Wireless Channel Selection with Reward-Observing Restless Multi-Armed Bandits

Julia Kuhn and Yoni Nazarathy

Abstract Wireless devices are often able to communicate on several alternative channels; for example, cellular phones may use several frequency bands and are equipped with base-station communication capability as well as typically with WiFi and Bluetooth communication. Automatic decision support systems in such devices need to decide which channels to use at any given time so as to maximize the long-run average throughput. A good decision policy needs to take into account that, due to cost, energy, technical, or performance constraints, the state of a channel is only sensed when it is selected for transmission. Therefore, the greedy strategy of always exploiting those channels assumed to yield the currently highest transmission rate is not necessarily optimal with respect to long-run average throughput. Rather, it may be favourable to give some priority to the exploration of channels of uncertain quality.

In this chapter we model such on-line control problems as a special type of Restless Multi-Armed Bandit (RMAB) problem in a partially observable Markov decision process framework. We refer to such models as Reward-Observing Multi-Armed Bandit (RORMAB) problems. These types of optimal control problems have been previously considered in the literature in the context of: (i) the Gilbert-Elliot (GE) channels (where channels are modelled as a two state Markov chain), and (ii) Gaussian autoregressive (AR) channels of order 1. A virtue of this chapter is that we unify the presentation of both types of models under the umbrella of our newly defined RORMAB. Further, since RORMAB is a special type of RMAB we also present an account of RMAB problems together with a pedagogical development of the Whittle index which provides an approximately optimal control method. Numerical examples are provided.

Julia Kuhn
The University of Queensland, University of Amsterdam, e-mail: j.kuhn@uq.edu.au

Yoni Nazarathy
The University of Queensland

1 Introduction

Communication devices are often configured to transmit on several alternative channels, which may differ in their type (e.g. WiFi versus cellular) or in their physical frequencies. Further, due to physical transmitter limitations, a device can only use a limited number of channels at any given time. Thus, the question arises which channels to select for transmission so as to maximize the throughput that is achieved over time.

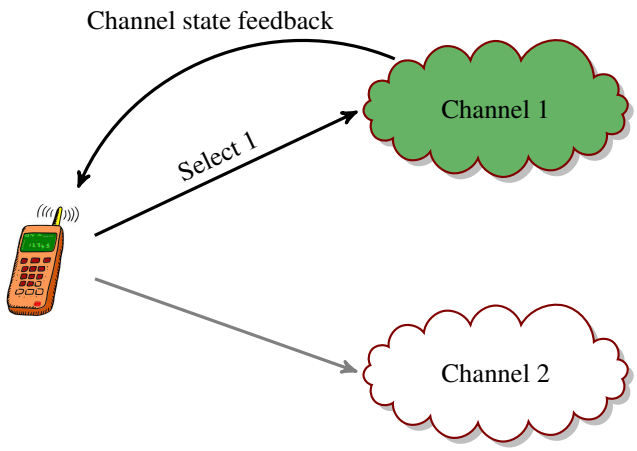


Fig. 1 A transmitting device needs to choose whether to transmit on Channel 1 or Channel 2. Transmitting on a channel results in immediate channel state feedback.

To illustrate the problem, we consider the scenario depicted in Figure 1. At every discrete time instance, the transmitter has the choice to use either channel 1 or channel 2. The channels cannot be used in parallel due to hardware limitations and/or energy constraints. The selected channel then yields an immediate reward based on the condition of the channel (e.g. the reward may be measured as the number of bits successfully transmitted). Consequently, an observation of the state of that channel is also obtained. The unselected channel on the other hand is not observed in this time instance.

Ideally, the transmitter would choose channels in a way that achieves the largest throughput over time. However, the nature of communication channels is often stochastic and thus the transmitter does not know the current state of each channel with certainty. A good prediction of the channel state can be obtained when there is strong dependence between the current state of a channel and its state in the (recent) past. Such *channel memory* can for example be caused by other users interfering on the same channel, multipath of physical signals, or other persistent disturbances.

In utilizing channel memory to make wise channel selection decisions, the transmitter needs to balance a trade-off between exploitation and exploration: On the one

hand, based on the controller's belief regarding the current state of each channel, it may make sense to choose the channel expected to transmit the highest number of bits over the next time slot. On the other hand, it may be sensible to check the condition of the other channel so as to decrease uncertainty regarding its current state. How should channels be selected, based on the information available, so as to maximize the long-run expected throughput?

A first step towards answering this control problem is to devise suitable *channel models*; that is, models that capture channel behaviour reasonably well and at the same time are simple enough to be tractable in practice. To capture such a dependency, channel states are often modelled as Markov processes. One such very simple process is the so-called *Gilbert-Elliot channel* (GE), where there are two possible states, 0 ("bad") and 1 ("good"), and transitions between states occur in a memoryless Markovian manner. The application of the GE model in channel selection and specifically opportunistic spectrum access is motivated by its ability to capture bursty traffic of primary users [7]. Due to its simplicity it has been very popular in modelling channel selection problems; refer to the literature review in Section 5.

Another class of models, which has only recently come to attention in the context of wireless channel selection [4, 20], are *Gaussian autoregressive processes of order* 1 (which we denote by AR). Here, the channel state is a continuous random variable following a normal distribution, and its evolution is determined by a simple linear recursion perturbed by Gaussian noise. It has been found that Gaussian autoregressions model the logarithmic signal-to-noise ratio of a channel reasonably well; for details see [1].

The virtue of both the GE and the AR model is that they are simple and tractable, yet allow to capture the exploration–exploitation trade-off that the controller faces. The models are simple in the sense that the *belief* which the controller maintains about the state of the channel is neatly summarized by sufficient statistics. In the GE case, this sufficient statistic is given by the conditional probability of being in the good state, given the information that is available to the controller at the time. In the AR case, it is sufficient to keep track of the conditional mean and variance of the state, which quantify the expected gain from exploitation and the need for exploration, respectively.

An optimal policy for the channel selection problem needs to balance this exploration–exploitation tradeoff. While such a policy could in principle be computed by dynamic programming, this is typically computationally infeasible in practice [33]. However, recognizing that the problem is essentially a Restless Multi-Armed Bandit (RMAB) problem, we may apply techniques from RMAB theory to find a (near-)optimal solution: the well-known Whittle index [41] (a generalization of the celebrated Gittins index [9, 10]).

Our main focus in this chapter is on the RMAB formulation of the problem. We call this special type of RMAB the *Reward-Observing Restless Multi-Armed Bandit* (RORMAB) problem. While the chapter does not contain new results, it is unique in that it provides a unified treatment of both the GE and the AR approaches for channel models, and considers also the channel selection problem in the mixed case where some of the channels are modelled as GE while the others are AR. This is of

interest in networks where some but not all of the channels may be subject to user interference.

The remainder of this chapter is structured as follows. In Section 2 we formulate the RORMAB problem, and present the GE and AR models in a unified manner. In Section 3 we motivate the use of index policies and in particular the Whittle index. The presentation can be used as a stand alone brief account of RMAB problems. In Section 4 we show how to use these channel models to evaluate the Whittle index numerically, and use it as a solution strategy for an example channel selection problem. In the latter section we also provide a number of performance comparisons. Section 5 contains a literature survey, and points out some open problems.

2 Reward-Observing Restless Multi-Armed Bandits

In this section we formulate the RORMAB problem within the context of wireless channel selection. This type of problem is a special case of a Partially Observable Markov Decision Process (POMDP) as considered in [36]. An MDP is partially observable if the decision maker does not know the current state of the system with certainty. In our setting, where we consider a network of d channels, the partially observable state of the system can be represented as d-dimensional, and corresponds to the joint state information of the individual channels.

