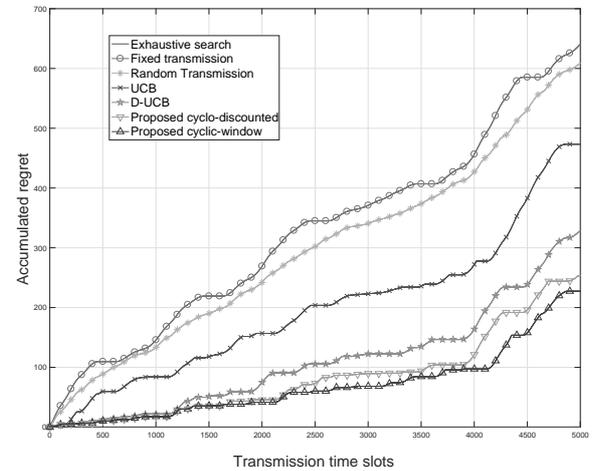


Fig. 6. Accumulated regret comparison between different relay selection policies.

these methods. Basic UCB and discounted-UCB algorithms provide a significant performance gain in terms of average reward and demonstrate lower regrets throughout the transmission. It can be seen that the proposed algorithms can further improve the performance and obtain more reward and lower regret compared to the existing algorithms. Although, at the beginning of the operation, the padding functions are still large for all the arms and therefore a uniform weight as in UCB outperforms the discounted weights of the proposed algorithms.

Fig. 7 illustrates the percentage of correct selection in different transmission policies. This percentage is calculated after the entire duration of the transmission, by comparing the selected relay nodes of each method to the selected relay nodes of the exhaustive search method as a measure of reference. It can be seen that the basic MAB algorithms result in a higher percentage of correct selection and the proposed MAB algorithms can achieve yet higher percentages.

The first proposed algorithm, namely the cyclo-discounted

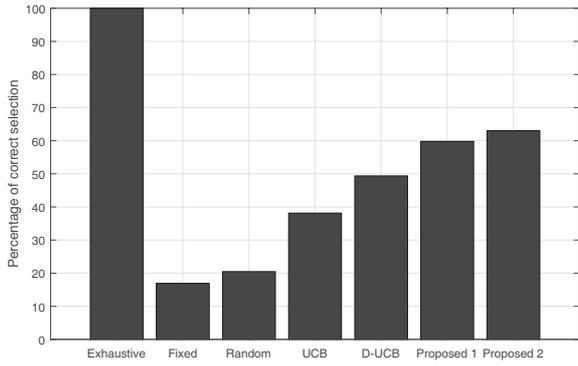


Fig. 7. Percentage of correct selection of different transmission policies.

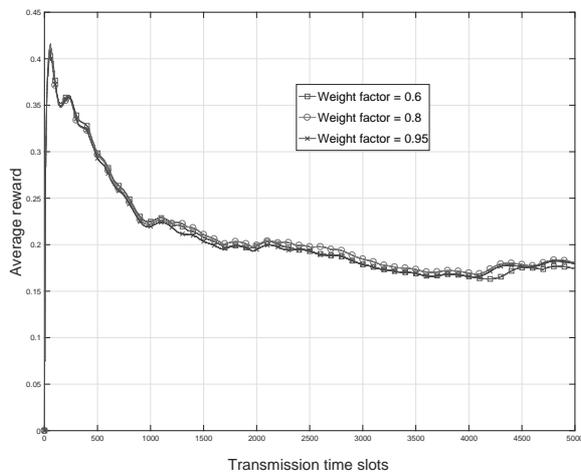


Fig. 8. Impact of weight factor in the proposed cyclo-discounted UCB algorithm.

UCB algorithm, uses a weight factor as a parameter of cyclic weights for the calculation of the UCB index, whereas the second proposed algorithm, namely the cyclic-window UCB algorithm, uses a window size parameter for the calculation of the UCB index. Figs. 8 and 9 demonstrates the impact of different weight factors and window sizes, respectively. Compared with Fig. 5 we can realize that the change in window size or weight factor can hugely affect the performance of the algorithms. Fig. 10 demonstrates the effect of the number of relays on the cyclic-window algorithm. It can be seen that in the same frame of time, with higher the number of relays it takes longer for the algorithm to adapt itself to the environment and the accumulated regret grows by the number of available relays.

