# Asymptotically optimal priority policies for indexable and non-indexable restless bandits

Ina Maria Maaike Verloop

# Asymptotically optimal priority policies for indexable and non-indexable restless bandits

I.M. Verloop
CNRS ; IRIT ; 2 rue C. Camichel, F-31071 Toulouse, France
Université de Toulouse ; INP ; *IRIT* ; F-31071 Toulouse, France

**Abstract**

We study the asymptotic optimal control of multi-class restless bandits. A restless bandit is a controllable stochastic process whose state evolution depends on whether or not the bandit is made active. Since finding the optimal control is typically intractable, we propose a class of priority policies that are proved to be asymptotically optimal under a global attractor property and a technical condition. We consider both a fixed population of bandits as well as a dynamic population where bandits can depart and arrive. As an example of a dynamic population of bandits, we analyze a multi-class *M/M/S+M* queue for which we show asymptotic optimality of an index policy.

We combine fluid-scaling techniques with linear programming results to prove that when bandits are indexable, Whittle's index policy is included in our class of priority policies. We thereby generalize a result of Weber and Weiss (1990) about asymptotic optimality of Whittle's index policy to settings with (i) several classes of bandits, (ii) arrivals of new bandits, and (iii) multiple actions.

Indexability of the bandits is not required for our results to hold. For non-indexable bandits we describe how to select priority policies from the class of asymptotically optimal policies and present numerical evidence that, outside the asymptotic regime, the performance of our proposed priority policies is nearly optimal.

**Keywords:** Restless bandits, asymptotic optimality, Whittle's index policy, arm-aquiring bandits, non-indexable bandits.

## 1 Introduction

Multi-armed bandit problems are concerned with the optimal dynamic activation of several competing *bandits*, taking into account that at each moment in time $\alpha$ bandits can be made active. A 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 overall cost associated to the states the bandits are in. In the by now classical multi-armed bandit model, [18], it is assumed that only active bandits can change state. In [50], 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 multi-armed restless bandit problem is a stochastic optimization problem that has gained popularity due to its multiple applications in, for example, sequential selection trials in medicine, sensor management, manufacturing systems, queueing and communication networks, control theory, economics, etc. We refer to [19, 31, 51] for further references, applications, and possible extensions that have been studied in the literature.

In 1979, Gittins [17] introduced index-based policies for the non-restless bandit problem. He associated to each bandit an index, which is a function of the state of the bandit, and defined the policy that activates $\alpha$ bandits with currently the largest indices. This policy is known as the Gittins index policy. It was first proved by Gittins that this policy is optimal in the case $\alpha = 1$ [17] for the time-average and discounted cost criteria. In the presence of restless bandits, finding an optimal control is typically intractable. In 1988, Whittle [50] proposed therefore to solve a relaxed optimization problem where the constraint of having at most $\alpha$ bandits active at a time is relaxed to a time-average or discounted version of the constraint. In addition, he defined the so-called indexability property, which requires to establish that as one increases the Lagrange multiplier of the relaxed optimization problem, the collection of states in which the optimal action is passive increases. Under this property, Whittle showed that an optimal solution to the relaxed

optimization problem can be described by index values. The latter, in turn, provide a heuristic for the original restless bandit problem, which is referred to as Whittle's index policy in the literature. It reduces to Gittins index policy when passive bandits are static (the non-restless case). Whittle's index policy is in general not an optimal solution for the original problem. In [46], Weber and Weiss proved Whittle's index policy to be asymptotically optimal.

In this paper we study the asymptotic optimal control of a general multi-class restless bandit problem. We consider both a fixed population of bandits as well as a dynamic scenario where bandits can arrive and depart from the system. The asymptotic regime is obtained by letting the number of bandits that can be simultaneously made active grow proportionally with the population of bandits.

In one of our main results we derive a set of priority policies that are asymptotically optimal when certain technical conditions are satisfied. In another main result we then prove that if the bandits are indexable, Whittle's index policy is contained in our set of priority policies. We thereby generalize the asymptotic optimality result of Weber and Weiss [46] to settings with (i) several classes of bandits, and (ii) arrivals of new bandits. Another extension presented in the paper is the possibility of choosing among multiple actions per bandit. This is referred to as "superprocess" in the literature [19]. Throughout the paper we discuss how our asymptotic optimality results extend to that scenario.

Efficient control of *non-indexable* restless bandits has so far received little attention in the literature. Non-indexable settings can however arise in problems of practical interest, see for example [25] in the context of a make-to-stock system. The definition of our set of priority policies does not rely on indexability of the system, and hence, provides asymptotically optimal heuristics for non-indexable settings. We describe how to select priority policies from this set and present numerical evidence that, outside the asymptotic regime, the performance of our proposed priority policies is nearly optimal.

