Exploration vs Exploitation with Partially Observable Gaussian Autoregressive Arms

Julia Kuhn
The University of Queensland,
University of Amsterdam
j.kuhn@uq.edu.au

Michel Mandjes
University of Amsterdam
m.r.h.mandjes@uva.nl

Yoni Nazarathy
The University of Queensland
y.nazarathy@uq.edu.au

ABSTRACT

We consider a restless bandit problem with Gaussian autoregressive arms, where the state of an arm is only observed when it is played and the state-dependent reward is collected. Since arms are only partially observable, a good decision policy needs to account for the fact that information about the state of an arm becomes more and more obsolete while the arm is not being played. Thus, the decision maker faces a tradeoff between exploiting those arms that are believed to be currently the most rewarding (i.e. those with the largest conditional mean), and exploring arms with a high conditional variance. Moreover, one would like the decision policy to remain tractable despite the infinite state space and also in systems with many arms. A policy that gives some priority to exploration is the Whittle index policy, for which we establish structural properties. These motivate a parametric index policy that is computationally much simpler than the Whittle index but can still outperform the myopic policy. Furthermore, we examine the many-arm behavior of the system under the parametric policy, identifying equations describing its asymptotic dynamics. Based on these insights we provide a simple heuristic algorithm to evaluate the performance of index policies; the latter is used to optimize the parametric index.

Categories and Subject Descriptors

I.2.8 [**Problem Solving, Control Methods, and Search**]: Control theory

Keywords

Restless bandits, partially observable, Whittle index, performance evaluation, asymptotic dynamics

1. INTRODUCTION

Inherent to the problem of decision making under partial observability is the tradeoff between exploration and exploitation: Should we collect new information, or opt for the immediate payoff? We investigate this question for a *reward observing restless multiarmed bandit problem*, where at every point in discrete time the decision maker wants to play a fixed number k out of d arms, with the objective of maximizing the expected (discounted or average) reward achieved over an infinite time horizon. The state of an arm is *restless* as it evolves also when the arm is not played, and *partially observable* as it can only be observed whenever the arm is played and the state-dependent reward is collected. This type of bandit problem has drawn much attention in recent literature, e.g. [3, 8, 9]. We call it *reward observing* to distinguish from other types of partially observable decision problems, for example those where the decision maker does not have fully accurate information due to measurement errors [11].

In view of the reward observing character of the problem, one would like a model that allows in a non-artificial way to keep track of the decision maker's current need for exploring an arm rather than exploiting others. As a starting point we assume in this paper that state processes are AR(1), Gaussian autoregressions of order 1. Since states are normally distributed, the objectives of exploitation and exploration naturally correspond to the conditional mean and variance of an arm, which at the same time contain all relevant information concerning its state (and thus fully describe the *belief state* of the arm). The AR(1) model has been found useful for example for modeling channels in wireless networks [1]. It seems that in the context of decision making under reward observability it has previously only been considered in [3], where the myopic (greedy) policy was compared numerically to an ad hoc randomized policy.

While in principle an optimal policy for Markovian restless bandit problems can be obtained with the aid of dynamic programming, in practice this is typically computationally infeasible [12]. Particularly when the system is large (with many arms), one therefore often resorts to the class of *index policies*, which remain tractable due to a decoupling of arms. For every arm, the information available is mapped to some real-valued priority index which does not depend on the state or history of any other arm. At every time slot the policy then activates those k arms that correspond to the k largest indices.

It has been shown in [15] and more recently in [14] that – under technical conditions that are not satisfied by the model considered in this paper – an index policy known as the *Whittle index* [16] is asymptotically optimal for restless

VALUETOOLS 2014, December 09-11, Bratislava, Slovakia

DOI 10.4108/icst.valuetools.2014.258207

bandit problems. These results hold as the number of arms tends to infinity while the ratio of played arms, k/d, tends to a constant ρ, a regime also considered in this paper. Furthermore, under the restrictive assumption that arms can be modeled as identically distributed two state Markov chains (Gilbert-Elliott), the authors of [9] derive non-asymptotic optimality results for the Whittle index for the reward observing decision problem. The key feature of the Gilbert-Elliott model is that due to its simplicity the Whittle index can be computed in closed form. For the AR(1), however, no optimal policy is known and a closed-form expression for the Whittle index does not appear to be available.

Our contributions are both structural and asymptotic. Considering the discounted reward case, we find structural properties of the one armed subsidy problem associated with the Whittle index. We establish convexity and monotonicity properties which, based on non-restrictive assumptions, imply the existence of a switching curve and the monotonicity of the related Whittle index. These properties motivate a simple parametric index which quantifies the virtue of exploration compared to exploitation in terms of variance and mean. For this index we analyze the mean-field behavior of the system in the average reward case. In particular, we put forward a deterministic measure-valued recursion that approximately describes the distribution of belief states when the number of arms is non-small. We merge these ideas into a performance evaluation and optimization procedure, which we illustrate to be asymptotically exact.

The paper is organized as follows. In Section 2 we formulate the decision problem. Sections 3 and 4 present our contributions with respect to structural and asymptotic analysis of the problem. We conclude in Section 5.

2. MODEL AND FRAMEWORK

The state processes $\boldsymbol{X}(t) := \big(X_1(t), \ldots, X_d(t)\big)$ are assumed to be independent, and satisfy the AR(1) recursion,

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

with $\big\{\varepsilon_i(t)\big\}_t$ denoting an i.i.d. sequence of $\mathcal{N}\big(0, \sigma^2\big)$ random variables. The parameters φ, σ are assumed to be known. We restrict our exposition to the case $\varphi \in (0,1)$, whence the processes are stable and observations are positively correlated over time.

At every point in time the decision maker may choose whether or not to play arm i, i.e., to observe its state and collect the reward. We denote the action of playing arm i by $a_i(t) = 1$ (active), while $a_i(t) = 0$ (passive) refers to the action of not playing. We require that exactly k arms have to be activated at each decision time, i.e., the action vector $\boldsymbol{a}(t) := \big(a_1(t), \ldots, a_d(t)\big)$ satisfies $\sum_{i=1}^d a_i(t) = k$.