This paper has provided results of a two-hop relay selection scheme in PLC networks. However, the use of learning algorithms can be further generalized to multi-hop relay selection problems as well. In the case of M -hop transmission, with N_i , $i \in \{1, \dots, M-1\}$ available relays at each hop, the number of available options raise to $\prod_{i=1}^{M-1} N_i$ which makes the time needed for finding the best path longer. Finding the best strategy for multi-hop relay selection can be regarded as a future

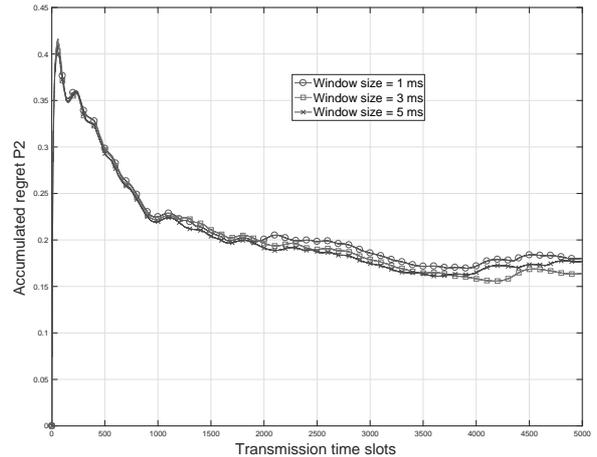


Fig. 9. Impact of window size in the proposed cyclic-window UCB algorithm.

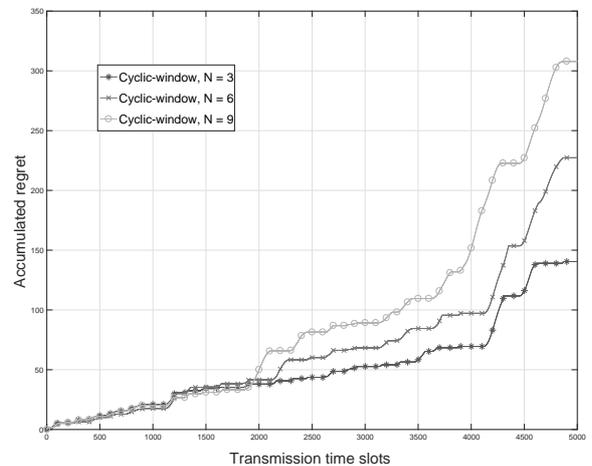


Fig. 10. The proposed cyclic-window algorithm for different number of available relays.

work for this problem. Moreover, a generalized version of the proposed algorithm can be developed to further increase the reliability of the transmission. In this case, a multi-relay selection approach based on the learning algorithm is applied and at the receiver an appropriate combining technique is performed to further exploit the combining diversity among the best chosen relays. This strategy can further improve the reliability and decrease the probability of transmission in a less optimal route.

VII. CONCLUSION

In this paper, we discussed the problem of relay selection in cooperative power line communication networks. The proper relay selection requires the availability of channel side information and the produced overhead of channel estimation makes it not realizable in a time-variant PLC environment. We introduced a class of machine learning algorithms, namely UCB algorithms in MAB model, to overcome this problem and improve

the chances of selecting the best relay for transmission without channel side information. Furthermore, we proposed two new algorithms to further improve the selection policy by taking advantage of the periodicity of the synchronous impulsive noise of power line channels. We demonstrated through simulation results the supremacy of the proposed algorithms to the other policies of relay selection.

ACKNOWLEDGMENTS

The authors thank S. Maghsudi for discussions on the use of upper confidence bound algorithm. The authors also thank anonymous reviewers for their constructive comments.