We consider channels $X_1(t),\dots,X_d(t)$, operating as independent Markov processes in discrete time $t \in \mathbb{N}_0$. We assume that the models and their parameters are known but do not have to be the same for each channel. At every time instance, the decision maker chooses a subset $\mathscr{C}(t)$ of k channels, $\mathscr{C}(t) \subset \{1,\dots,d\}$. For every selected channel $i \in \mathscr{C}(t)$ a reward $r_i\big(X_i(t)\big)$ is obtained and the value of $X_i(t)$ is observed, where each r_i is assumed to be a known, deterministic function from the underlying state space to $\mathbb{R}$. The other channels $i \notin \mathscr{C}(t)$ are not observed and do not yield a reward.

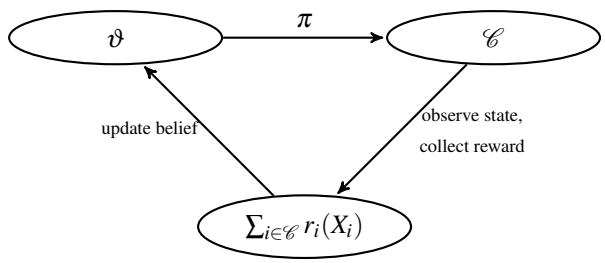
In an ideal situation, at every decision time the controller would choose those channels that yield the highest reward $\sum_{i\in\mathscr{C}(t)} r_i\big(X_i(t)\big)$. Unfortunately, this cannot generally be achieved because channel states are stochastic and the values of $X_1(t),\dots,X_d(t)$ are not known at decision time t.

However, because the channel states are sequentially dependent due to the channel memory, the controller can use information about the previous state of a channel to make predictions about its current state. The accuracy of the prediction depends on the *age* of the information, i.e. the number of time steps ago that a channel was last observed. This number is denoted by $\eta_i(t) := \min\{\tau \geq 1 : i \in \mathscr{C}(t-\tau)\}$, so that the last time channel i was chosen is given by $t-\eta_i(t)$. The information available to the transmitter at time t (prior to making its decision) can then be summarized and represented by $\mathscr{I}(t) := \big(\mathscr{I}_1(t),\dots,\mathscr{I}_d(t)\big)$, where for $i = 1,\dots,d$,

$$\mathscr{I}_i(t) = \Big(\eta_i(t), X_i\big(t-\eta_i(t)\big)\Big).$$

Based on $\mathscr{I}_i(t)$, the controller's *belief* about the state of channel i at time t is summarized by $F_i(x) := \mathbb{P}\big(X_i(t) \leq x \mid \mathscr{I}_i(t)\big)$, the conditional distribution of channel i given the information collected up to that time. For the channel models we consider, this probability distribution is characterised by scalar- or vector-valued sufficient statistics. That is, for channel i there exists a parameter $\vartheta_i(t)$ that fully specifies the probability distribution of $X_i(t)$ given the information $\mathscr{I}_i(t)$. In our first model (GE as described below), $F_i(\cdot)$ is a Bernoulli distribution, so $\vartheta_i(t)$ is the "success probability". In our second model (AR as described below), $F_i(\cdot)$ is a normal distribution, hence, $\vartheta_i(t)$ is a two-dimensional vector specifying the conditional mean and conditional variance. Using the terminology common in literature on POMDP, we refer to $\vartheta_i(t)$ as the *belief state* of channel i at time t – indeed $\vartheta_i(t)$ represents our belief concerning the state of the channel.

In summary, as time evolves from t to $t+1$, given the current belief state $\vartheta := (\vartheta_1, \ldots, \vartheta_d)$ and a channel selection policy π, the following chain of actions takes place:



Objective. Our aim is to find a policy π so as to maximize the accumulated rewards over an infinite time horizon as evaluated by the *average expected reward criterion*

$$G^{\pi}(\vartheta) := \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}^{\pi}_{\vartheta} \left[\sum_{t=0}^{T-1} \sum_{i\in\mathscr{C}(t)} r_i\big(X_i(t)\big) \right], \tag{1}$$

where the subscript indicates conditioning on $X_i(0)$ being distributed with parameter ϑ_i. Note that other reward criteria including finite horizon problems and/or discounted rewards/costs have been considered [34], but when optimizing throughput the average reward criterion (1) is more natural: Neither discounting nor imposing a restriction on the time horizon seems plausible in this context.

It can be proven formally [5, 36] that a POMDP as considered here, with partially observable states $X_i(t)$ and rewards $r_i\big(X_i(t)\big)$, is equivalent to a fully observable MDP with states $\vartheta_i(t)$ and rewards $\mathbb{E}_{\vartheta_i(t)}\left[r_i\big(X_i(t)\big)\right]$. Namely, the best throughput that we can achieve is the same for both, and it is achieved by the same (optimal) policy. This justifies that we consider the MDP with states $\vartheta_i(t)$ in the remainder of this chapter.

Belief State Evolution. For RORMAB, the decisions determined by a policy π affect the updating of the belief state based on the *observation update* mapping $\mathscr{O}_i(\cdot)$

and the *belief propagation* operator $\mathscr{T}_i(\cdot)$ as follows:

$$\vartheta_i(t+1) = \begin{cases} \mathscr{O}_i\big(X_i(t)\big), & \text{if } i \in \mathscr{C}(t), \\ \mathscr{T}_i\big(\vartheta_i(t)\big), & \text{if } i \notin \mathscr{C}(t). \end{cases} \tag{2}$$

The *observation update* mapping $\mathscr{O}_i(\cdot)$ determines how the belief state of channel i is updated when that channel is selected for transmission. In this case we observe $X_i(t)$, and hence its realization can be used by the observation update rule when implementing the controller. Further, for analytical, modelling and simulation purposes, when $i \in \mathscr{C}(t)$, the distribution of $X_i(t)$ is determined by the known value $\vartheta_i(t)$, so $X_i(t)$ can be replaced by a generic random variable coming from this distribution. This comes to emphasise that $X_i(t)$ is actually not part of the state of the MDP.

The *belief propagation* operator $\mathscr{T}_i(\cdot)$ defines the update of the belief state of a channel when it is not selected for transmission. Because in this case no new observation is obtained, the update is deterministic.

Since a channel may not be selected for several consecutive time slots, it is useful to also consider $\mathscr{T}_i^k(\cdot)$ (the k-step operator obtained by applying $\mathscr{T}_i(\cdot)$ k times) as well as attracting fixed points of the operator $\mathscr{T}_i(\cdot)$. As we describe below, in both the GE and the AR model, the k-step operator has an explicit form and converges to a unique attracting fixed point; this is useful for understanding the dynamics of the model.

We now specify the observation update and belief propagation operations in the context of each of our two channel models. For the sake of notational simplicity we omit the subscripts i where they are not relevant.

Gilbert-Elliot (GE) Channels. In this case $X_i(t)$ is a two state Markov chain on the state space $\{0,1\}$, where 0 represents a "bad" state and 1 is a "good" state of the channel. The transition matrix can be parametrized as

$$P_i = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = \begin{bmatrix} 1-\gamma\overline{\rho} & \gamma\overline{\rho} \\ \overline{\gamma}\overline{\rho} & 1-\overline{\gamma}\overline{\rho} \end{bmatrix},$$

where we denote $\overline{x} := 1-x$.One standard parametrization of this Markov chain uses transition probabilities $p_{01}, p_{10} \in [0,1]$ (and sets $p_{00} = \overline{p_{01}}$, $p_{11} = \overline{p_{10}}$). Alternatively we may specify the stationary probability of being in state 1, denoted by $\gamma \in [0,1]$, together with the second eigenvalue of P_i, denoted by $\rho \in \big[1-\min(\gamma^{-1}, \overline{\gamma}^{-1}),\ 1\big]$. Then ρ quantifies the time-dependence of the chain — when $\rho = 0$ the chain is i.i.d., otherwise there is memory. Specifically, when $\rho > 0$ there is positive correlation between successive channel states and when $\rho < 0$ that correlation is negative. The relationship between these parameterisations is given by $\gamma = p_{01}/(p_{01}+p_{10})$, and $\rho = 1 - p_{01} - p_{10}$. The parametrization with transition probabilities $p_{ij} \in [0,1]$ is standard. As opposed to that, our alternative parametrization in terms of γ and ρ has not been used much in the literature. Nevertheless, since it simplifies and gives

intuitive meaning to the expressions for $\mathscr{T}^k(\cdot)$, we advocate using it when modelling GE channels.

