# Asymptotic Optimal Control of Markov-Modulated Restless Bandits

Santiago Duran, Ina Verloop

# Asymptotic Optimal Control of Markov-Modulated Restless Bandits

Santiago Duran[1,2] and Ina Maria Verloop[3,4]

[1]CNRS, LAAS, 7 Avenue du Colonel Roche, Toulouse, France
[2]Univ. de Toulouse, LAAS, Toulouse, France
[3]CNRS, IRIT, 2 rue C. Camichel, Toulouse, France
[4]Univ. de Toulouse, INP, Toulouse, France

**Abstract**

This paper studies optimal control subject to changing conditions. This is an area that recently received a lot of attention as it arises in numerous situations in practice. Some applications being cloud computing systems where the arrival rates of new jobs fluctuate over time, or the time-varying capacity as encountered in power-aware systems or wireless downlink channels. To study this, we focus on a restless bandit model, which has proved to be a powerful stochastic optimization framework to model scheduling of activities. In particular, it has been extensively applied in the context of optimal control of computing systems. This paper is a first step to its optimal control when restless bandits are subject to changing conditions, the latter being modeled by Markov-modulated environments.

We consider the restless bandit problem in an asymptotic regime, which is obtained by letting the population of bandits grow large, and letting the environment change relatively fast. We present sufficient conditions for a policy to be asymptotically optimal and show that a set of priority policies satisfies these. Under an indexability assumption, an averaged version of Whittle's index policy is proved to be inside this set of asymptotic optimal policies. The performance of the averaged Whittle's index policy is numerically evaluated for a multi-class scheduling problem in a wireless downlink subject to changing conditions. While keeping the number of bandits constant, we observe that the average Whittle index policy becomes close to optimal as the speed of the modulated environment increases.

## 1 Introduction

Optimal control subject to changing conditions is an area that recently received a lot of attention as it arises in numerous situations in practice. For example, the arrival rate of new jobs in call centers, hospital ERs or cloud computing systems vary strongly over time. Another example is systems with time-varying capacity as encountered in power-aware systems where the speed is a controllable parameter [1] or in a wireless downlink channel where quality of the downlink channel is influenced by fading and interference effects [13].
In general, finding optimal solutions for stochastic scheduling problems in a changing environment is notoriously difficult and the results obtained are scarce. In this paper we will restrict our attention to optimal control in a restless bandit problem as this provides a powerful optimization framework to model dynamic scheduling of activities. In particular, regarding optimal control of computing systems, the restless bandit framework has been successfully applied in for example the context of wireless downlink scheduling [3, 7, 33], load balancing problems [26], systems with delayed state observation [19], broadcast systems [35], multi-channel access models [2, 27], stochastic scheduling problems [4, 23, 30] and scheduling in the presence of impatient customers [9, 22, 25, 31].

The restless bandit problem is concerned with the optimal dynamic activation of several competing bandits. Each bandit is a controllable stochastic process whose state evolution depends on whether or not the bandit is made active. The aim is to find a control that determines at each decision epoch which bandits to activate in order to minimize the cost over time associated to the states the bandits are in. In the by now classical multi-armed bandit model, [21], it is assumed that only active bandits can change state. In [39], Whittle introduced the so-called restless bandits, where a bandit can also change its state while being passive (that is, not active), possibly according to a different law from the one that applies when it is active. The restless bandit model gained popularity due to its multiple applications in, real-life examples.

The restless bandit problem is an important subclass of Markov decision problems, which, in general, are hard to solve [21, 28, 40]. Obtaining optimal solutions for Markov decision problems are typically out of reach for dimensions higher than two, the so-called "curse of dimensionality". In fact, finding an optimal solution for restless bandit problems is PSPACE-hard, hence infeasible. The elegance of the class of restless bandit problems lies in the fact that a powerful approximation framework exists, allowing to go beyond the dimension of two. This technique, as proposed by Whittle [39], consists of solving a relaxed version of the optimization problem where the sample-path constraint on the maximum number of active bandits is relaxed to its time-average version. This simplifies notably the problem as it allows to reduce the multi-dimensional control problem to several one-dimensional control problems. An optimal solution of the relaxed problem is described by index values, where each bandit is associated an index that only depends on its own state and its own transition rates. This provides a policy for the original problem, the so-called Whittle index policy that activates at each moment in time the bandits having currently the highest index values.

In this paper, we make a first step in the optimal control of the restless bandit problem *living in a changing environment*. Each bandit is associated a Markov modulated environment, whose state influences the transition rates of the bandits. No assumption is made on the correlation between the different random environments of bandits. In the case of independent distributed environments, this allows to model the effect of fluctuating parameters, for example, the arrival rate of new jobs as one may encounter in load balancing systems, or the abandonment rate of impatient customers. On the other hand, when environments are strongly correlated, one could model dependence between different bandits through "environmental effects" as one encounters for example in wireless downlink scheduling.

Our main focus will be on a rapidly varying environment, which allows us to find asymptotically optimal policies for Markov-modulated restless bandits. In addition, we assume one cannot observe the environment. Note that in practice one could use Bayesian analysis to infer the environment from the events that happened. However, as the environments vary relatively fast, it might be too costly to learn the environment and/or to change action each time the environment changes. We therefore focus on policies that are not trying to learn the state of the environments.

We will propose an index policy and prove it to be optimal in the asymptotic regime as the number of bandits grows large and the environment changes relatively fast. This regime is motivated by the seminal work in [38], where for the standard restless bandit problem, Whittle's index policy was shown to be optimal as the number of bandits that can be made active grows proportionally to the total number of bandits. Recently, in [24] approaches of [38] were set forth and extended to problems for which a bandit can have multiple activation levels. In [37] a different proof technique is used to include models with possible new arrivals of bandits, and asymptotic optimality is proved for a set of priority policies. Another recent result on asymptotic optimality is [33] where the authors considered a specific model representing downlink scheduling in wireless systems.

When *fixing* the policy, our model can be seen as a particle system living in a Markov modulated environment. Since the speed of the background processes will scale proportionally with the number of bandits, we can use convergence results as obtained in [10, 12] for particle systems living in a rapidly varying environment. In particular, they derived that in the limit the system is described by an ODE where the transitions rates of the bandits are averaged according to the steady-state

distribution of the modulated environments. The paper [12] considers a countable state space and each particle is associated its own modulated environment (no assumption is made on the joint evolution of the environments). On the other hand, in [10] a finite state space and *one common* environment is considered, resulting in less complex technical conditions.
The *novelty of our work* is the addition of the Markov modulated environment to the restless bandit model.

- In our *first main contribution*, we prove a set of priority policies to be asymptotically optimal.
- In our *second main contribution*, we introduce the averaged Whittle index policy and prove it to be inside the set of asymptotically optimal policies.
- The *numerical evaluation* of the averaged Whittle index policy shows optimal performance as the speed of the environment is fast enough. The numerical example considered is that of a multi-class scheduling problem in a wireless downlink. In particular, this example shows that the averaged Whittle index policy performs close to optimal even though the number of bandits remains constant.

The remainder of this paper is organized as follows. In Section 2 we describe related work on optimal scheduling in a changing environment. In Section 3, we define the multi-class restless bandit control problem and introduce the Markov modulated environments. Section 4 contains the asymptotic optimality results. In Section 5, we define a set of priority policies and prove them to be asymptotically optimal when the state space is finite. In Section 6 we define an averaged version of Whittle's index policy and prove it to be asymptotically optimal. Section 7 presents our numerical results.

## 2 Related work

In this section we describe several works related to the optimal control of stochastic scheduling problems living in a changing environment.
In the case of a(n) (partially) *unobservable* modulating environment, optimal stochastic control can be solved using Bayesian dynamic programming. The latter involves continuity of the state space and as such there is a lack of tractable solution methodologies [34]. For results regarding specific models, we mention here for example [5, 6, 29]. In [5] optimal load balancing is studied when the queue lengths of the servers are unobservable. A set of round-robin policies is proved to be optimal in a heavy-traffic many-server limiting regime. Optimal control in a multi-class queueing system where the server performance varies according to a Markov modulated random environment is studied in [6, 29]. The environment can be *partially observed* for the activated server. As the control decision influences the observation made, the authors search for policies that achieve maximum stability.
In the case of an *observable* environment, we refer to [7, 8, 17] where efficient controls are derived for a single server with an *observable* time-varying capacity. For limiting regimes, optimal policies in the *observable* environment setting have been obtained in [15] and [20]. In [20], a general particle system with one common underlying observable environment is studied. The authors show that the optimal cost and optimal policy converges, as the number of particles grow, to those of a discrete-time *deterministic* system with observable environment. The authors of [15] study a multi-class scheduling problem with one common observable environment in a heavily-loaded regime. For modulated arrival rates it is shown that the $c\mu$-rule (which is optimal in a standard multi-class queue [16]) is asymptotically optimal for fast changing environment, fixed environment, and slowly changing environment. In the case of modulated service rates, an averaged version of the $c\mu$-rule is shown to be asymptotically optimal only in the case of a fast changing environment.
In the context of learning bandits, we like to mention the work [36], where optimal control in a changing environment is studied. Here, the laws according to which the bandits evolve and receive

reward/cost, are *not known* to the decision maker [14]. In [36] it is assumed that the (unknown) laws depend on the modulated environment. The aim is to find an algorithm that finds the right trade-off between exploitation and exploration in order to converge to *the best* bandit, i.e., minimize the regret (the cost due to the fact that a globally optimal policy is not followed at all time). This as opposed to the restless bandit setting in the current paper, where the laws according to which the state of the bandits changes are known and the aim is to find a *dynamic* control that achieves close to optimal performance.

# 3 Model description

We consider a multi-class restless bandit problem in continuous time. There are $K$ classes of bandits and there are $N_k$ class-$k$ bandits present in the system. We further define $N := \sum_k N_k$ as the total number of bandits and define $\gamma_k := N_k/N$ as the fraction of class-$k$ bandits. At any moment in time, a class-$k$ bandit is in a certain state $j \in \{1, 2, \ldots, J_k\}$, with $J_k \leq \infty$. In particular, the state space can be countable infinite.

At any moment in time, a bandit can either be kept passive or active, denoted by $a = 0$ and $a = 1$, respectively. There is the restriction that at most $\alpha N$ bandits can be made active at a time, $\alpha \leq 1$. The transitions of the class-$k$ bandits depend on a background process described by the Markov process $D_k(t)$ that lives in a countable state space $\mathcal{Z} = \{1, \ldots, d, \ldots\}$ and is positive recurrent. We make no further assumptions on the distribution of the joint vector $\vec{D} = (D_1, \ldots, D_K)$. For example, it could be that there is one common environment for all classes of bandits, or instead, the environments per class are independently distributed. We further let $\phi(\vec{d})$ denote the stationary probability vector that the environment vector $\vec{D}$ is in state $\vec{d}$. We let $\phi_k(d)$ denote the marginal probability of environment $D_k$ to be in state $d$. We further assume that $\sum_{\vec{d}'} r(\vec{d}'|\vec{d}) < C_1$, for all $\vec{d}$, for some $C_1 < \infty$, with $r(\vec{d}'|\vec{d})$ the transition rate of $\vec{D}(t)$ from $\vec{d}$ to $\vec{d}'$.

When action $a$ is performed on a class-$k$ bandit in state $i$, $i = 1, \ldots, J_k$, and the environment of this class-$k$ bandit is in state $d$, it makes a transition to state $j$ with rate $\frac{1}{N} q_k^{(d)}(j|i,a)$, $j = 1, \ldots, J_k$, $j \neq i$. The scaling $1/N$ makes sure that the evolution of the state of a bandit is relatively slow compared to that of its environment, i.e., the environment changes relatively fast. We assume that the evolution of one bandit (given its action and the state of its environment) is independent of that of all the other bandits. We further define the averaged transition rate by $\overline{q}_k(j|i,a) := \sum_{d \in \mathcal{Z}} \phi_k(d) q_k^{(d)}(j|i,a)$. The fact that the state of a bandit might evolve even under the passive action explains the term of a *restless* bandit. Throughout the paper, we assume that the transition rates are uniformly bounded, i.e.,

$$\sum_{j=1}^{J_k} q_k^{(d)}(j|i,a) < C_2, \text{ for all } a, d, i, k, \tag{1}$$

for some $C_2 < \infty$.