We are in the partially observable setting, that is, the state of an arm is only observed when that arm is activated, while at every time slot all arms evolve to the next state. Thus, the states of the $d-k$ passive arms are unknown to the decision maker, and he has to rely on his belief concerning these states. The *belief state* of arm i at time t is given by the probability distribution over all of its possible states conditional on the information available at that time. Since this distribution is Normal, the belief state is fully characterized by the conditional mean and variance defined as

$$\begin{aligned}\mu_i(t) &:= \mathbb{E}\Big[X_i(t) \,\Big|\, X_i\big(t-\eta_i(t)\big), \eta_i(t)\Big] \qquad (1)\\ &= \varphi^{\eta_i(t)} X_i\big(t-\eta_i(t)\big),\\ \nu_i(t) &:= \mathrm{Var}\left(X_i(t) \,\middle|\, X_i\left(t-\eta_i(t)\right), \eta_i(t)\right) \qquad (2)\\ &= \sigma^2 \sum_{h=0}^{\eta_i(t)-1} \varphi^{2h} = \sigma^2 \frac{1-\varphi^{2\eta_i(t)}}{1-\varphi^2}.\end{aligned}$$

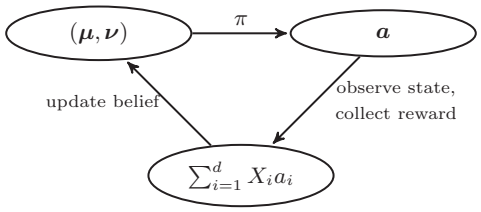
Here, $\eta_i(t) := \min\{h \geq 1 | a_i(t-h) = 1\}$ denotes the number of time steps ago arm i was last played and observed. We denote the joint (belief) state space of (μ_i, ν_i) by $\Psi := \Psi_1 \times \Psi_2$ so that $(\boldsymbol{\mu}, \boldsymbol{\nu}) \in \Psi^d$. It is worth noting that $\Psi_1 = \mathbb{R}$ and Ψ_2 is countable and bounded by $[\nu_{\min}, \nu_{\max}) := \big[\sigma^2, \sigma^2/(1-\varphi^2)\big)$. The conditional variance increases in η_i, i.e. while the arm is not being played.

The following evolutions show how the belief states are updated in a Markovian manner; these also appear in [3]. Let $Y_{\mu,\nu}$ denote a generic random variable with distribution $\mathcal{N}(\mu, \nu)$. If action $a_i(t)$ is chosen at time t, then at time $t+1$,

$$\big(\mu_i(t+1),\, \nu_i(t+1)\big) = \begin{cases} \big(\varphi\, \mu_i(t),\, \varphi^2\, \nu_i(t) + \sigma^2\big), & a_i(t) = 0,\\ \big(\varphi\, Y_{\mu_i(t),\, \nu_i(t)},\, \sigma^2\big), & a_i(t) = 1. \end{cases} \qquad (3)$$

Here, the realization of $Y_{\mu_i(t),\nu_i(t)}$ corresponds to the realization of the state process the decision maker observes when playing arm i at time t. On the other hand, if the arm is not played, then the previous belief state is updated in a deterministic fashion as no new observation of the state of arm i is made.

In summary, as time evolves from t to $t+1$, given the current belief states $(\boldsymbol{\mu}, \boldsymbol{\nu})$ and a *policy* $\pi : \Psi \to \{0,1\}^d$, the following chain of actions takes place:



The aim is to find a policy π so as to maximize the accumulated rewards over an infinite time horizon as evaluated by the *total expected discounted reward criterion*,

$$V^\pi(\boldsymbol{\mu}, \boldsymbol{\nu}) := \lim_{T\to\infty} \mathbb{E}^\pi_{\boldsymbol{\mu},\boldsymbol{\nu}} \left[\sum_{t=0}^{T} \beta^t \sum_{i=1}^{d} X_i(t)\, a_i(t)\right], \qquad (4)$$

where $\beta \in (0,1)$, and the subscript indicates conditioning on $\boldsymbol{X}(0) \sim \mathcal{N}\big(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\nu})\big)$, or the *average expected reward criterion*

$$G^\pi(\boldsymbol{\mu}, \boldsymbol{\nu}) := \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}^\pi_{\boldsymbol{\mu},\boldsymbol{\nu}} \left[\sum_{t=0}^{T-1} \sum_{i=1}^{d} X_i(t)\, a_i(t)\right]. \qquad (5)$$

Note that $X_i(t)$ in (4) and (5) can be replaced by $\mu_i(t)$.

LEMMA 2.1. *The function V^π is well-defined in the sense that the limit in (4) exists and is finite. Furthermore, the optimal value function $\sup_\pi V^\pi$ is finite.*

In view of computational tractability we restrict our exposition to policies from the class of index policies.

3. INDEX POLICIES

An index policy is a policy of the form

$$\pi_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) = \underset{\boldsymbol{a}:\sum_{i=1}^d a_i = k}{\arg\max} \left\{ \sum_{i=1}^d \gamma\left(\mu_i, \nu_i\right) a_i \right\}, \tag{6}$$

where the *index function* $\gamma : \Psi \to \mathbb{R}$ maps the belief state of each arm to some priority index. That is, at every point in time, π_γ activates those k arms that correspond to the k largest indices; ties are broken arbitrarily. Without loss of generality the index function can be written as

$$\gamma(\mu, \nu) = \mu + q(\mu, \nu) \tag{7}$$

for some known function $q : \Psi \to \mathbb{R}$. The basic example is the *myopic* index with $q \equiv 0$. The resulting policy always activates those k arms with the largest expected immediate reward. As it does not account for information growing obsolete (giving full priority to exploitation), the performance of the myopic policy deteriorates as $\beta \uparrow 1$. A more sophisticated index policy, the Whittle index, is surveyed in the next subsection.

3.1 Whittle Index

The Whittle index is a generalization of the Gittins index to the restless bandit case in which state processes (or belief states) evolve irrespective of whether an arm is being played or not – as opposed to the classical multiarmed bandit problem [6, 7] in which the states of unplayed arms do not evolve. In order to devise a heuristic policy for a restless multiarmed bandit problem, Whittle [16] considered arms separately (i.e., he considered d decoupled one-armed bandits), and introduced a subsidy paid for leaving the arm under consideration passive. Intuitively this subsidy can be thought of as a substitute for the rewards the decision maker could have obtained from playing other arms in the multiarmed setting; from a theoretical point of view it is a Lagrange multiplier associated with the relaxed constraint that k arms have to be activated *on average* rather than requiring that *exactly* k arms be activated. Due to this relaxation, the Whittle index policy is not generally optimal for small systems but, at least under certain conditions that do not apply here, it is asymptotically optimal as $k, d \to \infty$ with $k/d \to \rho$ [14, 15]. This is also the asymptotic regime we consider in Section 4.

