Dynamic pilot allocation over Markovian fading channels: A restless bandit approach

Maialen Larrañaga, Mohamad Assaad, Apostolos S Destounis, Georgios S Paschos

Dynamic pilot allocation over Markovian fading channels: A restless bandit approach

Maialen Larrañaga*, Mohamad Assaad*, Apostolos Destounis†, Georgios S. Paschos †
* Laboratoire des Signaux et Systemes (L2S, CNRS), CentraleSupelec, Gif-sur-Yvette, France.
†Huawei Technologies & Co., Mathematical and Algorithmic Sciences Lab, Boulogne Billancourt, France.

***Abstract*—We investigate a pilot allocation problem in wireless networks over Markovian fading channels. In wireless systems, the Channel State Information (CSI) is collected at the Base Station (BS) through either a feedback channel (FDD mode) or a pilot-aided channel estimation method (TDD mode). This paper focuses on the latter. Typically, there are less available pilots than users, hence at each slot the scheduler needs to decide an allocation of pilots to users with the goal of maximizing the long-term average throughput. A trade-off emerges between exploiting users with up-to-date CSI for immediate gains or, exploring users with outdated CSI for a potential larger future gain. As we show, the arising pilot allocation problem is a restless bandit problem and thus its optimal solution is out of reach. In this paper, we propose a Lagrangian relaxation approach to obtain a Whittle index policy, which represents a low-complexity heuristic solution with remarkably good performance.**

I. Introduction

In order to support applications with large data traffic rates in the downlink, future generations of communication networks will support technologies such as multiple input multiple output (MIMO) possibly with massive antenna installations, e.g., [1]. The performance of these techniques critically depends on acquiring accurate channel state information (CSI) at the transmitter, which is then used to precode the transmitting signals and null the interference at the receivers [1].

In practice wireless channels are highly volatile, and CSI needs to be acquired very frequently. Furthermore, in both FDD (Frequency Division Duplex) and TDD (Time Division Duplex) systems, only a minority of the users can be selected to provide CSI to the base station at each given time, since the resources used for CSI acquisition reduce the system efficiency. In this paper, we focus on pilot-aided CSI acquisition proposed for TDD systems. However, we mention that our framework can be applied directly to the CSI feedback context (i.e. FDD) as well.

For TDD systems downlink CSI is inferred by the uplink training symbols and the use of the reciprocity property of the channel; the process is as follows. The BS allocates the M available pilot sequences to M users out of the total N users in the system. The chosen users transmit the training symbols to the BS which provides uplink CSI information. Last, the base station estimates the downlink CSI exploiting the channel reciprocity. For the estimation to be successful, M needs to be small to avoid the pilot contamination issue. Hence in systems with a large number of users it is expected that $M < N$.

It has been observed that once a channel is measured and its CSI is acquired, the channel coefficients remain the same for some period of time termed *channel coherence time*. In fact, sophisticated transmission schemes can exploit this channel property to avoid requesting CSI constantly. Therefore, the problem under study here is to exploit the channel memory to optimize the allocation of pilots for CSI acquisition. To model the channel memory we consider channels that evolve according to a Markovian stochastic process and we study the pilot allocation problem over these channels. Markovian modeling of the wireless channel is commonly used in the literature to incorporate memory, e.g., to model the shadowing phenomenon, [2], [3].

The pilot allocation problem introduced above, with channels evolving in a Markovian fashion, can be formulated as a restless bandit problem (RBP). RBPs are a generalization of multi-armed bandit problems (MABPs) [4], sequential decision-making problems that can be seen as a particular case of Markov decision processes (MDPs). In a MABP, at each decision epoch, a scheduler chooses which bandit[1] to play, and a reward is obtained accordingly. The objective is to design a bandit selection policy that maximizes the average expected reward. In MABPs the bandits that have not been played remain at the same state and provide no reward. Gittins [4] proved that the optimal solution of a MABPs is characterized by a simple index, known today as Gittins index. In the more general framework of RBPs, the statistics of all bandits evolve even in slots that are not chosen, and the analysis becomes more difficult. As a result, obtaining an optimal solution is typically out of reach. In [5], Whittle, based on the Lagrangian relaxation approach, proposed a scheduling algorithm, the so-called Whittle's index policy, as a heuristic for solving RBPs. This has been the approach considered in this paper. We derive Whittles index policy for the above mentioned problem and we numerically observe its remarkably good performance.