The asymptotic optimality results obtained in this paper hold under certain technical conditions. For a fixed population of bandits, these conditions reduce to a differential equation having a global attractor, which coincides with the condition as needed by Weber and Weiss [46]. For a dynamic population of bandits, additional technical conditions are needed due to the infinite state space. To illustrate the applicability of the results, we present a large class of restless bandit problems for which we show the additional technical conditions to hold. This class is characterized by the fact that a bandit that is kept passive will eventually leave the system. This can represent many practical situations such as impatient customers, companies that go bankrupt, perishable items, etc. We then present a multi-class $M/M/S+M$ queue, which is a very particular example of the above described class. We describe a priority policy that satisfies the global attractor property and hence asymptotic optimality follows.

In this paper we consider a generalization of the standard restless bandit formulation: Instead of having at each moment in time exactly $\alpha$ bandits active, we allow strictly less than $\alpha$ bandits to be active at a time. We handle this by introducing so-called dummy bandits. In particular, we show that it is asymptotically optimal to activate those bandits having currently the largest, *but strictly positive*, Whittle's indices. Hence, whenever a bandit is in a state having a negative Whittle's index, this bandit will never be activated.

Our proof technique relies on a combination of fluid-scaling techniques and linear programming results: first we describe the fluid dynamics of the restless bandit problem, taking only into account the average behavior of the original stochastic system. The optimal equilibrium points of the fluid dynamics are described by an LP problem. We prove that the optimal value of the LP provides a lower bound on the cost of the original stochastic system. The optimal fluid equilibrium point is then used to describe priority policies for the original system whose fluid-scaled cost coincides with the lower bound, and are hence referred to as asymptotically optimal policies. In order to prove that Whittle's index policy is one of these asymptotically optimal policies, we then reformulate the relaxed optimization problem into an LP problem. An optimal solution of this LP problem is proved to coincide with that of the LP problem corresponding to the fluid problem as described above. This is a different proof approach than that taken in [46] and allows to include arrivals of bandits to the system, whereas the approach of [46] does not.

To summarize, the main contributions of this paper are the following:

- For a multi-class restless bandit problem (possibly non-indexable) with either a fixed or dynamic population of bandits, we determine a set of priority policies that are asymptotically optimal if the corresponding ODE has a global attractor and certain technical conditions (Condition 4.12) are satisfied (Proposition 4.14).

- We show that Condition 4.12 is satisfied for a large class of restless bandit problems. In particular, for a fixed population of bandits under a unichain assumption and for a dynamic population when

passive bandits will eventually leave the system (Proposition 4.13).

- In the case the bandits are indexable, we show that Whittle's index policy is inside our set of priority policies, both for a fixed population of bandits (Proposition 5.6) and for a dynamic population of bandits (Proposition 5.9).
- For non-indexable bandits we describe how to select priority policies from the class of asymptotically optimal policies (Section 8.1) and for a particular example we numerically show that outside the asymptotic regime their suboptimality gap is very small (Section 8.2).

The remainder of the paper is organized as follows. In Section 2 we give an overview of related work and in Section 3 we define the multi-class restless bandit problem. In Section 4 we define our set of priority policies and state the asymptotic optimality result, both for a fixed population as well as for a dynamic population of bandits. In Section 5 we define Whittle's index policy and prove it to be asymptotically optimal. In Section 6 we discuss the global attractor property required in order to prove the asymptotic optimality result. In Section 7 we present the *M/M/S+M* queue as an example of an indexable restless bandit and derive a robust priority policy that is asymptotically optimal. Section 8 focuses on the selection of asymptotically optimal priority policies for non-indexable bandits and numerically evaluates the performance.

# 2 Related work

For the non-restless bandit problem, optimality of Gittins index policy has been proved in [17], for the case $\alpha = 1$ and a time-average or discounted cost criteria. In [48, 49], the optimality result was extended to a dynamic population of bandits where new bandits may arrive over time, e.g., Poisson arrivals or Bernouilli arrivals. For $\alpha > 1$, the optimality results do not necessarily go through. In [38], sufficient conditions on the reward processes were given in order to guarantee optimality of the Gittins policy for the discounted cost criterion, when $\alpha > 1$.

For the restless bandit problem, the authors of [20] have extended Whittle's index heuristic to the setting where each restless bandit may choose from multiple actions, i.e., representing a divisible resource to a collection of bandits. Over the years, Whittle's index policy has been extensively applied and numerically evaluated in various application areas such as wireless downlink scheduling [5, 37], systems with delayed state observation [14], broadcast systems [40], multi-channel access models [1, 30], stochastic scheduling problems [2, 22, 35] and scheduling in the presence of impatient customers [7, 21, 29, 36].

