Exploiting Channel Memory for Joint Estimation and Scheduling in Downlink Networks – A Whittle's Indexability Analysis

Wenzhuo Ouyang, Sugumar Murugesan, Atilla Eryilmaz, Ness B. Shroff

***Abstract*—We address the problem of opportunistic multiuser scheduling in downlink networks with Markov-modeled outage channels. We consider the scenario in which the scheduler does not have full knowledge of the channel state information, but instead estimates the channel state information by exploiting the memory inherent in the Markov channels along with ARQ-styled feedback from the scheduled users. Opportunistic scheduling is optimized in two stages: (1) Channel estimation and rate adaptation to maximize the expected immediate successful transmission rate of the scheduled user; (2) User scheduling, based on the optimized immediate rate, to maximize the overall long term sum-throughput of the downlink. The scheduling problem is a partially observable Markov decision process with the classic 'exploitation vs exploration' trade-off that is difficult to quantify. We therefore study the problem in the framework of Restless Multi-armed Bandit Processes and perform a Whittle's indexability analysis. Whittle's indexability is traditionally known to be hard to establish and the index policy derived based on Whittle's indexability is known to have optimality properties in various settings. We show that the problem of downlink scheduling under imperfect channel state information is Whittle indexable and derive the Whittle's index policy in closed form. Via extensive numerical experiments, we show that the Whittle's index policy has near-optimal performance and is robust against imperfections in channel state feedback.**

Our work reveals that, under incomplete channel state information, exploiting channel memory for opportunistic scheduling can result in significant performance gains and that almost all of these gains can be realized using the easy-to-implement Whittle's index policy.

I. Introduction

The wireless channel is inherently time-varying and stochastic. It can be exploited for dynamically allocating resources to the network users, leading to the classic *opportunistic scheduling* principle (e.g., [1]). Understandably, the success of opportunistic scheduling depends heavily on reliable knowledge of the instantaneous channel state information (CSI) at the scheduler. Many sophisticated scheduling strategies have been developed with provably optimal characteristics (e.g., [2]-[6]) by assuming perfect CSI to be readily available, free of cost at the scheduler.

Wenzhuo Ouyang and Atilla Eryilmaz are with the Department of ECE, The Ohio State University (e-mails: ouyangw@ece.osu.edu, eryilmaz@ece.osu.edu). Sugumar Murugesan was with the Department of ECE, The Ohio State University and is currently with the School of ECEE, Arizona State University (e-mail: sugumar.murugesan@asu.edu). Ness B. Shroff holds a joint appointment in both the Department of ECE and the Department of CSE at The Ohio State University (e-mail: shroff@ece.osu.edu).

A preliminary version of this paper appeared in INFOCOM 2011.

This research was supported by NSF grants CAREER-CNS-0953515, CCF-0916664, CNS-0721236, CNS-0813000, DTRA Grant HDTRA 1-08-1-0016 and ARO MURI grant W911NF-08-1-0238.

In realistic scenarios, however, perfect CSI is rarely, if ever, available and never cost-free, i.e., a non-trivial amount of network resource, that could otherwise be used for data transmission, must be spent in estimating the CSI [2]. This calls for jointly designing channel estimation and opportunistic scheduling strategies – an area that has recently received attention when the channel state is modeled by *i.i.d.* processes across time (e.g., [7], [8]). The *i.i.d.* model has traditionally been a popular choice for researchers to abstract the fading channels, because of its simplicity and associated ease of analysis. On the other hand, this model fails to capture an important characteristics of the fading channels – the time-correlation or the *channel memory* [2].

In the presence of estimation cost, memory in the fading channels is an important resource that can be intelligently exploited for more efficient, joint estimation and scheduling strategies. In this context, Markov channel models have been gaining popularity as realistic abstractions of fading channels with memory (e.g., [9]-[11]).

In this paper, we study joint channel estimation and scheduling using channel memory, in downlink networks. We model the downlink fading channels as two-state Markov Chains with *non-zero achievable rate* in both states. The scheduling decision at any time instant is associated with two potentially contradicting objectives: (1) Immediate gains in throughput via data transmission to the scheduled user; (2) Exploration of the channel of a downlink user for more informed decisions and associated throughput gains in the future. This is the classic 'exploitation vs exploration' trade-off often seen in sequential decision making problems (e.g., [12], [13]). We model the joint estimation and scheduling problem as a Partially Observable Markov Decision Process (POMDP) and study the structure of the problem, by explicitly accounting for the estimation cost. Specifically, our contributions are as follows.

- We recast the POMDP scheduling problem as a Restless Multi-armed Bandit Process (RMBP) [14] and establish its *Whittle's indexability* [14] in Section IV and V. Even though Whittle's indexability is difficult to establish in general [15], we have been able to show it in the context of our problem.
- Based on a Whittle's indexability condition, we explicitly characterize the Whittle's index policy for the scheduling problem in Section VI. Whittle's index policies are known to have optimality properties in various RMBP processes and have been shown to be easy to implement (e.g., [15], [16]).
- Using extensive numerical experiments, we demonstrate in Section VII that Whittle's index policy in our setting has near-

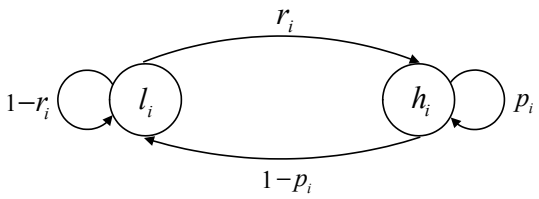

Fig. 1. A two state Markov Chain.

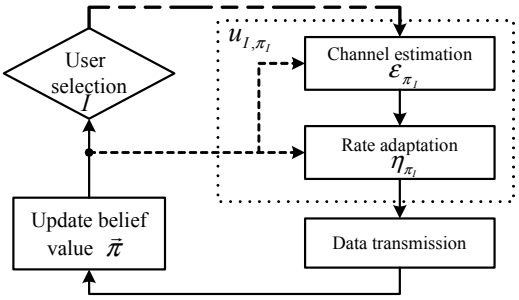

Fig. 2. Opportunistic scheduling with estimation and rate adaptation.

optimal performance and that significant system level gains can be realized by exploiting the channel memory for estimation and scheduling. Numerical experiments also suggest that Whittle's index policy is robust against imperfections in channel state feedback such as delays and errors. Also, the Whittle's index policy we derive is of polynomial complexity in the number of downlink users (contrast this with the PSPACE-hard complexity of optimal POMDP solutions [17]).

Our setup significantly differs from related works (e.g., [10] [11] [18]) in the following sense: In these works, the channels are modeled by ON-OFF Markov Chains with the OFF state corresponding to *zero*-achievable rate of transmission. There, once a user is scheduled, there is no need to estimate the channel of that user, since it is optimal to transmit at the constant rate allowed by the ON state irrespective of the underlying state. In contrast, in our model, the achievable rate at the lower state is, in general, non-zero and any rate above this achievable rate leads to *outage*. This extended model captures the realistic scenario when non-zero rates are possible with the use of sophisticated physical layer algorithms, even when the channel is bad. In this model, once a user is scheduled, the scheduler must estimate the channel of that user, with an associated cost, and adapt the transmission rate based on the estimate. The rate adaptation must balance between aggressive transmissions that lead to outage and conservative transmissions that lead to under-utilization of channels. The achievable rate expected from this process, in turn, influences the choice of the scheduled user. Thus the channel estimation and scheduling stages are tightly coupled, introducing several technical challenges to the problem, which we address in this paper.

II. System Model and Problem Statement

A. Channel Model

We consider a downlink system with one base station (BS) and N users. Time is slotted with the time slots of all users synchronized. The channel between the BS and each user is modeled as a two-state Markov chain, i.e., the state of the channels remains static within each time slot and evolves across time slots according to Markov chain statistics. The Markov channels are assumed to be independent and, in general, non-identical across users. The state space of channel C_i between the BS and user i is given by $S_i = \{l_i, h_i\}$. Each state corresponds to a maximum allowable data rate. Specifically, if the channel is in state l_i, there exists a rate δ_i, $0 \leq \delta_i < 1$, such that data transmissions at rates below δ_i succeed and transmissions at rates above δ_i fail, i.e., outage occurs. Similarly, state h_i corresponds to data rate 1. Note that fixing the higher rate to be 1 across *all* users does not impose any loss of generality in our analysis. This will be evident as we proceed.

The Markovian channel model is illustrated in Fig. 1. For user i, the two-state Markov channel is characterized by a 2×2 probability transition matrix

$$P_i = \begin{bmatrix} p_i & 1-p_i \\ r_i & 1-r_i \end{bmatrix},$$

where

$$p_i := \text{prob}\big(C_i[t]{=}h_i \mid C_i[t{-}1]{=}h_i\big),$$
$$r_i := \text{prob}\big(C_i[t]{=}h_i \mid C_i[t{-}1]{=}l_i\big).$$

B. Scheduling Model

We adopt the one-hop interference model, i.e., in each time slot, only one user can be scheduled for data transmission. At the beginning of the slot, the scheduler does not have exact knowledge of the channel state of the downlink users. Instead, it maintains a belief value π_i for channel i which is the probability that C_i is in state h_i at the time. We will elaborate on the belief values soon. Using these belief values, the scheduler picks a user, estimates its current channel state and subsequently transmits data at a rate adapted to the channel state estimate – all with an objective to maximize the overall sum-throughput of the downlink system. Specifically, in each slot, the scheduler jointly makes the following decisions: (1) Considering each user, the scheduler decides on the optimal channel estimator (that could involve the expenditure of network resources such as time, power, etc.) and rate adapter pair; (2) Based on the average rate of successful transmission promised for each user by the previous decision, the scheduler picks a user for channel estimation and subsequent data transmission. At the end of the slot, consistent with recent models (e.g., [10] [11] [18]), the scheduled user sends back accurate information on the state of the Markov channel in that slot. This accurate feedback is, in turn, used by the scheduler to update its belief on the channels, based on the Markov channel statistics. Note that these belief values are sufficient statistics to the past scheduling decisions and feedback [19]. Using ε_π to denote an arbitrary estimator and η_π to denote an arbitrary rate adapter, as functions of the belief value π, the basic operation is summarized in Fig. 2. The scheduling problem can be formulated as a partially observable Markov decision process [19], with the Markov channel states being the partially observable system states.

As noted in Section I, the scheduling decision in each slot involves two objectives: data transmission to the scheduled user and probing the channel of the scheduled user (through the accurate end-of-slot feedback). On one hand, the scheduler can transmit data to the user that promises the best achievable rate at the moment and hence realize immediate performance gains. On the other hand, the scheduler can schedule possibly another user and use the channel feedback from that user to gain a better understanding of the overall downlink system, which, in turn, could result in more informed future scheduling decisions with corresponding performance gains.

C. Formal Problem Statement

We now proceed to formally introduce the expected immediate reward. We let π_i denote the current belief value of the channel of user i, and let $u := \{\varepsilon, \eta\}$ denote an arbitrary estimator and rate adapter pair. Recall from the discussion on the scheduling model that, at the end of the slot, the scheduled user sends back accurate feedback on its Markov channel state in that slot. With this setup, once a user is scheduled, the choice of the channel estimator and rate adapter pair does not affect the future paths of the scheduling process. Thus, within each slot, it is optimal to design this pair to maximize the expected rate (of successful transmission) of the user scheduled in that slot. Henceforth, in the language of POMDPs, we call this maximized rate the *expected immediate reward*. We now proceed to formally introduce the expected immediate reward. We let π_i denote the current belief value of the channel of user i. The optimal estimator and rate adapter pair, u^*_{i,π_i}=$\{\varepsilon^*_{i,\pi_i}, \eta^*_{i,\pi_i}\}$, for user i, when the belief is π_i, is given by

$$u^*_{i,\pi_i} \quad = \quad \arg\max_{u} E_{C_i}[\gamma_i(C_i, u)], \tag{1}$$

where the quantity $\gamma_i(C_i, u)$ is the average rate of successful transmissions to user i when the channel is in state C_i and the estimator and rate adapter pair u is deployed. The expectation in (1) is taken over the underlying channel state C_i, with distribution characterized by belief value π_i, i.e.,

$$C_i = \begin{cases} h_i & \text{with probability } \pi_i, \\ l_i & \text{with probability } 1-\pi_i. \end{cases}$$

The expected immediate reward when user i is scheduled is thus given by

$$R_i(\pi_i) = E_{C_i}[\gamma_i(C_i, u^*_{i,\pi_i})]. \tag{2}$$

Note that our model is very general in the sense that we do not restrict to any specific estimation, data transmission structure or to any specific class of estimators. A typical estimation, data transmission structure, corresponding to the estimator and rate adapter pair u is illustrated in Fig. 3. Here a pilot-aided training[2]-based estimation is performed for a fraction of the time slots followed by data transmission at an adapted rate in the rest of the time slots.