As the Bernoulli distribution is fully specified by the success probability, it suffices to keep track of this parameter. Because we have

$$\vartheta_i(t) = \mathbb{P}\big(X_i(t) = 1 \,|\, \mathscr{I}_i(t)\big),$$

the belief state space of channel i, denoted by Ψ_i, is given by the interval $[0,1]$. Now the *observation update* operation is defined by:

$$\mathscr{O}_i(x) = \begin{cases} p_{01}, & \text{if } x = 0\,, \\ p_{11}, & \text{if } x = 1\,. \end{cases}$$

That is, if the observed channel was "bad" ($x = 0$), then the chance of a good channel is given by the entry p_{01}, and otherwise ($x = 1$) by p_{11}. The *belief propagation* operation is

$$\mathscr{T}_i(\vartheta) = \vartheta\, p_{11} + \overline{\vartheta}\, p_{01} = \rho\, \vartheta + \gamma\overline{\rho}\,.$$

This follows by evaluating the probability of $\{X_i(t+1) = 1\}$ based on $\vartheta_i(t)$ and the probability transition matrix P_i. It is a standard exercise for two state Markov chains (recurrence relations) to show that the k-step transition probability, and thus the k-step belief propagation operator, takes the form

$$\mathscr{T}_i^k(\vartheta) = \gamma + \rho^k(\vartheta - \gamma).$$

Note that γ is a fixed point of this operator, and the sequence $\mathscr{T}_i^k(\vartheta)$ converges to this fixed point. Further note that when $\rho > 0$ this sequence is monotonic, otherwise if $\rho < 0$ it oscillates about γ as it converges to it. The case of $\rho = 0$ is not of interest because in that case there is no channel memory.

Gaussian Autoregressive (AR) Channels. In this case the channel states follow an AR process of order 1, that is,

$$X_i(t) = \varphi_i X_i(t-1) + \varepsilon_i(t)\,,$$

with $\big\{\varepsilon_i(t) : t \in \mathbb{N}_0\big\}$ denoting an i.i.d. sequence of $\mathscr{N}\left(0, \sigma_i^2\right)$ random variables. We assume $|\varphi| < 1$ in which case the processes are stable. Note that if $\varphi \in (0,1)$ the states are positively correlated over time; for $\varphi \in (-1,0)$ the correlation is negative. The case $\varphi = 0$ may be neglected as it corresponds to observations being independent. Linear combinations of Gaussian random variables are still Gaussian, and hence, their conditional distribution at time t is fully described by the conditional mean $\mu_i(t)$ and the conditional variance $\nu_i(t)$. That is, the sufficient statistic (vector) for the state of channel i is:

$$\vartheta_i(t) = \big(\mu_i(t),\, \nu_i(t)\big)\,.$$

In this AR case, the *observation update* operation is:

$$\mathscr{O}_i(x) = \left(\varphi\, x,\, \sigma^2\right).$$

This is due to the fact that an observation of x at time t implies a predicted expected value of $\varphi\, x$ at time $t+1$ with variance σ^2. In contrast, the *belief propagation* operation is given by

$$\mathscr{T}_i(\mu_i,\, \nu_i) = \left(\varphi\, \mu_i,\, \varphi^2\, \nu_i + \sigma^2\right). \tag{3}$$

It is easy to show by recursion of the mean and the variance that the k-step belief propagation is:

$$\mathscr{T}_i^k(\mu_i,\, \nu_i) = \left(\varphi^k\, \mu_i,\, \varphi^{2k}\, \nu_i + \frac{1-\varphi^{2k}}{1-\varphi^2}\, \sigma^2\right). \tag{4}$$

The belief state space in this case is $\Psi_i = \mathbb{R} \times [\nu_{\min}, \nu_{\max})$ where $\nu_{\min} = \sigma^2$ and $\nu_{\max} = \sigma^2/(1-\varphi^2)$. The attracting fixed point of $\mathscr{T}_i(\cdot)$ is the mean-variance pair $(0,\, \nu_{\max})$. It is further interesting to note that the second coordinate of the belief state can only attain values in a countable subset of $[\nu_{\min}, \nu_{\max})$. This is because when the channel is selected, the conditional variance decreases to the value $\nu_{\min}$, and thus, ν_i in (4) is always proportional to σ^2, where the factor is given by a geometric series in φ^2. Observe further that, since $\nu < \nu_{\max}$ and because $|\varphi| < 1$, it always holds that the variance increases when updated with $\mathscr{T}_i(\cdot)$, that is, the decision maker's uncertainty regarding the state of the channel indeed grows as long as no new observation is obtained.

Mixed Model Example. Having specified the GE and AR channel models, we now consider a mixed model example, which is also used for numerical illustration in Section 4. Research papers in this field to date seem to have focussed on problems with channels of the same type (mostly GE, some AR); it is therefore interesting to investigate a mixed channel model example, where a proportion $q \in [0,1]$ of the channels is GE and the others are AR. This can occur in examples where the dominating phenomena of some of the channels is user interference (GE channels), while for other channels the key feature is slow-fading behaviour (AR channels).

Our model parameters are α_i, γ_i for $i = 1, \ldots, qd$ (GE channels) and φ_j, σ_j^2 for $j = qd+1, \ldots, d$ (AR channels); we assume qd is an integer.

For the purpose of exposition we consider the following stylised case of reward functions:

$$r_i(x_i) = \frac{x_i - \gamma_i}{\sqrt{\gamma_i\,(1-\gamma_i)}}, \quad \text{and} \quad r_j(x_j) = \frac{\sqrt{1-\varphi_j^2}}{\sigma_j} x_j, \tag{5}$$

where x_i is a value observed in GE channel i and x_j is the value observed in AR channel j. We specifically choose these functions so that the steady state values of rewards from both channels have zero-mean and unit-variance, hence making the channels equivalent in these terms. That is, in the case where the controller does not

have additional state information, the controller obtains the same mean and variance on any channel chosen.

The state space of the MDP with (joint belief) states $\vartheta = (\vartheta_1, \ldots, \vartheta_d)$, with scalars ϑ_i, $i = 1, \ldots, qd$, and 2-dimensional vectors ϑ_j, $j = qd+1, \ldots, d$, is given by:

$$\Psi := [0,1]^{qd} \times \{\mathbb{R} \times [\nu_{\min}, \nu_{\max})\}^{\bar{q}d}.$$

An optimal policy π for such an MDP is usually not available in closed form. It can then be computed approximately with the aid of dynamic programming algorithms, on a discretized and truncated state space. This is feasible with sufficient accuracy only if d is very small (and indeed this is carried out as part of the numerical examples provided in Section 4 for $d = 2$). With more channels, the computational task quickly becomes intractable – hence we resort to a sensible index heuristic (the Whittle index), which we present in the next section.

3 Index Policies and the Whittle Index

In this section we explain the idea behind the use of index policies and specifically the Whittle index, a generalization of the well-studied Gittins index [9, 41]. Whittle proposed this type of index as a heuristic solution to RMAB problems. We first describe a general form of RMAB problem so as to put our specific RORMAB problem in context.

Consider state processes $\vartheta_1(t), \ldots, \vartheta_d(t)$ subject to a control set $\mathscr{C}(t) \subset \{1, \ldots, d\}$ which selects k of the d processes at each time. In the (more general) context of RMAB, we refer to each of these processes as an *arm* of a bandit. Based on the control decisions captured in the control set $\mathscr{C}(\cdot)$, each of the processes evolves either according to an *active* mapping $\mathscr{A}_i(\cdot)$ if $i \in \mathscr{C}(t)$, or according to a *passive* mapping $\mathscr{P}_i(\cdot)$ otherwise. This can be represented as follows:

$$\vartheta_i(t+1) = \begin{cases} \mathscr{A}_i\big(\vartheta_i(t), U_i(t)\big), & \text{if} \quad i \in \mathscr{C}(t), \\ \mathscr{P}_i\big(\vartheta_i(t), U_i(t)\big), & \text{if} \quad i \notin \mathscr{C}(t). \end{cases} \tag{6}$$

Here, $\{U_i(t), i = 1, \ldots, d\}$ are independent i.i.d. (driving) sequences of uniform $(0,1)$ random variables. An alternative representation is to use Markovian transition kernels, one for the active operation and one for the passive operation.