As opposed to Gittins policy, Whittle's index policy is in general not an optimal solution for the original problem. For a fixed population of bandits, optimality has been proved though for certain settings. For example, in [1, 30] this has been proved for a restless bandit problem modeling a multi-channel access system. For a general restless bandit model, in [27] Whittle's index policy has been shown to be optimal for $\alpha = 1$ when (i) there is one dominant bandit or when (ii) all bandits immediately reinitialize when made passive. Other results on optimality of Whittle's index policy for a fixed population of bandits exist for *asymptotic regimes*. In [50], Whittle conjectured that Whittle's index policy is nearly optimal as the number of bandits that can be simultaneously made active grows proportionally with the total number of bandits in the system. In the case of *symmetric* bandits, i.e., all bandits are governed by the same transition rules, this conjecture was proved by Weber and Weiss [46] assuming that the differential equation describing the fluid approximation of the system has a global attractor. They further presented an example for which the conjecture does not hold. In [26], the approaches of [46] were set forth and extended to problems for which multiple activation levels are permitted at any bandit. Another recent result on asymptotic optimality can be found in [37] where the authors considered a specific model, as studied in [30], with two classes of bandits. They proved asymptotic optimality of Whittle's index policy under a recurrence condition. The latter condition replaces the global attractor condition needed in [46] and was numerically verified to hold for their model.

For a dynamic population of restless bandits, that is, when new bandits can arrive to the system, there exist few papers on the performance of index policies. We refer to [5, 6] and [7] where this has been studied in the context of wireless downlink channels and queues with impatient customers, respectively. In particular, in [5, 7], Whittle's index policy was obtained under the discounted cost criterion and numerically shown to perform well. In [6], it was shown that this heuristic is in fact maximum stable and asymptotically fluid optimal. We note that the asymptotic regime studied in [6] is different than the

one as proposed by Whittle [46]. More precisely, in [6] at most one bandit can be made active at a time (the fluid scaling is obtained by scaling both space and time), while in [46] (as well as in this paper) the number of active bandits scales.

Arrivals of new "entities" to the system can also be modelled by a fixed population of restless bandits. In that case a bandit represents a certain type of entities, and the state of a bandit represents the number of this type of entities that are present in the system. Hence, a new arrival of an entity will change the state of the bandit. In the context of queueing systems this has been studied for example in [2, 21, 29]. A Whittle's index obtained from the relaxation of this problem formulation can depend both on the arrival characteristics and on the state, i.e., the number of entities present in the system. This in contrast to the dynamic population formulation of the problem, as discussed in the previous paragraph, where the index will be independent of the arrival characteristics or number of bandits present. Asymptotic optimality results for a fixed population of bandits modeling arrivals of new "entities" have been obtained in, for example, [21] where Whittle's index was shown to be optimal both in the light-traffic and the heavy-traffic limit.

This paper presents heuristics for non-indexable bandits that are asymptotically optimal. Other heuristics proposed for non-indexable problems can be found in [9, 25]. In [9], the primal-dual index heuristic was defined and proved to have a suboptimality guarantee. In Remark 4.8 we will see that an adapted version of the primal-dual index heuristic is included in the set of priority policies for which we obtain asymptotic optimality results. Using fair charges, the authors of [25] proposed heuristics for non-indexable bandits in the context of a make-to-stock system. Numerically the heuristic was shown to perform well. It can be checked that their heuristic is not inside the set of priority policies for which we show asymptotic optimality results.

We conclude this related work section with a discussion on the use of LP techniques in the context of restless bandits. An LP-based proof approach was previously used in, e.g., [9, 33, 34]. In [33, 34], it allowed to characterize and compute indexability of restless bandits. In [9], a set of LP relaxations was presented, providing performance bounds for the restless bandit problem under the discounted-cost criterion.

# 3 Model description

We consider a multi-class restless bandit problem in continuous time. There are $K$ classes of bandits. New class-$k$ bandits arrive according to a Poisson process with arrival rate $\lambda_k \geq 0$, $k = 1, \ldots, K$. At any moment in time, a class-$k$ bandit is in a certain state $j \in \{1, 2, \ldots, J_k\}$, with $J_k < \infty$. When a class-$k$ bandit arrives, with probability $p_k(j)$ this bandit starts in state $j \in \{1, \ldots, J_k\}$.

At any moment in time, a bandit can either be kept passive or active, denoted by $a = 0$ and $a = 1$, respectively. When action $a$ is performed on a class-$k$ bandit in state $i$, $i = 1, \ldots, J_k$, it makes a transition to state $j$ after an exponentially distributed amount of time with rate $q_k(j|i, a)$, $j = 0, 1, \ldots, J_k$, $j \neq i$. Here $j = 0$ is interpreted as a departure of the bandit from the system. We further define $q_k(j|j, a) := -\sum_{i=0, i\neq j}^{J_k} q_k(i|j, a)$. The fact that the state of a bandit might evolve even under the passive action explains the term of a *restless* bandit.

