Nomikos, Nikolaos; Talebi, Sadegh; Wichman, Risto; Charalambous, Themistoklis

Bandit-based relay selection in cooperative networks over unknown stationary channels

Published in:
Proceedings of the 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing, MLSP 2020

DOI:
10.1109/MLSP49062.2020.9231604

Published: 01/09/2020

Document Version
Peer reviewed version

Please cite the original version:
Nomikos, N., Talebi, S., Wichman, R., & Charalambous, T. (2020). Bandit-based relay selection in cooperative networks over unknown stationary channels. In *Proceedings of the 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing, MLSP 2020* [9231604] (IEEE International Workshop on Machine Learning for Signal Processing). IEEE. https://doi.org/10.1109/MLSP49062.2020.9231604

BANDIT-BASED RELAY SELECTION IN COOPERATIVE NETWORKS OVER UNKNOWN STATIONARY CHANNELS

Nikolaos Nomikos, Sadegh Talebi, Risto Wichman, and Themistoklis Charalambous

ABSTRACT

In recent years, wireless node density has increased rapidly, as more base stations, users, and machines coexist. Exploiting this node density, cooperative relaying has been deployed to improve connectivity throughout the network. Such a configuration, however, often demands relay scheduling, which comes with increased channel estimation and signaling overheads. To reduce these overheads, in this paper, we propose low-complexity relay scheduling mechanisms with the aid of a multi-armed bandit (MAB) framework. More specifically, this MAB framework is used for relay scheduling, based only on observing the acknowledgements/negative-acknowledgements (ACK/NACK) of packet transmissions. Hence, a bandit-based opportunistic relay selection ($\mathrm{BB-0RS}$) mechanism is developed, recovering eventually the performance of classical opportunistic relay selection ($\mathrm{0RS}$) when channel state information (CSI) is available without requiring any CSI. In addition, a distributed implementation of $\mathrm{BB-0RS}$ is presented, herein called $\mathrm{d-BB-0RS}$, where distributed timers are used at the relays for relay selection, thus reducing the signaling overhead significantly. $\mathrm{BB-0RS}$ is compared to optimal scheduling with full CSI and the negligible performance gap is compensated by the low-complexity low-overhead implementation, while it surpasses the performance of $\mathrm{0RS}$ with outdated CSI.

Index Terms— Relay selection, machine learning, multi-armed bandits, upper confidence bound policies.

1. INTRODUCTION

Fifth generation (5G) networks comprise dense topologies where users and machines compete for wireless resources. In such environments, excessive signaling and feedback overheads threaten the network's performance [1], necessitating a shift towards distributed solutions.

Recently, machine learning techniques have been proposed for low-complexity coordination in wireless networks (see, for example, [2–4] and references therein). In [5], The MAB framework was adopted in several 5G cases. MAB enables a player (user) to pick an action from a given set of actions, aiming to maximize her cumulative expected reward. As MAB takes into account the uncertainties involved in the problem, it can be of great importance for distributed resource allocation, involving channels, relays, power, and energy.

Various works have studied user scheduling and channel access through the MAB framework. In a multi-user network, relay selection, as a stochastic MAB game, was pioneered by Maghsudi *et al.* [6], in which side information on the actions of other users was not available. Thus, selection and assignment problem was formulated as an adversarial multi-player MAB game and it was shown that the proposed selection strategy ensured that the empirical frequencies of the game converged to a correlated equilibrium. In the case where side information is available, the authors in [7] based selection on a calibrated forecaster, predicting the action of the other users. Furthermore, the exploration-exploitation trade-off was balanced, obtaining asymptotically, the maximum achievable accumulated reward.

The work in [8] studied cooperative spectrum sharing without information on the performance of the secondary users. The proposed solutions adopted Markovian MAB and the upper confidence bound (UCB) algorithm, achieving low-complexity and superior performance compared to exploration-exploitation ϵ-greedy algorithm. Then, the Markovian MAB has been deployed when secondary users independently access the spectrum [9] and an online learning policy considering channel quality and interference levels was developed, showing logarithmic order regret. Finally, MAB games have been studied for device-to-device (D2D) channel and mode selection in the seminal papers [10, 11]. Calibrated forecasting was used, while mode selection was modeled as a multi-player MAB game, where power consumption and throughput corresponded to the cost and the reward of each selection, respectively.