A *policy* determines at each *decision epoch* which $\alpha N$ bandits are made active. Decision epochs are moments when one of the $N$ bandits changes state. We focus on Markovian policies that base their decisions only on the current proportion of bandits present in the different states. Hence, this means that the decision maker cannot observe the state of the background process $\vec{D}(t)$. In addition, we assume the decision maker does not attempt to learn the state of the environment either. We write $\vec{x} := (x_{j,k}; k = 1, \ldots, K, j = 1, \ldots, J_k)$, where $x_{j,k}$ represents the proportion of class-$k$ bandits that are in state $j$, hence,

$$\vec{x} \in \mathcal{B} := \left\{ \vec{x} : 0 \leq x_{j,k} \leq 1 \ \ \forall j,k \text{ and } \sum_j x_{j,k} = \gamma_k \right\}.$$

Given policy $\pi$, we then define the function $y^{\pi,1} : \mathcal{B} \to [0,1]^{\sum_{k=1}^{K} J_k}$ that distinguishes the action chosen for the bandits. That is, given a policy $\pi$, $y^{\pi,1}_{j,k}(\vec{x})$ denotes the proportion of class-$k$ bandits in state $j$ that are activated when the proportion of bandits in each state is given by $\vec{x}$. Hence, $y^{\pi,1}(\cdot)$ satisfies $y^{\pi,1}_{j,k}(\vec{x}) \leq x_{j,k}$ and $y^{\pi,1}_{j,k}(\vec{x}) \leq \alpha, \quad \forall j,k$. We focus on the set of policies such that $y^{\pi,1}(\cdot)$ is continuous. We further define $y^{\pi,0}_{j,k}(\vec{x}) := x_{j,k} - y^{\pi,1}_{j,k}(\vec{x})$, as the proportion of class-$k$ bandits in state $j$ that are kept passive.
For a given policy $\pi$, we define $\vec{X}^{N,\pi}(t) := (X^{N,\pi}_{j,k}(t); k = 1, \ldots, K, j = 1, \ldots, J_k)$, with $X^{N,\pi}_{j,k}(t)$ the number of class-$k$ bandits that are in state $j$ at time $t$.
Our performance criteria are stability and long-run average holding cost. For a given policy $\pi$, we will call the system *stable* if the process $\vec{X}^{N,\pi}(t)$ has a unique invariant probability distribution. We will denote by $\vec{X}^{N,\pi}$ and $X^{N,\pi}_{j,k}$ the random variables following the steady state distributions, assuming they exist. In case the state space is finite, the process $\vec{X}^{N,\pi}(t)$ being unichain is a sufficient condition for stability of the policy $\pi$. For infinite state space, whether or not the system is stable can depend strongly on the employed policy. We will only be interested in the set of stable policies.
We denote by $C^{(d)}_k(j,a) \in \mathbb{R}$, $j = 1, \ldots, J_k$, the holding cost per unit of time for having a class-$k$ customer in state $j$ under action $a$ and when in environment $d$. We note that $C^{(d)}_k(j,a)$ can be negative, i.e., representing a reward. We define the holding cost averaged over the states of the environment as $\overline{C}_k(j,a) := \sum_{d \in \mathcal{Z}} \phi_k(d) C^{(d)}_k(j,a)$. We further introduce the following value functions for given policy $\pi$, and initial proportion of bandits $\dfrac{\vec{X}^{N,\pi}(0)}{N} = \vec{x} \in \mathcal{B}$:

$$V^{N,\pi}_{-}(\vec{x}) := \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}\left( \int_0^T \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} C^{(D_k(t))}_k(j,a) y^{\pi,a}_{j,k}\left( \frac{\vec{X}^{N,\pi}(t)}{N} \right) \mathrm{d}t \right)$$

and

$$V^{N,\pi}_{+}(\vec{x}) := \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\left( \int_0^T \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} C^{(D_k(t))}_k(j,a) y^{\pi,a}_{j,k}\left( \frac{\vec{X}^{N,\pi}(t)}{N} \right) \mathrm{d}t \right). \tag{2}$$

If $V^{N,\pi}_{-}(\vec{x}) = V^{N,\pi}_{+}(\vec{x})$ for all $\vec{x}$, then we define $V^{N,\pi}(\vec{x}) := V^{N,\pi}_{+}(\vec{x})$. We assume there exists a stable policy for which $V^{N,\pi}(\vec{x}) < \infty$. Our objective is to find a policy $\pi^*$ that is average optimal,

$$V^{N,\pi^*}_{+}(\vec{x}) \leq V^{N,\pi}_{-}(\vec{x}), \quad \text{for all } \vec{x} \text{ and for all policies } \pi, \tag{3}$$

under the constraint that at any moment in time at most $\alpha N$ bandits can be made active, that is,

$$\sum_{k=1}^{K} \sum_{j=1}^{J_k} y^{\pi,1}_{j,k}\left( \frac{\vec{X}^{N,\pi}(t)}{N} \right) \leq \alpha, \text{ for all } \ t. \tag{4}$$

# 4 Rapidly varying modulated environments

In this section we study the process as the background process is fast changing and the number of bandits scales. In Section 4.1 we prove asymptotic independence between the proportion of bandits in each state and the environment. In Section 4.2 we establish convergence to a fluid limit and use this to lower bound the performance for any policy. This lower bound allows to prove asymptotic optimality of certain policies, which is presented in Section 4.3.

## 4.1 Asymptotic independence

In the lemma below we prove that the bandits are asymptotically independent of the environment, i.e., when the amount of bandits $N$ tends to infinity. The proof can be found in the Appendix.

**Lemma 4.1** *Assume $\pi$ is a stable policy for any $N$. Then, for any subsequence of $N$ such that $\left(\vec{X}^{N,\pi}/N\right)_{N\in\mathbb{N}}$ converges in distribution, we have*

$$\lim_{N\to\infty} \mathbb{E}\left(\mathrm{e}^{-s_{11}\frac{X_{11}^{N,\pi}}{N}}\dots\mathrm{e}^{-s_{J_KK}\frac{X_{J_KK}^{N,\pi}}{N}}\mathbf{1}_{(\vec{D}=\vec{d})}\right) = \phi(\vec{d})\lim_{N\to\infty}\mathbb{E}\left(\mathrm{e}^{-s_{11}\frac{X_{11}^{N,\pi}}{N}}\dots\mathrm{e}^{-s_{J_KK}\frac{X_{J_KK}^{N,\pi}}{N}}\right). \quad (5)$$

## 4.2 Fluid control problem and lower bound

In this section we study the behavior of the system as $N$ grows large, that is, as the number of bandits grows large and the environments vary rapidly. We state convergence to the fluid limit and derive an associated fluid control problem. The latter allows to derive a lower bound on the average holding cost for any policy $\pi$.

Before presenting the fluid process to which the stochastic system converges, we first provide some intuition. Recall that the transition rate of a bandit in state $j$ to state $i$, when action $a$ is performed, is given by $\frac{1}{N}q_k^{(d)}(i|j,a)$. Since the rates of the background process do not scale with $N$, when we take $N\to\infty$, the bandit will perceive a rapidly changing environment. Before it can make a new transition, its environment has already changed infinitely many times. Its transition rate will therefore be the average over the states the environment can be in, that is $\overline{q}_k(j|i,a)=\sum_{d\in\mathcal{Z}}\phi_k(d)q_k^{(d)}(j|i,a)$. The fluid process then arises by taking into account only the mean drifts $\overline{q}$.

We denote by $u$ a fluid control and let $x^u(t)$ be the corresponding fluid process. Let $x_{j,k}^{u,a}(t)$ denote the proportion of class-$k$ fluid in state $j$ under action $a$ at time $t$ and let $x_{j,k}^u(t)=x_{j,k}^{u,0}(t)+x_{j,k}^{u,1}(t)$ be the proportion of class-$k$ fluid in state $j$.

We consider fluid controls $u(t)$ that base their actions only on the state of the fluid process $x(t)$. As such, policies for the stochastic process can be reduced to controls for the fluid problem. In particular, when given a policy $\pi$, the corresponding fluid control $u=\pi$ is defined as

$$x_{j,k}^{u,a}(t)=y_{j,k}^{\pi,a}(\vec{x}^{\pi}(t)). \quad (6)$$

We define the dynamics of $x^u(t)$ as follows:

$$\begin{aligned}\frac{\mathrm{d}x_{j,k}^u(t)}{\mathrm{d}t} &= \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k} x_{i,k}^{u,a}(t)\overline{q}_k(j|i,a)-\sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k} x_{j,k}^{u,a}(t)\overline{q}_k(i|j,a)\\ &= \sum_{a=0}^{1}\sum_{i=1}^{J_k} x_{i,k}^{u,a}(t)\overline{q}_k(j|i,a), \end{aligned} \quad (7)$$

where the last step follows from $\overline{q}_k(j|j,a):=-\sum_{i=1,i\neq j}^{J_k}\overline{q}_k(i|j,a)$. The constraint on the fluid control $u$ is that the total proportion of active fluid satisfies

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k} x_{j,k}^{u,1}(t)\leq\alpha,\ \text{for all } t\geq 0. \quad (8)$$

In the lemma below we formally state the convergence of the stochastic process $\vec{X}^{\pi}(t)/N$ of the proportions of bandits. This result will not be used in any of the proofs of our results, but is presented for completeness. The proof can be found in the appendix.

For the convergence to hold, it is assumed in Lemma 4.2 that the process describing the state of a class-$k$ bandit at time $t$, is tight [11] in $N$, that is, roughly speaking that the processes must not oscillate too wildly so that probability mass cannot disappear from compact sets. For either a finite state space ($J_k < \infty$) or when the possible transitions in each state is finite, tightness follows directly, see [12, page 12] for details.

**Lemma 4.2** *Assume policy $\pi$ is such that $y_{j,k}^{\pi,1}(\cdot)$ is uniformly Lipschitz continuous, i.e.,*

$$\sup_{j,k} |y_{j,k}^{\pi^*,1}(\vec{x}) - y_{j,k}^{\pi^*,1}(\vec{z})| \leq C \sup_{i,l} |x_{i,l} - z_{i,l}|, \quad \text{for all } \vec{x}, \vec{z}, \tag{9}$$

*with $C > 0$. For a given policy $\pi$, if the process describing the state of a class-$k$ bandit at time $t$ is tight (with respect to $N$), then the stochastic process $\frac{\vec{X}^{N,\pi}(Nt)}{N}$ converges to the deterministic process $x^u(t)$, with $u = \pi$ (as defined in (6) and (7)).*

We will be interested in finding an optimal equilibrium point $\vec{x}$ of the fluid dynamics that minimizes the holding cost averaged over the environments,

$$\sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \sum_{d \in \mathcal{Z}} \phi_k(d) C_k^{(d)}(j,a) x_{j,k}^a = \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \overline{C}_k(j,a) x_{j,k}^a.$$

Setting (7) equal to zero, this gives us the following linear optimization problem:

$$(LP) \qquad v^* := \min_{(x_{j,k}^a)} \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \overline{C}_k(j,a) x_{j,k}^a$$

$$\text{s.t.} \quad 0 = \sum_{a=0}^{1} \sum_{i=1}^{J_k} x_{i,k}^a \overline{q}_k(j|i,a), \quad \forall\ j,k, \tag{10}$$

$$\sum_{k=1}^{K} \sum_{j=1}^{J_k} x_{j,k}^1 \leq \alpha, \tag{11}$$

$$\sum_{j=1}^{J_k} \sum_{a=0}^{1} x_{j,k}^a = \gamma_k, \quad \forall\ k \tag{12}$$

$$x_{j,k}^a \geq 0, \quad \forall\ j,k,a, \tag{13}$$

Let $x^*$ and $v^*$ denote an optimal solution and optimal value of the (LP) problem, respectively. In Lemma 4.4 below, we use the LP formulation to find a lower bound on the original stochastic optimization problem. For that, we need the following condition.