Comparing (6) and (2) it is evident that our RORMAB channel selection problem is a special case of the RMAB problem. Channels and belief states of the RORMAB correspond to arms and states of the RMAB, respectively. In the RORMAB, the active mapping $\mathscr{A}_i(\vartheta, u)$ is replaced by $\mathscr{O}_i\big(F^{-1}(u; \vartheta)\big)$ where $F^{-1}(\cdot; \vartheta)$ is the inverse probability transform generating a random value of the state, distributed with parameter(s) ϑ; and the passive mapping $\mathscr{P}_i(\vartheta, u)$ does not depend on the random component and is replaced by $\mathscr{T}_i(\vartheta)$. In the remainder of this section we depart

from the RORMAB context and present a brief exposition of RMAB and the Whittle index.

The RMAB problem arose as a generalisation of the Multi-Armed Bandit (MAB) problem. In that case $\mathscr{P}_i(\vartheta, u) = \vartheta$; that is, unselected arms do not evolve. This problem is known to be solved optimally in great generality by the celebrated Gittins index [9]. Whittle's RMAB generalization allows for "restless" state processes, where arms keep evolving also while they are not used for transmission (although not necessarily according to the same transition kernel). This modelling framework is more realistic for the channel selection problem but also appears in a variety of other application areas; see [41] for further examples. For RMAB problems, index policies are typically not optimal. However, the Whittle index, which we present below, has in many cases proven to be asymptotically optimal with respect to the average reward criterion as the number of arms grows large [38, 40].

We now first describe index policies in general before we turn to Whittle's optimization problem and the associated Whittle index policy.

Index Policies. Index policies are defined in terms of functions $\iota_1, \dots, \iota_d$ such that ι_i maps the current state of arm i to a certain priority index, irrespective of the current state of any other arm.

Definition 1. Let $\vartheta := (\vartheta_1, \dots, \vartheta_d)$ denote the vector of states in a system with d arms. An index policy π_ι activates those k arms that correspond to the k largest indices,

$$\pi_\iota\big(\vartheta\big) = \operatorname*{arg\,max}_{\mathscr{C}:|\mathscr{C}|=k} \sum_{i\in\mathscr{C}} \iota_i\big(\vartheta_i\big)\,.$$

Ties are broken arbitrarily, but in compliance with the requirement that k arms have to be selected.

To get an intuitive justification as to why *index policies* may work well in large systems, consider the following: Pick an arbitrary arm and suppose we want to decide whether to select it as active or not (passive), based on the current state. Generally, we would make our decision dependent on the states of the remaining arms. In this way, our decision strategy is highly influenced by the proportion of arms that are in a certain state. In a large system (with many arms), however, this proportion can be expected to remain relatively stable over time. In this sense, the larger the system, the less important it is for us to consider other arms; we always find ourselves in roughly the same situation for decision making. In conclusion, in a system with many arms, little is lost if we make decisions for each arm solely based on its current state, disregarding the current state of any other arm in the system.

The question is still open how to best define the index functions ι_i. A simple example is the myopic index. In the context of RORMAB, where states are actually belief states, it is defined by $\iota_i^M(\vartheta_i) = \mathbb{E}_{\vartheta_i}\big[r_i(X_i)\big]$. Thus, under the myopic policy the transmitter greedily chooses those channels that promise the largest immediate rewards ("exploitation"). However, as one may expect, it turns out that such a policy is not necessarily optimal (see our numerical examples in Section 4, as well as the

literature survey in Section 5). It may be favourable to give some priority to "exploring" other channels in order to decrease the transmitter's uncertainty with respect to their current state.

Moving back to the more general RMAB, this motivates us to consider the more sophisticated Whittle index, which takes the possible need for considering future states (or "exploration" in the case of RORMAB) into account. To derive his heuristic, Whittle relaxed the constraint that exactly k arms have to be selected at each time point, and replaced it by the weaker requirement that k arms are selected *on average*. Since the latter constraint is weaker, the optimal throughput (value/gain of the MDP) under this constraint is an upper bound for the optimal throughput that can be achieved in the original problem. We shall see that this relaxation allows to formulate the decision making problem as a Lagrange optimization problem, from which Whittle obtained a rule for determining $\imath_i\big(\vartheta_i\big)$.

Whittle's Optimization Problem. For the sake of exposition we consider an MDP with a finite state space $\Psi = \Psi_1 \times \cdots \times \Psi_d$. We further remain in the setting where arms that are not selected do not yield a reward (this assumption is easily generalized so that the Whittle index is applicable much more broadly). Under suitable regularity conditions, the optimal long-run average throughput rate is independent of the initial state of the system (see e.g. [5]); in this section we assume that we are in such a setting.

Recall the definition of the average reward criterion, Eq. (1). For a time horizon T ($T \to \infty$) we sum up the rewards that are obtained from the selected arms at each time point. Equivalently, we could group selected arms according to their states, and keep track of how many arms were selected while being in a specific state, over the whole time horizon. That is, rather than considering each time step t separately and adding up rewards as obtained at each time step, we can consider how many arms were in a certain state when selected, and multiply this proportion with the reward that is obtained from an arm in that state. If we do so for all states, then the total approaches the value of the average reward as $T \to \infty$.

This is the viewpoint we are taking in this subsection; it is inspired by the exposition in [27]. Define $p_i(v)$ as the expected long-run fraction of time that arm i is selected when it currently is in state $v \in \Psi_i$; that is,

$$p_i(v) := \lim_{T\to\infty} \frac{1}{T} \mathbb{E}^{\pi} \left[\sum_{t=0}^{T} \mathbb{1}\big\{ i \in \mathscr{C}(t),\, \vartheta_i(t) = v \big\} \right].$$

Subject to Whittle's relaxation, we can then formulate the optimization problem as the Linear Programming (LP) problem:

$$G^W = \max_p \sum_{i=1}^{d} \sum_{v\in\Psi_i} r_i(v)\, p_i(v), \quad \text{subject to} \quad \sum_{i=1}^{d} \sum_{v\in\Psi_i} p_i(v) = k\,, \tag{7}$$

where $r_i(v)$ denotes the reward that is obtained from selecting arm i when its state is v (as before $r_i(\cdot)$ is a known, deterministic function). Formulating this as an equiva-

lent Lagrangian optimization problem we obtain:

$$\begin{aligned}\mathscr{L}(\lambda) &= \max_p \sum_{i=1}^d \sum_{v\in\Psi_i} r_i(v)\, p_i(v) - \lambda \left(\sum_{i=1}^d \sum_{v\in\Psi_i} p_i(v) - k \right) \\ &= \max_p \sum_{i=1}^d \sum_{v\in\Psi_i} \big(r_i(v)-\lambda\big) p_i(v) + \lambda k \\ &= \sum_{i=1}^d \mathscr{L}_i(\lambda) + \lambda k\,,\end{aligned} \tag{8}$$

with

$$\mathscr{L}_i(\lambda) = \max_{p_i} \sum_{v\in\Psi_i} \big(r_i(v)-\lambda\big) p_i(v)\,. \tag{9}$$

Since in (8) there is no longer a common constraint for the arm, each arm can be optimized separately through (9). By strong LP duality we know that there exists a Lagrange multiplier λ^* that yields $\mathscr{L}(\lambda^*) = G^W$. LP complementary slackness ensures that (assuming $\lambda^* \neq 0$) any optimal solution to (8) must satisfy the relaxed constraint, and is therefore also optimal for Whittle's relaxed problem (7). It was observed by Whittle [41] that we can interpret the Lagrange multiplier as a cost for selecting an arm (or equivalently, as a subsidy for not selecting an arm). That is, imposing a cost of λ^* on the selection of an arm causes the controller to select k arms on average under a policy that optimises G^W.