We now introduce the optimality equations for the scheduling problem. Let $\vec{\boldsymbol{\pi}}[t] = (\pi_1[t], \cdots, \pi_N[t])$ denote the vector of current belief values of the channels at the beginning of slot t. A stationary scheduling policy, Ψ, is a stationary mapping

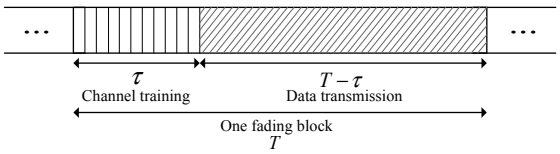

Fig. 3. A typical estimation - data transmission structure.

$\Psi : \vec{\boldsymbol{\pi}} \rightarrow I$ between the belief vector and the index of the user scheduled for data transmission in the current slot. Our performance metric is the infinite horizon, discounted sum-throughput of the downlink (henceforth simply the *expected discounted reward* in the language of POMDPs), formally defined next.

For a stationary policy Ψ, the expected discounted reward under initial belief $\vec{\boldsymbol{\pi}}$ is given by

$$V(\Psi, \vec{\boldsymbol{\pi}}) \quad = \quad \sum_{t=0}^{\infty} \beta^t E_{\vec{\boldsymbol{\pi}}[t]} R_{I[t]=\Psi(\vec{\boldsymbol{\pi}}[t])}(\pi_{I[t]}[t])$$

where $\vec{\boldsymbol{\pi}}[t]$ is the belief vector in slot t, $\pi_i[t]$ denotes the belief value of user i in slot t, $\vec{\boldsymbol{\pi}}[0] = \vec{\boldsymbol{\pi}}$, $I[t]$ denotes the index of the user scheduled in slot t. The discount factor $\beta \in [0, 1)$ provides relative weighing between the immediately realizable rates and future rates. For any initial belief $\vec{\boldsymbol{\pi}}$, the optimal expected discounted reward, $V(\vec{\boldsymbol{\pi}}) = \max_{\Psi} V(\Psi, \vec{\boldsymbol{\pi}})$, is given by the Bellman equation [20]

$$V(\vec{\boldsymbol{\pi}}) \quad = \quad \max_{I} \{R_I(\pi_I) + \beta E_{\vec{\boldsymbol{\pi}}^+}[V(\vec{\boldsymbol{\pi}}^+)]\}.$$

Here $\vec{\boldsymbol{\pi}}^+$ denotes the belief vector in the next slot when the current belief is $\vec{\boldsymbol{\pi}}$. The belief evolution $\vec{\boldsymbol{\pi}} \rightarrow \vec{\boldsymbol{\pi}}^+$ proceeds as follows:

$$\pi_i^+ = \begin{cases} p_i & \text{if } I = i \text{ and } C_i = h_i \\ r_i & \text{if } I = i \text{ and } C_i = l_i, \\ Q_i(\pi_i) & \text{if } I \neq i \end{cases} \tag{3}$$

where $Q_i(x) = xp_i + (1-x)r_i$ is the belief evolution operator for user i when it is not scheduled in the current slot. A stationary scheduling policy Ψ^* is optimal if and only if $V(\Psi^*, \vec{\boldsymbol{\pi}}) = V(\vec{\boldsymbol{\pi}})$ for all $\vec{\boldsymbol{\pi}}$ [20].

In the introduction, we briefly contrasted our setup with those in [10][11][18]. We provide a rigorous comparison here. The works [10][11][18] studied opportunistic scheduling with the channels modeled by ON-OFF Markov chains. In these works, the lower state is an 'OFF' state, i.e., it does not allow transmission at any non-zero data rate. Contrast this with our model where, at the lower state l_i, a possibly non-zero rate δ_i is achievable and outage occurs at any rate above δ_i. We now further explain how these two models are fundamentally different.

• In the ON-OFF channel model, the scheduler does not need a channel estimator and rate adapter pair. The scheduler can aggressively transmit at rate 1, since it has nothing to gain by transmitting at a lower rate – a direct consequence of the 'OFF' nature of the lower state. On the other hand, transmitting at a rate lesser than 1 can lead to losses due to under-utilization of the channel.

• In contrast, in our model, when $\delta > 0$, the scheduler must strike a balance between aggressive and conservative rates of transmission. An aggressive strategy (transmit at rate 1) can lead to losses due to outages, while a conservative strategy can lead to losses due to under-utilization of the channel. This underscores the importance of the knowledge of the underlying channel state and, therefore, the need for intelligent estimation and rate adaptation mechanisms.

• As a direct consequence of the preceding arguments, the expected immediate reward in our model is not a trivial δ-shift of the expected immediate reward when the rates supported by the channel states are 0 and $1-\delta$. Formally,

$$R^{\{\delta,1\}}(\pi) \neq R^{\{0,1-\delta\}}(\pi) + \delta = (1-\delta)\pi + \delta,$$

where $R^{\{x,y\}}(\pi)$ is the immediate reward when the channel state space is $\{x, y\}$ and belief value of the scheduled user is π. In fact, it can be shown that (in Lemma 1)

$$R^{\{\delta,1\}}(\pi) \leq R^{\{0,1-\delta\}}(\pi) + \delta = (1-\delta)\pi + \delta.$$

We believe that, our channel model, in contrast to the ON-OFF model, better captures realistic communication channels where, using appropriate physical layer algorithms, it is possible to transmit at a non-zero rate even at the lowest state of the channel model and the same physical layer algorithms may impose outage behavior when this allowed rate is exceeded.

III. Optimal Expected Transmission Rate – Structural Properties

In this section, we study the structural properties of the expected immediate reward, $R_i(\pi_i)$, defined in Equation (2). These properties will be crucial for our analysis in subsequent sections. For notational convenience, we will drop the suffix i in the rest of this section.

Lemma 1. *The expected immediate reward $R(\pi)$ has the following properties:*
(a) $R(\pi)$ is convex and increasing in π for $\pi \in [0,1]$
(b) $R(\pi)$ is bounded as follows:

$$\max\{\delta, \pi\} \leq R(\pi) \leq (1-\delta)\pi + \delta. \tag{4}$$

Proof: Let U^* be the set of optimal estimator and rate adapter pairs for all $\pi \in [0,1]$, i.e., $U^* = \{u^*_\pi, \pi \in [0,1]\}$. The expected immediate reward, provided in Equation (2), can now be rewritten as

$$\begin{aligned} R(\pi) &= \max_{u \in U^*} E_C[\gamma(C,u)] \\ &= \max_{u \in U^*} [\pi\gamma(h,u) + (1-\pi)\gamma(l,u)], \end{aligned}$$

where $\gamma(s,u)$ denotes the average rate of successful transmission when the channel state is $s \in \{l,h\}$. Note that, for fixed u, the average rate $\pi\gamma(h,u) + (1-\pi)\gamma(l,u)$ is linear in π. Thus, $R(\pi)$ is given as a point-wise maximum over a family of linear functions, which is convex [21]. $R(\pi)$ is therefore convex in π, establishing the convexity statement in (a).

We next proceed to derive the bounds to $R(\pi)$. From Equation (2),

$$R(\pi) = \max_u E_C[\gamma(C,u)] \geq \max_{\{u:u=\{\eta\}\}} E_C[\gamma(C,u)]$$

where $\{u : u = \{\eta\}\}$ indicates that we are considering rate adaptation without channel estimation. This explains the last inequality. Note that without the estimator, the rate adaptation is solely a function of the belief value π. Thus, the average rate achieved under the rate adapter, conditioned on the underlying channel state, can be expressed simply by indicator functions, as seen below:

$$\begin{aligned} &\max_{\{u:u=\{\eta\}\}} E_C[\gamma(C,u)] \\ &= \max_\eta [P(C=l)\eta \cdot \mathbf{1}(\eta \leq \delta) + P(C=h)\eta \cdot \mathbf{1}(\eta \leq 1)] \\ &= \max_\eta \eta[P(C=l) \cdot \mathbf{1}(\eta \leq \delta) + P(C=h) \cdot \mathbf{1}(\eta \leq 1)] \\ &= \max\{\delta, \pi\}. \end{aligned}$$

This establishes the lower bound in (b).

The upper bound in (b) corresponds to the expected immediate reward when *full* channel state information is available at the scheduler.

It is clear from the upper and lower bounds that $\delta \leq R(\pi) \leq 1$. Note that when $\pi=0$ or $\pi=1$, there is no uncertainty in the channel, hence $R(0)=\delta$ and $R(1)=1$. Using these properties, along with the convexity property of $R(\pi)$, we see that $R(\pi)$ is monotonically increasing in π, establishing the monotonicity of (a). The lemma thus follows. ■

Remark: Here we present some insights into the effect of the non-zero rate δ on the channel estimation and rate adaptation mechanisms by studying the upper and lower bounds to $R(\pi)$ provided in Lemma 1. The upper bound essentially corresponds to the case when perfect channel state information is available at the scheduler at the beginning of each slot. Here, no channel estimation and rate adaptation is necessary. The lower bound, on the other hand, corresponds to the case when the channel estimation stage is eliminated and rate adaptation is performed solely based on the belief value π of the scheduled user.

Fig. 4 plots the lower and upper bounds to $R(\pi)$ for different values of δ. Note that the lower bound approaches the upper bound in both directions, i.e., when $\delta \to 0$ or when $\delta \to 1$. This behavior can be explained as follows: (1) $\delta \to 1$ essentially means that the states of the Markov channel move closer to each other. This progressively reduces the channel uncertainty and hence the need for channel estimation (and, consequently, rate adaptation), essentially bringing the bounds closer. (2) As $\delta \to 0$, the channel uncertainty increases. At the same time, the impact of the channel estimator and rate adapter pair decreases. This is because, as $\delta \to 0$, the loss in immediate reward due to outage (transmitting at 1 when channel is in state δ) is less severe than the loss due to under-utilization of the channel (transmitting at rate δ when the channel is in state 1), essentially making it optimal for the rate adaptation scheme to be progressively more aggressive (transmit at rate 1). Thus channel estimation loses its significance as $\delta \to 0$. This brings the bounds closer as $\delta \to 0$.

It can be verified that the separation between the lower and upper bounds is at its peak when $\delta = 0.5$. This, along with the preceding discussion, indicates the potential for rate improvement when intelligent channel estimation and rate adaptation is performed under moderate values of δ.

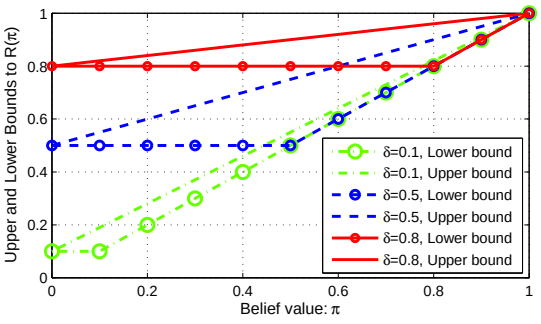

Fig. 4. Upper and lower bounds to the average rate of successful transmission.

IV. Restless Multi-Armed Bandit Processes, Whittle's Indexability and Index Policies

A direct analysis of the downlink scheduling problem appears difficult due to the complex nature of the 'exploitation vs exploration' tradeoff. We therefore establish a connection between the scheduling problem and the Restless Multiarmed Bandit Processes (RMBP) [14] and make use of the established theory behind RMBP in our analysis. We briefly overview RMBPs and the associated theory of Whittle's indexability in this section.

RMBPs are defined as a family of sequential dynamic resource allocation problems in the presence of several competing, independently evolving projects. In RMBPs, a subset of the competing projects are served in each slot. The states of all the projects in the system stochastically evolve in time based on the current state of the projects and the action taken. Once a project is served, a reward dependent on the states of the served projects and the action taken is accrued by the controller. Hence, the RMBPs are characterized by a fundamental tradeoff between decisions guaranteeing high immediate rewards versus those that sacrifice immediate rewards for better future rewards. Solutions to RMBPs are, in general, known to be PSPACE-hard [17].