**Condition 4.3** *Given a policy $\pi$,*

*a) the process $\frac{\vec{X}^{N,\pi}}{N}(t)$ has a unique invariant probability distribution $p^{N,\pi}$ $\forall N$.*

*b) the family $\left\{p^{N,\pi}\right\}_{N \in \mathbb{N}}$ is tight.*

*c) the family $\left\{p^{N,\pi}\right\}_{N \in \mathbb{N}}$ is uniform integrable.*

When bandits have a finite state space ($J_k < \infty$), the condition is true whenever $X^{N,\pi}(t)$ is unichain.

**Lemma 4.4** *Assume Condition 4.3 is satisfied. It then holds that the feasible set of the (LP) problem is non-empty and*

$$\liminf_{N \to \infty} V_-^{N,\pi}(\vec{x}) \geq v^*, \tag{14}$$

*with $v^*$ the optimal value of the (LP) problem.*

The proof can be found in the appendix.

### 4.3 Asymptotic optimality

In this section we present the asymptotic optimality results.

**Proposition 4.5** *Let $x^*$ be an optimal solution of the (LP) problem. Let $\pi^*$ be a policy for which the fluid process $x^{\pi^*}(t)$ converges to $x^*$ as $t \to \infty$, and $x^*$ is the unique equilibrium point (global attractor property). Assume $\pi^*$ satisfies Condition 4.3 and the assumptions made in Lemma 4.2. Then, $\pi^*$ is asymptotically optimal, that is, for all $\vec{x}$ and all policies $\pi$, it holds that*

$$\lim_{N\to\infty} V^{N,\pi^*}(\vec{x}) \leq \liminf_{N\to\infty} V_-^{N,\pi}(\vec{x}). \tag{15}$$

In Section 5 we will present a class of policies that satisfy (9). The global attractor property is verified numerically for the different examples presented in this paper, see Section 7.

**Proof of Proposition 4.5:** In [12, Theorem 2.3] the mean-field limit for a particle system in a rapidly varying environment is given for the stationary regime. Since we assumed tightness of $\frac{\vec{X}^{N,\pi^*}}{N}$ (Condition 4.3) and the fact that the fluid process $x^{\pi^*}(t)$ has a unique global attractor $x^*$, we can apply their result. Also recall the discussion in the proof of Lemma 4.2. Hence, from [12, Theorem 2.3] we have, $\lim_{N\to\infty} \mathbb{P}\left(\frac{\vec{X}^{N,\pi^*}}{N} = x^*\right) = 1$, for each state $j$ and class $k$. Together with Lemma 4.1, this gives

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{\vec{X}^{N,\pi^*}}{N} = x^*, \vec{D} = \vec{d}\right) = \phi(\vec{d}). \tag{16}$$

Recall that $\overline{C}_k(j,a) = \sum_{d\in\mathcal{Z}} \phi_k(d) C_k^{(d)}(j,a)$. Thus

$$\begin{aligned}
\lim_{N\to\infty} V_+^{N,\pi^*}(\vec{x}) &= \lim_{N\to\infty} \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \sum_{d\in\mathcal{Z}} \sum_{\vec{x}} C_k^{(d)}(j,a) y_{j,k}^{\pi^*,a}(\vec{x}) \cdot \mathbb{P}(\frac{\vec{X}^{N,\pi^*}}{N} = \vec{x}, \vec{D} = \vec{d}) \\
&= \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \sum_{d\in\mathcal{Z}} C_k^{(d)}(j,a) y_{j,k}^{\pi^*,a}(\vec{x}^*) \phi_k(d) \\
&= \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \overline{C}_k(j,a) x_{j,k}^{*,a} = v^*.
\end{aligned}$$

where the first step follows from the ergodicity theorem, the second step from uniform integrability of $\vec{X}/N$(interchange of limit and summations) and (16), the third step from the fact that $y_{j,k}^{\pi^*,a}(\vec{x}^*) = x_{j,k}^{*,a}$ (see (6) and $\lim_{t\to\infty} x^{\pi^*}(t) = x^*$), and the last step follows since $\overline{x}^*$ is an optimal solution of (LP). This concludes the proof of Proposition 4.5. □

# 5 Priority policies

In this section we will define an important class of priority policies for which we can prove asymptotic optimality results.

A priority policy is defined as follows. There is a predefined priority ordering on the states each bandit can be in. At any moment in time, a priority policy makes active a maximum number of bandits being in the states having the highest priority among all the bandits present. Hence, for a given priority policy *prio*, we would have that the proportion of class-$k$ bandits in state $j$ that see action 1 is given by

$$y_{j,k}^{prio,1}(\vec{x}) = \min\left(\left(\alpha - \sum_{(i,l)\in S_k^{prio}(j)} x_{i,l}\right)^+, x_{j,k}\right), \tag{17}$$

where $S_k^{prio}(j)$ denotes the set of pairs $(i,l), i = 1,\ldots,J_l, l = 1,\ldots,K$ such that class-$l$ bandits in state $i$ have higher priority than class-$k$ bandits in state $j$ under policy *prio*. In the lemma below we show that this function satisfies (9) when bandits have a finite state space. The proof is in the Appendix.

**Lemma 5.1** *If $J_k < \infty$, Equation* (9) *is valid for any priority policy.*

We now define a set of priority policies $\Pi^*$ that will play a key role in the paper. The priority policies are derived from (the) optimal equilibrium point(s) $x^*$ of the (LP) problem: for a given equilibrium point $x^*$, we consider all priority orderings such that the states that in equilibrium are never passive ($x_{j,k}^{*,0} = 0$) are of higher priority than states that receive some passive action ($x_{j,k}^{*,0} > 0$). In addition, states that in equilibrium are both active and passive ($x_{j,k}^{*,0} \cdot x_{j,k}^{*,1} > 0$) receive higher priority than states that are never active ($x_{j,k}^{*,1} = 0$). Further, if the full capacity is not used in equilibrium (that is, $\sum_k \sum_j x_{j,k}^{*,1} < \alpha$), then the states that are never active in equilibrium are never activated in the priority ordering. The set of priority policies $\Pi^*$ is formalized in the definition below. In particular, in the next section we will prove that an averaged version of the well-known Whittle's index policy is in fact inside this set of policies.

**Definition 5.2 (Set of priority policies $\Pi^*$)**
*We define*

$$X^* := \{x^* : x^* \text{ is an optimal solution of (LP) with } x_k(0) = X_k(0)\ \}.$$

*The set of priority policies $\Pi^*$ is defined as*

$$\Pi^* := \cup_{x^* \in X^*} \Pi(x^*),$$

*where $\Pi(x^*)$ is the set of all priority policies that satisfy the following rules:*

1. *A class-$k$ bandit in state $j$ with $x_{j,k}^{*,1} > 0$ and $x_{j,k}^{*,0} = 0$ is given higher priority than a class-$\tilde{k}$ bandit in state $\tilde{j}$ with $x_{\tilde{j},\tilde{k}}^{*,0} > 0$.*

2. *A class-$k$ bandit in state $j$ with $x_{j,k}^{*,0} > 0$ and $x_{j,k}^{*,1} > 0$ is given higher priority than a class-$\tilde{k}$ bandit in state $\tilde{j}$ with $x_{\tilde{j},\tilde{k}}^{*,0} > 0$ and $x_{\tilde{j},\tilde{k}}^{*,1} = 0$.*

3. *If $\sum_{k=1}^{K} \sum_{j=1}^{J_k} x_{j,k}^{*,1} < \alpha$, then any class-$k$ bandit in state $j$ with $x_{j,k}^{*,1} = 0$ and $x_{j,k}^{*,0} > 0$ will **never** be made active.*

We can now state the asymptotic optimality result for priority policies in the class $\Pi^*$. Intuitively, this result can be explained as follows. Let $\pi^*$ be some policy from the set $\Pi(x^*) \subset \Pi^*$. It is easily verified that $x^*$ is an equilibrium point of the fluid process $x^{\pi^*}(t)$. If in addition, the point $x^*$ is a global attractor, that is, for any initial point, the process $x^{\pi^*}(t)$ converges to $x^*$, then using Proposition 4.5 one obtains the result.

**Corollary 5.3** *Assume a finite state space, $J_k < \infty$, for all $k$. For a given policy $\pi^* \in \Pi(x^*) \subset \Pi^*$, assume $x^*$ is the global attractor of the fluid process $x^{\pi^*}(t)$. If in addition, the process $X^{N,\pi^*}(t)$ is unichain, then $\pi^*$ is asymptotically optimal, that is,* (15) *is satisfied.*

**Proof:** Lemma 5.1 gives that $\pi^*$ satisfies (9). Hence the result follows directly from Proposition 4.5. □

**Remark 5.4 (Infinite state space)** *The assumption of finite state space in order for the priority policy to be asymptotically optimal was made in order to assure uniformly Lipschitz continuity of the function $y_{j,k}^{\pi^*,1}(\cdot)$. In fact, when $J_k = \infty$, one can easily construct a setting in which* (9) *does*

*not hold. For example, for $K = 1$, take $\vec{x}^{(l)}$ s.t. $x_i^{(l)} = \alpha/l$ for $i < l$, $x_l^{(l)} = 1-\alpha$, and $x_i^{(l)} = 0$, for $i > l$. Take $\vec{z}^{(l)}$ s.t. $z_i^{(l)} = 0$ for $i < l$, $z_l^{(l)} = 1-\alpha$, and $z_i^{(l)} = \alpha/l$, for $i > l$. Then, $y_l^{prio,1}(\vec{x}^{(l)}) = 0$ and $y_l^{prio,1}(\vec{z}^{(l)}) = 1-\alpha$, where prio is the policy that prioritizes state 1 over state 2, and state 2 over state 3, etc. However, $\sup_j |x_j^{(l)} - z_j^{(l)}| = \alpha/l$. Since the state space is infinite, we can now take $l \to \infty$. We then directly see that (9) does not hold.*
*We however note that in our numerical example, where we consider an infinite state space, we do observe a very close to optimal performance of the priority policies. In future research, we plan to further investigate this, and find policies for which asymptotic optimality can be proved in the infinite state space setting.*

# 6 Averaged Whittle's index policy

In this section we introduce the averaged version of Whittle's index policy. This is a particular case of a priority policy. It is however only defined in case the system is so-called *indexable*, while our definition of the set of policies $\Pi^*$ is well-defined for both indexable and non-indexable systems.
Before stating our result, we first give a short introduction to Whittle's index policy for the *standard* restless bandit setting (that is, without modulated environment). In order to obtain a policy for the restless bandit model, Whittle introduced in 1988 a relaxation technique [39]. This approximation consists in relaxing the sample-path constraint into a time-average constraint. That is, the constraint that at most $\alpha N$ bandits can be active at a time, see (4), is replaced by its time-average version:

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_{t=0}^{T} \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{j,k}^{\pi,1}\left(\frac{\vec{X}^{N,\pi}(t)}{N}\right)\mathrm{d}t\right) \leq \alpha. \tag{18}$$

Using the Lagrangian approach, this relaxed-constraint problem (minimize (2) under constraint (18)) can then be written as minimizing

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_{0}^{T} \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k^{(D_k(t))}(j,a) y_{j,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}(t)}{N}\right)\mathrm{d}t \right.$$
$$\left. + W\Big(\int_{t=0}^{T} \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{j,k}^{\pi,1}\left(\frac{\vec{X}^{N,\pi}(t)}{N}\right)\mathrm{d}t - \alpha N\Big)\right),$$

where $W$ is the Lagrange multiplier (chosen such that the time-average constraint (18) holds). The latter can be decomposed into $K$ subproblems, one for each class of bandit, where in each subproblem one needs to minimize the cost term plus a cost $W$ whenever the bandit is made active:

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_{t=0}^{T} (C_k^{(D_k(t))}(S_k^\pi(t), A_l^\pi(t)) + W\mathbf{1}_{(A_k^\pi(t)=1)}\mathrm{d}t\right), \tag{19}$$

where $S_k^\pi(t)$ denotes the state of a class-$k$ bandit at time $t$ and $A_k^\pi(t)$ denotes the action chosen for the class-$k$ bandit under policy $\pi$ (for the one-dimensional process).
The above relaxation simplifies notably the problem, as it transforms the multi-dimensional stochastic problem to several one-dimensional stochastic problems. We can now define Whittle's index.