Indexability and the Whittle Index. In accordance with [41], we make the following reasonable regularity assumption.

Assumption: Channel i is *indexable*, that is, the set of states for which it is optimal to select arm i decreases monotonically from Ψ_i to $\emptyset$ as the cost λ increases from $-\infty$ to ∞. This property holds for every arm in the system.

While this assumption is intuitively appealing, it turns out that it does not generally hold [38, 41], and proving its validity can be surprisingly difficult [9]. It has been verified for the GE model as considered in [22], and numerical evidence suggests that it also holds for the AR model [20].

Indexability implies that for each arm i there exists a function of the current state, $\lambda_i(v)$, such that it is optimal to select the arm whenever $\lambda_i(v) > \lambda$ and to leave it passive otherwise (the decision maker is indifferent when $\lambda = \lambda_i(v)$). In this sense, $\lambda_i(v)$ measures the "value" of arm i when it is in state s. Furthermore, applying this policy to all arms in the case where $\lambda = \lambda^*$ (that is, selecting arm i whenever $\lambda_i(v) > \lambda^*$) results in a policy that is optimal for Whittle's relaxed problem (7)[1]. This motivates choosing the index function $\iota_i(\cdot)$ as $\iota_i^W(v) := \lambda_i(v)$ (as was proposed in [41]).

[1] When $\lambda_i(v) = \lambda^*$, one needs to decide for the action to be taken in state v in an appropriately randomized fashion that ensures that the relaxed constraint is satisfied [40, 41].

How do we find $\lambda_i(\cdot)$? Recall that the decision maker is indifferent when $\lambda = \lambda_i(v)$, and that we are interested in the case where the cost λ is chosen to be the optimal cost λ^* that causes the decision maker to select k arm on average. Furthermore, we saw that the Lagrangian (8) can be solved by considering arms one by one (in accordance with the intuition described at the beginning of this section, where it is argued that not much is lost by decoupling arms provided the system is large enough). In fact, (9) is the Lagrangian corresponding to a *one-arm sub-problem* in which there is only a single arm which can be selected or not, and where selecting the arm yields the state-dependent reward but also an associated cost λ. Now the optimal $\lambda_i(v)$ is the one that makes us indifferent between selecting or not selecting the arm when it is in state v. In summary, we define the Whittle index as follows (cf. [41]).

Definition 2. The Whittle index is the largest cost λ in (9) such that it is still optimal to select the arm in the one-arm sub-problem.

Intuitively, the Whittle index can perhaps best be thought of as an opportunity cost, to be paid for loosing the opportunity to select one of the other arms in the constrained system with multiple arms. Naturally, we then prioritize arms with higher opportunity cost.

Computing the Whittle Index. As stated in Definition 2, the Whittle index is derived from the optimal policy for the one-arm sub-problem. Thus, the computational complexity of the Whittle index increases only linearly with the number of arms: we need the optimal policy for at most d non-identical single-arm sub-problems. In contrast, the complexity of computing the optimal policy for the full system increases exponentially (the latter problem is in fact PSPACE hard [33]).

It is well-known [14, 34] that in great generality the optimal average reward G is constant (independent of the initial state), and satisfies Bellman's optimality equation. For the one-arm sub-problem associated with our arm selection problem this optimality equation reads as,

$$G + h(\vartheta_i) = \max\left\{ r_i(\vartheta_i) - \lambda_i + \mathbb{E}\Big[h\big(\mathscr{A}_i(\vartheta_i, U)\big)\Big],\ \mathbb{E}\Big[h\big(\mathscr{P}_i(\vartheta_i, U)\big)\Big] \right\}, \quad (10)$$

with $\vartheta_i \in \Psi_i$ and U a uniform $(0,1)$ random variable. Here, $r_i(\vartheta_i) - \lambda_i$ is the immediate reward obtained from deciding to use the arm, corrected by the opportunity cost λ_i. The function h accounts for the transient effect caused by starting at initial state ϑ rather than at equilibrium.

The optimal policy for this one-arm sub-problem is then to choose the action that maximizes the right-hand side of (10). It can be found from dynamic programming algorithms such as (relative) value or policy iteration. Then the Whittle index for state ϑ_i can be effectively computed by solving (10) for an increasing sequence of $\lambda_i(v)$ and finding the maximal $\lambda_i(v)$ for which selecting the arm is still optimal.

Note that for the RORMAB, the one armed subsidy problem (10) becomes:

$$G+h(\vartheta_i)=\max\left\{\mathbb{E}_{\vartheta_i}\big[r_i(X_i)\big]-\lambda_i+\mathbb{E}_{\vartheta_i}\Big[h\big(\mathscr{O}(X_i)\big)\Big],\,h\big(\mathscr{T}(\vartheta_i)\big)\right\}. \tag{11}$$

As before, an observation is obtained which is the realization of a random variable X with probability distribution determined by ϑ (as indicated by the subscript). If the arm is not used, then no reward is obtained and the belief is propagated using the operator $\mathscr{T}$. We solve this problem for GE and AR arms in the next section.

4 Numerical Illustration and Evaluation

We now return to RORMAB and compare the performance of the Whittle index policy to that of the myopic policy and, for small d, to the optimal policy. To evaluate the Whittle indices, we usually need the optimal policy associated with Whittle's one-armed problem with subsidy. We obtain the latter from relative value iteration (on a discretized state space) using the optimality equation (11). This can be written more explicitly using the reward functions from (5) as

$$G+h(\omega)=\max\left\{r_i(\omega)-\lambda+\omega\,h\big(p_{11}\big)+\overline{\omega}\,h(p_{01})\,,\,h\big(\omega p_{11}+\overline{\omega}p_{01}\big)\right\} \tag{12}$$

when the channel is GE (so that $\vartheta=\omega$), and

$$G+h(\mu,\nu)=\max\left\{r_j(\mu)-\lambda+\int_{-\infty}^{\infty}h(\varphi y,\sigma^2)\,\phi_{\mu,\nu}(y)\,dy,\,h\big(\varphi\mu,\varphi^2\nu+\sigma^2\big)\right\} \tag{13}$$

when the channel is AR (in which case $\vartheta=(\mu,\nu)$). Here $\phi_{\mu,\nu}$ denotes the normal density with mean μ and variance ν. Note, however, that in the case of GE channels the Whittle indices are in fact available in closed form[2] [22], so that we only need to perform these iterative computations for the AR channels.

Fig. 2 shows the optimal switching curve for a small mixed system with one AR and one GE channel. To the left of the curve, where ω is large in comparison to μ, the optimal policy is to select the GE channel. To the right of the curve selecting the AR channel is optimal. The curve shifts with the age of the AR channel: the more time has passed since the AR channel has last been observed, the more inclined the transmitter should be to select that channel in order to update the available information regarding its state. In other words, it is indeed optimal to give some priority to exploration if AR channels are present in the system. Note, however, that for "older" channels this effect is less pronounced because in that case the resulting change in the conditional variance ν is smaller (recall the belief propagation of ν defined by (3)).

Fig. 3 shows a comparison of the rewards that are obtained per channel on average under different policies. Here, $k=d/2$ channels are selected at a time in a

[2] We do not reproduce these closed form expressions here as the formulas are rather cumbersome.