Under an *average* constraint on the number of projects scheduled per slot, a low complexity index policy developed by Whittle [14], commonly known as Whittle's index policy, is optimal. Under stringent constraint on the number of users scheduled per slot, Whittle's index policy may not exist and if it does exist, its optimality properties are, in general, lost. However, Whittle's index policies, upon existence, are known to have near optimal performance in various RMBPs (e.g., [15] [16]). For an RMBP, Whittle's index policy exists if and only if the RMBP satisfies a condition known as *Whittle's indexability* [14], defined next.

Consider the following setup: for each project P in the system, consider a virtual system where, in each slot, the controller must make one of two decisions: (1) Serve project P and accrue an immediate reward that is a function of the state of the project. This reward structure reflects the one in the original RMBP for project P. (2) Do not serve project P, i.e., stay passive and accrue an immediate reward for passivity ω. The state of the project P evolves in the same fashion as it would in the original RMBP, as a function of its current state and current action (whether P is served or not in the current state). Let $D(\omega)$ be the set of states of project P in which it is optimal to stay passive, where optimality is defined based on the infinite horizon net reward.

Project P is Whittle indexable if and only if as ω increases from $-\infty$ to ∞, the set $D(\omega)$ monotonically expands from $\emptyset$ to S, the state space of project P. The RMBP is Whittle indexable if and only if all the projects in the RMBP are Whittle indexable.

For each state, s, of a project, Whittle's index, $W(s)$, is given by the value of ω in which the net reward after both the active and passive decisions are the same in the ω-subsidized virtual system. The notion of indexability gives a consistent ordering of states with respect to the indices. For instance, if $W(s_1)>W(s_2)$ and if it is optimal to serve the project at state s_1, then it is optimal to serve the project at s_2. This natural ordering of states based on indices renders the near-optimality properties to Whittle's index policy (e.g., [15], [16]).

The downlink scheduling problem we have considered is in fact an RMBP process. Here, each downlink user, along with the belief value of its channel, corresponds to a project in the RMBP, and the project is served when the corresponding user is scheduled for data transmission. Now, referring to our earlier discussion on the RMBPs, we see that Whittle's index policy is very attractive from an optimality point of view. The attractiveness of the index policy can be attributed to the natural ordering of states (and hence projects) based on indices, as guaranteed by Whittle's indexability. In the rest of the paper, we establish that this advantage carries over to the downlink scheduling problem at hand. As a first step in this direction, in the next section, we study the scheduling problem in Whittle's indexability framework and show that the downlink scheduling problem is, in fact, Whittle indexable.

V. Whittle's Indexability Analysis of the Downlink Scheduling Problem

In this section, we study the Whittle's indexability of our joint scheduling and estimation problem. To that end, we first describe the downlink scheduling setup:

At the beginning of each slot, based on the current belief value π (we drop the user index i in this section since only one user is considered throughout), the scheduler takes one of two possible actions: schedules data transmission to the user (action $a = 1$) or stays idle ($a = 0$). Upon an idle decision, a *subsidy* of ω is obtained. Otherwise, optimal channel estimation and rate adaptation is carried out, with a reward equal to $R(\pi)$ (consistent with the immediate reward seen in previous sections). The belief value is updated based on the action taken and feedback from the user (upon transmit decision). This belief update is consistent with that in the Section II. The optimal scheduling policy (henceforth, the ω-subsidy policy) maximizes the infinite horizon discounted reward, parameterized by ω. The optimal infinite horizon discounted reward is given by the Bellman equation [20]

$$V_\omega(\pi) = \max\{\big[R(\pi) + \beta\big(\pi V_\omega(p) + (1-\pi)V_\omega(r)\big)\big], \big[\omega + \beta V_\omega\big(Q(\pi)\big)\big]\}, \quad (5)$$

where, recall from Section II, $Q(\pi)$ is the evolution of the belief value when the user is not scheduled. The first quantity

inside the max operator corresponds to the infinite horizon reward when a *transmit* decision is made in the current slot and optimal decisions are made in the future slot. The second element corresponds to *idle* decision in the current slot and optimal decisions in all future slots.

We note that the indexability analysis in the rest of this section bears similarities to that in [18], where the authors studied indexability of a sequential resource allocation problem in a cognitive radio setting. This problem is mathematically equivalent to our downlink scheduling problem when $\delta = 0$. We have already discussed in detail (in Section II) that the structure of the immediate reward $R(\pi)$ when $\delta > 0$ is very different than when $\delta = 0$, due to the need for channel estimation and rate adaptation in the former case. Consequently, in the Whittle's indexability setup, the infinite horizon discounted reward $V_\omega(\pi)$ in our problem is different (and more general) than that in [18], underscoring the significance of our results.

As a crucial preparatory result, we now proceed to show that the ω-subsidy policy is *thresholdable*.

A. Thresholdability of the ω-subsidy policy

We first record our result on the convexity property of the infinite horizon discounted reward, $V_\omega(\pi)$, of (5) in the following proposition.

Proposition 1. *The infinite horizon discounted reward, $V_\omega(\pi)$ is convex in $\pi \in [0, 1]$.*

Proof: We first consider the discounted reward for finite horizon ω-subsidy problem. We let $\nu^1(\pi) = R(\pi)$ and $\nu^0(\pi) = \omega$ represent the immediate reward corresponding to active and idle decisions, respectively. The reward function associated with M-stage finite horizon process is expressed as

$$\widetilde{V}_M(\pi[0]) = \max_{\substack{a[t],\\ t=0,\ldots,M-1}} E\Big[\sum_{t=0}^{M} \beta^t \nu^{a[t]}(\pi[t])\Big|\pi[0]\Big]$$

Let $\widehat{V}_{\omega,t}(\pi)$ be the reward at time t with belief value $\pi[t] = \pi$. Hence $\widetilde{V}_M(\pi[0]) = \widehat{V}_{\omega,0}(\pi[0])$ and the last stage value function $\widehat{V}_{\omega,M-1}(\pi[M-1])$ is given by

$$\begin{aligned}\widehat{V}_{\omega,M-1}(\pi[M-1]) &= \max_{a[M-1]} \{\nu^{a[M-1]}(\pi[M-1])\} \\ &= \max\{\omega, R(\pi[M-1])\}.\end{aligned}$$

Therefore, $\widehat{V}_{\omega,M-1}(\pi)$ is convex with π since it is the maximum of a constant and a convex function. For any time $0 \le t < M-1$, the Bellman ([20]) equation can be written as

$$\widehat{V}_{\omega,t}(\pi[t]) = \max\{\widehat{V}^0_{\omega,t}(\pi[t]), \widehat{V}^1_{\omega,t}(\pi[t])\}.$$

where

$$\widehat{V}^0_{\omega,t}(\pi) = \omega + \beta \widehat{V}_{\omega,t+1}\big(Q(\pi)\big), \tag{6}$$

$$\widehat{V}^1_{\omega,t}(\pi) = R(\pi) + \beta\big(\pi \widehat{V}_{\omega,t+1}(p) + (1-\pi)\widehat{V}_{\omega,t+1}(r)\big). \tag{7}$$

Suppose now $\widehat{V}_{\omega,t+1}(\pi)$ is convex with π. If $a[t] = 1$, it is clear from (7) that $\widehat{V}^1_{\omega,t}(\pi)$ is convex function of π since it is a summation of a convex function and a linear function of π. If $a[t] = 0$, $\widehat{V}^0_{\omega,t}(\pi)$, expressed in (6), is also a convex function, because composition of convex function $\widehat{V}_{\omega,t+1}(\cdot)$ and linear function $Q(\pi)$ is convex [21]. Therefore $\widehat{V}_{\omega,t}(\pi)$ is convex with π as maximum of two convex functions. By induction, the the convexity of $V_{\omega,0}(\pi)$ is thus established.

Since $\widetilde{V}_M(\pi) = V_{\omega,0}(\pi)$, $\widetilde{V}_M(\pi)$ is convex with π. For discounted problem with bounded reward per slot, the infinite horizon reward is the limit of of finite horizon reward ([20]). Therefore $V_\omega(\pi) = \lim_{M\to\infty} V_{\omega,M}(\pi)$. Upon point-wise convergence, point-wise limit of convex functions is convex [21]. Hence $V_\omega(\pi)$ is a convex function of π. ■

In the next proposition, we show that the optimal ω-subsidy policy is a threshold policy.

Proposition 2. *The optimal ω-subsidy policy is thresholdable in the belief space π. Specifically, there exists a threshold $\pi^*(\omega)$ such that the optimal action a is 1 if the current belief $\pi > \pi^*(\omega)$ and the optimal action a is 0, otherwise. The value of the threshold $\pi^*(\omega)$ depends on the subsidy ω, partially characterized below.*

(i) If $\omega \ge 1$, $\pi^*(\omega) = 1$;
(ii) If $\omega \le \delta$, $\pi^*(\omega) = \kappa$ for some arbitrary $\kappa < 0$;
(iii) If $\delta < \omega < 1$, $\pi^*(\omega)$ takes value within interval $(0, 1)$.

Proof: Consider the Bellman equation (5), let $V^1_\omega(\pi)$ be the reward corresponding to transmit decision and $V^0_\omega(\pi)$ be the reward corresponding to idle decision, i.e.,

$$\begin{aligned}V^1_\omega(\pi) &= R(\pi) + \beta\big(\pi V_\omega(p) + (1-\pi)V_\omega(r)\big), \\ V^0_\omega(\pi) &= \omega + \beta V_\omega\big(Q(\pi)\big) = \omega + \beta V_\omega\big(\pi p + (1-\pi)r\big).\end{aligned}$$

It is clear from the Bellman equation (5) that the optimal action depends on the relationship between $V^1_\omega(\pi)$ and $V^0_\omega(\pi)$, presented as follows.

Case (i). If $\omega \ge 1$, since $R(\pi) \le 1$, in each slot, the immediate reward for being idle always dominates the reward for being active. Hence it will be optimal to always stay idle. We can thus set the threshold to 1.
Case (ii). If $\omega \le \delta$, then for any $\pi \in [0, 1]$, we have

$$\begin{aligned}V^0_\omega(\pi) =& \omega + \beta V_\omega(\pi p + (1-\pi)r) \\ \le& R(\pi) + \beta(\pi V_\omega(p) + (1-\pi)V_\omega(r)), \\ =& V^1_\omega(\pi),\end{aligned}$$

where the inequality is due to $\delta \le R(\pi)$ along with Jensen's inequality [21] due to the convexity of $V_\omega(\pi)$ from Proposition 2. Hence, it is optimal to stay active. Consistent with the threshold definition, we can set $\pi^*(\omega) = \kappa$ for any $\kappa < 0$.
Case (iii). If $\delta < w < 1$, then at the extreme values of belief,

$$\begin{aligned}V^0_\omega(0) &= \omega + \beta V_\omega(r) > \delta + \beta V_\omega(r) = V^1_\omega(0) \\ V^0_\omega(1) &= \omega + \beta V_\omega(p) < 1 + \beta V_\omega(p) = V^1_\omega(1)\end{aligned}$$

Note that the relationship of $V^0_\omega(\pi)$ and $V^1_\omega(\pi)$ is reversed at the end points 0 and 1, and they are both convex functions of π. Thus, there must exist a threshold $\pi^*(\omega)$ within $(0, 1)$ such that a equals 1 whenever $\pi > \pi^*(\omega)$. ■

B. Whittle's Indexability of Downlink Scheduling

Having established that the ω-subsidy policy is thresholdable in Proposition 2, Whittle's indexability, defined in Section IV, is re-interpreted for the downlink scheduling problem as follows: the downlink scheduling problem is Whittle indexable if the threshold boundary $\pi^*(\omega)$ is monotonically increasing with subsidy ω.

Using our discussion in Section IV, the index of the belief value π, i.e., $W(\pi)$ is the infimum value of the subsidy ω such that it is optimal to stay idle, i.e.,

$$\begin{aligned} W(\pi) &= \inf\{\omega : V_\omega^0(\pi) \geq V_\omega^1(\pi)\} \\ &= \inf\{\omega : \pi^*(\omega) = \pi\}. \end{aligned} \tag{8}$$

To establish indexability, we need to investigate the infinite horizon discounted reward $V_\omega(\pi)$, given by (5). We can observe from (5) that given the value of $V_\omega(p)$ and $V_\omega(r)$, $V_\omega(\pi)$ can be calculated for all $\pi \in [0,1]$. Let π^0 denote the steady state probability of being in state h. The next lemma provides a closed form expression for $V_\omega(p)$ and $V_\omega(r)$ and is critical to the proof of indexability.