We show how to obtain the Whittle index when the underlying states are AR(1), and provide structural results.

Consider a special one-armed bandit problem where at each time slot, the decision maker can either activate the arm ($a = 1$) or leave it passive ($a = 0$). When it is activated, then the decision maker observes the state and collects the corresponding reward. When the arm remains passive, he obtains a (possibly negative) subsidy λ. The objectives are analogous to (4) and (5), but our focus in this section is on (4). We call this problem *one-armed bandit problem with subsidy*. The Whittle index is then defined as the smallest subsidy for which it is optimal to leave the arm passive.

DEFINITION 3.1. *Let $\mathcal{P}_\lambda$ denote the passive set associated with the one-armed problem with subsidy λ,*

$$\mathcal{P}_\lambda := \{(\mu, \nu) \,|\, a = 0 \text{ is optimal action}\}.$$

Then the Whittle index associated with this arm and state (μ, ν) is given by $\omega(\mu, \nu) = \inf\{\lambda \,|\, (\mu, \nu) \in \mathcal{P}_\lambda\}$.

This way, the Whittle index is obtained from the optimal policy for the one-armed problem with subsidy, which can be derived from the optimal value function as outlined below. The Whittle index policy is sensible only if any arm rested under the subsidy λ remains passive under every subsidy $\widetilde{\lambda} > \lambda$. If this is the case, the one-armed bandit problem is called *indexable* [16]. Even if the state space is finite, proving indexability is likely to be highly involved [5]; we assume that it holds here as confirmed through extensive numerical experimentation (see for example Fig. 1).

The discount-optimal value function $V^\lambda := \sup_\pi V^{\lambda,\pi}$ for the one-armed problem with subsidy can be obtained using *value iteration* (see Prop. 3.2). For the average reward case one can formulate a similar result stating that $G^\lambda := \sup_\pi G^{\lambda,\pi}$ can be found from *relative value iteration* as defined for example in [13, Section 8.5.5]. First we introduce the operator $Tv := \max_{a\in\{0,1\}} T_a v$, where

$$T_a v\,(\mu, \nu) := \begin{cases} \lambda + \beta\, v(\varphi\,\mu,\, \varphi^2\nu + \sigma^2), & a = 0, \\ \mu + \beta \int_{-\infty}^{\infty} v\,(\varphi\, y,\, \sigma^2)\, \phi_{\mu,\nu}(y)\, dy, & a = 1, \end{cases}$$

with $\phi_{\mu,\nu}$ denoting the normal density with mean μ and variance ν. Proofs can be found in the appendix.

PROPOSITION 3.2. *For $V_0^\lambda \equiv 0$ the iteration*

$$V_n^\lambda = TV_{n-1}^\lambda \tag{8}$$

converges to a unique function $V^\lambda : \Psi \to \mathbb{R}$ as $n \to \infty$ that satisfies the Bellman equation,

$$V^\lambda = TV^\lambda.$$

This V^λ is the discount-optimal value function for the one-arm bandit problem with subsidy λ. An optimal policy for this problem maps (μ, ν) to action a if $V^\lambda(\mu, \nu) = T_a V^\lambda(\mu, \nu)$.

Structural Properties. Let us first consider monotonicity properties of the optimal value function V^λ.

LEMMA 3.3. *Let $\varphi \in (0, 1)$. Then $V^\lambda(\cdot, \nu)$ is convex, continuous, non-decreasing, and not constant; and $V^\lambda(\mu, \cdot)$ is non-decreasing.*

Fig. 1 and similar numerical experiments[1] suggest that the passive set $\mathcal{P}_\lambda$ and the active set $\mathcal{P}_\lambda^c$ are separated by a

[1]To execute the value iteration we truncate Ψ_1 to $[-6\sigma, 6\sigma]$, and consider $\lambda \in [-2\sigma, 2\sigma]$. Discretization is done in steps of size 0.01, which is preserved when truncating Ψ_2.

switching curve (defined on the countable space Ψ_2).

ASSUMPTION 1. *It holds that* $\overline{\mu} := \sup\{\mu \,|\, (\mu,\nu) \in \mathcal{P}_\lambda\} < \infty$ *and* $\underline{\mu} := \inf\{\mu \,|\, (\mu,\nu) \in \mathcal{P}_\lambda^c\} > -\infty$.

We conjecture that this assumption generally holds here; it states that if the expected reward obtained when playing is large enough, then it is optimal to indeed play the arm, while if it is very small, we should rather take the subsidy. To see that there is a switching curve, it remains to be proven that in between $\underline{\mu}$ and $\overline{\mu}$ there are no $\mu_1 < \mu_2$ such that $(\mu_1,\nu) \in \mathcal{P}_\lambda^c$ and $(\mu_2,\nu) \in \mathcal{P}_\lambda$.

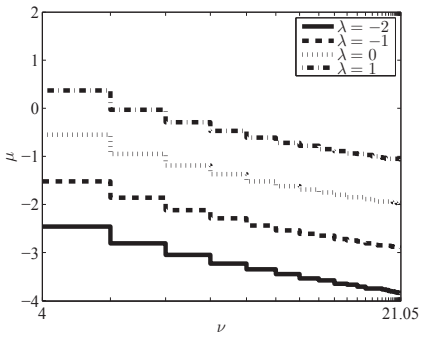


Figure 1: Switching curves: below the curve the optimal action is passive, above it is active. $\beta = 0.8$, $\varphi = 0.9$, $\sigma = 2$.

PROPOSITION 3.4. *If Assumption 1 holds, then a policy that achieves the optimal value function* V^λ *is a threshold policy: There exists a switching curve (sequence)* $\zeta_\lambda : \Psi_2 \to \Psi_1$ *such that* $\mathcal{P}_\lambda = \{(\mu,\nu) \,|\, \mu \le \zeta_\lambda(\nu)\}$.

Numerical evidence such as provided in Fig. 1 suggests that the switching curve is in fact strictly decreasing, i.e. it is optimal to give some priority to exploration.

ASSUMPTION 2. *The switching curve is non-increasing.*

It follows from the assumptions and Prop. 3.4 that the Whittle index is monotone.