**Definition 6.1 (Whittle's index)** *Whittle's index $W_k(j)$ is defined as the least value of $W$ for which it is optimal in* (19) *to make the class-$k$ bandit in state $j$ passive.*

Under a technical *indexability* condition – assuring structural properties on the optimal policy of the relaxed control problem – the optimal solution to the $K$ subproblems can be described by the *Whittle's indices*. We refer to [30] for a survey on indexability results.

**Definition 6.2 (Indexability)** *A bandit is indexable if the set of states in which passive is an optimal action in* (19) *is increasing in* $W$.

The solution for the relaxed subproblem (19) is to activate all bandits that are currently in a state $j$ such that there Whittle index $W_k(j)$ is larger than $W$. This solution is however not feasible for the original $K$-dimensional problem, since sometimes it may activate more than $\alpha N$ bandits at a time. For the original problem, the following heuristic was then proposed: active those $\alpha N$ bandits having currently the highest index value. This is referred to as *Whittle's index policy*. The relaxation technique as such provides a systematic approach to get a simple index policy.
In this paper, we study restless bandits living in a modulated environment. In the limiting regime, as the environment varies rapidly, we found that a bandit observes only the averaged (over the steady state of the environment) parameters, that is, $\overline{q}_k(\cdot)$ and $\overline{C}_k(\cdot,\cdot)$. This motivates us to define the averaged Whittle index policy.

**Definition 6.3 (Averaged Whittle's index policy)** *The* averaged Whittle's index $\overline{W}_k(j)$ *for our restless bandit problem living in a modulated environment is defined as the Whittle index that would result from the restless bandit problem with parameters* $\overline{q}_k(i|j,a)$, $\overline{C}_k(j,a)$, *and no modulating environment.*
*The* averaged Whittle index policy *activates those* $\alpha N$ *bandits having currently the highest averaged Whittle index value* $\overline{W}_k(S_{n,k}(t))$, *where* $S_{n,k}(t)$ *is the state of the* $n$*th class-*$k$ *bandit,* $k = 1,\ldots,K$, $n = 1,\ldots,N_k$.

**Proposition 6.4** *If for the averaged version of the restless bandit problem the process describing the state of a class-*$k$ *bandit is unichain, regardless of the policy employed, and in addition the averaged restless bandit problem is indexable, then there is an* $x^*$ *such that the averaged Whittle's index policy is inside the set of priority policies* $\Pi(x^*) \subset \Pi^*$.
*If in addition* $J_k < \infty$, *for all* $k$, *and* $x^*$ *is the global attractor of the fluid process* $x^{\pi^*}(t)$, *then the averaged Whittle index policy is asymptotically optimal.*

The above proposition extends the asymptotic optimality result of Whittle's index policy as obtained in [24, 37, 38] to that of restless bandits living in rapidly varying Markov modulated environments. The assumptions made in Proposition 6.4 are the same as those needed in [24, 37, 38].

**Proof of Proposition 6.4:** In [37, Proposition 5.6] it was proved that for the standard restless bandit problem, Whittle's index policy was inside some set of priority policies defined by some linear program problem $(\tilde{LP})$. Since we defined our averaged Whittle's index policy based on the standard restless bandit problem with averaged parameters, we directly have that the averaged Whittle index policy is inside the set of priority policies defined for $(\tilde{LP})$ based on the averaged parameters. However, it can be checked that the $(\tilde{LP})$ for the averagered parameters is equivalent to our (LP) problem, hence, the averaged Whittle index policy is inside $\tilde{\Pi}^*$, which is equal to $\Pi^*$. From the unichain assumption, we obtain that Condition 4.3 is satisfied when $J_k < \infty$. The asymptotic optimality result now follows directly from Proposition 4.5. □

# 7 Numerical evaluation

The objective of this section is to evaluate the performance of the averaged Whittle index policy (and other heuristics) outside the asymptotic regime. We do this for a multi-class scheduling problem in a wireless downlink channel. The model we consider is the following. There is one wireless downlink channel that is shared by two classes of users. For a given policy, let $Q_k^{\pi}(t)$ denote the number of class-$k$ users in the system. Each class of users is associated a Markov modulated environment, $D_k(t)$, $k = 1, 2$, which can be in two states. When $D_k(t) = d$, class-$k$ users arrive according to a Poisson process with parameter $\lambda_k^{(d)}$, $k = 1, 2$. At each moment in time,

the base station can send data to at most one of the classes. Given $D_k(t) = d$ and $Q_k^\pi(t) = n_k$, class $k$ (if served) has departure rate

$$\mu_k^{(d)} \frac{n_k}{n_k + 1}.$$

This mimics opportunistic scheduling, since the more class-$k$ users present in the system, the higher will be the maximum capacity available among the class-$k$ users. See for further details the discussion in [26, Section 6.1].
The scheduler now needs to decide at each moment in time which of the two classes to activate. We assume

$$\rho := \sum_{k=1}^{2} \frac{\overline{\lambda}_k}{\overline{\mu}_k} < 1,$$

with $\overline{\lambda}_k = \sum_{d=1}^{2} \lambda_k^{(d)} \phi_k(d)$ and $\overline{\mu}_k = \sum_{d=1}^{2} \mu_k^{(d)} \phi_k(d)$, so any work-conserving policy makes the system stable. The objective will be to minimize the mean number of users in the system.
The performance for a given policy will be computed using the Value Iteration approach [34] by writing the dynamic programming equation for the process

$$(Q_1^\pi(t), Q_2^\pi(t), D_1(t), D_2(t)).$$

Note that we only consider policies that cannot observe the environments when the scheduling decision is taken.
We describe the averaged Whittle index policy in Section 7.1. We calculate an optimal policy and numerically evaluate index policies for several settings: in Section 7.3 each class is associated its own environment in order to model Markov modulated arrival processes. In Section 7.4 we will study the effect of having one common environment that influences the departure rates of the classes. In Section 7.5 we discuss the observable case. Our overall conclusion is that the averaged Whittle index policy performs close to optimal when the modulated environments varies rapidly.

## 7.1 Averaged Whittle index

This model fits in the restless bandit framework with Markov modulated environments, as presented in this paper. In particular, there are two bandits, each bandit representing a class of users, and the state of the class-$k$ bandit representing its queue length $N_k$. Hence, when $D_k = d$, the class-$k$ bandit makes a transition from state $n_k$ to state $n_k + 1$ at rate $\lambda_k^{(d)}$, and from $n_k$ to $n_k - 1$ at rate $\mu_k^{(d)} \frac{n_k}{n_k+1} a$, where $a = 1$ when the class-$k$ bandit is served, and $a = 0$ otherwise. At most one bandit can be activated at a time.
Using the expression of Whittle's index as derived in [26], we obtain that the *averaged* Whittle index for the class-$k$ bandit in state $n$ (its queue length), is given by

$$\overline{W}_k(n) = \frac{\mathbb{E}(Q_k^n) - \mathbb{E}(Q_k^{n-1})}{\pi_k^n(n) - \pi_k^{n-1}(n-1)}, \ n = 1, 2, \ldots.$$

Here $Q_k^n$ denotes the stationary random variable of the one-dimensional birth-and-death process process with birth rates $\overline{\lambda}_k$ and death rates in state $n_k$ at rate $\mathbf{1}_{(n_k > n)} \overline{\mu}_k \frac{n_k}{n_k+1}$, and $\pi_k^n(\cdot)$ denotes the stationary measure of $Q_k^n$.
The averaged Whittle index policy serves at each moment in time the class of users having the highest index value $\overline{W}_k(N_k(t))$. From Proposition 6.4 we have that this policy is asymptotically optimal under certain conditions. One of the conditions concerns indexability, which is easily verified for this example using [26, Proposition 2]. The global attractor property is verified numerically for the different examples presented in this section. Another condition needed is that the state space is bounded, which is not the case for our example. The numerical results however do indicate a good performance.

| $\gamma$ | 1 | 5 | 10 | 25 | 50 | 100 | 500 | 750 | 1000 | 2500 | 5000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $g^{OPT}$ | 9.6 | 9.7 | 9.4 | 8.6 | 7.7 | 7.0 | 6.2 | 6.1 | 6.0 | 6.0 | 5.9 |
| $g^{\overline{W}}$ | 11.2 | 10.9 | 10.2 | 8.8 | 7.7 | 7.0 | 6.2 | 6.1 | 6.0 | 6.0 | 5.9 |
| $g^{\sum_d \phi(d)W^{(d)}}$ | 12.6 | 11.9 | 10.7 | 8.9 | 7.8 | 7.1 | 6.3 | 6.2 | 6.1 | 6.1 | 6.0 |
| $g^{W^{(1)}}$ | 10.6 | 10.5 | 10.1 | 8.8 | 7.8 | 7.1 | 6.2 | 6.1 | 6.1 | 6.0 | 6.0 |
| $g^{W^{(2)}}$ | 12.3 | 12.9 | 13 | 12.1 | 11.1 | 10.2 | 9.2 | 9.1 | 9.1 | 9.1 | 9.0 |
| $Rel(\overline{W})$ | 16.2 | 12.7 | 8.4 | 2.4 | 0.4 | 0.01 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 |
| $Rel(\sum_d \phi(d)W^{(d)})$ | 30.9 | 22.6 | 14 | 4.2 | 1.7 | 1.4 | 1.7 | 1.7 | 1.7 | 1.7 | 1.7 |
| $Rel(W^{(1)})$ | 10.5 | 9 | 6.7 | 2.9 | 1.3 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 1 |
| $Rel(W^{(2)})$ | 27.6 | 33.6 | 37.5 | 42.1 | 44 | 45 | 49.2 | 50.2 | 50.9 | 52.1 | 52.6 |

Table 1: Results for Example 1.

The optimality results in this paper are for the limiting regime where both the number of bandits (classes) as well as the speed of the modulated environments grow large. In the numerical examples we will instead keep the number of bandits equal to two, which allows to evaluate the performance of our policy outside the asymptotic regime. We then evaluate the performance of the averaged Whittle index policy, as well as other heuristics, for different speeds of the environments.

## 7.2 Optimal solution

The modulated environments are *unobservable*. In addition, their evolution does not depend on the state of the bandits. Since we assume no learning we can find optimal actions with the following Bellman equation:

$$V(\vec{n}) = n_1 + n_2 + \sum_{k=1}^{2}\sum_{d=1}^{2} \phi_k(d)\lambda_k^{(d)} V(\vec{n}+e_k) + \min_{a\in\{1,2\}} \left[ \sum_{k=1}^{2}\sum_{d=1}^{2} \mathbf{1}_{(a=k)}\phi_k(d)\mu_k^{(d)} V((\vec{n}-e_k)^+) \right], \tag{20}$$

that is, in every state $n$, the only information available to the decision maker is the steady-state distribution of the environment. In the next sections we compare the performance of our heuristics to that of the performance under the actions as defined in (20).