Here, the MAB framework is also adopted for low-overhead relay scheduling. The setup is similar to that of [7]. However, contrary to [7] in which the relays were always able to forward the data, in our case both hops are prone to

N. Nomikos is with the Department of Information and Communication Systems Engineering, University of the Aegean, Samos, Greece. *Email:* `nnomikos@aegean.gr`.

S. Talebi is with the Department of Computer Science, University of Copenhagen, Copenhagen, Denmark. *Email:* `sadegh.talebi@di.ku.dk`.

R. Wichman and T. Charalambous are with the School of Electrical Engineering, Aalto University, Espoo, Finland. *Emails:* `{name.surname}@aalto.fi`

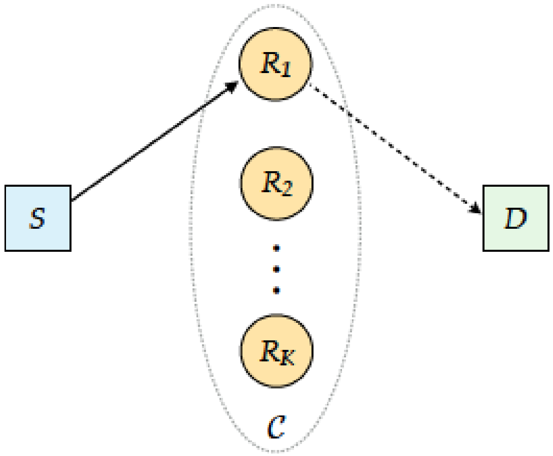

Fig. 1. The two-hop relay-assisted topology where a source S communicates with a single destination D via a cluster $\mathcal{C}$ of half-duplex decode-and-forward relays R_k, $k \in \mathcal{C}$.

outages. Our contributions are as follows:

- A bandit-based opportunistic relay selection (BB − 0RS) is proposed, based on ACK/NACK observations such that the nodes in the network (source and relays) do not need to estimate the forward channels.
- Network coordination through distributed timers at the relays is integrated into the MAB framework, reducing the complexity, compared to [6, 7], in which each source has knowledge of the action set;
- Comparisons of BB − 0RS based on different UCB policies and the optimal 0RS are presented, showing a promising performance-complexity trade-off.

The remainder of this paper is organized as follows. In Section 2, we introduce the system model. In Section 3, we provide in detail the MAB modeling of the selection process. The proposed bandit-based relay selection mechanism, BB − 0RS, is described in Section 4, while performance evaluation is provided in Section 5. Finally, conclusions and future directions are given in Section 6.

2. SYSTEM MODEL

We consider a relay-assisted two-hop network consisting of a source, S, a destination, D, and a cluster $\mathcal{C}$ of K half-duplex (HD) decode-and-forward (DF) relays $R_k \in \mathcal{C}$ $(1 \leq k \leq K)$, as depicted in Fig. 1. Due to severe fading, the direct source-destination ($\{S \to D\}$) link does not exist and communication is only established via relaying.

For modeling the evolution of the radio channels over time, we consider a time slotted system, where the duration of a slot corresponds to the transmission of a single packet duration (e.g., fixed-size packets). A time-slot can, in general, span more than one packet. At any arbitrary time-slot t, the wireless channel quality is degraded by additive white Gaussian noise (AWGN) and frequency non-selective Rayleigh block fading, according to a complex Gaussian distribution with zero mean and variance σ_{ij}^2 for the $\{i \to j\}$ link. AWGN is assumed to be normalized with zero mean and unit variance, the complex channel coefficient for the $\{i \to j\}$ link is denoted by h_{ij}, and the channel gain, $g_{ij} \triangleq |h_{ij}|^2$, is exponentially non-identically distributed, reflecting an asymmetric topology. Thermal noise variance at receiver j is denoted by η_j, and it is assumed to be AWGN and the same at all nodes. We assume that no channel knowledge is available and the channel conditions evolve over time according to an independent not necessarily identically distributed process whose average is initially unknown. This corresponds to scenarios where the average channel conditions evolve relatively slowly, in the sense that the link allocation can be updated several times before this average exhibits significant changes.