COROLLARY 3.5. *Provided Assumption 1 holds, the Whittle index* $\omega(\mu,\nu)$ *is non-decreasing in* μ. *If in addition Assumption 2 holds, then* $\omega(\mu,\nu)$ *is non-decreasing in* ν.

Consequently, the Whittle index policy assigns comparatively larger indices to arms that have not been activated for a longer time. In accordance with this observation, Fig. 2 shows that the correction term $q(\mu,\nu)$ is positive and increases in ν. It is larger for μ close to zero, which may be explained by noting that exploration is less important if $|\mu|$ is large, for in that case there is less uncertainty about the direction in which μ will evolve. Furthermore, we confirmed numerically that the slope of $\zeta_\lambda(\nu)$ increases as β increases as in this case exploration becomes more beneficial.

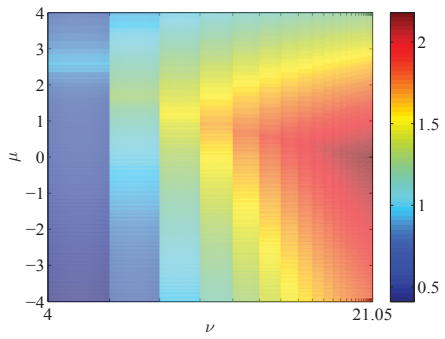


Figure 2: Correction term $q(\mu,\nu)$ **obtained for the Whittle index.** $\beta = 0.8$, $\varphi = 0.9$, $\sigma = 2$.

3.2 Parametric Index

As no closed form for the Whittle index is available, the Whittle indices have to be computed and stored for every belief state in Ψ, while the evaluation of the optimal value function for the one-armed problem with subsidy is computationally expensive. Therefore, instead of finding the index for the one-armed problem with subsidy that is optimal (the Whittle index), we propose to find the index that is optimal when restricting to a family of parametric functions. A simple example is obtained by picking a function $q(\mu,\nu)$ that is proportional to ν, the most obvious measure for the decision maker's uncertainty. This yields the *parametric index*

$$\gamma(\mu,\nu) = \mu + \theta\nu, \tag{9}$$

where $\theta \ge 0$ as motivated by Corollary 3.5. The correction term $\theta\nu$ allows to adjust the priority the decision maker wants to give to exploration. We denote the associated policy by π_θ.

The parametric index can be related to the Whittle index as follows. Numerical experiments (such as Fig. 1 for discounted, and related experiments for average rewards) suggest that the optimal switching curve may be well approximated by a linear function, the slope of which is negative but does not depend on λ. The position of the curve on the other hand does depend on λ. Such an approximation is given by $\zeta_\lambda(\nu) \approx -\theta\nu + \lambda + c$ with $\theta \ge 0$, $c \in \mathbb{R}$. As $\zeta_\lambda(\nu)$ takes some value $\mu \in \mathbb{R}$, solving for λ (which may correspond to the Whittle index) suggests using an index of the form (9), where without loss of generality we take $c = 0$.

In the next section we show that we can explicitly describe the asymptotic dynamics of the system induced by π_θ.

4. SYSTEM WITH MANY ARMS

We investigate the behavior of the system with many arms as $d \to \infty$ and $k_d/d \to \rho$. Section 4.1 outlines the main idea: the limiting proportion of belief states remains stable in an equilibrium system with infinitely many arms. In Section 4.2 we relate the equilibrium system to a single arm process, and use this connection to propose an algorithm for performance evaluation. This algorithm is used to optimize π_θ in Section 4.3.

4.1 Limiting Empirical Distribution

We first informally describe the intuition motivating this section. Consider the system with d arms as before, and to simplify the exposition suppose for now that the system is stationary. Let $\Gamma_i(t)$ denote the process of indices associated with arm i, that is, $\Gamma_i(t) := \gamma\big(\mu_i(t),\, \nu_i(t)\big)$. Note that the index processes $\Gamma_i(t)$ and $\Gamma_j(t)$, $i, j = 1, \ldots, d$, are generally dependent because the belief states of both arms depend on the action that was chosen, which in turn depends on the index of all arms in the system (as they are coupled by the requirement that those k arms with the largest indices are activated). Let us now focus on a single arm i in this system and suppose that its belief state evolves from ψ at time t to another belief state $\widetilde{\psi}$ at time $t+1$. While d is small, this certainly changes the proportion of arms with current belief state ψ considerably. However, if d is very large, we should be able to find another arm j whose new belief state at time t is (in close proximity to) ψ. It thus seems reasonable to expect that, as we add more arms to the system, it approaches a mean-field limit in which the proportion of arms associated with a certain belief state remains fixed. Thus, in the limit, the action chosen for a certain arm is independent of the current belief state of any other arm, as there is always the same proportion of arms associated with a certain belief state in the system.

Let us now more formally investigate the proportion of arms that are associated with a certain belief state at time t. We focus on parametric index functions as defined in (9). The empirical measure

$$M^d(C,t) := \frac{1}{d}\sum_{i=1}^{d} \mathbb{1}\Big\{\big(\mu_i(t),\, \nu_i(t)\big) \in C\Big\} \tag{10}$$

quantifies the proportion of arms in the d-dimensional system whose belief state falls into $C \in \mathcal{B}(\Psi)$ at time t, where $\mathcal{B}(\Psi)$ denotes the Borel σ-algebra on Ψ. It is related to the measure on indices,

$$\widetilde{M}^d(B,t) := \frac{1}{d}\sum_{i=1}^{d} \mathbb{1}\big\{\Gamma_i(t) \in B\big\} \tag{11}$$

with $B \in \mathcal{B}(\mathbb{R})$, through

$$\widetilde{M}^d(B,t) = M^d\bigg(\Big\{(\mu,\nu) \in \Psi \,|\, \mu + \theta\nu \in B\Big\},\, t\bigg). \tag{12}$$

We examine the dynamics of $M^d(C,t)$. To this end we enumerate the elements in Ψ_2, that is, $\nu^{(h)} = \sigma^2(1-\varphi^{2(h+1)})/(1-\varphi^2)$, $h = 0, 1, 2\ldots$, so that $h+1$ is the number of time steps since an arm was played last. We refer to h as the *age* of an arm. Then (10) can be written as, with $B \in \mathcal{B}(\mathbb{R})$,

$$\sum_{h=0}^{\infty} M_h^d(B,t) := \sum_{h=0}^{\infty} \frac{1}{d} \sum_{i:\, \nu_i(t)=\nu^{(h)}} \mathbb{1}\big\{\mu_i(t) \in B\big\}. \tag{13}$$