Decision epochs are defined as the moments when an event takes place, i.e., an arrival of a new bandit, a change in the state of a bandit, or a departure of a bandit. A *policy* determines at each decision epoch which bandits are made active, with the restriction that at most $\alpha$ bandits can be made active at a time. This is a generalization of the standard restless bandit formulation where at each moment in time *exactly* $\alpha$ bandits need to be activated, as will be explained in Remark 3.1.

Throughout this paper we will consider both a fixed population of bandits and a dynamic population of bandits:

- *Fixed population:* In this case there are no new arrivals of bandits, i.e., $\lambda_k = 0$, for all $k = 1, \ldots, K$, and there are no departures, i.e., $q_k(0|j, a) = 0$, for all $j, k, a$.
- *Dynamic population:* In this case there are new arrivals of bandits, i.e., $\lambda_k > 0$, for all $k = 1, \ldots, K$, and each bandit can depart from the system, i.e., for each class $k$ there is at least one state $j$ and one action $a$ such that $q_k(0|j, a) > 0$.

For a given policy $\pi$, we define $X^\pi(t) := (X^{\pi,a}_{j,k}(t); k = 1, \ldots, K, j = 1, \ldots, J_k, a = 0, 1)$, with $X^{\pi,a}_{j,k}(t)$ the number of class-$k$ bandits at time $t$ that are in state $j$ and on which action $a$ is performed. We

further denote by $X^{\pi}_{j,k}(t) := \sum_{a=0}^{1} X^{\pi,a}_{j,k}(t)$ the total number of class-$k$ bandits in state $j$ and $X^{\pi}_{k}(t) := \sum_{j=1}^{J_k} X^{\pi}_{j,k}(t)$ the total number of class-$k$ bandits.

Our performance criteria are stability and long-run average holding cost.

**Stability** For a given policy $\pi$, we will call the system *stable* if the process $X^{\pi}(t)$ has a unique invariant probability distribution. We further use the following weaker notions of stability: a policy is *rate-stable* if $\lim_{t\to\infty} \sum_{j,k} \frac{X^{\pi}_{j,k}(t)}{t} = 0$ almost surely and *mean rate-stable* if $\lim_{t\to\infty} \sum_{j,k} \frac{\mathbb{E}(X^{\pi}_{j,k}(t))}{t} = 0$. For a fixed population of bandits the state space is finite, hence the process $X^{\pi}(t)$ being unichain is a sufficient condition for stability of the policy $\pi$. In the case of a dynamic population of bandits the stability condition is more involved. Whether or not the system is stable can depend strongly on the employed policy. In Section 4 we will state necessary stability conditions for the dynamic restless bandit problem.

**Long-run average holding cost:** Besides stability, another important performance measure is the average holding cost. We denote by $C_k(j,a) \in \mathbb{R}$, $j = 1,\ldots,J_k$, the holding cost per unit of time for having a class-$k$ customer in state $j$ under action $a$. We note that $C_k(j,a)$ can be negative, i.e., representing a reward. We further introduce the following notations for long-run average holding costs under policy $\pi$ and initial state $x := (x_{j,k}; k = 1,\ldots,K, j = 1,\ldots,J_k)$:

$$V^{\pi}_{-}(x) := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}_x \left( \int_{t=0}^{T} \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k(j,a) X^{\pi,a}_{j,k}(t)\mathrm{d}t \right),$$

and

$$V^{\pi}_{+}(x) := \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_x \left( \int_{t=0}^{T} \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k(j,a) X^{\pi,a}_{j,k}(t)\mathrm{d}t \right).$$

If $V^{\pi}_{-}(x) = V^{\pi}_{+}(x)$, for all $x$, then we define $V^{\pi}(x) := V^{\pi}_{+}(x)$. We focus on Markovian policies, which base their decisions on the current state and time. Our objective is to find a policy $\pi^*$ that is average optimal, that is,

$$V^{\pi^*}_{+}(x) \le V^{\pi}_{-}(x), \quad \text{for all } x \text{ and for all policies } \pi, \tag{1}$$

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

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k} X^{\pi,1}_{j,k}(t) \le \alpha, \text{ for all } t. \tag{2}$$

**Remark 3.1** *The standard formulation for the restless bandit problem with a fixed population of bandits is to make* exactly $\alpha$ *bandits active at any moment in time. This setting can be retrieved from our formulation by replacing $C_k(j,0)$ with $C_k(j,0)+C$, for all $j,k$, where $C$ represents an additional cost of having a passive bandit. The average additional cost for having passive bandits in the system is equal to $(N-A)C$, with $N$ the total number of bandits in the system and $A$ the average number of active bandits in the system. When $C$ is large enough, an optimal policy will set $A$ maximal, that is $A = \alpha$. Hence, we retrieve the standard formulation.*