## 7.3 Markov modulated arrival processes

Our first numerical example studies Markov modulated arrival processes. That is, environments $D_1(t)$ and $D_2(t)$ are two independent Markov processes. Each environment can be in two states $\{1,2\}$ and environment $D_k(t)$ makes a transition from state $d$ to state $d'$ at rate $r_k(d'|d)$. When class $k$ sees environment $d$, the arrival rate is $\lambda_k^{(d)}$ (while the departure rates remain unchanged).
*Example 1:* We set the parameters as follows: $\lambda_1^{(1)} = 5$, $\lambda_1^{(2)} = 0.1$, $\lambda_2^{(1)} = 0.5$, $\lambda_2^{(2)} = 5$, and $\mu_1^{(d)} = 7.5$, $\mu_2^{(d)} = 9$, for $d = 1, 2$. We take $r_k(2|1) = 0.001 \cdot \gamma$, $r_k(1|2) = 0.009 \cdot \gamma$, $k = 1, 2$, and let $\gamma$ vary from 1 up to 5000, in order to study the effect of the speed of the modulated environments. We then have $\phi_1(1) = \phi_2(1) = 0.9$, hence $\rho \approx 0.67$.
In Table 1 we show the average performance under the averaged Whittle index policy and that of the optimal policy. We also show the performance under three other index policies: we consider the index policy $\sum_{d=1}^{2} \phi_k(d) W_k^{(d)}(n)$, which is the Whittle index averaged over the different environments, and we consider the index policy $W_k^{(d)}(n)$, which is the Whittle index in case the environment would be always in state $d$, $d = 1, 2$. We denote by $g^{OPT}$ the performance of the optimal policy as defined in Section 7.2 and let $g^\pi$ denote the average cost under policy $\pi$. We define by $Rel(\pi) := \frac{g^\pi - g^{OPT}}{g^{OPT}} * 100\%$ the suboptimality gap (in %). We observe that the averaged

| $\gamma$ | 1 | 5 | 10 | 50 | 100 | 500 | 750 | 1000 | 2500 | 5000 | 7500 | 10000 | 25000 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $g^{OPT}$ | 56.6 | 45.8 | 43.2 | 40.6 | 39.8 | 25.7 | 17.9 | 14.1 | 9.0 | 7.8 | 7.4 | 7.3 | 7.0 |
| $g^{\overline{W}}$ | 87.4 | 85.4 | 84.5 | 83.0 | 81.9 | 43.8 | 21.0 | 15.0 | 9.1 | 7.8 | 7.5 | 7.3 | 7.0 |
| $Rel(\overline{W})$ | 54.4 | 86.3 | 96 | 105 | 106 | 70.2 | 17.4 | 6.4 | 0.8 | 0.3 | 0.2 | 0.1 | 0.04 |

Table 2: Results for Example 2.

Whittle index policy $\overline{W}$ is 2.5% away from the lower bound for $\gamma = 25$, i.e., when the transition rates of the environment are $r_k(2|1) = 0.025$ and $r_k(1|2) = 0.225$. Hence, already for a normal scaled environment, the performance of the averaged Whittle index policy is very close to optimal. For slow speed, $\gamma = 1$, the averaged Whittle index policy is only 16% away from the optimal. The index policy $\sum_{d=1}^{2} \phi_k(d) W_k^{(d)}(n)$ gives slightly worse performance than that of $\overline{W}$.

For any speed of the environment, we observe that the index policy $W_k^{(1)}(n)$ outperforms the averaged Whittle index policy, while the index policy $W_k^{(2)}(n)$ gives very bad performance ranging between a suboptimality gap of 30% until 53%. This can be explained from the fact that the environment is 90% of the time in state 1.

## 7.4 One common environment affecting the departure rates

We now consider one common environment $D(t)$, which can be in two states $\{1, 2\}$, with transition rates $r(d'|d)$. This time, we let the arrival rates be independent of the environment. Instead, the state of the environment influences both the departure rate of class 1 and class 2.

*Example 2:* In this example, we chose the parameters such that, when in environment 1, class 1 has a high departure rate and class 2 a low departure rate, while in environment 2, we have the opposite. The values for the parameters are: $\lambda_1^{(d)} = 0.6$, $\lambda_2^{(d)} = 1.2$, for $d = 1, 2$, and $\mu_1^{(1)} = 4$, $\mu_1^{(2)} = 0.5$, $\mu_2^{(1)} = 0.1$, $\mu_2^{(2)} = 6$. We take $r(2|1) = 0.004 \cdot \gamma$, $r(1|2) = 0.006 \cdot \gamma$. We then have $\phi_1(1) = \phi_2(1) = 0.6$, hence $\rho \approx 0.72$.

Note that for these parameters, $\lambda_1^{(2)} > \mu_1^{(2)}$ and $\lambda_2^{(1)} > \mu_2^{(1)}$. That is, if $D(t) = 1$, then class 2 is in overload, while if $D(t) = 2$, then class 1 is in overload. In particular, this implies that the indices $W_k^{(d)}(n)$ are not well-defined. We therefore only simulate the performance under the averaged Whittle index policy $\overline{W}$.

The results can be found in Table 2. We observe that the averaged Whittle index policy is 6.4% away from the optimal policy when $\gamma = 1000$, i.e., $r(2|1) = 4$ and $r(1|2) = 6$. Hence, we observe that already for a normal scaled environment, the performance of the averaged Whittle index policy is very close to optimal. When $\gamma = 2500$, i.e., $r(2|1) = 10$ and $r(1|2) = 15$, the gap reduces to 0.8%. For $\gamma = 500$, i.e., $r(2|1) = 2$ and $r(1|2) = 3$, and smaller $\gamma$, the suboptimality gap becomes significantly large.

*Example 3:* In this example, we chose the parameters such that the departure rate of class 1 is always lower than that of class 2, in each environment. In addition, the departure rate for class 1 is considerably higher in environment 2 compared to its rate in environment 1.

The values for the parameters are: $\lambda_1^{(d)} = 1$, $\lambda_2^{(d)} = 3.5$, for $d = 1, 2$, and $\mu_1^{(1)} = 1.5$, $\mu_1^{(2)} = 10$, $\mu_2^{(1)} = 12$, $\mu_2^{(2)} = 11$. We take $r(2|1) = 0.002 \cdot \gamma$, $r(1|2) = 0.008 \cdot \gamma$. We then have $\phi_1(1) = \phi_2(1) = 0.8$, hence $\rho \approx 0.61$.

The results can be found in Table 3. We observe that the averaged Whittle index policy $\overline{W}$ is 7% away from the optimal policy for $\gamma = 5000$, i.e., the transition rates of the environment are $r_k(2|1) = 10$ and $r_k(1|2) = 40$. For slow speed, $\gamma = 1$, the averaged Whittle index policy is 29% away from the optimal. The index policy $\sum_{d=1}^{2} \phi_k(d) W_k^{(d)}(n)$ gives worse performance than that of $\overline{W}$.

For any speed of the environment, we observe that the index policy $W_k^{(1)}(n)$ outperforms the averaged Whittle index policy, while the index policy $W_k^{(2)}(n)$ gives very bad performance ranging

| $\gamma$ | 1 | 5 | 10 | 25 | 50 | 100 | 500 | 750 | 1000 | 2500 | 5000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $g^{OPT}$ | 25.9 | 17.0 | 13.2 | 9.4 | 7.6 | 6.3 | 4.8 | 4.6 | 4.4 | 4.2 | 4.0 |
| $g^{\overline{W}}$ | 33.3 | 21.4 | 16.7 | 12.0 | 9.5 | 7.8 | 5.6 | 5.2 | 5.0 | 4.5 | 4.3 |
| $g^{\sum_d \phi(d) W^{(d)}}$ | 33.3 | 21.5 | 16.8 | 12.2 | 9.8 | 8.1 | 5.8 | 5.4 | 5.2 | 4.8 | 4.6 |
| $g^{W^{(1)}}$ | 30.0 | 19.4 | 15.2 | 10.9 | 8.7 | 7.2 | 5.3 | 5.0 | 4.8 | 4.4 | 4.2 |
| $g^{W^{(2)}}$ | 34.1 | 25.4 | 20.8 | 15.4 | 12.4 | 10.1 | 6.8 | 6.3 | 6.0 | 5.3 | 5.0 |
| $Rel(\overline{W})$ | 28.5 | 26.2 | 26.6 | 26.7 | 25.7 | 23.7 | 16 | 13.9 | 12.5 | 8.8 | 6.9 |
| $Rel(\sum_d \phi(d) W^{(d)})$ | 28.9 | 26.9 | 27.8 | 29 | 29 | 27.9 | 21.4 | 19.5 | 18.2 | 14.7 | 13 |
| $Rel(W^{(1)})$ | 15.7 | 14.6 | 15.1 | 15.8 | 15.6 | 14.6 | 10.5 | 9.2 | 8.4 | 6.1 | 4.8 |
| $Rel(W^{(2)})$ | 31.8 | 49.9 | 57.7 | 63.5 | 63.6 | 60.1 | 42.2 | 37.5 | 34.5 | 27.5 | 24.4 |

Table 3: Results for Example 3.

between a suboptimality gap of 32% until 21%. Again, this difference can be explained from the fact that the environment is 80% of the time in environment 1.

### 7.5 Observable environment

In the numerical examples, we observed that already for a rather normal speed of the environment, the averaged Whittle index policy works well. Hence, in case the decision maker aims for a policy that is robust with respect to the environment, our heuristic seems to be a good choice. We leave it for future research to find efficient heuristics in case of an observable environment (or when the environment could be learned from the state transitions of the bandits). Regarding the latter, for Example 3 we have calculated the performance under the optimal policy when the environment can be observed, and where decision epochs are moments when one of the bandits changes state, or the environment changes state. Note that for the observable setting, the stability condition strongly depends on the policy employed. Numerically, we derived that being able to observe the environment gives an improvement of 16% when $\gamma = 1$, and of 60% when $\gamma = 50000$. The large improvement, especially when $\gamma = 5000$, comes from the fact that the policy will schedule in an opportunistic manner. Recall that the departure rate of class 1 is much larger in environment 2 (compared to environment 1). Hence, in environment 2, an optimal policy will give more preference to class 1. However, in environment 1, class 2 has a much higher departure rate compared to class 1, hence, more priority will be given to class 2. Observing the environment, and being able to change the action when the environment changes, makes this large improvement in the performance possible.

## 8 Conclusions and future work

This paper presents a first step to the optimal control of stochastic scheduling problems in a Markov modulated environment. We assumed the decision maker cannot observe the state of the environment. Since the environment varies in the limit very rapidly, bandits only experience the averaged behavior of the environments, and as such we proved that the averaged version of Whittle's index policy is asymptotically optimal.

Numerically we observed that the averaged Whittle index policy performs close to optimal when the speed of the environment grows large, even though the number of bandits remained fixed. Further insights remain to be derived for efficient control when the speed of the environment is of normal order, or arbitrarily slow. On a similar note, it would be interesting to investigate efficient control in the case of an *observable* environment. Then, it could be expected that certain environment-dependent index policies provide provably close to optimal performance.

For the problem as presented in this paper, there are many interesting threads to be further

developed. For example, it remains to be understood what would be efficient heuristics for the case of an infinite state space, since only in the case of a finite state space we were able to propose concrete policies that satisfy (9). As future work, we further plan to investigate what happens if the evolution of the environment can also depend on the population of bandits. For example, it might be the case that the environment represents the queue length, whose evolution depends on the state of a server (bandit). Other interesting extensions are to include arrivals of new bandits to the system, as was done in [37]. This would require an extension of the results as presented in [12].