Moreover, the source is assumed to be saturated and transmits with a fixed rate r_0. In general, a successful transmission from a transmitter i to its corresponding receiver j takes place when the signal-to-noise ratio (SNR) at the reception, denoted by Γ_j, is greater than or equal to the *capture ratio* γ_j. Therefore, we require that

$$\Gamma_j(P_i) \triangleq \frac{g_{ij} P_i}{\eta_j} \geq \gamma_j. \tag{1}$$

Link $\{i \to j\}$ is in outage if $\Gamma_j(P) < \gamma_j$, i.e., $\frac{g_{ij} P_i}{\eta_j} < \gamma_j$, and the probability of outage is given by

$$\bar{p}_{ij} = \mathbb{P}\left[g_{ij} < \frac{\gamma_j \eta_j}{P_i} \right]. \tag{2}$$

This framework is equivalent to the *capture model*. Hence, the instantaneous SNR from S to R_j when relay R_j is selected for reception is expressed as

$$\Gamma_{R_j}(P_S) = \frac{g_{SR_j} P_S}{\eta_{R_j}} \geq \gamma_{R_j} \ , \tag{3}$$

and, equivalently, the instantaneous SNR from R_j to D when relay R_j is selected for transmission is given by

$$\Gamma_D(P_{R_j}) = \frac{g_{R_j D} P_{R_j}}{\eta_D} \geq \gamma_D \ . \tag{4}$$

Re-transmissions rely on ACKs/NACKs with short-length error-free packets over a separate narrow-band channel.

3. MAB MODELING

3.1. 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 choosing the best decision more often to gain more reward. For a thorough discussion on the topic, see, for example, [12, 13]. In the classical stochastic MAB problem, introduced by Robbins [14], a player has access to a finite set of arms, and to each arm 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 $U_{j,t}$. In our setup, each arm corresponds to one of the $2K$ available links $\ell \in \mathcal{L}$, $|\mathcal{L}| = 2K$, in our network setup.

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 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 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 sequential relay selection policy that minimizes 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 their seminal paper, Lai and Robbins [15] 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.

3.2. Upper Confidence Bound Policies

A big class of policies for MAB problems, whose regret grows logarithmically with time horizon, are based on the *optimism in the face of uncertainty* principle (or for short, the *optimistic* principle) proposed by Lai and Robbins [15]. The underlying idea of 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., $n_{j,t} := \sum_{s=1}^{t} \mathbb{1}_{\{I_s=j\}}$, where $\mathbb{1}_A$ denotes the indicator function of the event A. We let $\widehat{q}_{j,t}$ represent the empirical average reward of arm j built using the observations from j up to t:

$$\widehat{q}_{j,t} = \frac{1}{n_{j,t}} \sum_{s=1}^{t} r_{j,s} \mathbb{1}_{\{I_s=j\}}, \quad (7)$$

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:

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

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.

3.2.1. UCB1 [16]

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:

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

3.2.2. KL-UCB [17]

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

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

[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 $\alpha > 0$.

where $\mathrm{kl}(x, y)$ is the Kullback-Leibler divergence between two Bernoulli distributions with means x and y: $\mathrm{kl}(x, y) := x \log\left(\frac{x}{y}\right) + (1 - x)\log\left(\frac{1-x}{1-y}\right)$. When the reward distribution of arms are Bernoulli distributions, KL-UCB achieves the problem-dependent lower bound (6), and is hence said to be *asymptotically optimal*[2]. We remark that computing $\overline{q}_{j,t}^{\texttt{KL-UCB}}$ amounts 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.

3.2.3. KL-UCB^{++} [18]

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

$$\overline{q}_{j,t}^{\texttt{KL-UCB}^{++}} = \sup\left\{\lambda \in [\widehat{q}_{j,t}, 1] : \mathrm{kl}(\widehat{q}_{j,t}, \lambda) \leq g(n_{j,t})/n_{j,t}\right\}, \tag{10}$$

where

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

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

4. ONLINE LEARNING FOR CHANNEL ALLOCATION

4.1. Online Learning Model

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

$$r_{j,t} = \begin{cases} 1, & \text{if packet received successfully,} \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Hence, the sequence $(r_{j,t})_{t\geq 1}$ of rewards of channel j follows a Bernoulli distribution, whose mean corresponds to the probability of successful transmission over j. The rewards are assumed to be independent across time and channels.

We consider a scenario with stationary success probabilities. In this case, success probabilities of various channels are assumed to be fixed but unknown. Hence, for each j, $(r_{j,t})_{t\geq 1}$ is a sequence of i.i.d. Bernoulli random variables with $\mathbb{E}[r_{j,t}|\mathcal{F}_{t-1}] = q_j$ for all t, where $\mathcal{F}_{t-1}$ denotes the set of channels chosen by the algorithm before round t, and their realized rewards.

4.2. Online Learning Algorithm

We are now ready to describe our learning algorithm. After the initial exploration phase, the channel with the best quality is exploited for minimizing the regret. Herein, the relay selection problem differs from the classical channel allocation problem, as explained next.