Previous papers that are related to our work ([2], [3], [6], [7]) study the Gilbert-Elliot channel model, the simplest Markovian channel having two states, where the channel is either in a GOOD or in a BAD state. The limitation of such binary models is that they fail to capture the complex nature

This work has been partly funded by Huawei Technologies France SASU.

[1]The notion of the bandit historically refers to a slot machine with an unknown reward distribution.

of the wireless channel. Instead, here we consider a multi-dimensional Markov process, which allows to model channels with multiple states corresponding to how the different modulation and coding techniques used in practice interact with the wireless channels to yield different level of rewards (=number of successfully transmitted bits). Thus, we have considered here a more challenging problem where channels are modeled by $K-$state Markov Chains, with K arbitrarily large. This represents a generalization of prior binary Markovian models.

The remainder of the paper is organized as follows. In Section II we describe the wireless downlink scheduling problem that has been considered. In Section III we introduce two approximations that can be solved using a Lagrangian relaxation approach. We derive a closed-form expression for the Whittle index and we define a heuristic for the original problem based on this index. Finally, in Section IV we evaluate the performance of Whittle's index policy and we compare it to the performance of a myopic policy and a randomized policy.

II. Model Description

We consider a wireless downlink scheduling problem with a single base station (BS) and N users. The channel between a user and the BS is modeled as a K-state Markov chain. Time is slotted and users are synchronized. Let us denote by $X_n(t)$ the state of channel of user n at time slot t. Then $X_n(t) \in \{h_1, h_2, \ldots, h_K\}$. The state of the channel remains the same during a time slot and evolves according to the probability transition matrix $P_n = (p_{n,ij})_{i,j\in\{1,\ldots,K\}}$, where $p_{n,ij} = \mathbb{P}(X_n(t+1) = h_j | X_n(t) = h_i)$. Channels are assumed to be independent and non identical across users.

We adopt the following scheduling model. We assume M different pilot sequences to be available. In the beginning of each time slot, the BS chooses M out of N users (typically, $M < N$). The selected users use the allocated pilots to send the uplink training symbols. This mechanism allows the BS to have perfect CSI during downlink data transmission of the selected users. Users that have not been selected cannot provide their current CSI. Therefore, the base station updates the belief channel state information for these users. Under Markovian channel models, the update of the belief channel state has an impact on the future decisions and on the expected average rewards.

Next we explain the belief channel state update for the pilot allocation problem introduced above. Let us define $\vec{b}_n^{\phi}(t)$ to be the belief state of user n during the t^{th} time slot under policy ϕ. The element $b_{n,j}^{\phi}(t)$ is the probability that user n is in state h_j in slot t given all the past channel state information. Let us denote by $a_n^{\phi}(\vec{b}_1^{\phi}(t), \ldots, \vec{b}_N^{\phi}(t)) \in \{0, 1\}$, the decision of the BS with respect to user n, where $a_n^{\phi}(\cdot) = 0$ if no pilot has been allocated to user n, and $a_n^{\phi}(\cdot) = 1$ if a pilot has been allocated to user n in slot t. For ease of notation we define $a_n^{\phi}(t) := a_n^{\phi}(\vec{b}_1^{\phi}(t), \ldots, \vec{b}_N^{\phi}(t))$. Since at most M pilots can be allocated we have $\sum_{n=1}^{N} a_n^{\phi}(t) \leq M$. Let us denote by $S^{\phi}(t) = \{n \in \{1, \ldots, N\} : a_n^{\phi}(t) = 1\}$ the set of users that have been selected in time slot t under policy ϕ. We then define $\vec{b}_n^{\phi}(t+1) := \vec{b}_n^{\phi}(t) P_n$ if $n \notin S^{\phi}(t)$, and $\vec{b}_n^{\phi}(t+1) := \vec{\pi}_{n,j}^{1}$, if $n \in S^{\phi}(t)$ and $X_n(t) = h_j$, to be the evolution of the belief states. In the latter equation $\vec{\pi}_{n,j}^{1} = (p_{n,j1}, \ldots, p_{n,jK})$ and $\vec{b}_n^{\phi}(t)$ take values in the countable state space $\Pi_n = \{\vec{\pi}_{n,j}^{\tau} \in \mathbb{R}^K : \vec{\pi}_{n,j}^{\tau} = \vec{e}_j P_n^{\tau}, \tau > 0\}$, where $\vec{e}_j$ is the vector with all entries 0 except the j^{th} entry which equals 1. We will use the notation $\vec{\pi}_{n,j}^{\tau} = (p_{n,j1}^{(\tau)}, \ldots, p_{n,jK}^{(\tau)})$ throughout the paper. Belief state $\vec{b}_n^{\phi}(t) = \vec{\pi}_{n,j}^{\tau}$ implies that user n has last been selected in slot $t - \tau$ and the observed channel state has been h_j. We note that $\vec{b}_n^{\phi}(t)$ is a sufficient statistic for the past scheduling and channel state information, see the proof in Smallwood et al. [8]. Next we make an assumption on $\vec{\pi}_{n,j}^{\tau}$ and we provide a sufficient condition for this assumption to hold.