## Acknowledgments

The authors would like to thank Charles Bordenave, for his valuable help he offered in order to apply his results in this work. Furthermore, the authors would like to thank the anonymous referees for their valuable comments and helpful suggestions.
Research partially supported by the french *Agence Nationale de la Recherche (ANR)* through the project *ANR-15-CE25-0004 (ANR JCJC RACON)*.

## References

[1] L.L.H. Andrew A. Wierman and A. Tang. Power aware speed scaling in processor sharing systems. In *Proceedings of IEEE INFOCOM*, 2009.

[2] S.H.A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari. Optimality of myopic sensing in multichannel opportunistic access. *IEEE Transactions on Information Theory*, 55:4040–4050, 2009.

[3] A. Anand and G. de Veciana. A Whittle's index based approach for QoE optimization in wireless networks. In *Proceedings of ACM SIGMETRICS*, Irvine, California, USA, 2018.

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

[5] J. Anselmi. Asymptotically optimal open-loop load balancing. *Queueing Systems*, 87:245–267, 2017.

[6] A. Asanjarani and Y. Nazarathy. The role of information in system stability with partially observable servers. *Arxiv report*, 1610.02781v1, 2016.

[7] U. Ayesta, M. Erausquin, and P. Jacko. A modeling framework for optimizing the flow-level scheduling with time-varying channels. *Performance Evaluation*, 67:1014–1029, 2010.

[8] U. Ayesta, M. Erausquin, and P. Jacko. Resource-sharing in a single server with time-varying capacity. In *Proceedings of 49th Annual Allerton Conference on Communication, Control and Computing*, 2011.

[9] U. Ayesta, P. Jacko, and V. Novak. A nearly-optimal index rule for scheduling of users with abandonment. In *Proceedings of IEEE INFOCOM*, Hong Kong, 2011.

[10] M. Benaïm and J-Y Le Boudec. A class of mean field interaction models for computer and communication systems. *Performance Evaluation*, 65:823–838, 2008.

[11] P. Billingsley. *Convergence of probability measures*. Wiley, New York NY, 1968.

[12] C. Bordenave, D. McDonald, and A. Proutiére. A particle system in interaction with a rapidly varying environment: Mean field limits and applications. *Networks and heterogeneous media*, 5(1):31–62, 2010.

[13] S.C. Borst. User level performance of channel aware scheduling algorithms in wireless data networks. *IEEE/ACM Transactions on Networking*, 13:636–647, 2005.

[14] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and trends in machine learning*, 5:1–122, 2012.

[15] A. Budhiraja, A. Ghosh, and X. Liu. Scheduling control for markov-modulated single-server multiclass queueing systems in heavy traffic. *Queueing Systems*, 78(1):57–97, 2014.

[16] C. Buyukkoc, P. Varaya, and J. Walrand. The $c\mu$ rule revisited. *Advances of Applied Probability*, 17:237–238, 1985.

[17] F. Cecchi and P. Jacko. Nearly-optimal scheduling of users with Markovian time-varying transmission rates. *Performance Evaluation*, 99–100:16–36, 2016.

[18] E. Çinlar. *Introduction to Stochastic Processes*. Prentice-Hall, New Jersey, 1975.

[19] N. Ehsan and M. Liu. On the optimality of an index policy for bandwidth allocation with delayed state observation and differentiated services. In *Proceedings of IEEE INFOCOM*, Hong Kong, 2004.

[20] N. Gast and B. Gaujal. A mean field approach for optimization in discrete time. *Discrete Event Dynamic Systems*, 21(1):63–101, 2011.

[21] J.C. Gittins, K.D. Glazebrook, and R.R. Weber. *Multi-Armed Bandit Allocation Indices*. Wiley, Chichester, 2011.

[22] K.D. Glazebrook, C. Kirkbride, and J. Ouenniche. Index policies for the admission control and routing of impatient customers to heterogeneous service stations. *Operations Research*, 57:975–989, 2009.

[23] K.D. Glazebrook and H.M. Mitchell. An index policy for a stochastic scheduling model with improving/deteriorating jobs. *Naval Research Logistics*, 49:706–721, 2002.

[24] D.J. Hodge and K. D. Glazebrook. On the asymptotic optimality of greedy index heuristics for multi-action restless bandits. *Advances in Applied Probability*, 47:652–667, 2015.

[25] M. Larrañaga, U. Ayesta, and I.M. Verloop. Index policies for multi-class queues with convex holding cost and abandonments. In *Proceedings of ACM SIGMETRICS*, Austin TX, USA, 2014.

[26] M Larrañaga, U Ayesta, and I.M. Verloop. Dynamic control of birth-and-death restless bandits: application to resource-allocation problems. *IEEE/ACM Transactions on Networking*, 24(6):3812–3825, 2016.

[27] K. Liu and Q. Zhao. Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56:5547–5567, 2010.

[28] A. Mahajan and D. Teneketzis. Multi-armed bandit problems. In *Foundations and Application of Sensor Management, eds. A.O. Hero III, D.A. Castanon, D. Cochran and K. Kastella.*, pages 121–308, Springer-Verlag, 2007.

[29] Y. Nazarathy, T. Taimre, A. Asanjarani, J. Kuhn, B. Patch, and A. Vuorinen. The challenge of stabilizing control for queueing systems with unobservable server states. In *IEEE Proceedings of the 5th Australian Control Conference*, 2015.

[30] J. Niño-Mora. Dynamic priority allocation via restless bandit marginal productivity indices. *TOP*, 15:161–198, 2007.

[31] J. Niño-Mora. Marginal productivity index policies for admission control and routing to parallel multi-server loss queues with reneging. *Lecture Notes in Computer Science*, 4465:138–149, 2007.

[32] J. R. Norris. *Markov chains*, volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. Reprint of 1997 original.

[33] W. Ouyang, A. Eryilmaz, and N.B. Shroff. Asymptotically optimal downlink scheduling over Markovian fading channels. In *Proceedings of IEEE INFOCOM*, Orlando Fl, USA, 2012.

[34] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York, 1994.

[35] V. Raghunathan, V. Borkar, M. Cao, and P.R. Kumar. Index policies for real-time multicast scheduling for wireless broadcast systems. In *Proceedings of IEEE INFOCOM*, 2008.

[36] A. Slivkins and E. Upfal. Adapting to a changing environment: The Brownian restless bandits. In *Proceedings of 21st Annual Conference on Learning Theory*, pages 343–354, 2008.

[37] I.M. Verloop. Asymptotic optimal control of multi-class restless bandits. *Annals of Applied Probability 26 (4), 1947-1995*, 2016.

[38] R.R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(03):637–648, 1990.

[39] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

[40] P. Whittle. *Optimal Control, Basics and Beyond*. John Wiley & Sons, 1996.

# Appendix

**Proof of Lemma 4.1:** For ease of notation, we remove the superscripts $\pi$ and $N$ in the proof. For the random vector $\frac{\vec{X}}{N}(t)$, we define the following Probability Generating Function (21) and the Moment Generating Function (22), conditioned on the environment vector $\vec{d}$:

$$\begin{aligned} g^{(\vec{d})}(\vec{z}) &:= \mathbb{E}\left( z_{11}^{\frac{X_{11}}{N}} z_{21}^{\frac{X_{21}}{N}} \dots z_{jk}^{\frac{X_{jk}}{N}} \dots z_{J_K K}^{\frac{X_{J_K K}}{N}} \mathbf{1}_{(\vec{D}=\vec{d})} \right) \\ &= \sum_{\vec{x}} z_{11}^{\frac{x_{11}}{N}} z_{21}^{\frac{x_{21}}{N}} \dots z_{jk}^{\frac{x_{jk}}{N}} \dots z_{J_K K}^{\frac{x_{J_K K}}{N}} \mathbb{P}\left( \frac{\vec{X}}{N} = \vec{x}, \vec{D} = \vec{d} \right), \end{aligned} \tag{21}$$

and

$$\tilde{g}^{(\vec{d})}(\vec{s}) = g^{(\vec{d})}(\mathrm{e}^{-\vec{s}}) = \mathbb{E}\left( \mathrm{e}^{-\vec{s}\frac{\vec{X}}{N}} \mathbf{1}_{(\vec{D}=\vec{d})} \right). \tag{22}$$

The balance equations for the Markov process $\left(\frac{\vec{X}(t)}{N}, \vec{D}(t)\right)$ state that for each $(\vec{x}, \vec{d})$, we have

$$\mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x}, \vec{D} = \vec{d}\right)\left[\sum_{\vec{d'}\in\mathcal{Z}^K} r(\vec{d'}|\vec{d}) + \sum_{k,a,i,j}\frac{1}{N}q_k^{(d_k)}(j|i,a)y_{i,k}^a(\vec{x})\cdot N\right] \tag{23}$$

$$= \sum_{\vec{d'}\in\mathcal{Z}^K} r(\vec{d}|\vec{d'})\mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x}, \vec{D} = \vec{d'}\right) + \sum_{\substack{k,a,j \\ i/x_{i,k}>0 \\ j/x_{j,k}<1}}\frac{1}{N}q_k^{(d_k)}(i|j,a)y_{j,k}^a(\vec{x} + \frac{\vec{e}_{j,k}}{N} - \frac{\vec{e}_{i,k}}{N})$$

$$\cdot N \cdot \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x} + \frac{\vec{e}_{j,k}}{N} - \frac{\vec{e}_{i,k}}{N}, \vec{D} = \vec{d}\right).$$

Note that we have to restrict the sum on the rhs in order to consider the boundary cases. Multiplying (23) by $z_{11}^{x_{11}}\dots z_{J_KK}^{x_{J_KK}}$, summing over $\vec{x}$ on both sides of the equality, we then obtain the following expression:

$$g^{(\vec{d})}(\vec{z})\sum_{\vec{d'}\in\mathcal{Z}^K} P(\vec{d'}|\vec{d}) \; + \sum_{\vec{x}} z_{11}^{x_{11}}\dots z_{J_KK}^{x_{J_KK}}\mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x}, \vec{D} = \vec{d}\right)\sum_{k,a,i,j} q_k^{(d_k)}(j|i,a)y_{i,k}^a(\vec{x}) \tag{24}$$

$$= \sum_{\vec{d'}\in\mathcal{Z}^K} P(\vec{d}|\vec{d'})\; g^{(\vec{d'})}(\vec{z}) \quad + \sum_{\vec{x}} z_{11}^{x_{11}}\dots z_{J_KK}^{x_{J_KK}} \sum_{\substack{k,a,j \\ i/x_{i,k}>0 \\ j/x_{j,k}<1}} q_k^{(d_k)}(i|j,a)y_{j,k}^a(\vec{x} + \frac{\vec{e}_{j,k}}{N} - \frac{\vec{e}_{i,k}}{N})$$

$$\cdot \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x} + \frac{\vec{e}_{j,k}}{N} - \frac{\vec{e}_{i,k}}{N}, \vec{D} = \vec{d}\right).$$

Below we will show that the second terms in both the LHS and the RHS of (24) are equal when taking limits. We begin rewriting the second term in the LHS. By changing variables $\vec{z} \to \mathrm{e}^{-\vec{s}}$ (since in the limit we have that $\frac{\vec{X}}{N}$ is a continuous variable), we have that it is equal to

$$\sum_{\vec{x}} \mathrm{e}^{-s_{11}x_{11}}\dots\mathrm{e}^{-s_{J_KK}x_{J_KK}} \sum_{k,a,i,j} q_k^{(d_k)}(j|i,a)y_{i,k}^a(\vec{x})\cdot\mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x}, \vec{D} = \vec{d}\right). \tag{25}$$