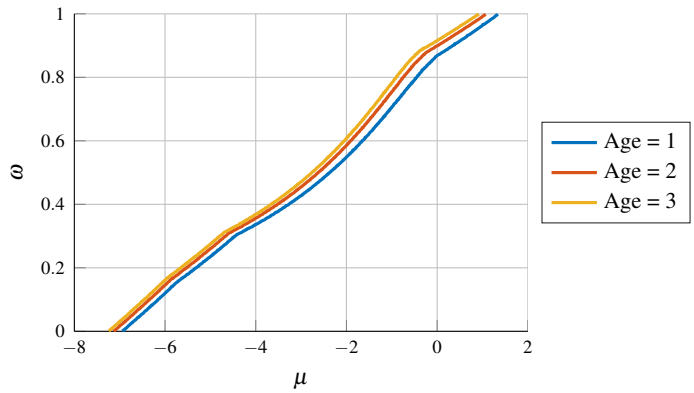


Fig. 2 Optimal switching curves for a system with $d = 2$ channel: one AR channel with $\varphi = 0.8$ and $\sigma = 2$, and one GE channel with $\rho = 0.5$ and $\gamma = 0.8$. This figure shows the switching curves on the ω, μ plane, one curve per age $\eta \in \{1, 2, 3\}$.

system with d channels, where half of the channels are GE and the other half is AR. All of the AR channels are with $\varphi = 0.8$ and $\sigma = 2$ (as in Fig. 6). The GE channels on the other hand are heterogeneous, with $\gamma = 0.8$ and $\rho_i \in [0.2, 0.8]$ evenly spaced such that $0.8 = \rho_1 > \cdots > \rho_{d/2} = 0.2$. Depicted are the average rewards per arm obtained under the Whittle and the myopic index policy, and, as an upper bound, we also computed the average rewards that could be obtained in a fully observable system under the myopic policy. Due to the high computational complexity, the optimal policy is only evaluated for $d = 2$.

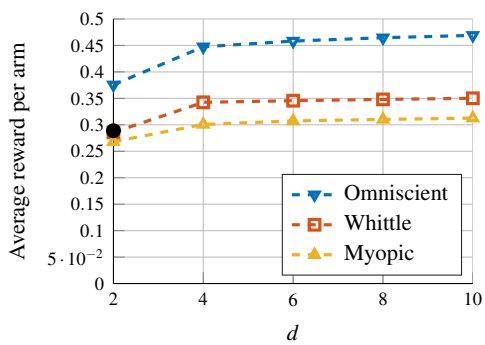


Fig. 3 Comparison of Whittle and myopic index policies for increasing number of channels d when half of the channels are GE and the other half is AR. For $d = 2$, the average reward obtained under the optimal policy is indicated by a black dot. We compare to the average reward that could be obtained if both arms were observed at each time point (that is in the fully observable setting).

All policies seem to approach a certain steady performance of average reward per arm rather quickly as the number of channels grows large while the ratio k/d

remains fixed. Note that without utilising the channel memory the average reward per arm is 0. So from a communication channel perspective (with this choice of objective function) the improvement in throughput is at the order of 35% when using the Whittle index.

It can further be seen that the Whittle index performs better than the myopic policy hence exploration pays off and improves the average reward per arm by roughly another 10% in this case. Compared with other literature results, this is in contrast to scenarios where all channels are GE and stochastically identical. In the latter case it can be shown that (under certain assumptions on the transition probabilities) the Whittle and the myopic index policy are equivalent. We give further details in (i) below.

Next, we investigate the Whittle indices $\iota^W(\omega)$ obtained for GE channels with various parameter combinations (Fig. 4). We observe the following properties of $\iota_W(\omega)$:

(i) The index function $\iota^W(\omega)$ increases monotonically; the larger the conditional probability that the channel is in a good state, the more priority should be given to that channel. This implies that the Whittle index is equivalent to the myopic policy in systems with *identical* channels, as we mentioned above.
(ii) $\iota_W(\omega)$ is linear in $\left[0, \min\{p_{01}, p_{11}\}\right]$ and $\left[\max\{p_{01}, p_{11}\}, 1\right]$, and changes slope at γ.

These properties have been proven in [22] for GE channels with reward function $r(\omega)$ given by the identity function.

We further note that the Whittle indices are overall smaller if γ is larger because in this case the rewards smaller (as r_j defined by (5) is decreasing in γ).

In Fig. 5 we show the difference between the Whittle and the myopic index function. It can be seen that the index functions are identical on $\left[0, \min\{p_{01}, p_{11}\}\right]$ and $\left[\max\{p_{01}, p_{11}\}, 1\right]$: In these regions exploration is not essential as it is rather certain that the state will evolve towards γ. Consequently, we see that ι^W and ι^M differ around γ, and on a larger interval to the left of γ.

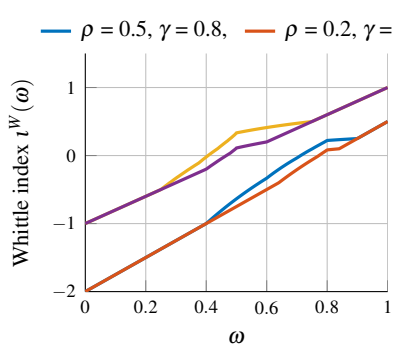


Fig. 4 Whittle indices for GE channels parametrized by α and γ.

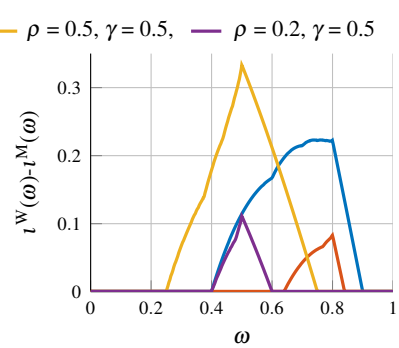


Fig. 5 Difference between Whittle and myopic index function.

In Fig. 6 we show the difference between Whittle and myopic indices as obtained for an AR channel. The obtained indices increase with μ because the expected immediate reward is larger. Note that for increasing age the Whittle indices increase relative to the myopic indices. Again this suggests that exploration pays off. Furthermore, for high ages the difference between the Whittle and the myopic indices is largest around zero, which corresponds to the unconditional mean reward of the channel. Similarly to the GE case, this may be explained by noting that exploration is more important if μ is close to the unconditional mean as it is less clear in which direction the belief state will evolve. If μ is far away from the unconditional mean on the other hand, then it is likely that the updated conditional mean will be closer to unconditional mean. However, when the age is close to zero, then due to the positive correlation of the channel it is also important that μ was large just an instance ago. Thus, while for small ages the Whittle indices are generally close to the myopic indices, the largest difference can be seen for positive μ (however not too far away from the unconditional mean of the channel).

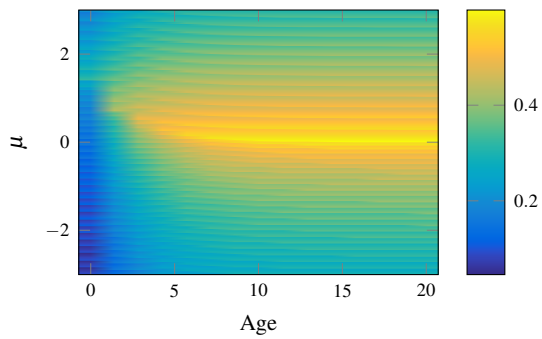


Fig. 6 Contour plot of $\gamma_W(\mu,\nu) - r(\mu)$, the difference of Whittle and myopic indices, for an AR channel with $\varphi = 0.8$, $\sigma = 2$.

5 Literature Survey

There is a vast body of literature on MDP as well as topics related to (restless) multiarmed bandits. Here, we focus on the RORMAB formulation of the basic channel selection problem as formulated in this chapter, with GE or AR channels. Other (approximate) solutions to this MDP problem have been put forward [15], but are not considered here.

GE Channels. The Gilbert-Elliot model was proposed in [8] for the purpose of modelling burst-noise telephone circuits. It was the first non-trivial channel model with memory. Since the 1990's, the model and its generalizations have been used for

modelling flat-fading channels in wireless communication networks. Its application in the context of opportunistic spectrum access (OSA) is motivated by the bursty traffic of primary users [17, 45]. For an account on the history of the GE model we refer to [35].

Due to its simplicity, the GE model is mathematically tractable and has been analysed extensively in the context of channel selection in wireless networks. We survey a number of papers that model the problem as a RORMAB with GE channels. Unless otherwise stated, channels are assumed to be independent and stochastically identical.