**Remark 3.2 (Multi actions)** *In the model description we assumed there are only two possible actions per bandit: $a = 0$ (passive bandit) and $a = 1$ (active bandit). A natural generalization is to consider multiple actions per bandit, that is, for a class-$k$ bandit in state $j$ the scheduler can chose from any action $a \in \{0,\ldots,A_k(j)\}$ and at most $\alpha$ bandits can be made active at a time, i.e., $\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=1}^{A_k(j)} X^{a}_{j,k}(t) \le \alpha$. This is referred to as "superprocess" in the literature [19]. For the* non-restless *bandit problem with $\alpha = 1$, an index policy is known to be optimal in the case each state has a dominant action, that is, if an optimal policy selects a class-$k$ bandit in state $j$ to be made active, it always chooses the same action $a_k(j)$, with $a_k(j) \in \{1,\ldots,A_k(j)\}$. A less strict condition is given in [19, Condition D].*

*In this paper we focus on the setting $A_k(j) = 1$, however all results obtained will go through in the multi-action context when the definition of the policies are modified accordingly, see Remark 4.7 and Remark 5.4.*

# 4 Fluid analysis and asymptotic optimality

In this section we present a fluid formulation of the restless bandit problem and show that its optimal fluid cost provides a lower bound on the cost in the original stochastic model. Based on the optimal fluid solution we then derive a set of priority policies for the original stochastic model that we prove to be asymptotically optimal.

In Section 4.1 we introduce the fluid control problem. In Section 4.2 we define the set of priority policies and the asymptotic optimality results can be found in Section 4.3.

## 4.1 Fluid control problem and lower bound

The fluid control problem arises from the original stochastic model by taking into account only the mean drifts. For a given control $u(t)$, let $x^{u,a}_{j,k}(t)$ denote the amount of class-$k$ fluid in state $j$ under action $a$ at time $t$ and let $x^u_{j,k}(t) = x^{u,0}_{j,k}(t) + x^{u,1}_{j,k}(t)$ be the amount of class-$k$ fluid in state $j$. The dynamics is then given by

$$\begin{aligned}\frac{\mathrm{d}x^u_{j,k}(t)}{\mathrm{d}t} &= \lambda_k p_k(j) + \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k} x^{u,a}_{i,k}(t) q_k(j|i,a) - \sum_{a=0}^{1}\sum_{i=0,i\neq j}^{J_k} x^{u,a}_{j,k}(t) q_k(i|j,a) \\ &= \lambda_k p_k(j) + \sum_{a=0}^{1}\sum_{i=1}^{J_k} x^{u,a}_{i,k}(t) q_k(j|i,a), \qquad (3)\end{aligned}$$

where the last step follows from $q_k(j|j,a) := -\sum_{i=0,i\neq j}^{J_k} q_k(i|j,a)$. The constraint on the total amount of active fluid is given by

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

We are interested in finding an optimal equilibrium point of the fluid dynamics that minimizes the holding cost. Hence, we pose the following linear optimization problem:

$$\begin{aligned}(LP) \qquad \min_{(x^a_{j,k})} & \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k(j,a) x^a_{j,k} \\ \text{s.t. } & 0 = \lambda_k p_k(j) + \sum_{a=0}^{1}\sum_{i=1}^{J_k} x^a_{i,k} q_k(j|i,a), \quad \forall\, j,k, \qquad (4) \\ & \sum_{k=1}^{K}\sum_{j=1}^{J_k} x^1_{j,k} \leq \alpha, \qquad (5) \\ & \sum_{j=1}^{J_k}\sum_{a=0}^{1} x^a_{j,k} = x_k(0), \text{ if } \lambda_k = 0, \quad \forall\, k, \qquad (6) \\ & x^a_{j,k} \geq 0, \quad \forall\, j,k,a, \qquad (7)\end{aligned}$$

where the constraint (6) can be seen as follows: if $\lambda_k = 0$, then $q_k(0|i,a) = 0$ for all $i$. Hence, from (3) we obtain $\sum_{j=1}^{J_k} \frac{\mathrm{d}}{\mathrm{d}t} x^u_{j,k}(t) = 0$.

We denote by $x^*$ an optimal solution of the above problem (LP), assuming it exists. For a fixed population, an optimal solution depends on $x_k(0)$. However, for ease of notation, this dependency is not stated explicitly. We further denote the optimal value of the (LP) by

$$v^*(x(0)) := \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k(j,a) x^{*,a}_{j,k}.$$

We can now state some results concerning the optimization problem (LP). The proof of the first lemma may be found in Appendix A.

**Lemma 4.1** *If there exists a policy $\pi$ such that the process $X^{\pi}(t)$ has a unique invariant probability distribution with finite first moments, then the feasible set of (LP) is non-empty and $v^*(x) < \infty$, for any $x$.*

As a consequence of Lemma 4.1 we get a necessary condition under which there exists a policy that makes the system stable and has finite first moments.