We make the same change of variables in the second term in the RHS. Furthermore, the sums in $k, a, i, j$ are bounded, because of the hypothesis (1), the fact that $y_{j,k}^a(.)$ is a proportion and $\mathbb{P}$ is a probability. Thus, as a consequence of Fubini's theorem, we can interchange the order of summations:

$$\sum_{k,a,i,j}\sum_{\substack{\vec{x}/x_{i,k}>0 \\ x_{j,k}<1}} \mathrm{e}^{-s_{11}x_{11}}\dots\mathrm{e}^{-s_{J_KK}x_{J_KK}} q_k^{(d_k)}(i|j,a)\cdot y_{j,k}^a(\vec{x} + \frac{\vec{e}_{j,k}}{N} - \frac{\vec{e}_{i,k}}{N})$$

$$\cdot \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x} + \frac{\vec{e}_{j,k}}{N} - \frac{\vec{e}_{i,k}}{N}, \vec{D} = \vec{d}\right).$$

For each fixed $k, i$ and $j$, we also make the change of variables $\vec{x} \to \vec{y} + \frac{e_{i,k}}{N} - \frac{e_{j,k}}{N}$,

$$\sum_{k,a,i,j} \sum_{\substack{\vec{y}/y_{i,k}<1 \\ y_{j,k}>0}} \mathrm{e}^{-s_{11}y_{11}} \dots \mathrm{e}^{-s_{J_K K}y_{J_K K}} \frac{\mathrm{e}^{-\frac{\sum_k s_{i,k}}{N}}}{\mathrm{e}^{-\frac{\sum_k s_{j,k}}{N}}} \cdot q_k^{(d_k)}(i|j,a) y_{j,k}^a(\vec{y}) \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{y}, \vec{D} = \vec{d}\right)$$

$$= \sum_{\vec{y}} \mathrm{e}^{-s_{11}y_{11}} \dots \mathrm{e}^{-s_{J_K K}y_{J_K K}} \frac{\mathrm{e}^{-\frac{s_i}{N}}}{\mathrm{e}^{-\frac{s_j}{N}}} \cdot \sum_{\substack{k,a,j \\ i/y_{i,k}<1 \\ j/y_{j,k}>0}} q_k^{(d_k)}(i|j,a) y_{j,k}^a(\vec{y}) \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{y}, \vec{D} = \vec{d}\right). \quad (26)$$

Comparing (25) and (26) there are two aspects that remain different. The first one is the restriction of suming only when $y_{j,k} > 0$ and $y_{i,k} < 1$ in (26), but we can add the boundary cases which only sum 0: note that when $y_{j,k} = 0$, there would be a $y_{j,k}^a(\vec{y}) = 0$ multiplying, and when $y_{i,k} = 1$, as it's a proportion in class $k$ bandits, this means that $y_{j,k} = 0$ and again there would be a $y_{j,k}^a(\vec{y}) = 0$ multiplying. The second aspect is the factor $\frac{\mathrm{e}^{-\frac{\sum_k s_{i,k}}{N}}}{\mathrm{e}^{-\frac{\sum_k s_{j,k}}{N}}}$, which converges to 1 as $N \to \infty$. Since $y_{j,k}^a(\vec{x}) \leq 1$, in the limit both terms are the same.
So we can conclude from (24) that

$$\lim_{N\to\infty} \tilde{g}^{(\vec{d})}(\vec{s}) \sum_{\vec{d'} \in \mathcal{Z}^K} r(\vec{d'}|\vec{d}) = \sum_{\vec{d'} \in \mathcal{Z}^K} r(\vec{d}|\vec{d'}) \lim_{N\to\infty} \tilde{g}^{(\vec{d'})}(\vec{s}), \quad (27)$$

that is, $\lim_{N\to\infty} \tilde{g}^{(\vec{d})}(\vec{s})$ satisfies the balance equations for $\vec{D}$. Thus, $\lim_{N\to\infty} \tilde{g}^{(\vec{d})}(\vec{s}) = c(\vec{s})\phi(\vec{d})$, where $c(\vec{s})$ does not depend on $\vec{d}$. When summing both sides over $\vec{d}$, we obtain $c(\vec{s}) = \lim_{N\to\infty} \mathbb{E}\left(\mathrm{e}^{-\frac{\vec{s}\vec{X}}{N}}\right)$, that is, Equation (5) holds true. □

**Proof of Lemma 4.2** In [12], the mean-field limit is obtained for a particle system living in a rapidly varying environment. In particular, in [12, Theorem 2.2] the convergence result in the transient regime is obtained, which would prove our lemma. Hence, to use the results obtained in [12] we need to verify several the assumptions **A0-A8** as made in that paper. We will do so below.
Each bandit represents a particle, and a transition of a class-$k$ bandit/particle from state $j$ to state $i$ when in environment $d$ occurs at rate

$$\frac{1}{N} \sum_{a=0}^{1} q_k^{(d)}(i|j,a) y_{j,k}^{\pi,a}\left(\frac{\vec{X}(t)}{N}\right), \quad (28)$$

where $y_{j,k}^{\pi,a}\left(\frac{\vec{X}(t)}{N}\right)$ needs to be interpreted as the probability for a class-$k$ bandit in state $j$ to see action $a$. Expression (28) is the equivalence of [12, Equation (3)].
In [12] a discrete-time setting is considered. In the model we consider, the transition rates are uniformly bounded[1], hence we can uniformize our system to obtain a discrete-time model [32, Section 2.6]. Furthermore, we consider a multi-class particle system, while the results in [12] are for an exchangeable system. As noted in [12, page 38], this is easily generalized to a multi-class model, when adding a class description to the state of the bandit and having the class of the vector of bandits at time 0 be determined by an exchangeable random vector.
We are left with verifying Assumptions **A0-A8** as stated in [12]. Most of them are true by definition, except for **A0** and **A2**, which we discuss in the remainder of the proof.

[1]The transition rate out of state $(\vec{X}, \vec{D})$ are $\sum_{i,k} X_{j,k} q_k^{(d)}(i|j,a)/N + \sum_{\vec{d'}} r(\vec{d'}|\vec{D}) \leq C_1 + \max_{j,k} \sum_i q_k^{(d)}(i|j,a) < C_1 + C_2$, where we use the assumption made on the environment in Section 3 and (1).

**[A0.]** states a weak correlation between the bandits' transitions. That is, let $A_{n,k}(t)$ denotes the event of the $n$-th class-$k$ bandit to have a transition at time $t$. Then in A0 of [12] it is assumed that $\mathbb{P}\left(A_{n_1,k_1}(t)A_{n_2,k_2}(t)\right) \leq \rho(N)/N$, with $\rho(N) \to 0$. For our model this is satisfied, since bandits behave independently from each other, hence this probability is equal to zero.
**[A2.]** states that the transition rates of the bandits are uniformly Lipschitz, using the total variation between two empirical measures. That is, there exists a constant $C$ such that for every $\vec{x}, \vec{z}$,

$$\begin{aligned}&\sup_{j,k,d}\left\{\sum_{i=1}^{J_k}|q_k^{(d)}(i|j,0)y_{j,k}^{\pi,0}(\vec{x}) + q_k^{(d)}(i|j,1)y_{j,k}^{\pi,1}(\vec{x}) - q_k^{(d)}(i|j,0)y_{j,k}^{\pi,0}(\vec{z}) - q_k^{(d)}(i|j,1)y_{j,k}^{\pi,1}(\vec{z})|\right\}\\&\leq C\sup_{j,k}|x_{j,k}-z_{j,k}|.\end{aligned} \tag{29}$$

The LHS is equal to

$$\begin{aligned}&\sup_{j,k,d}\left\{\sum_{i=1}^{J_k}|q_k^{(d)}(i|j,0)\left(y_{j,k}^{\pi,0}(\vec{x}) - y_{j,k}^{\pi,0}(\vec{z})\right) + q_k^{(d)}(i|j,1)\left(y_{j,k}^{\pi,1}(\vec{x}) - y_{j,k}^{\pi,1}(\vec{z})\right)|\right\}\\&\leq \sup_{j,k,d}\left\{\sum_{i=1}^{J_k} q_k^{(d)}(i|j,0)|y_{j,k}^{\pi,0}(\vec{x}) - y_{j,k}^{\pi,0}(\vec{z})| + q_k^{(d)}(i|j,1)|y_{j,k}^{\pi,1}(\vec{x}) - y_{j,k}^{\pi,1}(\vec{z})|\right\}\\&\leq C_2\sup_{j,k,d}\left\{|y_{j,k}^{\pi,0}(\vec{x}) - y_{j,k}^{\pi,0}(\vec{z})| + |y_{j,k}^{\pi,1}(\vec{x}) - y_{j,k}^{\pi,1}(\vec{z})|\right\}\\&\leq C_2\sup_{j,k,d}\left\{|x_{j,k}-z_{j,k}| + 2|y_{j,k}^{\pi,1}(\vec{x}) - y_{j,k}^{\pi,1}(\vec{z})|\right\}\\&\leq C_2(2C+1)\sup_{j,k}|x_{j,k}-z_{j,k}|,\end{aligned}$$

where we used (1) and (9). That is, (29) is satisfied. □

**Proof of Lemma 4.4:** Let $\pi$ be a policy that satisfies Condition 4.3. We first prove that the feasible set of the (LP) problem is non-empty.
Note that we have relative compactness for the sequence of random variables $\left(\vec{X}^{N,\pi}/N\right)_{N\in\mathbb{N}}$. In case $J_k < \infty$ for every class $k$, this is valid because the random vectors live on a compact space. In case $J_k = \infty$ for some $k$, relative compactness comes from Condition 4.3 and [11, Theorem 6.1]. As a consequence, we can consider a subsequence of $N$ (for ease of notation still denoted as $N$) such that $\vec{X}^{N,\pi}/N$ converges in distribution when $N \to \infty$. Together with Condition 4.3 we can then define the following limiting point $\vec{y} := (y_{j,k}^{\pi,a})$, where

$$y_{j,k}^{\pi,a} := \mathbb{E}\left(\lim_{N\to\infty}\liminf_{T\to\infty}\frac{1}{T}\int_0^T y_{j,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}(t)}{N}\right)\mathrm{d}t\right).$$

It is important to note that this limit can depend on the subsequence of $N$ considered. For ease of notation, we are not writing this dependence. We will first prove that $\vec{y}^{\pi}$ is a feasible solution of the (LP) problem. Since at most $N$ bandits are in the system, we have

$$\lim_{t\to\infty}\frac{X_{j,k}^{N,\pi}(t)}{t} = 0, \text{ for all } j,k. \tag{30}$$

Note that $\int_0^t y_{j,k}^{\pi,a}(\frac{\vec{X}^{N,\pi}(s)}{N})\cdot N\mathrm{d}s$ is the total aggregated amount of time spent on action $a$ on class-$k$ bandits in state $j$ during the interval $(0,t]$. Hence, we can write the following sample-path

construction of the process $X_{j,k}^{N,\pi}(t)$:

$$X_{j,k}^{N,\pi}(t) = X_{j,k}^{N,\pi}(0) + \sum_{a=0}^{1}\sum_{i\neq j}\sum_{d\in\mathcal{Z}} N^{\frac{q_k^{(d)}(j|i,a)}{N}}\left(\int_0^t \mathbf{1}_{(D_k(s)=d)} y_{i,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}(s)}{N}\right)\cdot N\mathrm{d}s\right)$$
$$- \sum_{a=0}^{1}\sum_{i\neq j}\sum_{d\in\mathcal{Z}} N^{\frac{q_k^{(d)}(i|j,a)}{N}}\left(\int_0^t \mathbf{1}_{(D_k(s)=d)} y_{j,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}(s)}{N}\right)\cdot N\mathrm{d}s\right), \tag{31}$$

where $N^{q_k^{(d)}(j|i,a)/N}(t)$ are independent Poisson processes having as rates $q_k^{(d)}(j|i,a)/N$, $i,j = 1,\ldots,J_k$, $k=1,\ldots,K$, $a=0,1$. By the ergodic theorem [18] and because of $\vec{X}^N$ having an invariant distribution (see Condition 4.3, item a), we obtain that $\frac{1}{t}\int_0^t \mathbf{1}_{(D_k(s)=d)} y_{j,k}^{\pi,a}(\frac{\vec{X}^{N,\pi}(s)}{N})\cdot N\mathrm{d}s$ converges to $\mathbb{E}\left(\mathbf{1}_{(D_k=d)} y_{j,k}^{\pi,a}(\frac{\vec{X}^{N,\pi}}{N})\cdot N\right) < \infty$ as $t\to\infty$, for all $j,k,a,d,N$. Because of Lemma 4.1 we further have

$$\mathbb{E}\left(\lim_{N\to\infty}\mathbf{1}_{(D_k=d)} y_{j,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}}{N}\right)\right) = \phi_k(d) y_{j,k}^{\pi,a}. \tag{32}$$

