Exploiting Channel Memory for Joint Estimation and Scheduling in Downlink Networks

Wenzhuo Ouyang, Sugumar Murugesan, Atilla Eryilmaz, Ness B. Shroff
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH, 43210

***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 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 index policy has near-optimal performance.**

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 an easy-to-implement 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 heavily depends on reliable knowledge of the instantaneous channel state information (CSI) at the scheduler. Assuming perfect CSI to be readily available, free of cost, at the scheduler, many sophisticated scheduling strategies have been developed with provably optimal characteristics (e.g. [2]-[6]).

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 a joint design of channel estimation and opportunistic scheduling strategies - an area that has recently received attention when the channel state is modeled as independent and identically distributed (*i.i.d*) across time (e.g. [7][8]). The *i.i.d* model has traditionally been a popular choice for researchers to abstract the fading channels, thanks to its simplicity and the 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 must be intelligently exploited for more efficient, joint estimation and scheduling strategies. Markov channel models have been gaining popularity as realistic abstractions of fading channels with memory (e.g. [9] [16] [17]).

In this paper, we study joint channel estimation - 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. [10] [11]). 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 below:

- We recast the POMDP scheduling problem as a Restless Multi-armed Bandit Process ([12]) and establish its *Whittle's indexability* ([12]) in Section IV and V. Even though Whittle's indexability is difficult to establish in general [13], we have been able to show Whittle's indexability in the context of our problem.
- Based on 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 ([13] [14]). Further, index policies are usually easy to implement and do not require history information.
- Using extensive numerical experiments, we demonstrate in section VII that the proposed index policy has near-optimal performance and that significant system level gains can be realized by exploiting the channel memory for estimation and

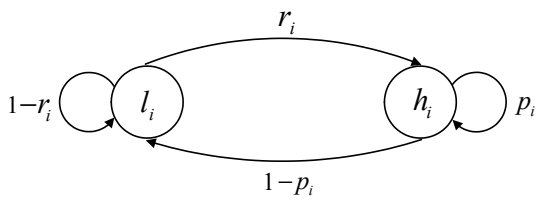

Fig. 1. Two state Markov Chain model.

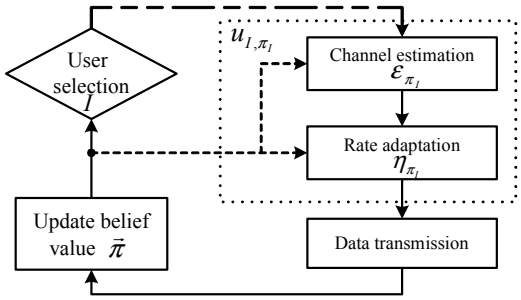

Fig. 2. Opportunistic scheduling with estimation and rate adaptation.

scheduling. The index policy we propose is of polynomial complexity in the number of downlink users (contrast this with the PSPACE-hard complexity of optimal POMDP solutions ([15])) and is amenable for distributed implementation in a multi-hop network setting.

Our setup differs from related works ([16]-[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. In this model, 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. 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 rate of transmission based on the estimate. The achievable rate expected from this process, in turn, influences the choice of scheduled user. Thus the channel estimation and scheduling stages are tightly coupled, introducing several technical challenges to the problem, which we address in the rest of the 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 remain static within each time slot and evolve 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, one user is 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 that is given by the probability the 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) - 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 ([16]-[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. This process is shown in Fig. 2. Note that these belief values are sufficient statistics to the past scheduling decisions and feedback ([19]). The scheduling problem can be formulated as a Partially Observable Markov Decision Process (POMDP) ([19]), with the Markov channel states being the partially observable system states.

Note that, as noted in Section I, 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 gains in the sum throughput. 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 gains in sum throughput.

C. Formal Problem Statement

We first discuss the estimator and rate adapter pair for the scheduled user, see Fig. 2. 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 for future use. 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 average 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. We let $u{=}\{\varepsilon, \eta\}$ denote an arbitrary estimator and rate adapter pair. The optimal estimator and rate adapter pair, $u^*_{i,\pi_i}{=}\{\varepsilon^*_{i,\pi_i}, \eta^*_{i,\pi_i}\}$, 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 expectation is over the channel state C_i, with distribution characterized by belief value π_i,

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

The quantity $\gamma_i(C_i, u)$ is the average rate of successful transmission to User i when the channel is in state C_i and the estimator and rate adapter pair u is deployed. 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 is illustrated in Fig. 2, where a pilot-aided training ([2]) based estimation is performed for a fraction of the time slot followed by data transmission at an adapted rate in the rest of the time slot.