Assumption 1 (A1). *Let $\vec{\pi}_{n,j}^{\tau}$ and $\vec{\pi}_{n,j}^{\tau'} \in \Pi_n$. We assume that $\max_i p_{n,ji}^{(\tau)} \geq \max_i p_{n,ji}^{(\tau')}$, for all j, if $\tau \leq \tau'$.*

Remark 1. *If P_n is doubly stochastic then Assumption 1 holds.*

If the Markov chain is irreducible, and P_n doubly stochastic, the belief channel vector approaches the uniform distribution as τ increases.

A. Throughput maximization problem

The objective of the present work is to efficiently allocate the available pilots to the users in the system in order to maximize the *long-run expected average throughput*. That is, find ϕ such that

$$\liminf_{T\to\infty} \frac{1}{T} \mathbb{E}\left(\sum_{n=1}^{N} \sum_{t=1}^{T} R_n(X_n(t), \vec{b}_n^{\phi}(t), a_n^{\phi}(t)) \right), \quad (1)$$

is maximized, with $R_n(\cdot, \cdot, \cdot)$ the immediate throughput. We have assumed that if a pilot has been allocated to a user, then the BS obtains full CSI of that particular user, and therefore, the immediate reward that corresponds to that user is independent of the belief state $R_n(h, \vec{\pi}_{n,j}^{\tau}, 1) = R_n(h, 1)$ for all $\vec{\pi}_j^{\tau} \in \Pi_n$ and $h \in \{h_1, \ldots, h_K\}$. Due to A1, we make the following natural assumption on the reward for not selected users.

Assumption 2 (A2). *Let R_n^1 and $R_n(\vec{\pi}_j^{\tau}, 0)$ be the average immediate rewards of user n under active and passive actions, respectively. Then, we assume $R_n^1 \geq R_n(\vec{\pi}_{n,j}^{\tau}, 0) \geq R_n(\vec{\pi}_{n,j}^{\tau'}, 0)$, for all $\tau' \geq \tau$.*

The latter implies that the more outdated the CSI of a user is, the less the average reward accrued by that user will be.

While (1) being a typical performance measure, it is not obvious at all to deal with it. In many existing works, a discounted reward function is usually used. In this work, we deal with (1) as follows. We first consider the *discounted reward over the infinite horizon:* find ϕ such that

$$\mathbb{E}\left(\sum_{n=1}^{N} \sum_{t=1}^{\infty} \beta^{t-1} R_n(X_n(t), \vec{b}_n^{\phi}(t), a_n^{\phi}(t)) \right), \quad (2)$$

is maximized, with $0 \leq \beta < 1$ the discount factor. We then retrieve the solution of (1) as a limit of the discounted reward

model (i.e., letting the discount factor $\beta \to 1$). This limit is not straightforward since certain conditions on Equation (2), [9, Chap. 8.10] must be verified. The proof can be found in the extended version of this paper [10].