Many-arms Asymptotics. As motivated at the beginning of this section, it is reasonable to believe that the limiting proportion of arms associated with a certain belief state evolves deterministically, and thus, that the dynamics of the limiting system can be described by non-random measures $m_h(\cdot,t)$. For brevity we write $m_h(x,t)$ for $m_h\big((-\infty,x],t\big)$, and denote by $\Phi_{\mu,\nu}$ the normal distribution function with mean μ and variance ν. We define $m_h(\cdot,t)$ by the recursion

$$m_h\big(x, t+1\big) \tag{14}$$
$$= \begin{cases} \sum_{h=0}^{\infty} \int_{\ell_h(t)}^{\infty} \Phi_{z,\nu^{(h)}}\left(\frac{x}{\varphi}\right) m_h(dz,\, t), & h = 0, \\ m_{h-1}\Big(\min\big\{\frac{x}{\varphi},\, \ell_{h-1}(t)\big\},\, t\Big), & h \geq 1, \end{cases}$$

where $\ell_h(t) := \ell(t) - \theta\nu^{(h)}$ with $\ell(t)$ defined by

$$\ell(t) = \sup\left\{\ell \,\Big|\, \sum_{h=0}^{\infty} \widetilde{m}_h\big(\,[\ell,\infty),\, t\big) = \rho\right\}. \tag{15}$$

Here, $\widetilde{m}_h$ denotes the measure on indices, i.e.

$$\widetilde{m}_h(B,\, t) = m_h\Big(\big\{\mu \in \Psi_1 \,|\, \mu + \theta\nu^{(h)} \in B\big\},\, t\Big), \tag{16}$$

cf. Eqn. (12). Note that $\ell_h(t)$ is a threshold such that at time t the policy π_θ activates all arms that are of age h and have conditional mean $\mu(t) \geq \ell_h(t)$, $h \geq 0$. Obviously, if the policy is myopic, then $\ell_h(t) = \ell(t)$ does not depend on the age of an arm and the above expressions can be simplified. Recursion (14) is obtained based on the dynamics of the belief states as given in (3). The evolution of $m_0(\cdot,t)$ is determined by the evolution of the belief state of all arms that have been played in the previous time slot. If $h > 0$ on the other hand, we use that arms of age h must have been of age $h-1$ at the previous decision time; and since they have not been activated, their mean must have been below the threshold $\ell_h(t-1)$.

For the (pre-limit) empirical processes M^d it obviously holds that $M^d(\Psi,t) = 1$ as well as $M_0^d(\Psi_1,t) = k_d/d$. These properties carry over to the limiting measure.

LEMMA 4.1. *If the sequence $\{m_h(B,0)\}_h$ satisfies*

$$m_0\big(\Psi_1,0\big) = \rho, \quad \textit{and} \quad \sum_{h=0}^{\infty} m_h\big(\Psi_1,0\big) = 1, \tag{17}$$

then the same holds for $m_h(B,t)$ for all $t > 0$.

This is easily proven by induction using (14)–(16). We believe that (14) indeed describes the mean-field behavior of the dynamical system:

CONJECTURE 1. *Assume that $M_h^d(B,0)$ converges weakly to $m_h(B,0)$ for all $h \geq 0$,*

$$M_h^d(B,0) \xrightarrow{w} m_h(B,0),$$

as $d \to \infty$ while $\lim_{d\to\infty} k_d/d = \rho$. Then, for all $t, h \geq 0$,

$$M_h^d(B,t) \xrightarrow{w} m_h(B,t).$$

Long-run Equilibrium. Note from (14) that for $h \leq t$ we can express $m_h(B,t)$ in terms of $m_0(B,t)$,

$$m_h\big(x,\, t\big) = m_0\left(\min_{j=1,\ldots,h}\left\{\frac{x}{\varphi^h},\, \frac{\ell_{h-j}(t-j)}{\varphi^{h-j}}\right\},\, t-h\right).$$

Then the fixed-point equation corresponding to (14) is given by

$$m_0^*(x) = \sum_{h=0}^{\infty} \int_{\ell_h^*}^{\infty} \Phi_{z,\nu^{(h)}}\left(\frac{x}{\varphi}\right) m_h^*(dz)$$
$$= \sum_{h=0}^{\infty} \int_{\frac{\ell_h^*}{\varphi^h}}^{\min_j \frac{\ell_{h-j}^*}{\varphi^{h-j}}} \Phi_{\varphi^h z,\nu^{(h)}}\left(\frac{x}{\varphi^h}\right) m_0^*(dz)$$

where $j = 1, \ldots, h$. Here, $\ell_h^* = \ell^* - \theta\nu^{(h)}$ where the steady state ℓ^* is defined by

$$\ell^* = \sup\left\{\ell \,\Big|\, \sum_{h=0}^{\infty} \widetilde{m}_h^*\Big([\ell,\infty)\Big) = \rho\right\}, \tag{18}$$

and $\widetilde{m}^*$ again denotes the measure on indices, cf. (16).

The above system of equations describes possible equilibrium points of the measure valued dynamical system. It is intricate due to the coupling of ℓ^*, $\widetilde{m}^*$ and the measures m_h, $h \geq 0$. Nevertheless, the system is elegant in that its solution can potentially be described through a single measure, namely m_0^*.

For the special case of $\theta = 0$ (myopic) we verified numerically that with arbitrary initial choice $\{m_h(\cdot, 0)\}$ satisfying (17) an equilibrium point satisfying (18) is indeed attracting. Furthermore, when d and t are large enough, the proportion of arms associated with a certain belief state in a simulated system with d arms is indeed fixed and well approximated by the solution to (18) when operated under the myopic policy.

4.2 The Equilibrium Index Process

We now relate the system with many arms operated under π_θ to a special *one-armed process with threshold*. For this process the arm is activated whenever the index exceeds a specified threshold ℓ, i.e. $a(t) = \mathbb{1}\{\mu(t) + \theta\nu(t) \geq \ell\}$. Because the evolution of the belief state and thus the evolution of the index depends on ℓ, we denote the associated stochastic process of indices by $\Gamma^\ell(t) := \mu(t) + \theta\nu(t)$.