REFERENCES

- [1] G. Bumiller, "Single frequency network technology for medium access and network management," in *Proc. IEEE ISPLC*, 2002.
- [2] G. Bumiller, L. Lampe, and H. Hrasnica, "Power line communication networks for large-scale control and automation systems," *IEEE Commun. Mag.*, pp. 106–113, 2010.
- [3] L. Lampe, R. Schober, and S. Yiu, "Distributed space-time coding for multihop transmission in power line communication networks," *IEEE J. Sel. Areas Commun.*, pp. 1389–1400, 2006.
- [4] A. Tonello, F. Versolatto, and S. D' Alessandro, "Opportunistic relaying in in-home PLC networks," in *Proc. IEEE GLOBECOM*, 2010.
- [5] V. Balakirsky and A.J.H. Vinck, "Potential performance of PLC systems composed of several communication links," in *Proc. IEEE ISPLC*, 2005.
- [6] L. Lampe and A.J.H. Vinck, "On cooperative coding for narrow band PLC networks," *AEU Int. J. Electron. Commun.*, pp. 681–687, 2011.
- [7] L. Lampe and A.J.H. Vinck, "Cooperative multihop power line communications," in *Proc. IEEE ISPLC*, 2012.
- [8] R. Agrawal, "Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem," *Advances in Applied Probability*, pp. 1054–1078, 1995.
- [9] L. Kocsis and C. Szepesvári, "Discounted UCB," in *Proc. PASCAL Challenges Workshop*, 2006.
- [10] A. Garivier and E. Moulines, "On upper-confidence bound policies for non-stationary bandit problems," *Arxiv preprint*, 2008.
- [11] S. Maghsudi and S. Stanczak, "Relay selection problem in wireless networks: A solution concept based on stochastic bandits and calibrated forecasters," in *Proc. SPAWC*, pp. 385–389, 2013.
- [12] S. Maghsudi and S. Stanczak, "Relay selection with no side information: An adversarial bandit approach," in *Proc. WCNC*, pp. 715–720, 2013.
- [13] B. Nikfar, S. Maghsudi, and A.J.H. Vinck, "Multi-armed Bandit Channel Selection for Power Line Communication," in *Proc. IEEE SmartGrid-Comm*, 2015.
- [14] F. Gruber and L. Lampe, "On PLC channel emulation via transmission line theory," in *Proc. IEEE ISPLC*, 2015.
- [15] H.C. Ferreira, L. Lampe, J. Newbury, and T. Swart, "Power line communications: Theory and applications for narrowband and broadband communications over power lines," *Wiley*, 2011.
- [16] M. Katayama, T. Yamazato, and H. Okada, "A mathematical model of noise in narrowband power line communication systems," *IEEE J. Sel. Areas Commun.*, pp. 1267–1276, 2006.
- [17] A. Nosratinia, T. Hunter, and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Commun. Mag.*, pp. 74–80, 2004.
- [18] J. Laneman, D. Tse, and G. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, pp. 3062–3080, 2004.
- [19] J. Boyer, D. Falconer, and H. Yanikomeroglu, "Multihop diversity in wireless relaying channels," *IEEE Trans. Commun.*, pp. 1820–1830, 2004.
- [20] O. Oyman, J. Laneman, and S. Sandhu, "Multihop relaying for broadband wireless mesh networks: From theory to practice," *IEEE Commun. Mag.*, 2007.



Babak Nikfar received his B.Sc. degree in Electrical Engineering with a major in Telecommunications from Iran University of Science and Technology, Tehran, Iran in 2008 and the M.Sc. degree in Digital Communications from the University of Kiel, Kiel, Germany in 2011. He is currently working toward the Ph. D degree with the Digital Communications Group, Department of Digital Signal Processing of the University of Duisburg-Essen, Duisburg, Germany.



A. J. Han Vinck is a Senior Professor in Digital Communications at the University of Duisburg-Essen, Germany, specialising in Information and Communication theory, Coding and Network aspects in digital communications. He held positions of professional responsibility including the Director of the Institute for Experimental Mathematics in Essen, founding Chairman of the IEEE German Information Theory chapter, President of the IEEE Information Theory Society (2003) and President of the Shannon, Leibniz and Gauss foundations for the stimulation of research in

the field of Information theory and Digital Communications. He received a number of accolades including the election by the IEEE as Fellow for his "Contributions to Coding Techniques", appointed as Distinguished Lecturer for the Information Theory Society as well as for the Communications Society of the IEEE, the IEEE ISPLC 2006 Achievement award for contributions to Power Line Communications, the SAIEE annual award for the best paper published in the SAIEE Africa Research Journal in 2008 and the 2015 Aaron D. Wyner Distinguished Service Award for longstanding contributions to the IEEE Information Theory society. He was instrumental in the organisation of research forums including the IEEE Information Theory workshops and symposia Japan-Benelux workshops on Information theory (now Asia-Europe workshop on "Concepts in Information Theory") and the International Winterschool on Coding, Cryptography and Information theory in Europe. He is Author of the book "Coding Concepts and Reed-Solomon Codes".