Lemma 2. *The discounted rewards $V_\omega(p)$ and $V_\omega(r)$ can be expressed as:*

Case 1: $p > r$ (positive correlation)

$$V_\omega(p)=\begin{cases} \frac{R(p)+\beta(1-p)V_\omega(r)}{1-\beta p} & \text{if } \pi^*(\omega) < p \\ \frac{\omega}{1-\beta} & \text{if } \pi^*(\omega) \geq p \end{cases}$$

$$V_\omega(r)=\begin{cases} \sum_{k=0}^{\infty} \beta^k R(\frac{r-(p-r)^{k+1}r}{1+r-p}) & \text{if } \pi^*(\omega)<r \\ \Theta & \text{if } r\leq\pi^*(\omega)<\pi^0 \\ \frac{\omega}{1-\beta} & \text{if } \pi^*(\omega) \geq \pi^0 \end{cases}$$

Case 2: $p \leq r$ (negative correlation)

$$V_\omega(p)=\begin{cases} \sum_{k=0}^{\infty} \beta^k R(\frac{r+(p-r)^{k+1}(1-p)}{1+r-p}) & \text{if } \pi^*(\omega)<p \\ \frac{\omega+\beta R(Q(p))+\beta^2(1-Q(p))V_\omega(r)}{1-\beta^2 Q(p)} & \text{if } p\leq\pi^*(\omega)<Q(p) \\ \frac{\omega}{1-\beta} & \text{if } \pi^*(\omega)\geq Q(p) \end{cases}$$

$$V_\omega(r)=\begin{cases} \frac{R(r)+\beta r V_\omega(p)}{1-\beta(1-r)} & \text{if } \pi^*(\omega)<r \\ \frac{\omega}{1-\beta} & \text{if } \pi^*(\omega)\geq r \end{cases}$$

The expression of Θ is given by Equation (9), where Q^n denotes n^{th} iteration of Q and $L(\pi, \pi^*(\omega))$ is a function of π and $\pi^*(\omega)$. Their expressions are given in Appendix A. From the above expressions, the closed form $V_\omega(p)$ and $V_\omega(r)$ can be readily obtained. The explicit expression is space-consuming and therefore is moved to Appendix A.

Proof: The derivation of $V_\omega(p)$ and $V_\omega(r)$ follows from substituting p and r in Equation (5). Together with the expression of $Q(\pi)$ given by in Section II, the expression of $V_\omega(p)$ and $V_\omega(r)$ can be obtained. For details, please refer to Appendix A. ■

We note that the value function expression depends on the correlation type of the Markov chain, because the transition function $Q(\pi)$ given in Section II will behave differently with the correlation type of the chain.

The closed form expression of the value function given by the previous lemma serves as a useful tool for us to establish indexability, which is given by the next proposition.

Proposition 3. *The threshold value is strictly increasing with ω. Therefore, the problem is Whittle indexable.*

Proof: The proof of indexability follows the lines of [18]. Details are provided in Appendix B. ■

VI. Whittle's Index Policy

A. Whittle's Index Policy

In this section, we explicitly characterize Whittle's index policy for the downlink scheduling problem. For user i, let π_i^0 denote the steady state probability of being in state h_i, and let $V_{i,\omega}(\pi_i)$ denote the reward function for its ω-subsidy problem in (5). We first characterize the Whittle's index as follows.

Proposition 4. *For user i, the index value at state π_i, i.e., $W_i(\pi_i)$ is characterized as follows,*

Case 1. Positively correlated channel ($p_i > r_i$)

$$W_i(\pi_i)=\begin{cases} R_i(\pi_i) & \text{if } \pi_i \geq p_i \\ \frac{\beta\pi_i R_i(p_i)+(1-\beta p_i)R_i(\pi_i)}{1+\beta\pi_i-\beta p_i} & \text{if } \pi_i^0 \leq \pi_i < p_i \\ [R_i(\pi_i)-\beta R_i(Q_i(\pi_i))]+\beta[\pi_i-\beta Q_i(\pi_i)]V_{i,W_i(\pi_i)}(p_i) & \\ +\beta[(1-\pi_i)-\beta(1-Q_i(\pi_i))]V_{i,W_i(\pi_i)}(r_i) & \text{if } \pi_i < \pi_i^0 \end{cases}$$

Case 2. Negatively correlated channel ($p_i \leq r_i$)

$$W_i(\pi_i)=\begin{cases} R_i(\pi_i) & \text{if } \pi_i \geq r_i \\ \frac{(1-\beta)[R_i(\pi_i)+\beta(1-\pi_i)V_{i,W_i(\pi_i)}(r_i)]}{1-\beta\pi_i} & \text{if } Q_i(p_i)\leq\pi_i<r_i \\ (1-\beta)\big[R_i(\pi_i)+\beta[\pi_i V_{i,W_i(\pi_i)}(p_i)+(1-\pi_i)V_{i,W_i(\pi_i)}(r_i)]\big] & \\ & \text{if } \pi_i^0\leq\pi_i<Q_i(p_i) \\ [R_i(\pi_i)-\beta R_i(Q_i(\pi_i))]+\beta[\pi_i-\beta Q_i(\pi_i)]V_{i,W_i(\pi_i)}(p_i) & \\ +\beta[(1-\pi_i)-\beta(1-Q_i(\pi_i))]V_{i,W_i(\pi_i)}(r_i) & \text{if } \pi_i < \pi_i^0 \end{cases}$$

Proof: The derivation of the index value follows from substituting the expression of $V_{i,\omega_i}(p_i)$ and $V_{i,\omega}(r_i)$ (given in Lemma 2) into Equation (5). Details of the proof are provided in Appendix C. ■

Remark: Notice that Proposition 4 does not give the closed form expression for $W_i(\pi_i)$. However, since the closed form expression of the value function $V_{i,W_i(\pi_i)}(p_i)$ and $V_{i,W_i(\pi_i)}(r_i)$ are derived in Lemma 2, closed form expressions of $W_i(\pi_i)$ can be easily calculated and is given in Appendix C. We now introduce Whittle's index policy.

Whittle's Index Policy: *In each slot, with belief values $\pi_1, \ldots, \pi_N$, the user I with the highest index value $W_i(\pi_i)$ is scheduled for transmission, i.e., $I = \arg\max_i W_i(\pi_i)$.*

Note that, from the definition of indexability, the index value $W_i(\pi_i)$ monotonically increases with π_i. Therefore, when

$$\Theta = \frac{(1-\beta^{L(r,\pi^*(\omega))})\omega + (1-\beta)\beta^{L(r,\pi^*(\omega))}[R(Q^{L(r,\pi^*(\omega))}(r)) + \beta Q^{L(r,\pi^*(\omega))}(r)V_\omega(p)]}{(1-\beta)[1-\beta^{L(r,\pi^*(\omega))+1}(1-Q^{L(r,\pi^*(\omega))}(r))]} \tag{9}$$

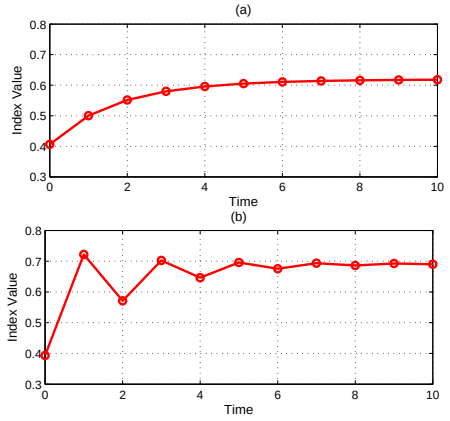

Fig. 5. Index value evolution of user i, with $\pi_i[0] = 0.3$. (a) Positive correlation, p_i=0.8, r_i=0.2; (b) Negative correlation, p_i=0.2, r_i=0.8.

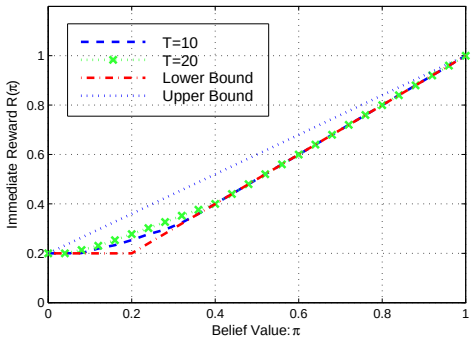

Fig. 6. Immediate reward versus π.

the Markovian channels have the same Markovian structure and vary independently across users (hence the state-index mappings are the same across users), Whittle's index policy essentially becomes the *greedy* policy – schedule the user with the highest belief value.

Fig. 5 plots an example of the index value evolution for the case of positively correlated and negatively correlated channels when they stay idle, i.e., not scheduled for transmission. We see that, for the positively correlated channel, the index value behaves monotonically, while, for the negatively correlated channel, the index value shows oscillation. This resembles the evolution of the belief values, which, as proven in Lemma 3 in Appendix A, approaches steady state monotonically for the positively correlated channel, and with oscillation for the negatively correlated channel. This resemblance in Fig. 5 is expected since, from Proposition 3, we can infer that the index value monotonically increases with the belief value. Thus, in essence, from Proposition 3 and Fig. 5, we see that the index value captures the underlying dynamics of the Markovian channel.

VII. Numerical Performance Analysis

A. Model for Simulation

In this section, we study, via numerical evaluations, the performance of Whittle's index policy, henceforth simply the index policy, for joint estimation and scheduling in our downlink system. We consider the specific class of estimator and rate adapter structure, with pilot-aided training, discussed in Section II and illustrated in Fig. 3. We consider a fading

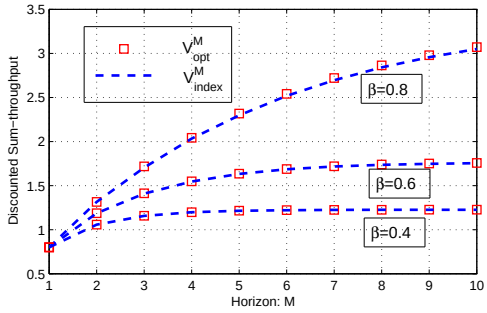

Fig. 7. Performance of the index policy in comparison with that of the optimal policy. System parameters used: N=5, $\{p_1$=0.2, r_1=0.75$\}$, $\{p_2 = 0.6, r_2 = 0.25\}$, $\{p_3 = 0.8, r_3 = 0.3\}$, $\{p_4 = 0.4, r_4$=0.7$\}$, $\{p_5 = 0.65, r_5 = 0.55\}$; Fading block length: $T = 20$.

channel with the fading coefficients quantized into two levels to reflect the two states of the Markov chain. Additive noise is assumed to be white Gaussian. The channel input-output model is given by $Y = hX + \epsilon$, where X, Y correspond to transmitted and received signals, respectively, h is the complex fading coefficient and ϵ is the complex Gaussian, unit variance additive noise. Conditioned on h, the Shannon capacity of the channel is given by $R = \log(1 + |h|^2)$. We quantize the fading coefficients such that the allowed rate at the lower state, $\delta = 0.2$ for all users. The channel state, represented by the fading coefficient, evolves as Markov chain with fading block length T.

We consider a class of Linear Minimum Mean Square Error (LMMSE) estimators [22] denoted as Φ. LMMSE estimators are attractive because with additive white Gaussian noise, they can be characterized in closed form [22] and, hence, can be conveniently used in simulation. Let ϕ_π denote the optimal LMMSE estimator with prior $\{\pi, 1-\pi\}$. We let Φ denote the set of LMMSE estimators optimized for various values of π.

B. Immediate Reward Structure

We now study the structure of the immediate reward $R(\pi)$. Note that $R(\pi)$ is optimized over the class of estimators Φ. Fig. 6 illustrates $R(\pi)$, in comparison with the upper and lower bounds derived in Lemma 2, for two values of block length T. As established in Lemma 2, $R(\pi)$ shows a convex increasing structure and takes values within the bounds. Note that $R(\pi)$ also increases with T, since a larger T provides more channel uses for channel probing and data transmission.

C. Near-optimal Performance of Whittle's Index Policy

We proceed to evaluate the performance of the index policy and compare it with the optimal policy. In Fig. 7, we compare the expected rewards V_{opt}^M and V_{index}^M that, respectively, correspond to the optimal finite M-horizon policy and the index policy, for increasing horizon length M and randomly generated system parameters. The value of V_{opt}^M is obtained via brute-force search over the finite horizon. Fig. 7 illustrates the near optimal performance of the index policy. Also, as expected, the higher the value of β, the higher the expected reward.