Suppose that ℓ is picked in such a way that we activate with probability ρ; denote it by $\overline{\ell}$. Then a policy $\pi_\theta^{\overline{\ell}}$ that chooses action $a_i(t) = \mathbb{1}\left\{\mu_i(t) + \theta\nu_i(t) \geq \overline{\ell}\right\}$ for every arm i in an unconstrained system with d arms is a policy which activates $\rho\, d$ arms *on average* (this is essentially the idea behind Whittle's relaxation [16]). Thus, as $d \to \infty$, the policy $\pi_\theta^{\overline{\ell}}$ activates approximately a proportion ρ of arms at every decision time.

We believe that in steady state (as $t \to \infty$ or under stationarity) the equilibrium of the measure-valued dynamical system is directly related to the one-armed process with this particular threshold $\overline{\ell}$, and further $\overline{\ell}$ equals ℓ^* of (18).

Conjecture 2. *Assume that the index is parametric, and that $\Gamma^\ell(t)$ is stationary. Then the equation*

$$\mathbb{P}\left(\Gamma^\ell(t) \geq \ell\right) = \rho \tag{19}$$

has a unique solution ℓ^*, which satisfies Eqn. (18), and

$$\mathbb{P}\left(\Gamma^{\ell^*}(t) \in B\right) = \sum_{h=0}^{\infty} \widetilde{m}_h^*(B),\ \forall\, B \in \mathcal{B}(\mathbb{R}).$$

A practical implication of Conj. 1 and 2 combined is that in the limit, as $d \to \infty$ and $t \to \infty$, a parametric index policy π_θ is equivalent to the policy that activates arm i in an unconstrained system whenever $\Gamma_i(t) \geq \ell^*$, where ℓ^* is defined by (19). This motivates the following simple algorithm for performance evaluation.

Algorithm Performance evaluation.

1: For large T determine $\widehat{\ell}^*$ such that $T^{-1}\sum_{t=0}^{T} a_i(t) = \rho$ is achieved for a policy $\pi_\theta^{\widehat{\ell}^*}$.
2: Use the sample path of Step 1 to obtain an estimate $\overline{G}$ for the expected average reward of the one-armed system.
3: Output $\overline{G}_d := d\,\overline{G}$ as an approximation of the expected average reward of the multiarmed system with d arms operated under π_θ.

The virtue of this algorithm is that the behavior of the many-armed system is approximated by simulating a much simpler one-armed problem.

4.3 Optimized Parametric Index

The algorithm can be used to approximate the best parameter values for a parametrized index policy. We approximate $\theta^* := \arg\max_\theta G_d(\theta)$ by $\overline{\theta}^* := \arg\max_\theta \overline{G}(\theta)$, where $G_d(\theta)$ is the average reward obtained under π_θ for the problem with d arms, and $\overline{G}(\theta)$ is the estimator for $G(\theta)$ as obtained from Step 2 of the algorithm. Fig. 3 depicts the estimated expected average reward $\overline{G}(\theta)$ as a function of θ. The figure suggests that for large φ, the myopic policy (which corresponds to $\theta = 0$) can be improved significantly.

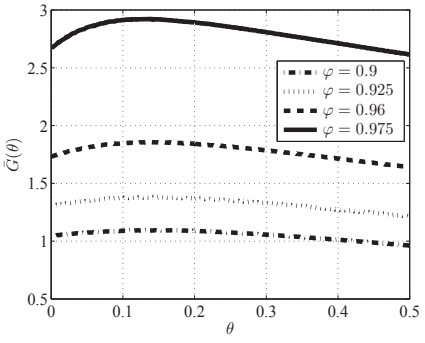


Figure 3: Expected average reward $\overline{G}(\theta)$ computed by the algorithm as a function of θ. $\sigma = 2$, $\varphi \in \{0.9, 0.925, 0.95, 0.975\}$, $\rho = 0.4$, $T = 2 \times 10^6$.

We now examine the performance of π_θ when the parameter is chosen to be $\overline{\theta}^*$. In contrast to the approximation $\overline{G}_d(\theta)$ that is obtained from the algorithm, we denote the estimated average reward obtained by Monte Carlo simulation of the d-armed system by $\widehat{G}_d(\theta)$. We define $\widehat{\theta}_d^* := \arg\max_\theta \widehat{G}_d(\theta)$. Accordingly, $\widehat{G}_d\left(\overline{\theta}^*\right)$ and $\widehat{G}_d\left(\widehat{\theta}_d^*\right)$ are the average rewards

obtained when simulating the system under π_θ, where θ is chosen as $\overline{\theta}^*$ and $\widehat{\theta}_d^*$ respectively. In Fig. 4 we compare these quantities to the average rewards obtained when simulating the system under the Whittle index and the myopic policy. Unsurprisingly, the Whittle index policy outperforms the other index policies – in fact, we believe it to be asymptotically optimal. However, the parametrized index does considerably better than the myopic.

Importantly, we note from Fig. 4 that $\widehat{\theta}_d^*$ is indeed well approximated by $\overline{\theta}^*$. Thus, instead of optimizing the parameter by simulating the multidimensional d-armed system, we can approximate the best θ-value directly from the one-armed process with threshold for any value of d (such that $k_d = \lfloor \rho\, d \rfloor$).

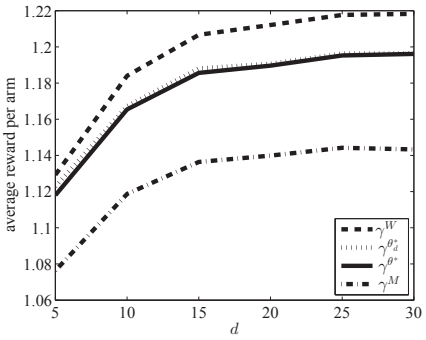


Figure 4: Comparison of average rewards achieved per arm under the Whittle, the parametric index (9) and the myopic policy. The parameter θ is found by optimizing (i) the problem with d arms (dotted), and (ii) the one-armed problem. $\varphi = 0.9$, $\sigma = 2$, $\rho = 0.4$, $T = 100,000$.