One of the first papers in this context appears to be [19]. The paper is motivated by the problem of allocating bandwidth of a shared wireless channel between a base station and multiple homogeneous mobile users. Thus, from an engineering perspective, the set-up slightly differs from the problem considered in this chapter; the model and the mathematical analysis, however, apply directly to the channel selection scenario (where simply "users" are replaced by "channels").

In [19], the noisiness of the link for the users is modelled using the GE model. At any point in time a user may either be connected to the base station or not. The current state of a user is only observed when a packet is transmitted to that user. Rewards are given by the number of successful transmissions. The analysis is with respect to the *discounted* reward criterion over an infinite time horizon. The authors show that the myopic policy is generally optimal for the case of $d = 2$ users. For the case $d > 2$ and positively correlated channels, it is proven that the myopic policy is optimal if the discount factor is small enough (Condition (A) in [19]). Furthermore, in the positively correlated scenario the myopic policy is seen to be equivalent to a "persistent round robin" policy where the link is dedicated to each user in a cyclic fashion according to their initial probability of being in a good state, and packets are transmitted to the same user until a packet fails to be transmitted correctly.

Following this work, the GE channel model has been analysed extensively in a surge of research on OSA, which goes back to [18]. The aim of this branch of research is to find secondary user policies that efficiently exploit transmission opportunities created by the bursty usage patterns of licensed primary users in wireless networks.

One of the first to formulate the RORMAB with GE channels in the context of OSA were Zhao et al. in 2005 [44]. The authors compare the transmission rate achieved by the myopic policy to the optimal policy in numerical examples.

This work was the starting point of a sequence of papers analysing the performance of the myopic policy. In [42], optimality is proven for the case of choosing one out of two channels, with respect to expected total discounted rewards over finite as well as infinite time horizon.

The scenarios in which the myopic policy is optimal are then generalized in a sequence of papers. Javidi et al. [16] consider the case of selecting 1 out of d channels and prove optimality of the myopic policy under the discounted reward criterion for positively correlated channels provided the discounted factor satisfies a certain inequality with respect to the transition probabilities. Under the additional ergodicity criterion

$$|p_{11} - p_{00}| < 1, \tag{14}$$

the myopic policy is further shown to be optimal under the average reward criterion (cf. (1)). The work of [16] is extended in [21] to the case of selecting $d-1$ out of d channels.

In [43] for the case of choosing 1 out of d channels the result of [19] is confirmed that the myopic policy is a persistent round robin scheme if channels are positively correlated. It is further shown that if correlation is negative, then the myopic policy is a round robin scheme, where the circular order is reversed in every time slot (and as for the positively correlated case, the user switches to the next channel as soon as the currently used channel signals has transitioned to the bad state). For the case $d = 2$, the myopic policy is shown to be optimal in general, as had already been established in [19]. Furthermore, it is shown that the performance of the myopic policy is determined by the stationary distributions of a higher-order countable-state Markov chain. The stationary distribution is known in closed form for the case $d = 2$. For the case $d > 2$, lower and upper bounds are established.

For negatively correlated channels and the case of selecting 1 out of d channels, the finite and infinite horizon discount-reward optimality of the myopic policy is proven in [3], provided that either $d \in \{2,3\}$ or the discount factor is less than half. These results also hold under average rewards under the additional ergodicity condition (14). For the finite-horizon discounted reward criterion, the results of [3] are generalized in [2] to the case of selecting k channels.

In 2014, Liu et al. [24] provide a unifying framework of the optimality conditions for the myopic policy that resulted from the OSA-motivated research of the channel selection problem with GE channels. The problem formulation in [24] is more general as it one to sense k out of d (identically distributed) channels but select only $l \leq k$ of those channels for transmission, based on the outcome of the sensing. The authors provide a set of unifying sufficient conditions under which the myopic policy is optimal. It is shown that the optimal policy is not generally myopic if $l < k$. (This is intuitive because the user is allowed to explore channels without having to use them.)

The Whittle index policy has also been studied both for the bandwidth allocation problem that was put forward in [19], and also in the context of OSA. As opposed to [19] a paper by Niño-Mora [28] handles the problem of bandwidth allocation when users are *heterogeneous*. The author proves that the problem is indexable and provides closed-form expressions for the index function.

For the basic RORMAB with GE channels, Liu and Zhao [22] prove that the Whittle index and the myopic policy are equivalent for positively correlated identical channels, thus, yielding the optimality of the Whittle index in this case. In [32], the indexability and closed-form expression for the Whittle index in the case of discounted rewards are derived for a more general model where the achievable transmission rate (the reward) for a channel in the bad state is, in general, non-zero and any rate above this achievable rate leads to outage.

Apart from the index policies proposed in this line of research, also algorithms for approximating an optimal policy have been investigated. See, for example, [11, 13], where algorithms for the more general model with correlated channels are proposed and investigated regarding their performance.

In the context of GE channels a number of generalizations of the basic model considered in this chapter have been considered. For example, a paper by Niño-Mora [29] allows for non-identical channels with sensing errors/measurement noise. Imperfect sensing was also considered in [23, 39]. In [32] the authors consider a problem where in both states, good and bad, transmission may fail with a certain non-zero probability, and it is only observed whether transmission was successful or not. Another recent paper with imperfect sensing is [26]. In this paper (co-authored by us) we focus on stability issues of queues associated with channel (server) selection in the context of imperfect sensing.

The paper [25] deals with random delay of feedback arrivals. Correlated channels were considered in [11, 12, 13]. Action-dependency of channel model parameters is taken into account in [37]. A very substantial paper is [38], which considers an RMAB in continuous time, and allows for non-identical channels, a time-dependent number of channels, and multiple actions. In this paper, a more general class of index policies is considered, which includes the Whittle index if the bandit problem is indexable. Asymptotic optimality for this class is proven for systems with many channels.

AR Channels. The AR channel model has only recently come to attention in the context of channel selection, and consequently the mathematical analysis is still at its starting point. The first to propose the application of this model for channel selection were Avrachenkov *et al.* [4] in 2012. This is motivated by empirical studies [1], showing that the AR model captures the logarithmic signal-to-noise ratio (SNR) of the channels reasonably well.

In [4], the authors compare the performance of the myopic and an ad-hoc randomized policy to the optimal policy by means of numerical examples. It is concluded that the myopic policy is "nearly optimal" when all channels are similarly correlated, with respect to the long-run average reward criterion. In contrast, the randomized policy appears to perform better when there is a significant difference in the magnitude of the correlation of the channels.

Subsequently, the authors show how to model the problem when two transmitters are present that can possibly interfere with each other. In this case the SNR is replaced by the signal-plus-interference-to-noise-ratio (SINR) to model the states of the channels. The scenario is formalized as a competitive MDP (also called a stochastic game) – an MDP in which the instantaneous rewards for each player and the transition probabilities among the states are controlled by the joint actions of the players in each state. Then, similar to the single user case, a randomized and a myopic policy are suggested (now based on the SINR).

A second paper that deals with channel selection with AR channels is [20]. In this paper (co-authored by us) we investigate structural properties of the Whittle

index with respect to expected total discounted rewards. The monotonicity and convexity of the value function associated with the one-channel sub-problem is proven. Furthermore, numerical evidence for the indexability of the one-armed problem is provided, and the Whittle index policy is shown to outperform the myopic policy in numerical examples.

Then, a parametric index is proposed that is as simple as the myopic index but allows to give some priority to exploration, and therefore yields a better performance than the latter. For this parametric index, we put forward recursive equations that identify the asymptotic behaviour of the network in the setting with many channels. In addition, a simple heuristic algorithm is proposed to evaluate the performance of index policies; the latter is used to optimize the parametric index.

We also note that a related body of literate to AR channels deals with the problem of optimal sensing of Kalman filters. A key paper in this line of research is [31]. A related paper is [30] as well as the recent [6] which appears to provide an indexabillity proof using a new novel method. It is possible that ideas put forward in these papers dealing with the Whittle index and simple Gaussian processes may be fruitful for the RORMAB problem with AR channels. This avenue of research remains to be explored.