III. Lagrangian relaxation

The model introduced above falls in the framework of RBP problems. Each user $n \in \{1, \ldots, N\}$ present in the system can be seen as bandit or arm. The state of each arm represents the belief channel state of the user. RBPs have been shown to be PSPACE-hard, see Papadimitriou et al. [11]. A well established method for solving RBPs is the Lagrangian relaxation introduced by Whittle in [5].

The Lagrangian relaxation technique consists in relaxing the constraint on the available resources, by letting it be satisfied on average and not in every time slot, that is,

$$\sum_{n=1}^{N} a_n^{\phi}(t) \leq M \Rightarrow \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\sum_{t=1}^{T}\sum_{n=1}^{N} a_n^{\phi}(t)\right) \leq M, \quad (3)$$

in the expected average reward model. Objective function (1) together with the relaxed constraint (3) constitute a Partially Observable Markov Decision Process (POMDP), and we will refer to it as the relaxed POMDP throughout the paper. Observe that, under constraint (3), users become independent from each other and the relaxed POMDP can be decomposed into N uni-dimensional optimization problems. We will refer to the latter as the single-arm POMDP. The solution of the relaxed POMDP can then be obtained by combining the solution of the single-arm POMDPs. This is known as the Whittle index policy (see Section III-C). In the remainder of this section we drop the user index from the notation since we will focus on the single-arm POMDPs.

A general recipe to compute Whittle's index is to: (1) prove some structure on the solution of the single-arm POMDP (usually optimality of threshold policies), (2) show that the indexability property holds (which ensures Whittle's index to exist), (3) derive an explicit expression for Whittle's index and (4) define Whittle's index policy. For this particular problem, proving threshold type of policies to be optimal has shown to be extremely challenging, except in the 2-state Markov channel systems (Gilbert-Elliot model), see Albright [12] and Lovejoy [13]. To the best of our knowledge, all the research work done in this area has focused on either i.i.d. channel model or the Gilbert-Elliot channel model. In the present work, we have considered an approximation that allows to obtain Whittle's index for a K-state Markov Chain channel model. Let us define $p^a(\vec{\pi}_i^{\tau}, \vec{\pi}_j^{\tau'})$ the transition probability from belief state $\vec{\pi}_i^{\tau}$ to belief state $\vec{\pi}_j^{\tau'}$ under action $a \in \{0, 1\}$. In the original model we have

$$p^0(\vec{\pi}_i^{\tau}, \vec{\pi}_j^{\tau'}) = \begin{cases} 1 & \text{if } j = i \text{ and } \tau' = \tau + 1, \\ 0 & \text{otherwise,} \end{cases} \text{ and}$$

$$p^1(\vec{\pi}_i^{\tau}, \vec{\pi}_j^{\tau'}) = \begin{cases} p_{ij}^{(\tau)} & \text{if } \tau' = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let us now denote by $p^{a,app}(\vec{\pi}_i^{\tau}, \vec{\pi}_j^{\tau'})$, the transition probability from belief state $\vec{\pi}_i^{\tau}$ to belief state $\vec{\pi}_j^{\tau'}$ for the approximation that we introduce next.

Approximation: If the user is not selected then the original transitions are kept, i.e., $p^{0,app}(\vec{\pi}_i^{\tau}, \vec{\pi}_j^{\tau'}) = p^0(\vec{\pi}_i^{\tau}, \vec{\pi}_j^{\tau'})$. If the user is selected then $p^{1,app}(\vec{\pi}_i^{\tau}, \vec{\pi}_j^{1}) = p_i^s$, where p_i^s is the steady-state probability of channel h_i. In Section IV-A we numerically evaluate the accuracy of this approximation.

A. Threshold policies

We will prove that the single-arm POMDP can be solved by threshold type of policies for the approximation above. We next give a formal definition of threshold policies.