5. CONCLUDING REMARKS

This paper provides a starting point for a rigorous investigation of the structural properties and performance of index policies in partially observable restless bandit problems with AR(1) arms. This incorporates (i) the analysis of the Whittle index as a likely candidate for an asymptotically optimal policy as $d \to \infty$ while $k_d/d \to \rho$, and (ii) insights into the behavior of the system in this asymptotic regime. In addition to our conjectures above, we also believe that some form of asymptotic independence holds for the index processes as the number of arms grows large. In this context we mention that Γ_i, $i = 1, \ldots, d$, are exchangeable [2]. This may yield a path for proving asymptotic independence. The recursions on measures defining the limiting dynamical system can perhaps be treated along the lines of [10].

Furthermore, many of the ideas in this paper can be generalized. For example, the results obtained in Section 3.1 for discounted rewards similarly hold in the average reward case, and the assumptions made in that section generally hold for the problem we consider. Beyond that, we can extend the treatment to AR processes of higher order, heterogeneous arms and bandit problems with correlated arms.

6. ACKNOWLEDGMENTS

This project is supported by the Australian Research Council grant DP130100156.

7. REFERENCES

[1] R. Aguero, M. Garcia, and L. Munoz. Bear: A bursty error auto-regressive model for indoor wireless environments. In *PIMRC*, pages 1–5, 2007.

[2] D. J. Aldous. *Exchangeability and related topics*. Springer, 1985.

[3] K. Avrachenkov, L. Cottatellucci, and L. Maggi. Slow fading channel selection: A restless multi-armed bandit formulation. In *ISWCS*, pages 1083–1087, 2012.

[4] N. Bäuerle and U. Rieder. *Markov Decision Processes with Applications to Finance*. Springer, 2011.

[5] F. Cecchi and P. Jacko. Scheduling of users with markovian time-varying transmission rates. In *ACM SIGMETRICS*, pages 129–140, 2013.

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

[7] K. Glazebrook, D. Hodge, C. Kirkbride, and R. Minty. Stochastic scheduling: A short history of index policies and new approaches to index generation for dynamic resource allocation. *J. of Scheduling*, pages 1–19, 2013.

[8] K. Liu, R. Weber, and Q. Zhao. Indexability and whittle index for restless bandit problems involving reset processes. In *CDC-ECC*, pages 7690–7696, 2011.

[9] K. Liu and Q. Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *Trans. on Info. Theory*, 56:5547–5567, 2010.

[10] R. McVinish and P. Pollett. The limiting behaviour of a stochastic patch occupancy model. *J. Math. Biol.*, 67:693–716, 2013.

[11] J. Niño-Mora. A restless bandit marginal productivity index for opportunistic spectrum access with sensing errors. In *Network Control and Optimization*, pages 60–74. Springer, 2009.

[12] C. Papadimitriou and J. Tsitsiklis. The complexity of optimal queuing network control. *Math. of Oper. Res.*, 24:293–305, 1999.

[13] M. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. Wiley, 1994.

[14] I. Verloop. Asymptotic optimal control of multi-class restless bandits. *CNRS Technical Report, hal-00743781*, 2014.

[15] R. Weber and G. Weiss. On an index policy for restless bandits. *J. of Appl. Prob.*, 27:637–648, 1990.

[16] P. Whittle. Restless bandits: Activity allocation in a changing world. *J. of Appl. Prob.*, 25:287–298, 1988.

APPENDIX

Proof of Lemma 2.1. For each arm i we have

$$\begin{aligned}\mathbb{E}_{\mu_i,\nu_i}|X_i(t)| \leq & \mathbb{E}_{\mu_i,\nu_i}\left[\varphi^t\,|X_i(0)| + \sum_{j=0}^{t-1}\varphi_i^j\,|\varepsilon_i(t-j)|\right] \\ \leq & \sqrt{\frac{2\nu_i}{\pi}} + |\mu_i| + \sqrt{\frac{2}{\pi}}\frac{1}{1-\varphi} =: B(\mu_i,\nu_i),\end{aligned}$$

which is finite, and thus,

$$\sup_{\pi} \sum_{t=0}^{\infty} \beta^t \mathbb{E}^{\pi}_{\mu_i,\nu_i} \left[\sum_{i=1}^{d} |X_i(t) a_i(t)| \right] \leq \frac{d \max_i B(\mu_i, \nu_i)}{1-\beta} < \infty,$$

whence $\sup_\pi V^\pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ is finite, and $\sum_{t=0}^{\infty} \beta^t \sum_{i=1}^{d} |X_i(t) a_i(t)|$ converges almost surely to a finite limit. The variables $Z_T := \sum_{t=0}^{T} \beta^t \sum_{i=1}^{d} X_i(t) a_i(t)$ thus converge almost surely as $T \to \infty$ and are dominated by the absolute sum which has finite mean. Hence, by dominated convergence $\mathbb{E}^{\pi}_{\boldsymbol{\mu},\boldsymbol{\nu}} Z_T \to \mathbb{E}^{\pi}_{\boldsymbol{\mu},\boldsymbol{\nu}} \left[\sum_{t=0}^{\infty} \beta^t \sum_{i=1}^{d} X_i(t) a_i(t) \right]$. □

Proof of Proposition 3.2. Define $b(\mu) := 1+\mu^2$. Then, because $|\max\{m, \mu\}| \leq \big(1 + |m|\big) b(\mu)$, the absolute value of the expected immediate reward under both actions (passive, active) is bounded by $\big(1 + |m|\big) b(\mu)$ for any belief state in Ψ. Furthermore,

$$\int_{-\infty}^{\infty} b(\varphi y) \phi_{\mu,\nu}(y) dy \leq \big(1 + \varphi^2 \nu_{\max}\big)\, b(\mu)$$

and hence, b is an upper bounding function in the sense of [4, Def. 7.1.2]. The implication of this is that the space $\mathcal{V}$ of measurable functions $v : \Psi \to \mathbb{R}$ with finite weighted supremum norm defined by

$$\|v\|_b := \sup_{\mu,\nu} \frac{|v(\mu,\nu)|}{b(\mu)} < \infty$$

contains the optimal value function V. We apply [4, Thm. 7.2.1]. To verify the main condition of the latter, define the operator Q by

$$Qv(\mu,\nu) := \max \left\{ \beta v(\varphi\mu, \varphi^2\nu + \sigma^2), \beta \int_{-\infty}^{\infty} v(\varphi y, \sigma^2) \phi_{\mu,\nu}(y) dy \right\}.$$