If the success probability of various channels q_j were known, one could use the timer-based mechanism introduced in [19], which provides a distributed solution to the relay selection problem. In this mechanism each relay R_j sets its local timer as

$$\tau_{j,t} = \frac{\lambda}{q_j}, \tag{12}$$

where λ is a constant shared among all relays.

When success probabilities are unknown, one cannot implement the aforementioned time-based mechanism. To accommodate this situation, one may use the empirical estimate of q_j. This approach may however fail to balance exploration-exploitation trade-off, and a sub-optimal channel may thus be played most of the time. As a result, the regret will grow *linearly* with time. In order to come up with a solution with sublinear regret, we propose the following estimate for the timer:

$$\tau_{j,t} = \frac{\lambda}{\overline{q}_{j,t}}, \tag{13}$$

where $\overline{q}_{j,t}$ denotes a UCB for q_j.

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

$\mathrm{BB-0RS}$ based on distributed channel access is given in Algorithm 1.

Remark 1. *Note that $K - 1$ relays that were not selected can enter to discontinuous reception (DRX) mode during the two time slots when the transmission over the selected links $\{S \to R_j\}$ and $\{R_j \to D\}$ takes place. After the transmission has completed, all relays start their timers and wake up to listen to the control channel to select the relay for the next round. Discontinuous reception is an important energy saving feature in long-term evolution (LTE) mobile networks [20].*

5. PERFORMANCE EVALUATION

Here, comparisons are presented in terms of average throughput and relay selection over time. Two $\mathrm{BB-0RS}$ versions are included, based on UCB1 [16] and kl-UCB^{++} [18]. As a performance upper bound, the best relay selection (BRS) with full CSI is considered [19]. Also, BRS with outdated CSI (oCSI) with accuracy, characterized by $\rho = 0.5$ [21]

[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 $v \geq u$.

```
Algorithm 1 Timer-based channel access r
Input: constant value for timer setup $\lambda$.
for $t = 1, 2, \ldots$ do
    compute $\widehat{q}_{j,t}$ (7) and then $\bar{q}_{j,t}$ according
     start timer $\tau_{j,t}$ (13)
     if $\tau_{j,t} \neq 0$ and timer is running then
        listen for signals
         if signal is received then
            | freeze $\tau_{j,t}$ and back off
        end
    else if $\tau_{j,t} = 0$ then
        send flag (so that other relays $i$ free
         receive packet from $S$ and transmi
         $n_{j,t+1} \leftarrow n_{j,t} + \mathbb{1}_{\{I_t=j\}}$ for all $j$
         if transmission is successful then
            | $r_{j,t} = 1$
        end
    end
end
```

and random selection are examined. Fo
zon over which the relays are evaluate
ward is equal to 10^5 time-slots for each
Moreover, a fixed transmission rate r_0
sidered in a topology with $K = 3$ re
wireless environment, a scenario corres
stochastic bandits where channel statis
for the whole transmission duration is
more, the two-hop relay-assisted topol
highly asymmetric with one relay prov
nificantly higher channel gain compare
relays.

Fig. 2 includes average throughput results for the stationary case. It can be observed that random relay scheduling offers the worst performance as often the relays providing weak channels are selected. After, BRS with outdated CSI provides reduced throughput, since it performs similarly to single relaying, as the transmit SNR increases [21]. More importantly, the two $\mathrm{BB}-0\mathrm{RS}$ policies exhibit a small performance gap with respect to BRS, with the kl-UCB^{++} version slightly providing improved throughput, as it adapts better to erroneous decisions.

Next, Fig. 3 shows the relay selection process over the transmission duration for the two $\mathrm{BB}-0\mathrm{RS}$ versions for a transmit SNR of 20 dB. It can be seen that both UCB policies converge in selecting R_1 providing the highest channel gain. It is noteworthy that kl-UCB^{++} admits less relay switchings, occurring mostly in the initial time-slots. As a result, kl-UCB^{++} is suitable for scenarios with significantly smaller game horizon, thus having higher practical interest.