Table I presents the performance of the index policy in a larger perspective. Here, with randomly generated system

N	β	V_{opt}	V_{index}	V_{nofb}	%gain
4	0.6337	1.6289	1.6289	1.4887	100 %
4	0.5896	1.5977	1.5866	1.2888	96.4045 %
4	0.6673	1.6537	1.6319	1.4342	90.0500 %
5	0.4537	0.9854	0.9854	0.9299	100 %
5	0.6082	1.6132	1.6072	1.4777	95.5518 %
5	0.6537	2.3728	2.3725	2.1494	99.8697 %
5	0.5397	1.6330	1.6330	1.5961	100 %

TABLE I
ILLUSTRATION OF THE GAINS ASSOCIATED WITH EXPLOITING CHANNEL MOMORY.

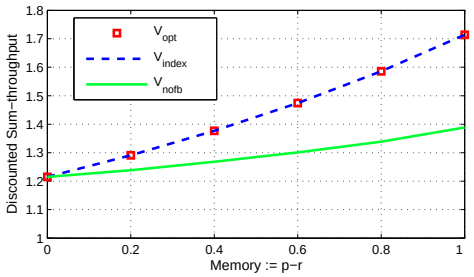

Fig. 8. Illustration of the influence of channel 'memory', $(p-r)$, on the performance of the index policy and baseline policies, when $\beta = 0.6$.

parameters, the infinite horizon reward under the index policy is compared with those of the optimal policy and a policy that 'throws away' the feedback from the scheduled user. Let V_{nofb} denote the reward under this 'no feedback' policy. The infinite horizon rewards are obtained as limits of the finite horizon until 1% convergence is achieved. The high values of the quantity $\%\text{gain}=\frac{V_{index}-V_{nofb}}{V_{opt}-V_{nofb}}\times 100\%$, in addition to underscoring the near-optimality of the index policy, also signifies the high system level gains from exploiting the channel memory using the end-of-slot feedback.

In Fig. 8 we study the effect of the channel 'memory' on the performances of various baseline policies. We consider five users with statistically identical but independently varying channels. Thus $p_i = p$, $r_i = r$, $i \in \{1, \cdots, 5\}$. We define the channel 'memory' as the difference $p - r$ and increase the memory by increasing p from 0.5 to 1 and maintaining $r = 1 - p$. Note that, with this approach, $p + r = 1$. Under this condition, the steady state probability that a channel is in the higher state h is kept constant under varying channel memory. This, essentially, provides a degree of fairness between systems with different channel memories. Fig. 8 compares the rewards V_{opt}, V_{index} and V_{nofb} that respectively correspond to the rewards under the optimal policy, the index policy, and the 'no feedback' policy introduced earlier, for increasing channel memory. Note that when $p = r$, the channel of each user evolves *i.i.d.* across time, with no information contained in the channel state feedback. Thus the policy that throws away this feedback achieves the same performance as the optimal policy that optimally uses this feedback, i.e., $V_{nofb} = V_{opt}$ when $p = r$. Also, since the channels are *i.i.d.* across users, when $p = r$, the index policy simplifies to a 'randomized' policy that schedules randomly and uniformly across users, in effect mirroring the 'no feedback' policy in this setting. This explains $V_{index} = V_{nofb}$ when $p = r$. As the channel memory increases, the significance of the channel state feedback increases, resulting in an increasing gap between the policies that use this feedback (optimal and Whittle's index policies) and the 'no feedback' policy.

Fig. 8, along with Table I, shows that exploiting channel memory for opportunistic scheduling can result in significant performance gains, and almost all of these gains can be realized using the easy-to-implement index policy.

D. Impact of Imperfections in Channel State Feedback

In realistic scenarios, the channel state feedback is subject to various imperfections such as random delays and errors, in turn, resulting from imperfections in the feedback generating mechanism and the feedback channel. In this section, we illustrate, via numerical experiments, that the index policy is robust against feedback imperfections, i.e., numerical results suggest that the index policy performs very close to the optimal [1] policy.

We first investigate the impact of feedback *delay* on the performance of the index policy. We consider the scenario where, once a user is scheduled, the corresponding channel state feedback is subject to a random delay which is *i.i.d.* across users. The delay in the feedback channel is an important consideration that cannot be overlooked in realistic scenarios. The effect of feedback delay on channel resource allocation has been studied under various settings in the past (e.g., [23]-[26]). While these works assume deterministic delay, we consider random, *i.i.d.* feedback delay. An instance when the feedback delay can be *i.i.d.* and non-negligible is when the delay is of the order of the scheduling slot length, resulting from channel propagation time of the feedback signal, and when the feedback channel environment changes drastically due to high mobility of users – a possibility in reality. We let $P_D(d)$ denote the probability that a channel state feedback experience d slot delays, where $d \in \{0, \cdots, d_{max}\}$. Here $d = 0$ indicates an end-of-slot feedback and d_{max} indicates the maximum delay that the feedback can experience. We assume the channel state feedback is time-stamped. Thus the scheduler takes this information into account when it updates the belief values upon receipt of a (possibly delayed) feedback signal. Specifically, at time slot t, if the latest (possibly delayed) feedback from user i corresponds to the channel state τ slots ago, then

$$\pi_i[t] = \begin{cases} Q^\tau(1) & \text{if } C_i[t-\tau] = h_i, \\ Q^\tau(0) & \text{if } C_i[t-\tau] = l_i. \end{cases}$$

where, recall that, Q^τ stands for the τ^{th} iteration of the function Q.

Now, with delay taken into account in the belief value updates, the performance of the original index policy is compared with that of the optimal policy in Table II with $d_{max} \in \{1, 2\}$ and $N \in \{3, 4\}$ and randomly generated system parameters (i.e., $P_D(d), p_i, r_i$, and $\vec{\boldsymbol{\pi}}[0]$). Note that the optimal policy takes into account the stochastic of the feedback delay and is implemented by exhaustive brute-force search, as before.

[1] optimal subject to the feedback imperfections

N	β	$[P_D(0), \cdots, P_D(d_{max})]$	(p_i, r_i), $i = 1, \cdots, N$	V_{index}	V_{opt}	%opt
3	0.6308	[0.8539,0.1461]	(0.5896,0.3478), (0.2703,0.6754), (0.6978,0.1376)	1.6644	1.6647	99.9863 %
4	0.5587	[0.4317,0.5683]	(0.8538,0.4462), (0.6529, 0.5207), (0.1792, 0.7268), (0.2877,0.9321)	1.7550	1.7560	99.9405 %
3	0.6536	[0.3682,0.5216,0.1102]	(0.9138,0.3075), (0.2298,0.7946), (0.3574,0.7851)	2.1044	2.1045	99.9963 %
4	0.5873	[0.6239,0.2589,0.1541]	(0.2513, 0.7258), (0.6285, 0.3801), (0.1676, 0.8245), (0.3058,0.6822)	1.4764	1.4790	99.8250 %

TABLE II
PERFORMANCE OF THE INDEX POLICY WITH UNDER RANDOMLY GENERATED SYSTEM PARAMETERS.

N	β	$[P_D(0), \cdots, P_D(d_{max})]$	V_{index}	V_{opt}	%opt
3	0.6	[1,0,0]	1.8941	1.8937	99.9802 %
3	0.6	[2/3,1/3,0]	1.8441	1.8440	99.9918 %
3	0.6	[1/3,1/3,1/3]	1.7883	1.7882	99.9943 %
3	0.6	[0,1/3,2/3]	1.7274	1.7267	99.9544 %
3	0.6	[0,0,1]	1.7195	1.7177	99.8969 %

TABLE III
PERFORMANCE OF THE INDEX POLICY UNDER VARIOUS DELAY DISTRIBUTIONS. SYSTEM PARAMETERS USED: $\{(p_i, r_i)_i\} = \{(0.2513, 0.7258), (0.6285, 0.3801), (0.1676, 0.8245), (0.3058, 0.6822)\}$

$N = 3$, $\beta = 0.6$, Delay=[2/3,1/3], $\vec{\boldsymbol{\pi}}[0] = [0.8, 0.3, 0.65]$ $\{(p_i, r_i)\}_i$=$\{(0.77, 0.25), (0.34, 0.90), (0.81, 0.30)\}$				$N = 4$, $\beta = 0.6$, Delay=[0.6,0.4], $\vec{\boldsymbol{\pi}}[0] = [0.8, 0.3, 0.65]$ $\{(p_i, r_i)\}_i$=$\{(0.85, 0.25), (0.6, 0.35), (0.2, 0.7), (0.3, 0.8)\}$			
ϵ	V_{index}	V_{opt}	%opt	ϵ	V_{index}	V_{opt}	%opt
0	1.8440	1.8441	99.9918 %	0	1.8085	1.8091	99.9703 %
0.25	1.7731	1.7757	99.8541 %	0.25	1.7528	1.7530	99.9864 %
0.5	1.7304	1.7304	100 %	0.5	1.7333	1.7333	100 %
0.75	1.7731	1.7757	99.8541 %	0.75	1.7528	1.7530	99.9864 %
1	1.8440	1.8441	99.9918 %	1	1.8085	1.8091	99.9703 %

TABLE IV
PERFORMANCE OF THE INDEX POLICY WITH KNOWN PROBABILITY OF ERROR IN CHANNEL STATE FEEDBACK.

$N = 3$, $\beta = 0.6$, Delay=[2/3,1/3], $\vec{\boldsymbol{\pi}}[0] = [0.8, 0.3, 0.65]$ $\{(p_i, r_i)\}_i$=$\{(0.77, 0.25), (0.34, 0.90), (0.81, 0.30)\}$				$N = 4$, $\beta = 0.6$, Delay=[0.6,0.4], $\vec{\boldsymbol{\pi}}[0] = [0.8, 0.3, 0.65]$ $\{(p_i, r_i)\}_i$=$\{(0.85, 0.25), (0.6, 0.35), (0.2, 0.7), (0.3, 0.8)\}$			
ϵ	V_{index}	V_{opt}	%opt	ϵ	V_{index}	V_{opt}	%opt
0	1.1.8440	1.8441	99.9918 %	0	1.8085	1.8091	99.9703 %
0.25	1.7722	1.7746	99.8695 %	0.25	1.7349	1.7397	99.7203 %
0.5	1.7031	1.7056	99.8520 %	0.5	1.6792	1.6849	99.6579 %
0.75	1.6305	1.6326	99.8743 %	0.75	1.6096	1.6147	99.6833 %
1	1.5692	1.5711	99.8794 %	1	1.5692	1.5711	99.8794 %

TABLE V
PERFORMANCE OF THE INDEX POLICY WITH UNKNOWN PROBABILITY OF ERROR IN CHANNEL STATE FEEDBACK.

The high value of the quantity $\%opt := V_{index}/V_{opt} \times 100\%$ indicates that Whittles index policy has a performance very close to that of the optimal policy under the delayed feedback setup. In Table III, the performance comparison is made under more controlled choice of delay, i.e., with $d_{\max}$ fixed at 2, and the tail of the delay mass function is gradually made heavy. We observe that as the delay tail grows heavier, the performances of both the optimal and the index policy decrease. This is expected because, with the delay tail growing heavier, the received channel state feedback progressively tends to become outdated, and hence the value of information contained in the feedback decreases, essentially reducing the performances of both the optimal and the index policies that use this feedback. In summary, Tables II and III illustrate that the index policy derived for the original system without feedback delay, performs very close to the optimal policy in the system *with* feedback delay, essentially indicating the robustness of the index policy.

We now study the performance of the index policy in the presence of random errors in the channel state feedbacks. This error could have initiated at the feedback generating mechanism at the user or during propagation in the feedback channel. Let $\mathcal{F}_i[t] \in \{l_i, h_i\}$ be the feedback received at the scheduler that corresponds to actual channel state $C_i[t]$. The channel state feedback error is characterized by the mismatch probability ϵ defined as follows:

$$\epsilon := \text{prob}\big(\mathcal{F}_i[t]{=}l_i \big| C_i[t]{=}h_i\big) = \text{prob}\big(\mathcal{F}_i[t]{=}h_i \big| C_i[t]{=}l_i\big).$$

We first assume that the error probability, ϵ, is *known* at the scheduler and compare the throughput performances in Table IV. Observe that, for various values of error probabilities, the index policy still has performance very close to the optimal

policy, essentially suggesting its robustness against feedback errors. As also observed from the table, the performances of both the optimal and Whittle's index policies are symmetric around $\epsilon = 0.5$ that corresponds to the worst rewards. This is expected since, when $\epsilon = 0.5$, the feedback contains no information about the channel state, in turn, resulting in zero gain from exploiting channel memory.