Take $b(\mu,\nu) = b(\mu)$ as defined above and observe that

$$Q^n b(\mu,\nu) \leq \beta^n \Big(1 + \varphi^2 \nu_{\max} + \varphi^{2n} \big(\nu + \mu^2 \big) \Big),$$

whence $Q^n b \to 0$ as $n \to \infty$. Noting that the further regularity conditions of [4, Thm. 7.2.1] are satisfied, we obtain that with initial choice $V_0 \equiv 0$ the value iteration converges to an optimal value function, and an optimal policy exists; namely, it is optimal to take the action that maximizes the right-hand side of the Bellman equation. The uniqueness of the value function can be seen as follows. Let v and w be two fixed points of T. Then

$$T^n v(\mu,\nu) = T\big(T^{n-1} v(\mu,\nu)\big) = \cdots = v(\mu,\nu),$$

for every $n \in \mathbb{N}$. Because $Q^n \to 0$ we know that for every (μ,ν) there exists n_μ such that

$$\frac{|T^{n_\mu} v(\mu,\nu) - T^{n_\mu} w(\mu,\nu)|}{b(\mu,\nu)} \leq \alpha \sup_{\mu,\nu} \frac{|v(\mu,\nu) - w(\mu,\nu)|}{b(\mu,\nu)}$$

for some $\alpha \in (0,1)$. Hence,

$$\|v - w\|_b = \sup_{\mu,\nu} \frac{|T^{n_\mu} v(\mu,\nu) - T^{n_\mu} w(\mu,\nu)|}{b(\mu,\nu)} \leq \alpha \|v - w\|_b,$$

which implies that $v \equiv w$. □

Proof of Lemma 3.3. The proof is by induction on (8), and consists of three parts. Part (a) refers to the convexity assertion, which implies continuity. In Part (b) we prove the monotonicity properties of $V^m(\cdot,\nu)$ for fixed ν, whereas in Part (c) we show monotonicity of $V^m(\mu,\cdot)$ with μ fixed.

(a) Suprema, expectations, compositions of convex and increasing functions as well as linear combinations with non-negative weights of convex functions are convex. Then the result follows from (8) by induction.

(b) For $V_0^m \equiv 0$, we have that $V_1^m(\cdot,\nu)$ is non-decreasing and thus we may assume that $V_n^m(\cdot,\nu)$ is non-decreasing for some n. If $\mu_1 \leq \mu_2$, then by the stochastic ordering of $Y_{\mu_i,\nu} \sim \mathcal{N}(\mu_i, \nu)$ it holds that $\mathbb{E}\left[V_n^m\left(\varphi\, Y_{\mu_1,\nu},\, \sigma^2\right)\right] \leq \mathbb{E}\left[V_n^m\left(\varphi\, Y_{\mu_2,\nu},\, \sigma^2\right)\right]$. It follows by induction that V_n^m is non-decreasing in μ for all $n \in \mathbb{N}$, and thus their limit V^m is non-decreasing in μ. Furthermore, since a lower bound for V^m is given by the value obtained when always playing active, $\mu/(1-\varphi\beta)$, which is strictly increasing in μ, it is evident that V^m cannot be constant in μ.

(c) Let $V_0^m \equiv 0$. Then $V_1^m(\mu,\cdot)$ is constant and thus non-decreasing. Assume that $V_n^m(\mu,\cdot)$ is non-decreasing. We prove below that $\mathbb{E}\left[V_n^m\left(\varphi\, Y_{\mu,\nu},\, \sigma^2\right)\right]$ is non-decreasing in ν. Then it follows by induction that V_n^m is non-decreasing in ν for every n, and thus, so is V^m.

For brevity we assume that $V^m(\cdot,\nu)$ is differentiable. Define $g(y) := V_n^m(\varphi\sqrt{\nu} y + \mu, \sigma^2)$; this is increasing and convex as it is a composition of a convex and a monotone increasing function. Applying Jensen's inequality we obtain

$$\begin{aligned}
&\frac{\partial}{\partial \nu} \mathbb{E}\left[V_n^m\left(\varphi\, Y_{\mu,\nu},\, \sigma^2\right)\right] \\
&\quad = \frac{\varphi}{2\sqrt{\nu}} \int_{-\infty}^{\infty} y\, V_n^{m\prime}\left(\varphi\sqrt{\nu} y + \mu,\, \sigma^2\right) \phi_{0,1}(y)\, dy \\
&\quad = \frac{\varphi}{2\sqrt{\nu}} \int_{-\infty}^{\infty} y\, g'(y)\, \phi_{0,1}(y)\, dy \\
&\quad \geq \frac{\varphi}{2\sqrt{\nu}} \int_{-\infty}^{\infty} \big(g(y) - g(0)\big)\, \phi_{0,1}(y)\, dy \geq 0
\end{aligned}$$

because g is convex (and thus $g(b) - g(a) \leq g'(b)(b-a)$ for all a, b in $\mathbb{R}$, the domain of g). □

Proof of Proposition 3.4. $V^m(\cdot,\nu)$ is non-decreasing and convex by Lemma 3.3, and thus, the same holds for $T_a V^m$, $a = 0, 1$. By Assumption 1 both passive and active set are non-empty, whence $T_0 V^m(\cdot,\nu)$ and $T_1 V^m(\cdot,\nu)$ intersect. Since both functions are convex and increasing, they can intersect at most twice (if their paths are equal on a connected set of points, we refer to this as a single intersection). But since by assumption $T_0 V^m(\mu,\nu) \geq T_1 V^m(\mu,\nu)$ for all $\mu < \underline{\mu}$ whereas $T_1 V^m(\mu,\nu) > T_0 V^m(\mu,\nu)$ for all $\mu > \overline{\mu}$, they can only intersect exactly once, whence there is a unique switching point, which we denote by $\zeta_m(\nu)$. This argument applies for every $\nu \in \Psi_2$. □

Proof of Corollary 3.5. Let $\mu_1 \leq \mu_2$, then Prop. 3.4 implies $(\mu_1,\nu) \in \mathcal{P}_{\omega(\mu_2,\nu)}$. Hence, $\omega(\mu_2,\nu) \in \{m \,|\, (\mu_1,\nu) \in \mathcal{P}_m\}$ and thus, $\omega(\mu_1,\nu) \leq \omega(\mu_2,\nu)$ by definition of the Whittle index as an infimum. The monotonicity of $\omega(\mu,\cdot)$ follows along similar lines using Assumption 2. □