Definition 1. *We say that ϕ is a threshold type of policy if it prescribes action $a \in \{0, 1\}$ in all states $\vec{\pi}_j^{\tau}$ such that $\tau \leq \Gamma_j$ and prescribes action $a' \in \{0, 1\}$ with $a' \neq a$ for all $\vec{\pi}_j^{\tau}$ where $\tau > \Gamma_j$, $j \in \{1, \ldots, K\}$ and $\vec{\Gamma} = (\Gamma_1, \ldots, \Gamma_K)$. Such a threshold policy will be referred to as policy $\vec{\Gamma}$.*

As mentioned above we will focus on the discounted reward model. The Bellman optimality equation, [9, Ch. 6], writes

$$V_{\beta}^{app}(\vec{\pi}_j^{\tau}) = \max\{R(\vec{\pi}_j^{\tau}, 0) + W + \beta V_{\beta}^{app}(\vec{\pi}_j^{\tau+1}); \\ R^1 + \beta \sum_{k=1}^{K} p_k^s V_{\beta}^{app}(\vec{\pi}_k^1))\}, \quad (4)$$

where W is the Lagrange multiplier, which can be understood as a *subsidy for passivity*. The Bellman optimality equation is often used to obtain optimal solutions for MDPs. The function V_{β}^{app} is the value function for the single-arm POMDP, and although not made explicit in the notation it also depends on W.

In the next theorem we prove that threshold type of policies are an optimal solution for (4). The proof can be found in Appendix A in the extended version [10].

Theorem 1. *Assume that Assumption 1 holds and let W be fixed. Then there exist $\Gamma_1, \ldots, \Gamma_K \in \{0, 1, \ldots\}$ such that the threshold policy $\vec{\Gamma} = (\Gamma_1, \ldots, \Gamma_K)$ is an optimal solution for problem* (4)*.*

Having proven the structure of the optimal policy, the explicit expression of V_{β}^{app} can be obtained. The latter enables to prove conditions 8.10.1- 8.10.4' in Puterman [9], see Appendix B in [10]. It then can be shown that the single-arm long-run expected average reward, under the approximation, equals $\lim_{\beta\to 1}(1-\beta)V_{\beta}^{app}$, see [9, Th. 8.10.7]. Moreover, it can be seen that threshold type of policies are an optimal solution of the average reward model too (the proof can be obtained again as a limit $\beta \to 1$).

In the next section we derive Whittle's index for the average reward model.

B. Indexability and Whittle's index

In this section we prove the problem to be indexable. Indexability is the property that ensures Whittle's index to exist. It establishes that as the Lagrange multiplier W increases, the

set of states in which the optimal action is the passive action increases. In the following we formally define this property.

Definition 2. *Let $\vec{\Gamma}(W)$ be an optimal threshold policy for a fixed subsidy W. We define the set $\mathcal{L}(W) := \{\vec{\pi}_j^{\tau} \in \Pi : \tau \leq \Gamma_j(W)$ for all $j \in \{1,\dots,K\}\}$, i.e., the set of all belief states in which passive action is prescribed by policy $\vec{\Gamma}(W)$.*

Definition 3. *Let $\mathcal{L}(W) \subseteq \Pi$ be as defined in Definition 2. Then a bandit is said to be indexable if $\mathcal{L}(W) \subseteq \mathcal{L}(W')$ for all $W < W'$, i.e., the set of belief states in which passive action is prescribed by an optimal policy of the relaxed problem increases as W increases. A RBP is indexable if all bandits are indexable.*

Although indexability seems a natural property not all problems satisfy this condition; a few examples are given in Hodge et al. [14] and Whittle [5]. In Appendix C in [10] we prove that for the problem under study all users are indexable. Having proven indexability Whittle's index can be defined as follows.

Definition 4. *Whittle's index in state π_j^{τ} is defined as the smallest value of W such that an optimal policy of the single-arm POMDP is indifferent of the action taken in π_j^{τ}.*

We can now proceed to solve Whittle's index. Let us define $\mathcal{T}(\vec{\Gamma}) = \{\vec{\Gamma}' = (\Gamma'_1,\dots,\Gamma'_K)$ with $\Gamma'_i \in \mathbb{N}\cup\{0\}$ for all $i : \vec{\Gamma}' > \vec{\Gamma}\}$, that is, the set of all threshold policies that are *greater* than $\vec{\Gamma}$ (*i.e.,* $\vec{\Gamma}' \geq \vec{\Gamma} \Leftrightarrow \Gamma'_j \geq \Gamma_j$ for all j). In particular, we denote $\mathcal{T}(0) = \{\vec{\Gamma}' = (\Gamma'_1,\dots,\Gamma'_K)$ with $\Gamma'_i \in \mathbb{N}\cup\{0\}$ for all $i : \vec{\Gamma}' > (0,\dots,0)\}$. Let $\alpha^{\vec{\Gamma}}(\vec{\pi}_j^{\tau})$ be the steady-state probability of being in state $\vec{\pi}_j^{\tau}$ under policy $\vec{\Gamma}$, and let $b^{\vec{\Gamma}}$ the steady-state belief state under policy $\vec{\Gamma}$. It then can be shown that