**Corollary 4.2** *If there exists a policy $\overline{\pi}$ such that the system is stable with finite first moments, then*

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k} y^{*1}_{j,k} \leq \alpha,$$

*with $y^* := \arg\min\{\sum_{k=1}^{K}\sum_{j=1}^{J_k} x^1_{j,k} : x$ satisfies (4), (6) and (7)$\}$.*

**Proof:** Assume there exists a policy $\overline{\pi}$ such that the process $X^{\overline{\pi}}(t)$ has a unique invariant probability distribution with finite first moments. By Lemma 4.1 the feasible set of (LP) is non-empty. That is, there exists an $(x^a_{j,k})$ such that (4), (6) and (7) hold and $\sum_{k=1}^{K}\sum_{j=1}^{J_k} x^1_{j,k} \leq \alpha$. Hence, by definition of the optimal solution $y^*$ we obtain $\sum_{k=1}^{K}\sum_{j=1}^{J_k} y^{*1}_{j,k} \leq \sum_{k=1}^{K}\sum_{j=1}^{J_k} x^1_{j,k} \leq \alpha$. This concludes the proof. □

The optimal solution of the fluid control problem (LP) serves as a lower bound on the cost of the original stochastic optimization problem, see the following lemma. The proof can be found in Appendix B.

**Lemma 4.3** *For a fixed population of bandits we have that for any policy $\pi$,*

$$V^{\pi}_{-}(x) \geq v^*(x). \tag{8}$$

*For a dynamic population of bandits, Relation (8) holds if*

- *policy $\pi$ is stable, or,*
- *policy $\pi$ is (mean) rate-stable and $C_k(j,a) > 0$, for all $j,k,a$.*

## 4.2 Priority policies

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. In addition, the policy can prescribe that certain states are never made active.

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 ($x^{*,0}_{j,k} = 0$) are of higher priority than states that receive some passive action ($x^{*,0}_{j,k} > 0$). In addition, states that in equilibrium are both active and passive ($x^{*,0}_{j,k} \cdot x^{*,1}_{j,k} > 0$) receive higher priority than states that are never active ($x^{*,1}_{j,k} = 0$). Further, if the full capacity is not used in equilibrium (that is, $\sum_k \sum_j x^{*,1}_{j,k} < \alpha$), 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.

**Definition 4.4 (Set of priority policies)** *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 $x^{*,1}_{j,k} > 0$ and $x^{*,0}_{j,k} = 0$ is given higher priority than a class-$\tilde{k}$ bandit in state $\tilde{j}$ with $x^{*,0}_{\tilde{j},\tilde{k}} > 0$.*

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

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

We emphasize that in order to define the set of priority policies $\Pi^*$, we do not require the bandits to be indexable, as defined in Definition 5.2. This is in contrast to the definition of Whittle's index policy, which is only well defined in the case the system is indexable. We note that Whittle's index policy is included in $\Pi^*$ for indexable systems as will be proved in Section 5.3.

If there exists a policy such that the system is stable and has finite first moments, then the feasible set of (LP) is non-empty (Lemma 4.1) and hence the set $\Pi^*$ is non-empty. Note that the set $\Pi^*$ can consist of more than one policy. When selecting a policy it might be of practical importance to aim for a policy that is *robust* in the arrival characteristics, the number of bandits that can be made active, and the number of bandits in each class.

**Definition 4.5 (Robust policy)** *A priority policy is called* robust *if the priority ordering does not depend on $\alpha$, $\lambda_k$ and $X_k(0)$, $k = 1, \ldots, K$.*

In the case the system is indexable, Whittle's index policy is a robust element of $\Pi^*$, see Section 5.1.2. For a non-indexable system, the set $\Pi^*$ might no longer contain a robust policy. In Section 8, we explain how to select in that case priority policies from the set $\Pi^*$.

Before continuing we first give an example of Definition 4.4.

**Example 4.6** *Assume $K = 2$ and $J_k = 2$. Let $x^*$ be such that for class 1 we have $x^{*,0}_{1,1} = 0$, $x^{*,0}_{2,1} = 4$, $x^{*,1}_{1,1} = 3$, $x^{*,1}_{2,1} = 1$ and for class 2 we have $x^{*,0}_{1,2} = 2$, $x^{*,0}_{2,2} = 0$, $x^{*,1}_{1,2} = 0$, $x^{*,1}_{2,2} = 5$ and $\alpha = 10$. The priority policies associated to $x^*$ in the set $\Pi(x^*)$, as defined in Definition 4.4, satisfy the following rules: By point 1.): class-1 bandits in state 1 and class-2 bandits in state 2 are given the highest priority. By point 3.): since $x^{*,1}_{1,1} + x^{*,1}_{2,1} + x^{*,1}_{1,2} + x^{*,1}_{2,2} = 9 < \alpha$, class-2 bandits in state 1 are never made active. Let the pair $(j,k)$ denote a class-$k$ bandit in state $j$. The set $\Pi(x^*)$ contains two policies: either give priority according to $(1,1) \succ (2,2) \succ (2,1)$ or give priority according to $(2,2) \succ (1,1) \succ (2,1)$. In neither policy, state $(1,2)$ is never made active.*