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 $\Psi : \vec{\boldsymbol{\pi}} \to I$ 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). For a stationary policy Ψ, the expected discounted reward under initial belief $\vec{\boldsymbol{\pi}}$ is given by

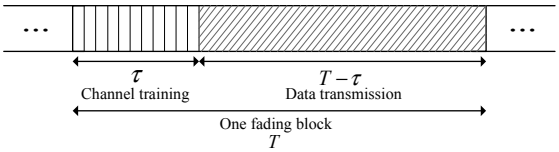

Fig. 3. Typical opportunistic scheduling with pilot-aided training.

$$V(\Psi, \vec{\boldsymbol{\pi}}) \quad = \quad \sum_{t=0}^{\infty} \beta^t E_{\vec{\boldsymbol{\pi}}[t]} R_{I[t]=\Psi(\vec{\boldsymbol{\pi}}[t])}(\pi_{I[t]}[t]) \tag{3}$$

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])

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

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

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

where $Q_i(x) = xp_i + (1-x)r_i$ is the belief evolution operator for User i, when User i 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]).

We briefly contrasted our setup with those in [16]-[18], in the introduction. We provide a rigorous comparison here. [16]-[18] studied opportunistic scheduling with the channels modeled by ON-OFF Markov Chains. Here, 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 provide more points on how these two models are fundamentally different.

- In the ON-OFF channel model, the scheduler does not need a channel estimator - 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,

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

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 (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 - rate adapter pairs for all $\pi \in [0,1]$, i.e., $U^* = \{u^*_\pi, \pi \in [0,1]\}$. The expected immediate reward, provided in (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 (a).

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

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

where $\{u : u = \{\eta\}\}$ denotes 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:

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

This establishes the lower bound in (4).

The upper bound in the lemma 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 and 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. The lemma thus follows. ■

Remark: Here we present some insights into the effect of the non-zero rate δ on the channel estimation/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 block. 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 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 - 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 $\delta \to 0$.

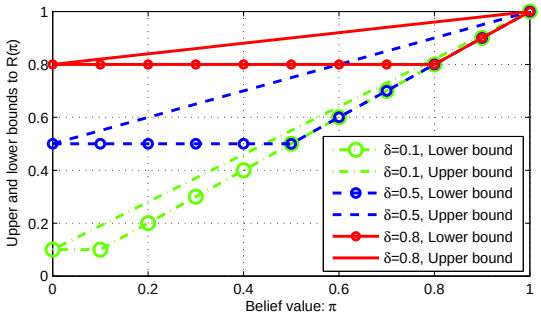

Fig. 4. Upper and lower bounds to average rate of successful transmission.

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.

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

A direct analysis of the downlink scheduling problem appears difficlt 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) ([12]) 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. They are characterized by a fundamental tradeoff between decisions guaranteeing high immediate rewards versus those that sacrifice immediate rewards for better future rewards - as seen in the downlink scheduling problem at hand. In RMBPs, in each slot, a subset of the competing projects are served and a reward dependent on the states of the served projects and the action taken is accrued by the controller. The states of all the projects in the system evolve in time based on the current state of the projects and the action taken. Solutions to RMBPs are known to be PSPACE-hard, in general ([15]).

Under an *average* constraint on the number of projects scheduled per slot, a low complexity index policy developed by Whittle ([12]), 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. [13] [14]). For an RMBP, Whittle's index policy exists if and only if the RMBP satisfies a condition known as *Whittle's indexability* ([12]), 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. [13] [14]).

The downlink scheduling problem we have considered is in fact an RMBP process. From the preceding discussion, we see that the 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 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 User i is considered throughout), the scheduler takes one of two possible actions: schedule data transmission to the user (action $a = 1$) or stay 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])

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

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 the case that 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 2. *The infinite horizon discounted reward, $V_\omega(\pi)$ is convex in $\pi \in [0,1]$.*

Proof Outline: The proof of the convexity of $V_\omega(\pi)$ involves two steps. First, we prove the convexity of discounted reward for finite horizon ω-subsidy problem using backward induction. The convexity of $V_\omega(\pi)$ is then established by extending the finite horizon problem to infinite horizon using results theory from [20]. Details of are available in [23]. ■