We now consider the case when the scheduler is *unaware* that there is a (possible) error in the channel state feedback, and study its impact on the performances of the index and the optimal policies. In this scenario, the scheduler simply trusts the feedback to be accurate when making scheduling decisions. The performances of both policies under various values of ϵ are recorded in Table V. Once again, the high value of $\%opt$ suggests the robustness of the index policy against feedback errors even when the scheduler is *unaware* of the possible presence of errors. Also, as expected, when the error probability, ϵ, increases, the performances of both policies decrease monotonically. This phenomenon contrasts to the case in Table IV, when the scheduler is aware of the presence of errors and its stochastic, i.e., the value ϵ.

E. Impact of Incorrect Transmission Rate knowledge

We now study the robustness of the index policy under mismatch between the supportable lower state transmission rate assumed at the scheduler and the actual lower state transmission rate. Recall that δ_i denotes the allowable transmission rate at the lower state l_i of user i. Let δ_i' denote the lower state transmission rate *assumed* at the scheduler for this channel. We also assume that the scheduler is unaware of the presence of this mismatch which could have resulted from errors in the initial rate estimate or when the actual underlying rate has shifted from the initial value over time. We consider the case when the actual lower state transmission rate and the assumed rate at the scheduler are identical across users, i.e., $\delta_i = \delta_j$ and $\delta_i' = \delta_j'$ for all $i, j \in \{1, \cdots, N\}$. We assume $\delta = 0.5$ and compare the performance of the index policy with that of the optimal policy in Table VI, for various values of δ'. Note that as before, the optimal policy is defined within the context of the imperfection (δ mismatch, in the present case). It can be observed that for various values of δ', the index policy closely tracks the performance of the optimal policy, indicating its robustness against transmission rate mismatch. Also, it can be expected that when $\delta' < \delta$, both Whittle's index and optimal policies under-utilize the available channel rate, resulting in reduced performances. On the other hand, when $\delta' > \delta$, both policies aggressively transmit, leading to outage, thus resulting in reduced performances. This can be observed in Table VI where the performances drop monotonically as δ' deviates from δ. Also, the drop in performance appears to be more severe when $\delta' > \delta$, suggesting that aggressive transmission and the resulting outages can be more detrimental than conservative transmission and the associated channel under-utilization.

δ'	V_{index}	V_{opt}	%opt
0.1	1.82698	1.82698	100 %
0.3	1.83417	1.83421	99.99820 %
0.5	1.87696	1.87746	99.97328 %
0.7	1.66211	1.66412	99.87927 %
0.9	1.01255	1.01514	99.74553 %

TABLE VI
PERFORMANCE OF THE INDEX POLICY UNDER IMPERFECT KNOWLEDGE OF LOWER STATE TRANSMISSION RATE. SYSTEM PARAMETERS USED: $\delta = 0.5$, $N = 3$, $\beta = 0.6$, $\{(p_i, r_i)_i\}_i = \{(0.38, 0.05), (0.16, 0.95), (0.86, 0.12)\}$.

VIII. CONCLUSION

In this paper, we studied downlink multiuser scheduling under Markov-modeled channels. We considered the scenario where the channel state information is not perfectly known at the scheduler, essentially requiring a joint design of user selection, channel estimation and rate adaptation. This calls for a two-stage optimization: (1) Within each slot, the channel estimation and rate adaptation is optimized to obtain an optimal transmission rate in the scheduling slot; (2) Across scheduling slots, users are selected to maximize the infinite horizon discounted reward. We formulated the scheduling problem as a partially observable Markov decision process with the classic 'exploitation versus exploration' trade-off. We then linked the problem to a restless multiarmed bandit processes and conducted a Whittle's indexability analysis. By obtaining structural properties of the optimal reward within the indexability setup, we showed that the downlink scheduling problem is Whittle indexable. We then explicitly characterized the the index policy and studied the performance of this policy using extensive numerical experiments, which suggest that the index policy has near optimal performance and that significant system level gains can be realized by exploiting the channel memory for joint channel estimation and scheduling. Numerical experiments also suggest that the index policy is robust against various imperfections in channel state feedback.

APPENDIX A
PROOF OF LEMMA 2

We first establish structural properties of the belief update when a user stays idle. Suppose a user has the initial belief value $\pi[0]$ and stays idle at all times, the belief value at t^{th} slot is then given by $\pi[t] = Q^t(\pi_i[0])$, where Q^t is the t^{th} iteration of function Q, given by

$$Q^t(\pi) = \frac{r - (p-r)^t\big(r - (1+r-p)\pi\big)}{1+r-p}. \qquad (10)$$

We let π^0 be the steady state distribution of the two-state channel being at the higher state, i.e.,

$$\pi^0 = \frac{r}{1+r-p}.$$

It is clear that $\pi^0 = \lim_{t\to\infty} Q^t(\pi)$. An example of the belief evolution when a user stays idle is depicted in Fig. 9. This figure shows that, when staying idle, the belief value approaches steady state monotonically for positively correlated channel and approaches steady state with oscillation for negatively correlated channel. The structural properties of

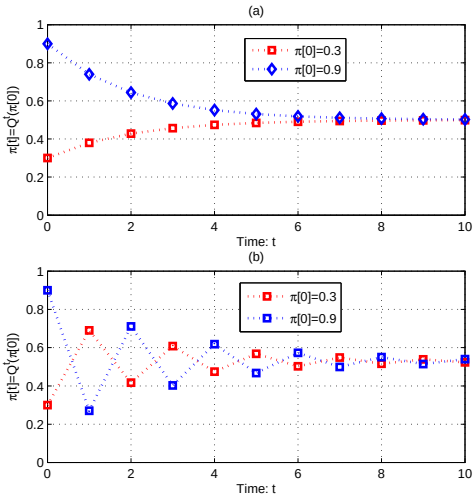

Fig. 9. Evolution of belief values under consecutive idle decisions. (a) Positive correlation, $p = 0.8$, $r = 0.2$; (b) Negative correlation, $p = 0.2$, $r = 0.8$.

$Q^t(\pi_i[0])$ is critical to the rest of the proof and is recorded in the following lemma.

Lemma 3.
(i) For positively correlated channel (i.e., $p > r$), $\pi[t]$ converges to steady state π^0 monotonically. For negatively correlated channel (i.e., $p \leq r$), $\pi[t]$ converges to steady state π^0 with oscillation and a monotonically converging envelope. (ii) $\min\{p, r\} \leq Q^t(\pi_i[0]) \leq \max\{p, r\}$ for all $t = 1, 2, \cdots$ and $\pi_i[0] \in [0, 1]$.

Proof: (i) Since we have $0 < p - r \leq 1$ for positively correlated channel and $-1 \leq p - r \leq 0$ for negatively correlated channel, it is clear from the expression of (10) that $\pi[t]$ converges to steady state π^0 monotonically and approaches steady state π^0 with oscillation and a monotonically converging envelop.
(ii) Since we have established part (i), it suffices to check that the first step transition satisfies: $\min\{p, r\} \leq Q[\pi] \leq \max\{p, r\}$, for all π, as shown below.

$$Q(\pi) = \frac{r - (p - r)\big(r - (1 + r - p)\pi\big)}{1 + r - p}.$$

For positively correlated channel, since $p - r > 0$

$$Q(\pi) \geq \frac{r - (p - r)r}{1 + r - p} = r.$$

$$Q(\pi) \leq \frac{r - (p - r)\big(r - (1 + r - p)\big)}{1 + r - p} = \frac{p(1 - p + r)}{1 + r - p} = p.$$

For negatively correlated channel, since $p - r \leq 0$,

$$Q(\pi) \leq \frac{r - (p - r)r}{1 + r - p} = r.$$

$$Q(\pi) \geq \frac{r - (p - r)\big(r - (1 + r - p)\big)}{1 + r - p} = \frac{p(1 - p + r)}{1 + r - p} = p.$$

The lemma is thus proved. ■

We then define $L(\pi, \pi^*)$ as the time needed for belief value of a user to exceed π^* from below, starting from initial value π. Formally,

$$L(\pi, \pi^*) = \min_t \{Q^t(\pi) > \pi^*\}$$

Using Lemma 3 and expression (10), $L(\pi, \pi^*)$ can be calculated as follows.

- Positive correlation ($p > r$)

$$L(\pi, \pi^*) = \begin{cases} 0 & \text{if } \pi > \pi^* \\ \lfloor log_{p-r} \frac{r-(1+r-p)\pi^*}{r-(1+r-p)\pi} \rfloor + 1 & \text{if } \pi \leq \pi^* < \pi_0 \\ \infty & \text{if } \pi \leq \pi^* \quad \text{and } \pi^* \geq \pi_0 \end{cases}$$

- Negative correlation ($p \leq r$)

$$L(\pi, \pi^*) = \begin{cases} 0 & \text{if } \pi > \pi^*; \\ 1 & \text{if } \pi \leq \pi^* \text{ and } Q(\pi) > \pi^*, \\ \infty & \text{if } \pi \leq \pi^* \text{ and } Q(\pi) \leq \pi^*. \end{cases}$$

We shall refer to the 'active set' as the set of belief values for which the optimal decision is to transmit. The 'idle set' denotes the set of belief values for which the optimal decision is to stay idle. We proceed to derive the value functions $V_\omega(p)$ and $V_\omega(r)$ based on the value of $\pi^*(\omega)$.

(1) Positive correlation ($p > r$).

- When $\pi^*(\omega) \geq p$, the belief value p is thus in the 'idle set'. From Lemma 3(ii), if $\pi[0] = p$, the system stays idle. Hence the reward function is expressed as

$$V_\omega(p) = \omega + \beta\omega + \beta^2\omega + \cdots = \frac{\omega}{1 - \beta}.$$

- When $\pi^*(\omega) < p$, the belief value p is then in the 'active set'. Hence from the Bellman equation in (5),

$$V_\omega(p) = R(p) + \beta\big(pV_\omega(p) + (1 - p)V_\omega(r)\big).$$

Rearranging the terms yields,

$$V_\omega(p) = \frac{R(p) + \beta(1 - p)V_\omega(r)}{1 - \beta p}.$$

- When $\pi^*(\omega) < r$, the value r is then in 'active set'. From Lemma 3(ii), regardless of the scheduling decision, the belief values $\pi[t]$, starting from $\pi[0] = r$, stays in the 'active set'. Therefore

$$V_\omega(r) = \sum_{t=0}^{\infty} \beta^t R(Q^t(r)) = \sum_{t=0}^{\infty} \beta^t R\Big(\frac{r-(p-r)^{t+1}r}{1+r-p}\Big).$$

- When $\pi^*(\omega) \geq \pi^0$, since $\pi^0 \geq r$, the belief value r is in 'idle set'. From Lemma 3(i), the belief values $\pi[t]$, starting from $\pi[0] = r$, stays in 'idle set'. Hence

$$V_\omega(r) = \omega + \beta\omega + \beta^2\omega + \cdots = \frac{\omega}{1 - \beta}.$$

- When $r < \pi^*(\omega) < \pi^0$, the belief value r is therefore in 'idle set'. Since the channel is positively correlated, from Lemma 3, starting from $\pi[0] = r$, the user remains idle for a duration of $L(r, \pi^*(\omega))$ slots. Therefore

$$V_\omega(r) = \frac{1-\beta^{L(r,\pi^*(\omega))}}{1-\beta}\omega + \beta^{L(r,\pi^*(\omega))} V_\omega^1\big(Q^{L(r,\pi^*(\omega))}(r)\big). \quad (11)$$

where

$$V_\omega^1\big(Q^{L(r,\pi^*(\omega))}(r)\big) = R\big(Q^{L(r,\pi^*(\omega))}(r)\big)+ \beta\big(Q^{L(r,\pi^*(\omega))}(r)V_\omega(p) + (1-Q^{L(r,\pi^*(\omega))}(r))V_\omega(r)\big)$$

Substituting the above expression in (11), we can obtain $V_\omega(r) = \Theta$ as given in expression (9) in the lemma.

(2) Negative correlation ($p \le r$).

The derivation of $V_\omega(p)$ and $V_\omega(r)$ for negative correlation case follows an approach similar to that for the case of positive correlation. Details are, therefore, omitted here. ■

Note that the expressions of the value functions $V_\omega(p)$ and $V_\omega(r)$ in Lemma 2 are not in closed form. However, the closed form expressions for $V_\omega(p)$ and $V_\omega(r)$ can be easily calculated based on the expressions in Lemma 2, recorded below.

Case (1) Positive correlation ($p > r$). First we give the closed form expression of $V_\omega(p)$.