$$\lim_{\beta\to 1}(1-\beta)V_{\beta}^{app}(\cdot)$$
$$= g^{\vec{\Gamma}}(W) = \mathbb{E}(R(b^{\vec{\Gamma}}, a^{\vec{\Gamma}}(b^{\vec{\Gamma}}))) + W\sum_{k=1}^{K}\sum_{i=1}^{\Gamma_k}\alpha(\vec{\pi}_k^i),$$

where $g^{\vec{\Gamma}}(W)$ is the average reward under policy $\vec{\Gamma}$ when the subsidy for passivity equals W. Whittle's index for the average reward problem can then be computed as explained in the next theorem. The proof can be found in Appendix E in [10].

Theorem 2. *Assume that an optimal solution of the single-arm POMDP is of threshold type and that $\sum_{k=1}^{K}\sum_{r=1}^{\Gamma_k}\alpha^{\vec{\Gamma}}(\vec{\pi}_k^r)$ is non-decreasing in $\vec{\Gamma}$. Then the problem is indexable and Whittle's index for user n is computed as follows (we omit the dependence on n from the notation):*

Step i: Compute

$$W_i = \inf_{\vec{\Gamma}\in\mathcal{T}(\vec{\Gamma}^{i-1})} \frac{\mathbb{E}(R(b^{\vec{\Gamma}^{i-1}}, a^{\vec{\Gamma}^{i-1}}(b^{\vec{\Gamma}^{i-1}}))) - \mathbb{E}(R(b^{\vec{\Gamma}}, a^{\vec{\Gamma}}(b^{\vec{\Gamma}})))}{\sum_{j=1}^{K}\left(\sum_{r=1}^{\Gamma_j}\alpha^{\vec{\Gamma}}(\vec{\pi}_j^r) - \sum_{r=1}^{\Gamma_j^{i-1}}\alpha^{\vec{\Gamma}^{i-1}}(\vec{\pi}_j^r)\right)}$$

for all $i \geq 0$, where $\vec{\Gamma}^{-1} = \vec{0}$. Denote by $\vec{\Gamma}^i$ the largest minimizer for all $i > 0$. We define $W(\vec{\pi}_j^{\tau}) := W_i$ for each j, such that $\Gamma_j^{i-1} < \tau \leq \Gamma_j^i$. If $\vec{\Gamma}_j^i = \infty$ for all j then stop, otherwise go to Step i+1. When the algorithm stops the Whittle index for all π_j^{τ} has been obtained and is given by $W(\pi_j^{\tau})$.

In the following lemma and corollary we derive an explicit expression for Whittle's index. The proofs can be found in Appendix F in [10].

Lemma 1. *If in Step i of Theorem 2 for an $i> 0$, the minimizer $\vec{\Gamma}^i$ is such that $\sum_{j=1}^{K}\Gamma_j^i = (\sum_{j=1}^{K}\Gamma_j^{i-1}) + 1$ and $\Gamma_j^i \geq \Gamma_j^{i-1}$ for all $j \in \{1,\dots,K\}$, then, for u such that $\Gamma_u^i = \Gamma_u^{i-1} + 1$,*

$$W_i = R^1 + \sum_{k=1}^{K}\sum_{j=1}^{\Gamma_k^{i-1}} R(\vec{\pi}_k^j, 0)p_k^s - R(\vec{\pi}_u^{\Gamma_u^i}, 0)\sum_{k=1}^{K}(\Gamma_k^{i-1}+1)p_k^s,$$