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

Proposition 3. *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 $\omega \geq 1$, $\pi^*(\omega) = 1$;
(ii) If $\omega \leq \delta$, $\pi^*(\omega) = \kappa$ 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 $V_\omega^1(\pi)$ be the reward corresponding to transmit decision and $V_\omega^0(\pi)$ be the reward corresponding to idle decision, i.e.,

$$V_\omega^1(\pi) = R(\pi) + \beta\big(\pi V_\omega(p) + (1-\pi)V_\omega(r)\big),$$
$$V_\omega^0(\pi) = \omega + \beta V_\omega\big(Q(\pi)\big) = \omega + \beta V_\omega\big(\pi p + (1-\pi)r\big).$$

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

Case (i). If $\omega \geq 1$, since $R(\pi) \leq 1$, 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 $\omega \leq \delta$, then for any $\pi \in [0,1]$, we have

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

where the inequality is due to $\delta \leq R(\pi)$ 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,

$$V_\omega^0(0) = \omega + \beta V_\omega(r) > \delta + \beta V_\omega(r) = V_\omega^1(0)$$
$$V_\omega^0(1) = \omega + \beta V_\omega(p) < 1 + \beta V_\omega(p) = V_\omega(^1 1)$$

Note that the relationship of $V_\omega^0(\pi)$ and $V_\omega^1(\pi)$ 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 3, 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 the 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.,

$$W(\pi) = \inf\{\omega : V_\omega^0(\pi) \geq V_\omega^1(\pi)\} \quad (6)$$
$$= \inf\{\omega : \pi^*(\omega) = \pi\}. \quad (7)$$

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 4. *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 (8), 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 [23]. 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 our online report [23].

Proof Outline: The derivation of $V_\omega(p)$ and $V_\omega(r)$ follows from substituting p and r in (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 our online report [23]. ■

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 5. *The threshold value is monotonically increasing with ω. Therefore, the problem is Whittle indexable.*

Proof Outline: The proof of indexability involves a careful study of (5) and follows the lines of [18]. Specifically, the monotonicity of threshold value is obtained by studying its derivation with respective to the ω-subsidy problem. See [23] for details. ■

VI. Whittle's Index Policy and Numerical Performance Analysis

A. Whittle's Index Policy

In this section, we explicitly derive the Whittle's index policy for the downlink scheduling problem and study its performance via numerical studies. For User i, let π_i^0 denote the steady state probability of being in state h, 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 in the following proposition.

Proposition 6. *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 ($p_i > r_i$)

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

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

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

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 4) into Equation (5). Details of the proof are provided in [23]. ■

Remark: Notice that Proposition 6 does not give the closed form expression for $W_i(\pi_i)$. However, since the value function $V_{i,W_i(\pi_i)}(p_i)$ and $V_{i,W_i(\pi_i)}(r_i)$ are all linear in $W_i(\pi_i)$, closed form expressions of $W_i(\pi_i)$ can be easily found and is given in our online report [23]. 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 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.

$$\Theta = \frac{(1-\beta^{L(r,\pi^*(\omega))})\omega + (1-\beta)\beta^{L(r,\pi^*(\omega))}[R(Q^{L(r,\pi^*(\omega))}(r)) + \beta Q^{L(r,\pi^*(\omega))}(r)V_\omega(p)]}{(1-\beta)[1-\beta^{L(r,\pi^*(\omega))+1}(1-Q^{L(r,\pi^*(\omega))}(r))]} \tag{8}$$

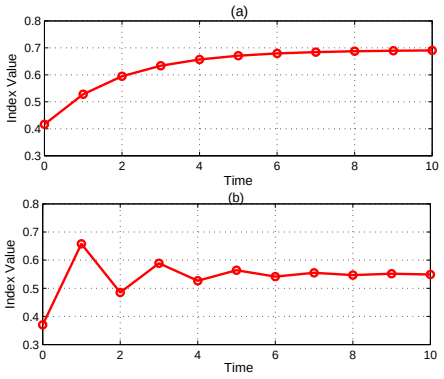

Fig. 5. Index value evolution of User i, with $\pi_i[0] = 0.3$. (a) Positive correlation, p_i=0.8, r_i=0.2; (b) Negative correlation, p_i=0.2, r_i=0.8.

Fig. 5 plots the index value evolution for the case of positively correlated and negatively correlated channels when they stay idle, i.e., not scheduled for transmission. Note that, for the positively correlated channel, the index value behaves monotonically, while, for the negatively correlated channel, the index value shows oscillation. The evolution of index values is in accordance with the evolution of belief values, which, monotonically approaches steady state for positively correlated channel, and approaches steady state with oscillation for negatively correlated channel. Thus, Fig. 5 shows that the index value essentially captures the quality of underlying Markovian channel.