- If $\pi^*(\omega) < \pi^0$,
$$V_\omega(p) = \sum_{t=0}^{\infty} \beta^t R\Big(\frac{r+(p-r)^{t+1}(1-p)}{1+r-p}\Big).$$
- If $\pi^0 \le \pi^*(\omega) < p$,
$$V_\omega(p) = \frac{\beta(1-p)\omega + (1-\beta)R(p)}{(1-\beta)(1-\beta p)}.$$
- If $\pi^*(\omega) \ge p$, $V_\omega(p) = \omega/(1-\beta)$.

We proceed to give the closed form expression of $V_\omega(r)$.

- If $\pi^*(\omega) < r$,
$$V_\omega(r) = \sum_{t=0}^{\infty} \beta^t R\Big(\frac{r-(p-r)^{t+1}r}{1+r-p}\Big).$$
- If $r \le \pi^*(\omega) < \pi^0$, $V_\omega(r)$ is given in equation (12).
- If $\pi^*(\omega) \ge \pi^0$, $V_\omega(r) = \omega/(1-\beta)$.

Case (2) Negative correlation ($p \le r$). In this case, the closed form expression of $V_\omega(r)$ is given as follows.

- If $\pi^*(\omega) \ge r$, then $V_\omega(r) = \omega/(1-\beta)$.
- If $Q(p) \le \pi^*(\omega) < r$, then
$$V_\omega(r) = \frac{\beta r\omega + (1-\beta)R(r)}{(1-\beta)(1-\beta(1-r))}.$$
- If $p \le \pi^*(\omega) < Q(p)$, we have
$$V_\omega(r) = \frac{\beta r\omega + \beta^2 r R(Q(p)) + (1-\beta^2 Q(p))R(r)}{(1-\beta(1-r))(1-\beta^2 Q(p)) - \beta^3 r(1-Q(p))}.$$
- If $\pi^*(\omega) < p$, then
$$V_\omega(r) = \sum_{t=0}^{\infty} \beta^t R\Big(\frac{r-(p-r)^{t+1}r}{1+r-p}\Big).$$

Then we give the closed form expression of $V_\omega(p)$.

- If $\pi^*(\omega) < p$, then
$$V_\omega(p) = \sum_{t=0}^{\infty} \beta^t R\Big(\frac{r+(p-r)^{t+1}(1-p)}{1+r-p}\Big).$$
- If $p \le \pi^*(\omega) < Q(p)$, we have
$$V_\omega(p) = \frac{\big(1-\beta(1-r)\big)\big[\omega+\beta R\big(Q(p)\big)\big]+\beta^2\big(1-Q(p)\big)R(r)}{\big(1-\beta(1-r)\big)\big(1-\beta^2 Q(p)\big) - \beta^3 r\big(1-Q(p)\big)}.$$
- If $\pi^*(\omega) \ge Q(p)$, then $V_\omega(p) = \omega/(1-\beta)$.

Appendix B
Proof of Proposition 3

We prove that the problem is Whittle indexable by showing that $\pi^*(\omega)$ monotonically increases with ω. It is clear from Proposition 2 that $\pi^*(\omega) = \kappa$ for $\omega \in [0, \delta)$. So it suffices to show that $\pi^*(\omega)$ is strictly increasing for $\omega \in [\delta, 1]$. The proof technique follows along the lines of [18] and is presented next. We first proceed with the following lemma.

Lemma 4. *If for all $\omega \in [\delta, 1]$, we have*

$$\frac{dV_\omega^1(\pi)}{d\omega}\Big|_{\pi=\pi^*(\omega)} < \frac{dV_\omega^0(\pi)}{d\omega}\Big|_{\pi=\pi^*(\omega)}, \tag{13}$$

then $\pi = \pi^*(\omega)$ is strictly increasing with ω for $\omega \in [\delta, 1]$.

Proof: The lemma is proven by contradiction. Suppose there exists $\omega_0 \in [\delta, 1]$, such that $\pi^*(\omega)$ is decreasing (i.e., non-increasing) at ω_0, hence it is decreasing in a neighborhood of ω_0, say, $[\omega_0, \omega_0 + \triangle\omega]$. Since $V^1_{\omega_0+\triangle\omega}\big(\pi^*(\omega_0+\triangle\omega)\big) = V^0_{\omega_0+\triangle\omega}\big(\pi^*(\omega_0+\triangle\omega)\big)$ and $\pi^*(\omega)$ is decreasing at ω_0, $\pi^*(\omega_0)$ is within the 'active set' for the $(\omega_0+\triangle\omega)$-subsidy problem. Therefore we have $V^1_{\omega_0+\triangle\omega}\big(\pi^*(\omega_0)\big) \ge V^0_{\omega_0+\triangle\omega}\big(\pi^*(\omega_0)\big)$. Besides, from the definition of threshold value $\pi^*(\omega_0)$, $V^1_{\omega_0}\big(\pi^*(\omega_0)\big) = V^0_{\omega_0}\big(\pi^*(\omega_0)\big)$. Therefore,

$$\begin{aligned}\frac{dV_\omega^1(\pi)}{d\omega}\Big|_{\pi=\pi^*(\omega)} &= \lim_{\triangle\omega\to 0} \frac{V^1_{\omega_0+\triangle\omega}\big(\pi^*(\omega_0)\big) - V^1_{\omega_0}\big(\pi^*(\omega_0)\big)}{\triangle\omega} \\ &\ge \lim_{\triangle\omega\to 0} \frac{V^0_{\omega_0+\triangle\omega}\big(\pi^*(\omega_0)\big) - V^0_{\omega_0}\big(\pi^*(\omega_0)\big)}{\triangle\omega} \\ &= \frac{dV_\omega^0(\pi)}{d\omega}\Big|_{\pi=\pi^*(\omega)},\end{aligned}$$

which contradicts with the assumption. ■

Therefore, to establish indexability, it suffices to prove the inequality (13), i.e., $\frac{dV_\omega^1(\pi)}{d\omega}\big|_{\pi=\pi^*(\omega)} < \frac{dV_\omega^0(\pi)}{d\omega}\big|_{\pi=\pi^*(\omega)}$. Let $D_\omega(\pi)$ be the discounted time the ω-subsidy process, with initial belief π, is made passive, i.e.,

$$D_\omega(\pi) = \sum_{t=0}^{\infty} \beta^t \mathbb{1}(a[t] = 0).$$

$$V_\omega(r) = \frac{(1-\beta^{L(r,\pi^*(\omega))})\omega+(1-\beta)\beta^{L(r,\pi^*(\omega))}\big[R(Q^{L(r,\pi^*(\omega))}(r))+\beta Q^{L(r,\pi^*(\omega))}(r)\sum_{t=0}^{\infty}\beta^t R\big(\frac{r+(p-r)^{t+1}(1-p)}{1+r-p}\big)\big]}{(1-\beta)\big[1-\beta^{L(r,\pi^*(\omega))+1}(1-Q^{L(r,\pi^*(\omega))}(r))\big]} \tag{12}$$

It follows from [14] that $D_\omega(\pi)=\frac{dV_\omega(\pi)}{d\omega}$. Taking derivative of both sides of the Bellman equation in (5) with respect to ω, the objective (13) now becomes

$$\beta\big(\pi^*(\omega)D_\omega(p)+(1-\pi^*(\omega))D_\omega(r)\big)<1+\beta D_\omega\big(Q(\pi^*(\omega))\big). \tag{14}$$

Case (1) If $0\le\pi^*(\omega)<\min\{p,r\}$, from Lemma 3(ii), starting from the initial belief value $\pi[0]=r$ or $\pi[0]=p$, the believe value $\pi[t]$ never evolves below $\pi^*(\omega)$, hence the project is active at all times under optimal control. Therefore $D_\omega(p)=D_\omega(r)=D_\omega(Q(\pi^*(\omega)))=0$. Equation (14) thus holds.

Case (2) If $\pi_0\le\pi^*(\omega)\le 1$, starting from initial belief $\pi[0]=Q(\pi^*(\omega))$, the belief value $\pi[t]$ always stays within the 'idle set', i.e., $D_\omega(Q(\pi^*(\omega)))=\frac{1}{1-\beta}$. Equation (14) holds since $D_\omega(p)\le 1+\beta+\beta^2+\cdots=\frac{1}{1-\beta}$ and, similarly, $D_\omega(r)\le\frac{1}{1-\beta}$.

Case (3) If $\min\{p,r\}\le\pi^*(\omega)\le\pi_0$, from Lemma 3(ii), $Q(\pi^*(\omega))$ is in 'active set'. Since

$$\begin{aligned}&V_\omega(Q(\pi^*(\omega)))\\ =&R(Q(\pi^*(\omega))+\beta[Q(\pi^*(\omega))V_\omega(p)+(1-Q(\pi^*(\omega)))V_\omega(r)],\end{aligned}$$

we have

$$\begin{aligned}&D_\omega(Q(\pi^*(\omega)))\\ =&\beta[Q(\pi^*(\omega))D_\omega(p)+(1-Q(\pi^*(\omega)))D_\omega(r)].\end{aligned} \tag{15}$$

We then discuss Equation (15) separately for negatively and positively correlated channels.

• Negatively correlated channel ($p\le r$). Since $r>\pi^0>\pi^*(\omega)$, the belief value r is in the 'active set', hence

$$V_\omega(r)=R(r)+\beta(rV_\omega(p)+(1-r)V_\omega(r)).$$

Therefore, we have

$$D_\omega(r)=\beta(rD_\omega(p)+(1-r)D_\omega(r)). \tag{16}$$

Substituting equation (15) and (16) in (14), we get

$$\frac{\beta}{1-\beta(1-r)}D_\omega(p)(1-\beta)(\beta r+\pi^*(\omega)-\beta Q(\pi^*(\omega)))<1.$$

Following the same technique as in [18], the above inequality can be verified by substituting $\pi^*(\omega)$ by π^0 and $D_\omega(p)$ by $\frac{1}{1-\beta}$.

• Positively correlated channel ($p>r$). In this case, p is in the 'active set', hence

$$V_\omega(p)=R(p)+\beta\big(pV_\omega(p)+(1-p)V_\omega(r)\big).$$

Taking derivative with respect to ω we have,

$$D_\omega(p)=\beta\big(pD_\omega(p)+(1-p)D_\omega(r)\big). \tag{17}$$

Substituting equations (15) and (17) in (14), we have

$$\beta D_\omega(p)(1-\beta)(1-\frac{\pi^*(\omega)-\beta Q(\pi^*(\omega))}{1-\beta p})<1.$$

By applying the same technique as in [18], it can be checked that the above inequality indeed holds.

Therefore the inequality (14) is justified and hence indexability holds. ■

APPENDIX C
PROOF OF PROPOSITION 4

For ω-subsidy problem of user i, from indexability, we know that $\pi_i^*(\omega)$ strictly increases from 0 to 1 as ω increases from δ_i to 1. Hence the index value, from its definition in (8), is the subsidy value for which the active and idle decisions are equally attractive. We can hence derive index value $W_i(\pi_i)$ by equating $V_{i,\omega}^1(\pi_i)$ and $V_{i,\omega}^0(\pi_i)$ and solve for ω as a function of π_i, i.e.,

$$\begin{aligned}&W_i(\pi_i)+\beta V_{i,W_i(\pi_i)}(Q_i(\pi_i))\\ =&R(\pi_i)+\beta[\pi_i V_{i,W_i(\pi_i)}(p_i)+(1-\pi_i)V_{i,W_i(\pi_i)}(r_i)].\end{aligned} \tag{18}$$

Note that the expressions of $V_{i,\omega}(p_i)$ and $V_{i,\omega}(r_i)$ have been given by Lemma 2. Substituting in (18) the values of $V_{i,\omega}(p_i)$ and $V_{i,\omega}(r_i)$, we obtain the index value expressions, explained in the following.

Case (1). Positively correlation ($p_i>r_i$).

• If $\pi_i\ge p_i$, the belief value $Q_i(\pi_i)$, p_i, r_i are in the 'idle set' and, starting from initial belief $\pi_i[0]=Q_i(\pi_i)$ or $\pi_i[0]=p_i$, or $\pi_i[0]=r_i$, $\pi_i[t]$ will stay in the 'idle set'. Hence

$$V_{i,\omega}(Q_i(\pi_i))=V_{i,\omega}(p_i)=V_{i,\omega}(r_i)=\frac{\omega}{1-\beta}.$$

Substituting the above expressions in (18) we obtain that $W_i(\pi_i)=R(\pi_i)$.