Corollary 1. *Let us define $u^0 = \arg\max_{u\in\{1,\dots,K\}} R(\vec{\pi}_u^1, 0)$, and $\vec{\Gamma}^0 = \vec{e}_{u^0}$, with $\vec{e}_{u^0}$ the vector with all entries 0 except the u^0th element which equals 1. Define $u^i = \arg\max_{u\in\{1,\dots,K\}} R(\pi_u^{\Gamma_u^{i-1}+1}, 0)$, and, $\vec{\Gamma}^i = \left\{\sum_{r=0}^{i} 1_{\{u^r=1\}},\dots,\sum_{r=0}^{i} 1_{\{u^r=K\}}\right\}$, for all $i > 0$. Then*

$$W(\vec{\pi}_{u^j}^{\Gamma^j}) = R^1 + \sum_{k=1}^{K}\sum_{r=1}^{\Gamma_k^{j-1}} R(\vec{\pi}_k^r, 0)p_k^s - R(\vec{\pi}_{u_j}^{\Gamma^j u^j}, 0)\sum_{k=1}^{K}(\Gamma_k^{j-1}+1)p_k^s, \text{ for all } j \geq 0.$$

Whittle's index, $W(\vec{\pi}_k^{\tau})$, is non-decreasing in τ for all k.

Whittle's index being non-decreasing in τ implies that, the longer a user has not been selected for channel sensing the more attractive it becomes to select him/her. The exploration vs. exploitation trade-off is captured by this latter property.

C. Whittle's index policy

In this section we explain how the Whittle index can be used in order to define a heuristic for the original unrelaxed problem, as in Equation (1).

Definition 5. *Assume the state of user n at time t to be $\vec{\pi}_{j_n}^{\tau_n}$. The Whittle index policy prescribes to allocate a pilot to the M users with the highest $W_n(\vec{\pi}_{j_n}^{\tau_n})$.*

Whittle's index policy is an optimal solution for the relaxed POMDP. It has been proven to perform strikingly well in various scenarios Verloop [15] and Ouyang et al. [16].

IV. Numerical analysis

We provide in this section some numerical results to assess the performance of the Whittle's index policy (WIP) and show some scenarios in which the approximation is accurate. One can refer to [10] for more details.

A. Accuracy of the approximation

As mentioned in Section III, we evaluate the approximation that has been considered throughout the paper. To do so, we compare the optimal solution obtained using a Value Iteration algorithm for both, the original system and the approximation.

TABLE I
RELATIVE (%) SUBOPTIMALITY GAP

	App. 1 pilot	App. 3 pilots
Rel. err. ex. 1	0.0798	0.0527
Rel. err. ex. 2	0.0149	0.0393

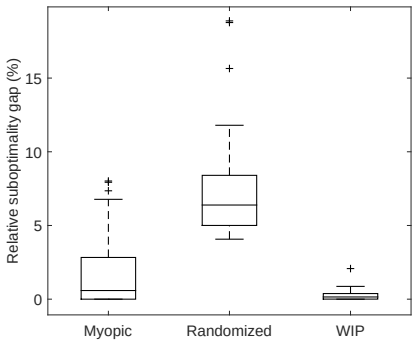

Fig. 1. Suboptimality gap (%) of a myopic policy, a randomized policy and WIP, for 40 randomly generated examples.

Example: Let us assume a system with a BS and four users. We assume users to be in three possible channel states h_{n1}, h_{n2}, h_{n3}. Let the transition matrices to be doubly stochastic and to be different for all four users. We further consider systems for which the steady-state belief vector for all four users is $(1/3, 1/3, 1/3)$. Therefore, the immediate average reward for user i if a pilot has been allocated to it is assumed to be $R_i^1 = \frac{1}{3}\sum_{k=1}^{3}\log_2(1+SNR),\ i \in \{1,\dots,K\}$. If user i has not been selected, the average immediate reward is considered to be $R_i(\vec{\pi}_j^{\tau},0) = \rho_i \frac{1}{3}\sum_{k=1}^{3}\log_2(1+SNR)$, where $\rho_i = \max_r\{p_{jr}^{(\tau)}\}$, that is, the highest probability channel state for user i, when its belief state is $\vec{\pi}_j^{\tau}$, and $\hat{h}_i = h_{i\sigma}$ where $\sigma = \arg\max_r\{p_{jr}^{(\tau)}\}$. We first assume that a single pilot is available to the system, and later on we assume that three pilots are available. The relative errors of the approximation w.r.t. the original problem can be found in Table I for two different examples (ex.1 and ex.2). We can observe in Table I that the error of the approximation is extremely small.