Now, when dividing both sides in (31) by $t$, taking $t\to\infty$, and using that $N^\theta(at)/t \to a\theta$ and (30) hold, we obtain

$$0 = \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k}\sum_{d\in\mathcal{Z}} \frac{q_k^{(d)}(j|i,a)}{N}\mathbb{E}\left(\mathbf{1}_{(D_k=d)} y_{i,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}}{N}\right)\cdot N\right)$$
$$- \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k}\sum_{d\in\mathcal{Z}} \frac{q_k^{(d)}(i|j,a)}{N}\mathbb{E}\left(\mathbf{1}_{(D_k=d)} y_{j,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}}{N}\right)\cdot N\right)$$
$$= \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k}\sum_{d\in\mathcal{Z}} q_k^{(d)}(j|i,a)\mathbb{E}\left(\mathbf{1}_{(D_k=d)} y_{i,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}}{N}\right)\right)$$
$$- \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k}\sum_{d\in\mathcal{Z}} q_k^{(d)}(i|j,a)\mathbb{E}\left(\mathbf{1}_{(D_k=d)} y_{j,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}}{N}\right)\right). \tag{33}$$

Letting $N\to\infty$, together with (32), we then obtain

$$0 = \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k}\sum_{d\in\mathcal{Z}} q_k^{d}(j|i,a)\phi_k(d) y_{i,k}^{\pi,a} - \sum_{a=0}^{1}\sum_{i=0,i\neq j}^{J_k}\sum_{d\in\mathcal{Z}} q_k^{d}(i|j,a)\phi_k(d) y_{j,k}^{\pi,a},$$
$$= \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k} \overline{q}_k(j|i,a) y_{i,k}^{\pi,a} - \sum_{a=0}^{1}\sum_{i=0,i\neq j}^{J_k} \overline{q}_k(i|j,a) y_{j,k}^{\pi,a}, \tag{34}$$

a.s., that is, $\vec{y}^\pi$ satisfies Equation (10). By definition, $\vec{y}^\pi$ satisfies $\sum_{k,j} y_{j,k}^{\pi,1} \le \alpha$ and $y_{j,k}^{\pi,a} \ge 0$. Hence, $\vec{y}^\pi$ is a feasible solution of (LP). Note that the interchange of limit and summations to go from (33) to (34) is possible because of the uniform integrability given by Condition 4.3.
It is left to prove Inequality (14). We have,

$$\liminf_{N\to\infty} V_-^{N,\pi}(x)$$
$$= \liminf_{N\to\infty}\liminf_{T\to\infty}\frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k^{(D_k(t))}(j,a)\cdot y_{j,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}(t)}{N}\right)\mathrm{d}t\right)$$
$$\ge \mathbb{E}\left(\liminf_{N\to\infty}\liminf_{T\to\infty}\frac{1}{T}\int_0^T \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k^{(D_k(t))}(j,a)\cdot y_{j,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}(t)}{N}\right)\mathrm{d}t\right),$$

where the inequality holds because of Fatou's Lemma. So it would be enough to prove that

$$\liminf_{N\to\infty}\liminf_{T\to\infty}\frac{1}{T}\int_0^T\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}C_k^{(D_k(t))}(j,a)y_{j,k}^{\pi,a}(\frac{\vec{X}^{N,\pi}(t)}{N})\mathrm{d}t\geq v^*, \tag{35}$$

almost surely.
Consider a fixed realization $\omega$ of the process. We first assume that the LHS in (35) is finite. We then obtain that

$$\begin{aligned}
&\liminf_{N\to\infty}\liminf_{T\to\infty}\frac{1}{T}\int_0^T y_{j,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}(t)}{N}\right)\mathbf{1}_{(D_k(t)=d)}\mathrm{d}t\\
&=\liminf_{N\to\infty}\mathbb{E}\left(\mathbf{1}_{(D_k=d)}y_{j,k}^{\pi,a}\left(\frac{\vec{X}^{N,\pi}}{N}\right)\right) \qquad (36)\\
&=\lim_{N_i\to\infty}\mathbb{E}\left(\mathbf{1}_{(D_k=d)}y_{j,k}^{\pi,a}\left(\frac{\vec{X}^{N_i,\pi}}{N_i}\right)\right)\\
&=\mathbb{E}\left(\lim_{N_i\to\infty}\mathbf{1}_{(D_k=d)}y_{j,k}^{\pi,a}\left(\frac{\vec{X}^{N_i,\pi}}{N_i}\right)\right)\\
&=\phi_k(d)y_{j,k}^{\pi,a}, \qquad (37)
\end{aligned}$$

with $N_i$ the subsequence corresponding to the liminf sequence and such that $\frac{\vec{X}^{N,\pi}}{N}$ converges in distribution. In the third step we used that $\vec{X}^{N,\pi}/N$ is uniformly integrable and in the fourth step we used (32) (this equation holds for any weakly converging subsequence of $\frac{\vec{X}^{N,\pi}}{N}$). As a consequence we obtain that

$$\begin{aligned}
&\liminf_{N\to\infty}\liminf_{T\to\infty}\frac{1}{T}\int_0^T\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}C_k^{(D_k(t))}(j,a)y_{j,k}^{\pi,a}(\frac{\vec{X}^{N,\pi}(t)}{N})\mathrm{d}t\\
&=\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}C_k^{(D_k(t))}(j,a)\liminf_{N\to\infty}\liminf_{T\to\infty}\frac{1}{T}\int_0^T y_{j,k}^{\pi,a}(\frac{\vec{X}^{N,\pi}(t)}{N})\mathrm{d}t\\
&=\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\sum_{d\in\mathcal{Z}}C_k^{d}(j,a)\phi_k(d)y_{j,k}^{\pi,a}\\
&=\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\overline{C}_k(j,a)y_{j,k}^{\pi,a}\geq v^*,
\end{aligned}$$

where the last inequality holds because $\vec{y}$ is a feasible solution of the (LP) problem.
In particular, the above shows that $v^*<\infty$, since we assumed in Section 3 that there exists a policy for which the LHS in (35) is finite.
Assume now that the LHS of (35) is infinite. Then inequality (35) follows directly, since $v^*<\infty$. This concludes the proof. □

**Proof of Lemma 5.1:** We consider four possible cases:
**(1)** If $\left(\alpha-\sum_{(i,l)\in S_k^\pi(j)}x_{i,l}\right)<0$, then by definition (17) we have $y_{j,k}^{prio,1}(\vec{x})=0$.
If as well $\left(\alpha-\sum_{(i,l)\in S_k^\pi(j)}z_{i,l}\right)<0$, then the LHS of (9) equals zero, hence (9) holds. If instead

$\left(\alpha - \sum_{(i,l)\in S_k^{\pi}(j)} z_{i,l}\right) \geq 0$, then by definition (17),

$$\begin{aligned}
|y_{j,k}^{prio,1}(\vec{x}) - y_{j,k}^{prio,1}(\vec{z})| &= \min\left(\alpha - \sum_{(i,l)\in S_k^{prio}(j)} z_{i,l}, z_{j,k}\right) \\
&\leq \alpha - \sum_{(i,l)\in S_k^{prio}(j)} z_{i,l} \\
&< \sum_{(i,l)\in S_k^{prio}(j)} (z_{i,l} - x_{i,l}) \\
&\leq \sum_{k=1}^{K} J_k \sup_{i,l} |x_{i,l} - z_{i,l}|.
\end{aligned}$$

In the remaining three cases, we can now assume that for both $\vec{u} = \vec{x}, \vec{z}$, it holds that

$$\left(\alpha - \sum_{(i,l)\in S_k^{prio}(j)} u_{i,l}\right) \geq 0.$$

**(2)** If $x_{j,k} \leq \alpha - \sum_{(i,l)\in S_k^{\pi}(j)} x_{i,l}$ and $z_{j,k} \leq \alpha - \sum_{(i,l)\in S_k^{prio}(j)} z_{i,l}$, then we obtain directly the result $|y_{j,k}^{prio,1}(\vec{x}) - y_{j,k}^{prio,1}(\vec{z})| = |x_{j,k} - z_{j,k}| \leq \sup_{i,l} |x_{i,l} - z_{i,l}|$.
**(3)** If $x_{j,k} \leq \alpha - \sum_{(i,l)\in S_k^{prio}(j)} x_{i,l}$ and $z_{j,k} \geq \alpha - \sum_{(i,l)\in S_k^{prio}(j)} z_{i,l}$, then

$$|y_{j,k}^{prio,1}(\vec{x}) - y_{j,k}^{prio,1}(\vec{z})| = \left| x_{j,k} - \alpha + \sum_{(i,l)\in S_k^{prio}(j)} z_{i,l} \right|. \tag{38}$$

If, in addition, $x_{j,k} > \alpha - \sum_{(i,l)\in S_k^{prio}(j)} z_{i,l}$, then since $x_{j,k} \leq \alpha - \sum_{(i,l)\in S_k^{prio}(j)} x_{i,l}$, we have

$$\begin{aligned}
|y_{j,k}^{prio,1}(\vec{x}) - y_{j,k}^{prio,1}(\vec{z})| &= x_{j,k} - \alpha + \sum_{(i,l)\in S_k^{prio}(j)} z_{i,l} \\
&\leq \sum_{(i,l)\in S_k^{prio}(j)} (z_{i,l} - x_{i,l}) \\
&\leq \sum_{k=1}^{K} J_k \sup_{i,l} |x_{i,l} - z_{i,l}|,
\end{aligned}$$

since $\sum_{k=1}^{K} J_k$ is the number of states $(i,l)$ bandits can be in. Instead, if $x_{j,k} \leq \alpha - \sum_{(i,l)\in S_k^{prio}(j)} z_{i,l}$, then

$$\begin{aligned}
|y_{j,k}^{prio,1}(\vec{x}) - y_{j,k}^{prio,1}(\vec{z})| &= \alpha - \sum_{(i,l)\in S_k^{prio}(j)} z_{i,l} - x_{j,k} \\
&\leq z_{j,k} - x_{j,k} \\
&\leq \sup_{i,l} |x_{i,l} - z_{i,l}|.
\end{aligned}$$

**(4)** If $x_{j,k} \geq \alpha - \sum_{(i,l)\in S_k^{prio}(j)} x_{i,l}$ and $z_{j,k} \geq \alpha - \sum_{(i,l)\in S_k^{prio}(j)} z_{i,l}$, then

$$\begin{aligned} |y_{j,k}^{prio,1}(\vec{x}) - y_{j,k}^{prio,1}(\vec{z})| &= \left| \sum_{(i,l)\in S_k^{prio}(j)} (z_{i,l} - x_{i,l}) \right| \\ &\leq \sum_{k=1}^{K} J_k \sup_{i,l} |x_{i,l} - z_{i,l}|. \end{aligned}$$

Setting $C = \sum_{k=1}^{K} J_k$, we proved (9). □