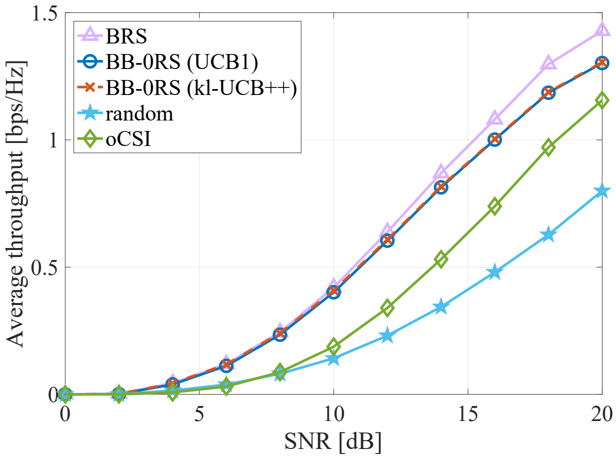

Fig. 2. Average throughput comparison for the stationary case.

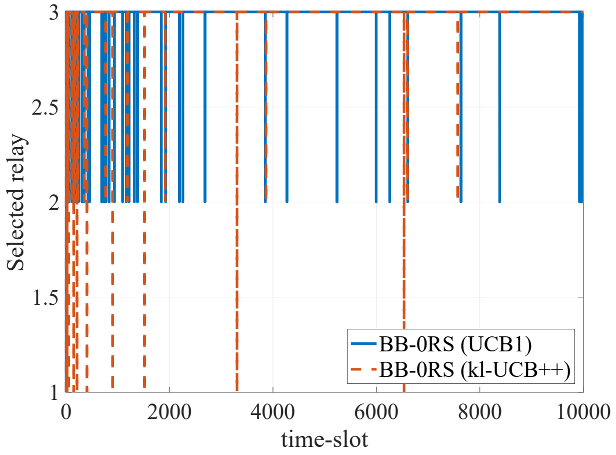

Fig. 3. Relay selection over time for the stationary case.

6. CONCLUSIONS AND FUTURE DIRECTIONS

6.1. Conclusions

Relay selection is an important problem in dense wireless networks, introducing significant coordination overheads. Aiming to facilitate this process, we have adopted the MAB framework on a stochastic setting where the channel conditions are stationary and developed relevant centralized and distributed algorithms. The learning process relied only on ACK/NACK observations, determining the best relay to establish end-to-end connectivity. Performance evaluation showed that the proposed algorithms follow closely the scheduling with full channel state information knowledge for different wireless environments.

6.2. Future Directions

Part of ongoing research includes adversarial scenarios where no assumptions are made regarding the evolution of channel conditions.

As a future direction, cases where channels have memory and, hence, the state of the machines advances to a new one, according to a Markov chain with rewards depending on the current state will be investigated. Such a framework can be studied using restless bandits, in which the the states of non-played arms can also evolve over time; see, e.g., [22].

7. REFERENCES

[1] 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, Thirdquarter 2019.

[2] 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.

[3] 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.

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

[5] 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.

[6] S. Maghsudi and S. Stańczak, "Relay selection with no side information: An adversarial bandit approach," in *Proc., IEEE Wireless Commun. and Netw. Conf.*, April 2013, pp. 715–720.

[7] ——, "Relay selection problem in wireless networks: A solution concept based on stochastic bandits and calibrated forecasters," in *Proc., IEEE Signal Proc. Adv. in Wireless Commun. (SPAWC)*, June 2013, pp. 385–389.

[8] M. López-Martínez, J. J. Alcaraz, L. Badia, and M. Zorzi, "A superprocess with upper confidence bounds for cooperative spectrum sharing," *IEEE Trans. Mobile Comput.*, vol. 15, no. 12, pp. 2939–2953, Dec. 2016.

[9] N. Modi, P. Mary, and C. Moy, "QoS driven channel selection algorithm for cognitive radio network: Multi-user multi-armed bandit approach," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 1, pp. 49–66, March 2017.

[10] 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, March 2015.

[11] S. Maghsudi and D. Niyato, "On transmission mode selection in D2D-enhanced small cell networks," *IEEE Wireless Commun. Lett.*, vol. 6, no. 5, pp. 618–621, Oct. 2017.

[12] S. Bubeck, N. Cesa-Bianchi *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[13] T. Lattimore and C. Szepesvári, "Bandit algorithms," 2020 (accessed May 2, 2020). [Online]. Available: https://tor-lattimore.com/downloads/book/book.pdf.

[14] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

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

[16] 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.

[17] 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.

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

[19] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 659–672, March 2006.

[20] E. Dahlman, S. Parkvall, and J. Sköld, *4G LTE/LTE-Advanced for Mobile Broadband*. Academic Press, 2011.

[21] 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, June 2009.

[22] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5588–5611, Aug. 2012.