B. Numerical Performance Analysis

In this section, we study, via numerical evaluations, the performance of the index scheduling policy. We consider the specific class of estimator and rate adapter structure, with pilot-aided training, discussed in Section II (Fig. 3). We consider a fading 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]). Let ϕ_π denote the optimal LMMSE estimator with prior $\{\pi, 1-\pi\}$. Φ consists of the set of LMMSE estimators optimized for various discretized values of π. Formally, Φ=$\{\phi_\pi : \pi \in [0, \rho, 2\rho, \ldots, 1]\}$.

We now study the structure of the immediate achieved rate $R(\pi)$. Note that $R(\pi)$ is optimized over the class of

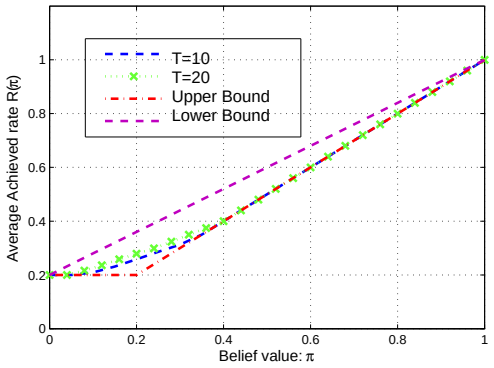

Fig. 6. Average achieved rate versus π.

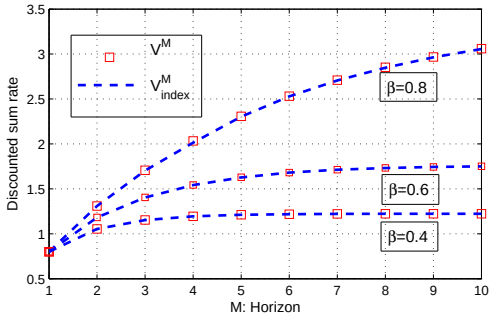

Fig. 7. Comparison of index policy and optimal finite horizon policy. N=5, $\{p_1$=0.2, r_1=0.75$\}$, $\{p_2$=0.6, r_2=0.25$\}$, $\{p_3$=0.8, r_3=0.3$\}$, $\{p_4$=0.4, r_4=0.7$\}$, $\{p_5$=0.65, r_5=0.55$\}$, $\vec{\boldsymbol{\pi}}[0]$=[0.6, 0.7, 0.3, 0.4, 0.8].

estimators Φ. Fig. 6 illustrates $R(\pi)$, in comparison with the upper and lower bounds derived in Lemma 1. For two values of block length T, compared with the upper and lower bounds. As established in Lemma 1, $R(\pi)$ shows convex increasing structure and takes value within the upper and lower bounds. As T increases, $R(\pi)$ also increases. This is because a larger T will bring more flexibility to the design of the optimal estimator. We then fix $T = 20$ and compare the expected throughput V^M and V_{index} that respectively correspond to the optimal finite M-horizon policy and the index policy in Fig. 7, with a growing horizon length M. This figure reveals the near optimal performance of the index policy. Also, as expected, the higher the value of β, the higher the average achieved throughput.

We then consider 5 users with statistically identical but independently varying channels. The infinite horizon rewards are obtained as limits of the finite horizon until 1% convergence is achieved. We fix $p_i + r_i = 1$ for all i. With this setting, the channel for every user will have the same steady state distribution of 0.5 to stay in state 1. We now increase p_i from 0.5 to 1, the steady state distribution of each channel will not change but the system becomes increasingly correlated over time. We define the system 'memory' as the difference of $p_i - r_i$. We also define the randomized policy as a policy that, at each slot, chooses each user with uniform probability, and hence does not exploit system memory. Fig. 8 compares the performance in terms of expected throughput

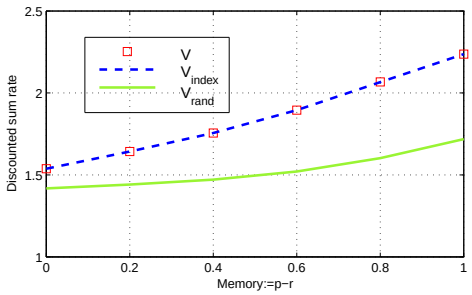

Fig. 8. Comparison of index policy and optimal policy with memory: $= p - r$. $\vec{\boldsymbol{\pi}}[0] = [0.6, 0.7, 0.55, 0.75, 0.8]$.