Acknowledgement: YN is supported by Australian Research Council (ARC) grants DP130100156 and DE130100291. JK is supported by DP130100156. The authors are indebted to Aapeli Vuorinen for his contribution to the numerical computations as well as to Michel Mandjes and Thomas Taimre for their comments.

References

1. R. Aguero, M. Garcia, and L. Munoz. BEAR: A bursty error auto-regressive model for indoor wireless environments. In *18th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 1–5. IEEE, 2007.
2. S. H. A. Ahmad and M. Liu. Multi-channel opportunistic access: A case of restless bandits with multiple plays. In *47th Annual Allerton Conference on Communication, Control, and Computing*, pages 1361–1368. IEEE, 2009.
3. T. W. Archibald, D. Black, and K. D. Glazebrook. Indexability and index heuristics for a simple class of inventory routing problems. *Operations research*, 57(2):314–326, 2009.
4. K. Avrachenkov, L. Cottatellucci, and L. Maggi. Slow fading channel selection: A restless multi-armed bandit formulation. In *International Symposium on Wireless Communication Systems (ISWCS)*, pages 1083–1087. IEEE, 2012.
5. D. P. Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, 1995.
6. C. R. Dance and T. Silander. When are kalman-filter restless bandits indexable? *arXiv preprint arXiv:1509.04541*, 2015.
7. D. Duchamp and N. Reynolds. Measured performance of a wireless LAN. In *17th Conference on Local Computer Networks*, pages 494–499. IEEE Press, 1992.
8. E. N. Gilbert. Capacity of a burst-noise channel. *Bell System Technical Journal*, 39(5):1253–1265, 1960.
9. J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed Bandit Allocation Indices*. Wiley Online Library, 2 edition, 2011.

10. J. C. Gittins. Bandit Processes and Dynamic Allocations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.
11. S. Guha and K. Munagala. Approximation algorithms for partial-information based stochastic control with Markovian rewards. In *48th Annual Symposium on Foundations of Computer Science (FOCS'07)*, pages 483–493. IEEE, 2007.
12. S. Guha, K. Munagala, and P. Shi. Approximation algorithms for restless bandit problems. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2009.
13. S. Guha, K. Munagala, and P. Shi. Approximation algorithms for restless bandit problems. *Journal of the ACM (JACM)*, 58(1):3, 2010.
14. O. Hernández-Lerma and J. B. Lasserre. *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media, 2012.
15. A. Itai and Z. Rosberg. A golden ratio control policy for a multiple-access channel. *IEEE Transactions on Automatic Control*, 29(8):712–718, 1984.
16. T. Javidi, B. Krishnamachari, Q. Zhao, and M. Liu. Optimality of myopic sensing in multichannel opportunistic access. In *International Conference on Communications (ICC'08)*, pages 2107–2112. IEEE, 2008.
17. L. A. Johnston and V. Krishnamurthy. Opportunistic file transfer over a fading channel: A POMDP search theory formulation with optimal threshold policies. *IEEE Transactions on Wireless Communications*, 5(2):394–405, 2006.
18. R. Knopp and P. A. Humblet. Information capacity and power control in single-cell multiuser communications. In *International Conference on Communications (ICC'95)*, volume 1, pages 331–335. IEEE, 1995.
19. G. Koole, Z. Liu, and R. Righter. Optimal transmission policies for noisy channels. *Operations Research*, 49(6):892–899, 2001.
20. J. Kuhn, M. Mandjes, and Y. Nazarathy. Exploration vs exploitation with partially observable Gaussian autoregressive arms. In *8th International Conference on Performance Evaluation Methodologies and Tools (Valuetools)*, 2014.
21. K. Liu and Q. Zhao. Channel probing for opportunistic access with multi-channel sensing. In *Asilomar Conference on Signals, Systems and Computers*. IEEE, 2008.
22. K. Liu and Q. Zhao. Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, 2010.
23. K. Liu, Q. Zhao, and B. Krishnamachari. Dynamic multichannel access with imperfect channel state detection. *IEEE Transactions on Signal Processing*, 58(5):2795–2808, 2010.
24. Y. Liu, M. Liu, and S. H. A. Ahmad. Sufficient conditions on the optimality of myopic sensing in opportunistic channel access: A unifying framework. *IEEE Transactions on Information Theory*, 60(8):4922–4940, 2014.
25. S. Murugesan, P. Schniter, and N. B. Shroff. Multiuser scheduling in a Markov-modeled downlink using randomly delayed arq feedback. *IEEE Transactions on Information Theory*, 58(2):1025–1042, 2012.
26. Y. Nazarathy, T. Taimre, A. Asanjarani, J. Kuhn, P. Brendan, and A. Vuorinen. The challenge of stabilizing control for queueing systems with unobservable server states. In *Australian Control Conference, AUCC, to appear*, 2015.
27. J. Niño-Mora. Dynamic priority allocation via restless bandit marginal productivity indices. *Top*, 15(2):161–198, 2007.
28. J. Niño-Mora. An index policy for dynamic fading-channel allocation to heterogeneous mobile users with partial observations. In *Next Generation Internet Networks (NGI)*, pages 231–238. IEEE, 2008.
29. J. Niño-Mora. A restless bandit marginal productivity index for opportunistic spectrum access with sensing errors. In R. Núñez-Queija and J. Resing, editors, *Network Control and Optimization*, volume 5894 of *Lecture Notes in Computer Science*, pages 60–74. Springer, Berlin, 2009.
30. J. Nio-Mora and S. S. Villar. Multitarget tracking via restless bandit marginal productivity indices and kalman filter in discrete time. In *Decision and Control, 2009 held jointly with*

the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on, pages 2905–2910. IEEE, 2009.

31. J. L. Ny, E. Feron, M. Dahleh, et al. Scheduling continuous-time kalman filters. *Automatic Control, IEEE Transactions on*, 56(6):1381–1394, 2011.
32. W. Ouyang, S. Murugesan, A. Eryilmaz, and N. B. Shroff. Exploiting channel memory for joint estimation and scheduling in downlink networks. In *30st Annual International Conference on Computer Communications (INFOCOM)*, pages 3056–3064. IEEE, 2011.
33. C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.
34. M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*, volume 414. Wiley, New York, 2009.
35. P. Sadeghi, R. A. Kennedy, P. B. Rapajic, and R. Shams. Finite-state Markov modeling of fading channels – a survey of principles and applications. *IEEE Signal Processing Magazine*, 25(5):57–80, 2008.
36. R. D. Smallwood and E. J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.
37. J. A. Taylor and J. L. Mathieu. Index policies for demand response. *IEEE Transactions on Power Systems*, 2013.
38. I. M. Verloop. Asymptotically optimal priority policies for indexable and non-indexable restless bandits. *Submitted*, Retrieved 07/09/2015 from https://hal.archives-ouvertes.fr/hal-00743781.
39. K. Wang, L. Chen, Q. Liu, and K. Al Agha. On optimality of myopic sensing policy with imperfect sensing in multi-channel opportunistic access. *IEEE Transactions on Communication*, 61(9):3854–3862, 2013.
40. R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.
41. P. Whittle. Restless bandits: activity allocation in a changing world. *Journal of Applied Probability*, Special Vol. 25A:287–298, 1988. A celebration of applied probability.
42. Q. Zhao and B. Krishnamachari. Structure and optimality of myopic sensing for opportunistic spectrum access. In *International Conference on Communications (ICC'07)*. IEEE, 2007.
43. Q. Zhao, B. Krishnamachari, and K. Liu. On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance. *IEEE Transactions on Wireless Communications*, 7(12):5431–5440, 2008.
44. Q. Zhao, L. Tong, and A. Swami. Decentralized cognitive mac for dynamic spectrum access. In *First International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*. IEEE, 2005.
45. M. Zorzi, R. R. Rao, and L. B. Milstein. Error statistics in data transmission over fading channels. *IEEE Transactions on Communications*, 46(11):1468–1477, 1998.