**Remark 4.7 (Multi actions)** *In this remark we explain how to define the set of priority policies $\Pi^*$ in the case of multiple actions per bandit. Similar to the non-restless bandit problem (see Remark 3.2), we are interested in priority policies such that if a class-$k$ bandit in state $j$ is chosen to be active, it will always be made active in a fixed mode $a_k(j) \in \{0, 1, 2, \ldots, A_k(j)\}$. We therefore need to restrict the set $X^*$ to optimal solutions of (LP) that satisfy $x^{*,a}_{j,k} x^{*,\tilde{a}}_{j,k} = 0$, for all $a, \tilde{a} \in \{1, \ldots, A_k(j)\}$. The latter condition implies that for all activation modes $a = 1, \ldots, A_k(j)$ one has $x^{*,a}_{j,k} = 0$, with the exception of at most one active mode, denoted by $a_k(j)$. The set $\Pi(x^*)$ is then defined as in Definition 4.4, replacing the action $a = 1$ by $a = a_k(j)$. All results obtained in Section 4 remain valid (replacing $a = 1$ by $a = a_k(j)$).*

**Remark 4.8** *In [9], a heuristic is proposed for the multi-class restless bandit problem for a fixed population of bandits: the so-called primal-dual heuristic. This is defined based on the optimal (primal and dual) solution of an LP problem corresponding to the discounted-cost criterion. In fact, if the primal-dual heuristic would have been defined based on the problem (LP), it can be checked that it satisfies the properties of Definition 4.4, and hence is included in the set of priority policies $\Pi^*$.*

In order to prove asymptotic optimality of a policy $\pi^* \in \Pi^*$, as will be done in Section 4.3, we investigate its fluid dynamics. Denote by $S^{\pi^*}_k(j)$ 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 $\pi^*$. Denote by $I^{\pi^*}$ the set of all states that will never be made active under policy $\pi^*$. The fluid dynamics under

policy $\pi^*$ can now be written as follows:

$$\frac{\mathrm{d}x_{j,k}^{\pi^*}(t)}{\mathrm{d}t} = \lambda_k p_k(j) + \sum_{a=0}^{1}\sum_{i=1}^{J_k} x_{i,k}^{\pi^*,a}(t) q_k(j|i,a), \tag{9}$$

$$\text{with } x_{j,k}^{\pi^*,1}(t) = \min\left(\left(\alpha - \sum_{(i,l)\in S_k^{\pi^*}(j)} x_{i,l}^{\pi^*}(t)\right)^+, \; x_{j,k}^{\pi^*}(t)\right), \quad \text{if } (j,k)\notin I^{\pi^*},$$

$$x_{j,k}^{\pi^*,1}(t) = 0, \quad \text{if } (j,k)\in I^{\pi^*},$$

$$x_{j,k}^{\pi^*,0}(t) = x_{j,k}^{\pi^*}(t) - x_{j,k}^{\pi^*,1}(t).$$

It follows directly that an optimal solution $x^*$ of (LP) is an equilibrium point of the process $x^{\pi^*}(t)$.

**Lemma 4.9** *Let $\pi^* \in \Pi^*$ and let $x^*$ be a point such that $\pi^* \in \Pi(x^*)$. Then $x^*$ is an equilibrium point of the process $x^{\pi^*}(t)$ as defined in* (9).

**Proof:** Since $x^*$ is an optimal solution of (LP), it follows directly from the definition of $\Pi(x^*)$ that $x^*$ is an equilibrium point of the process $x^{\pi^*}(t)$. □

In order to prove asymptotic optimality of a policy $\pi^*$, we will need that the equilibrium point $x^*$ is in fact a global attractor of the process $x^{\pi^*}(t)$, i.e., all trajectories converge to $x^*$. This is not true in general, which is why we state it as a condition for a policy to satisfy. In Section 6 we will further comment on this condition.

**Condition 4.10** *Given an equilibrium point $x^* \in X^*$ and a policy $\pi^* \in \Pi(x^*) \subset \Pi^*$. The point $x^*$ is a global attractor of the process $x^{\pi^*}(t)$. That is, for any initial point, the process $x^{\pi^*}(t)$ converges to $x^*$.*

## 4.3 Asymptotic optimality of priority policies

In this section we present the asymptotic optimality results for the set of priority policies $\Pi^*$. In particular, we obtain that the priority policies minimize the *fluid-scaled* average holding cost.