N	β	V	V_{index}	V_{nofb}	%opt
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 GAIN ASSOCIATED WITH INDEX POLICY.

V, V_{index} and V_{rand} that respectively corresponds to optimal policy, index policy, and the randomized policy, with growing system 'memory.' The figure shows that exploiting channel memory for opportunistic scheduling can result in significant performance gains, and index policy has near-optimal performance. As the memory increases, the significance of the channel feedback on the performance the index policy increases, resulting in an increasing gap between the index policy and the randomized policy.

Table I presents the performance of the index policy in a larger perspective. Here, with randomly generated system parameters, index policy performance is compared with the optimal scheduling policy and the policy that 'throws away' the feedback from the scheduled user. The high values of the quantity $\%opt = \frac{V_{index} - V_{nofb}}{V - V_{nofb}} \times 100\%$, in addition to underscoring the near-optimality of the index policy, also signifies the gains achieved by exploiting the channel memory using the end-of-slot feedback. Table I, along with Fig. 8, shows that exploiting channel memory for opportunistic scheduling can result in significant performance gains, and almost all of these gains can be realized using an easy-to-implement index policy.

VII. Conclusion

In this paper, we have studied downlink multiuser scheduling under a Markov-modeled channel. We considered the scenario in which 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 throughput. We formulated the scheduling problem as a partially observable Markov decision processes 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 Whittle's index policy and studied the performance of this policy using extensive numerical experiments, which suggest that the index policy has near optimal performance and significant system level gains can be realized by exploiting the channel memory for joint channel estimation and scheduling.

References

[1] R. Knopp, P. A. Humblet, "Information capacity and power control in single cell multiuser communications," *IEEE ICC*, 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 Trans. on Inform. Theory,* 1997

[4] M. Neely, E. Modiano, C. Rohrs, "Power Allocation and Routing in Multi-Beam Satellites with Time Varying Channels," *IEEE Trans. Net.,* 2003.

[5] S. Shakkottai, A. L. Stolyar, "Scheduling for Multiple Flows Sharing a Time-Varying Channel: The Exponential Rule," *App. Prob.*, 2002.

[6] A. Eryilmaz, R. Srikant, "Fair Resource Allocation in Wireless Networks using Queue-length based Scheduling and Congestion Control," *IEEE INFOCOM,* 2005.

[7] M. J. Neely, "Max weight learning algorithms with application to scheduling in unknown environments," *arXiv:0902.0630v1,* 2009.

[8] C. Thejaswi, J. Zhang, S. Pun, V. H. Poor, "Distributed Opportunistic Scheduling with Two-Level Channel Probing," *IEEE INFOCOM,* 2009.

[9] L. A. Johnston, V. Krishnamurthy, "Opportunistic file transfer over a fading channel: a POMDP search theory formulation with optimal threshold policies," *IEEE Trans. Wireless Comm.,* 2006.

[10] C. Safran, C. G. Chute, "Exploration and exploitation of clinical databases", *International Journal of Bio-Medical Computing,* 1995.

[11] L.P. Kaelbling, M.L. Littman, A.W. Moore, "Reinforcement learning: a survey," *Journal of Artificial Intelligence Research,* 1996.

[12] P. Whittle, "Restless Bandits: Activity Allocation in a Changing World," *Journal of Applied Probability,* 1988.

[13] 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,* 2005

[14] 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,* 2003

[15] C. Papadimitriou, J.N. Tsitsiklis " The complexity of optimal queueing network control ," *Mathematics of Operation Research,* 1999.

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

[17] S. H. Ahmad, M. Liu, T. Javidi, Q. Zhao, B. Krishnamachari, "Optimality of myopic sensing in multi-channel opportunistic access," *IEEE Trans. on Inform. Theory*, 2009.

[18] K. Liu and Q. Zhao "Indexability of Restless Bandit Problems and Optimality of Whittle's Index for Dynamic Multichannel Access," submitted to IEEE Trans. on Inform. Theory, 2008.

[19] E. J. Sondik, *"The optimal control of partially observable Markov Decision Processes,"* PhD thesis, Stanford University, Palo Alto, 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,"* 2004.

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

[23] W. Ouyang, S. Murugesan, A. Eryilmaz, N. B. Shroff, "Exploiting Channel Memory for Downlink Scheduling with Estimation and Rate Adaptation," Tech. Rep., (www.ece.osu.edu/∼ouyangw/infocom11.pdf).