B. Performance of Whittle's index policy

To assess the performance of WIP we generate 40 examples with randomly generated doubly stochastic transition probability matrices. We generate the channel vectors for each user randomly from a zero-mean complex Gaussian distribution. The latter allows heterogeneity among users. The throughput obtained by each user under both passive (no pilot has been allocated) and active actions (pilot has been allocated) are considered to be as in Section IV-A. We have computed the relative error of all 40 examples, using the WIP policy a myopic policy, and a randomized policy, w.r.t. the optimal solution. The results can be found in Figure 1, where the horizontal line refers to the average relative error, the upper and lower edges of the box are the 25th and 75th percentiles and the crosses are outliers. We observe that the relative error of Whittle's index policy is extremely small, whereas the myopic and the randomized policies can have big relative errors. WIP being remarkably simple to apply, captures very closely the optimal exploration vs. exploitation trade-off.

V. CONCLUSIONS

We investigate the challenging problem of pilot allocation in wireless networks over Markovian fading channels where typically, there are less available pilots than users. At each time, the BS can know the current CSI of users to whom a pilot has been assigned. A channel belief state is estimated for other users. The problem can be cast as a restless multi-armed bandit problem for which obtaining an optimal solution is out of reach. We have proposed a Lagrangian relaxation approach to obtain a Whittle index policy, that has a low complexity, and has shown to perform remarkably well. Future work include obtaining explicit performance bounds for the approximation considered in this paper.

REFERENCES

[1] T. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," in *IEEE Transactions on Wireless Communications*, 2010.

[2] K. Liu and Q. Zhao, "Indexability of retless bandit problems and optimality of whittle index for dynamic multichannel access," vol. 56, no. 11, 2010, pp. 5547–5567.

[3] W. Ouyang, S. Murugesan, A. Eryilmaz, and N. Shroff, "Exploiting channel memory for joint estimation and scheduling in downlink networks-a Whittle's indexability analysis," *IEEE Transcations on Information Theory*, vol. 61, no. 4, pp. 1702–1719, 2015.

[4] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed Bandit Allocation Indices*. Wiley, 2011.

[5] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of Applied Probability*, vol. 25, pp. 287–298, 1988.

[6] P. Jacko and S. Villar, "Opportunistic schedulers for optimal scheduling of flows in wireless systems with ARQ feedback," *24th International Teletraffic Congress*, 2012.

[7] K. Liu, Q. Zhao, and B. Krishnamachari, "Dynamic multichannel access with imperfect channel state detection," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2795–2808, 2010.

[8] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable markov processes over a finite horizon," *Operations Research*, vol. 21, pp. 1071–1088, 1973.

[9] M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2005.

[10] M. Larrañaga, M. Assaad, A. Destounis, and G. Paschos, "Dynamic pilot allocation over markovian fading channels: A restless bandit approach." *Extended version*, HAL report (hal-01349104).

[11] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queuing network control," *Mathematics of Operations Research*, vol. 24, no. 2, pp. 293–305, 1999.

[12] S. C. Albright, "Structural results for partially observable markov decision processes," *Operations Research*, vol. 27, no. 5, pp. 1041–1053, 1979.

[13] W. S. Lovejoy, "Some monotonicity results for partially observed markov decision processes," *Operations Research*, vol. 35, no. 5, pp. 736–743, 1987.

[14] D. Hodge and K. D. Glazebrook, "Dynamic resource allocation in a multi-product make-to-stock production system," *Queueing Systems*, vol. 67, no. 4, pp. 333–364, 2011.

[15] I. Verloop, "Asymptotically optimal priority policies for indexable and non-indexable restless bandits," *To appear in Annals of Applied Probability*, 2016.

[16] W. Ouyang, A. Eryilmaz, and N. Shroff, "Asymptotically optimal downlink scheduling over Markovian fading channels," *Proceedings of IEEE INFOCOM*, pp. 1–9, 2012.