• If $\pi_i^0\le\pi_i<p_i$, then p_i is in 'active set', and starting from initial belief $\pi_i[0]=r_i$ or $\pi_i[0]=Q_i(\pi_i)$, $\pi_i[t]$ stays within 'idle set' at all times. Hence

$$V_{i,\omega}(Q_i(\pi_i))=V_{i,\omega}(r_i)=\frac{\omega}{1-\beta}.$$

Substituting the above expressions and the expression of $V_{i,\omega}(p_i)$ (given in Lemma 2) in equation (18), we get

$$W_i(\pi_i)=\frac{\beta\pi_i R(p_i)+(1-\beta p_i)R(\pi_i)}{1+\beta\pi_i-\beta p_i}.$$

• If $\pi_i<\pi_i^0$, then the value $Q_i(\pi_i)$ is in the 'active set'. Therefore,

$$\begin{aligned}V_{i,\omega}\big(Q_i(\pi_i)\big)=&R(Q_i(\pi_i))+\\ &\beta[Q_i(\pi_i)V_{i,\omega}(p_i)+(1-Q_i(\pi_i))V_{i,\omega}(r_i)].\end{aligned}$$

Again, substituting the expression of $V_{i,\omega}\big(Q_i(\pi_i)\big)$ in equation (18), we have

$$\begin{aligned}W_i(\pi_i)=&[R(\pi_i)-\beta R(Q_i(\pi_i))]+\beta[\pi_i-\beta Q_i(\pi_i)]V_{i,W_i(\pi_i)}(p_i)\\ &+\beta[(1-\pi_i)-\beta(1-Q_i(\pi_i))]V_{i,W_i(\pi_i)}(r_i).\end{aligned}$$

Case (2). Negative correlation ($r_i\ge p_i$).

Using the similar approach as in the positive correlation case, the expressions of the index value can be derived for the case of negative correlation, which are given in the Proposition 4. Details are, therefore, omitted here. ■

Note that the expressions given in Proposition 4 are not in closed form. However, the closed form expression for the index value $W_i(\pi_i)$ can be easily calculated based on these expressions provided in Proposition 4, recorded as follows.

Case (1). Positively correlated channel ($p_i > r_i$).

- If $\pi_i \geq p_i$, then the index value $W_i(\pi_i) = R_i(\pi_i)$.
- If $\pi_i^0 \leq \pi_i < p_i$, then

$$W_i(\pi_i) = \frac{\beta\pi_i R_i(p_i) + (1-\beta p_i)R_i(\pi_i)}{1+\beta\pi_i - \beta p_i}$$

- If $r_i \leq \pi_i < \pi_i^0$, $W_i(\pi_i)$ is given in equation (19), where

$$\Gamma_i = (1-\beta)\big[1-\beta^{L(r_i,\pi_i)+1}\big(1-Q_i^{L(r_i,\pi_i)}(r_i)\big)\big],$$
$$\Lambda_i = (1-\beta)\beta^{L(r_i,\pi)}\big[R_i(Q_i^{L(r_i,\pi_i)}(r_i)) + \beta Q^{L(r_i,\pi_i)}(r_i)\sum_{t=0}^{\infty}\beta^t R\big(\frac{r_i+(p_i-r_i)^{t+1}(1-p_i)}{1+r_i-p_i}\big)\big].$$

- If $\pi_i < r_i$, the index value $W_i(\pi_i)$ is given in equation (20).

Case (2). Negatively correlated channel ($p_i \leq r_i$.)

- If $\pi_i \geq r_i$, we have $W_i(\pi_i) = R_i(\pi_i)$.
- If $Q_i(p_i) \leq \pi_i < r_i$, then

$$W_i(\pi_i) = \frac{(1-\beta)[1-\beta(1-r_i)]R(\pi_i) + \beta(1-\beta)(1-\pi_i)R(r_i)}{[1-\beta\pi_i][1-\beta(1-r_i)] - \beta^2(1-\pi_i)r_i}.$$

- If $\pi_i^0 \leq \pi_i < Q_i(p_i)$, the index value is expressed as

$$W_i(\pi_i) = \frac{(1-\beta)R(\pi_i)\Delta_i + \beta(1-\beta)\pi_i\Omega_i + \beta(1-\beta)(1-\pi_i)\Upsilon_i}{\Delta_i - \beta(1-\beta)(1-\beta(1-r_i))\pi_i - (1-\beta)\beta^2 r_i(1-\pi_i)},$$

where

$$\Delta_i = \big(1-\beta(1-r_i)\big)\big(1-\beta^2 Q_i(p_i)\big) - \beta^3 r_i\big(1-Q_i(p_i)\big), \quad (21)$$
$$\Omega_i = \beta\big(1-\beta(1-r_i)\big)R_i(Q_i(p_i)) + \beta^2\big(1-Q_i(p_i)\big)R_i(r_i), \quad (22)$$
$$\Upsilon_i = \beta^2 r_i R_i(Q_i(p_i)) + (1-\beta^2 Q_i(p_i))R_i(r_i). \quad (23)$$

- If $p_i \leq \pi_i < \pi_i^0$, $W_i(\pi_i)$ is given in equation (24), where Δ_i, Ω_i and Υ_i are given by (21)-(23), respectively.

- If $\pi_i < p_i$, the index value $W_i(\pi_i)$ is given in equation (25).

REFERENCES

[1] R. Knopp, P. A. Humblet, "Information capacity and power control in single cell multiuser communications," in *IEEE International Conference on Communications*, 1995.

[2] D. Tse, P. Viswanath, *"Fundamentals of wireless communication,"* Cambridge University Press, 2005.

[3] L. Tassiulas, "Scheduling and performance limits of networks with constantly changing topology," *IEEE Transactions on Information Theory,* 1997.

[4] M. Neely, E. Modiano, C. Rohrs, "Dynamic power allocation and routing for time varying wireless networks," *IEEE Journal on Selected Areas in Communications,* vol. 23, pp. 89–103, 2005.

[5] S. Shakkottai, A. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: the exponential rule," *American Mathematical Society Translations,* vol. 207, pp. 185–202., 2002.

[6] A. Eryilmaz, R. Srikant, "Fair Resource Allocation in Wireless Networks using Queue-length based Scheduling and Congestion Control," *IEEE/ACM Transactions on Networking,* vol. 15, pp. 1333–1344, 2007.

[7] M. J. Neely, "Max weight learning algorithms with application to scheduling in unknown environments," *arXiv preprint:0902.0630,* 2009.

[8] C. Thejaswi, J. Zhang, S. Pun, V. H. Poor, "Distributed Opportunistic Scheduling with Two-Level Channel Probing," *IEEE/ACM Transactions on Networking,* vol. 18, pp.1464–1477, 2009.

[9] L. A. Johnston, V. Krishnamurthy, "Opportunistic file transfer over a fading channel: a POMDP search theory formulation with optimal threshold policies," *IEEE Transactions on Wireless Communications,* vol.5, pp. 394–405, 2006.

[10] S. Murugesan, P. Schniter, N. B. Shroff, "Multiuser Scheduling in a Markov-modeled Downlink using Randomly Delayed ARQ Feedback," *arXiv preprint:1002.3312*, 2010.

[11] S. H. Ahmad, M. Liu, T. Javidi, Q. Zhao, B. Krishnamachari, "Optimality of myopic sensing in multi-channel opportunistic access," *IEEE Transactions on Information Theory,* vol. 55, pp. 4040–4050, 2009.

[12] C. Safran, C. G. Chute, "Exploration and exploitation of clinical databases", *International Journal of Bio-Medical Computing,* vol. 39, pp. 151–156, 1995.

[13] L.P. Kaelbling, M.L. Littman, A.W. Moore, "Reinforcement learning: a survey," *Journal of Artificial Intelligence Research,* vol. cs.AI/9605, pp. 237–285, 1996.

[14] P. Whittle, "Restless Bandits: Activity Allocation in a Changing World," *Journal of Applied Probability,* vol. 25, pp. 287–298. 1988.

[15] K.D. Glazebrook, H.M. Mitchell, P.S. Ansell "Index policies for the maintenance of a collection of machines by a set of repairmen," *European Journal of Operational Research,* vol. 165, pp. 267–284, 2005.

$$W_i(\pi_i) = \frac{\big[R_i(\pi_i) - \beta R_i(Q_i(\pi_i)) + \beta\big(\pi_i - \beta Q_i(\pi_i)\big)\sum_{t=0}^{\infty}\beta^t R\big(\frac{r_i+(p_i-r_i)^{t+1}(1-p_i)}{1+r_i-p_i}\big)\big]\Gamma_i + \beta\big[(1-\pi_i) - \beta(1-Q_i(\pi_i))\big]\Lambda_i}{\Gamma_i - \beta(1-\beta^{L(r_i,\pi)})[(1-\pi_i) - \beta(1-Q_i(\pi_i))]} \quad (19)$$

$$W_i(\pi_i) = [R_i(\pi_i) - \beta R_i(Q_i(\pi_i))] + \beta\big(\pi_i - \beta Q_i(\pi_i)\big)\cdot\sum_{t=0}^{\infty}\beta^t R\Big(\frac{r_i + (p_i-r_i)^{t+1}(1-p_i)}{1+r_i-p_i}\Big) + \beta[(1-\pi_i) - \beta(1-Q_i(\pi_i))]\cdot\sum_{t=0}^{\infty}\beta^t R\Big(\frac{r-(p-r)^{t+1}r}{1+r-p}\Big). \quad (20)$$

$$W_i(\pi_i) = \frac{\big[R_i(\pi_i) - \beta R_i(Q_i(\pi_i))\big]\cdot\Delta_i + \beta[\pi_i - \beta Q_i(\pi_i)]\cdot\Omega_i + \beta[(1-\pi_i) - \beta(1-Q_i(\pi_i))]\cdot\Upsilon_i}{\Delta_i - \beta(\pi_i - \beta Q_i(\pi_i))(1-\beta(1-r_i)) - \beta^2 r_i[(1-\pi_i) - \beta(1-Q_i(\pi_i))]} \quad (24)$$

$$W_i(\pi_i) = \big[R(\pi_i) - \beta R_i(Q_i(\pi_i))\big] + \beta\big[\pi_i - \beta Q_i(\pi_i)\big]\sum_{t=0}^{\infty}\beta^t R\Big(\frac{r_i + (p_i-r_i)^{t+1}(1-p_i)}{1+r_i-p_i}\Big) + \beta\big[(1-\pi_i) - \beta\big(1-Q_i(\pi_i)\big)\big]\sum_{t=0}^{\infty}\beta^t R_i\Big(\frac{r_i - (p_i-r_i)^{t+1}r_i}{1+r_i-p_i}\Big). \quad (25)$$

[16] P. S. Ansell, K. D. Glazebrook, J. Nino-Mora, M. O'Keeffe "Whittle's index policy for a multi-class queueing system with convex holding costs," *Mathematical Methods of Operations Research,* vol. 57, pp. 21–39, 2003.

[17] C. Papadimitriou, J.N. Tsitsiklis " The complexity of optimal queueing network control ," *Mathematics of Operation Research,* vol. 24, pp. 293–305, 1999.

[18] K. Liu and Q. Zhao "Indexability of Restless Bandit Problems and Optimality of Whittle's Index for Dynamic Multichannel Access," *IEEE Transactions on Information Theory,* vol. 56, pp. 5547-5567, 2010.

[19] E. J. Sondik, *"The optimal control of partially observable Markov Decision Processes,"* PhD thesis, Stanford University, 1971.

[20] D. P. Bertsekas *"Dynamic Programming and Optimal Control, vol. 1 and 2"* Athena Scentific, Belmount, Massachusetts,2005.

[21] S. Boyd, L. Vandenberghe, *"Convex optimization,"* Cambridge University Press, 2004.

[22] T. Kailath, A. Sayed, B. Hassibi, *"Linear estimation,"*, Prentice Hall, 2000.

[23] H. Viswanathan, "Capacity of Markov channels with receiver CSI and delayed feedback," *IEEE Transactions on Information Theory*, vol. 45, No. 2, pp. 761-771, Mar. 1999.

[24] L. Ying and S. Shakkottai, "On Throughput Optimality with Delayed Network-State Information," *Information Theory and Applications Workshop*, 2008.

[25] K. Kar, X. Luo, and S. Sarkar, "Throughput-optimal scheduling in multichannel access point networks under infrequent channel measurements," *IEEE Transactions on Wireless Communications*, vol. 7, pp. 2619–2629, 2008.

[26] V. S. Annapureddy, D. V. Marathe, T. R. Ramya, and S. Bhashyam, "Outage probability of multiple-input and single-output (MISO) systems with delayed feedback," *IEEE Transactions on Communications*, vol. 57, pp. 319-326, 2009.