We will consider the restless bandit problem in the following fluid-scaling regime: we scale by $r$ both the arrival rates and the number of bandits that can be made active. That is, class-$k$ bandits arrive at rate $\lambda_k \cdot r$, $k = 1, \ldots, K$, and $\alpha \cdot r$ bandits can be made active at any moment in time. We let $X_{j,k}^r(0) = x_{j,k} \cdot r$, with $x_{j,k} \geq 0$. For a given policy $\pi$, we denote by $X_{j,k}^{r,\pi,a}(t)$ the number of class-$k$ bandits in state $j$ experiencing action $a$ at time $t$ under scaling parameter $r$.

We make the important observation that the set of policies $\Pi^*$ is invariant to the scaling parameter. This follows since an optimal solution of (LP) scales with the parameter $r$: if $x^*$ is an optimal solution, then so is $x^* r$ for the (LP) with parameters $\alpha \cdot r$, $x(0) \cdot r$ and $\lambda_k \cdot r$. By Definition 4.4, the set of priority policies does therefore not depend on $r$.

We will be interested in the process after the fluid scaling, i.e., space is scaled linearly with the parameter $r$, $\frac{X_{j,k}^{r,\pi,a}(t)}{r}$. We further define for a given initial state $x$,

$$V_-^{r,\pi}(x) := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}_{r\cdot x}\left(\int_0^T \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k(j,a)\frac{X_{j,k}^{r,\pi,a}(t)}{r}\mathrm{d}t\right),$$

and

$$V_+^{r,\pi}(x) := \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_{r\cdot x}\left(\int_0^T \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k(j,a)\frac{X_{j,k}^{r,\pi,a}(t)}{r}\mathrm{d}t\right).$$

If $V_-^{r,\pi}(x) = V_+^{r,\pi}(x)$ for all $x$, then we define $V^{r,\pi}(x) := V_+^{r,\pi}(x)$.

Our goal is to find policies that minimize the cost of the stochastic model after fluid scaling. We therefore call a policy $\pi^*$ asymptotically optimal when the fluid-scaled version of (1) holds.

**Definition 4.11 (Asymptotic optimality)** *A policy $\pi^*$ is asymptotically optimal if*

$$\limsup_{r\to\infty} V_+^{r,\pi^*}(x) \le \liminf_{r\to\infty} V_-^{r,\pi}(x), \quad \text{for all } x \text{ and all policies } \pi \in G,$$

*where $G$ is a set of admissible policies.*

In our asymptotic optimality result, the set $G$ will consist of all policies for the fixed population of bandits, while it will consists of all policies that are stable, rate-stable or mean rate-stable for the dynamic population of bandits, see Proposition 4.14.

In order to prove asymptotic optimality of priority policies in the set $\Pi^*$, we need the following technical condition.

**Condition 4.12** *Given a policy $\pi^* \in \Pi^*$.*

*a) The process $\frac{X^{r,\pi^*}(t)}{r}$ has a unique invariant probability distribution $p^{r,\pi^*}$, which has a finite first moment, for all $r$.*

*b) The family $\{p^{r,\pi^*}, r\}$ is tight.*

*c) The family $\{p^{r,\pi^*}, r\}$ is uniform integrable.*

For a fixed population of bandits, the state space of $X^{r,\pi^*}(t)$ is finite, hence Conditions b) and c) are satisfied. A sufficient condition for Condition 4.12 a) to hold is the Markov process $X^{r,\pi^*}(t)$ to be unichain, for any $r$, [43].

For a dynamic population of bandits we present a large class of restless bandit problems for which Condition 4.12 is satisfied. More precisely, we consider problems in which bandits that are kept passive will eventually leave the system. For many real-life situations this assumption arises naturally. For example, customers that become impatient and abandon the queue/system, companies that go bankrupt, perishable items, etc. The proof of the proposition may be found in Appendix C.

**Proposition 4.13** *Assume that the state 0 is positive recurrent for a class-$k$ bandit that is kept passive. For any priority policy $\pi$ for which $X^{r,\pi}(t)$ is irreducible, Condition 4.12 is satisfied.*

Another class of problems satisfying Condition 4.12 would be those in which only active bandits are allowed in the system, i.e., $q_k(0|i,0) = \infty$, for all $k, i$. This could describe for example the hiring process where new candidates are modeled by new arriving bandits, room occupation in a casualty departments where patients require direct attention, or a loss network. When $q_k(0|i,0) = \infty$, for all $k, i$, at most $\alpha$ bandits are present in the system, hence, due to the finite state space, Condition 4.12 follows directly from a unichain assumption.

We can now state the asymptotic optimality result.

**Proposition 4.14** *For a given policy $\pi^* \in \Pi(x^*) \subset \Pi^*$, assume Condition 4.10 and Condition 4.12 are satisfied. Then,*

$$\lim_{r\to\infty} V^{r,\pi^*}(x) = v^*(x), \quad \text{for any } x.$$

*In particular, we have*

$$\liminf_{r\to\infty} V_-^{r,\pi}(x) \ge \lim_{r\to\infty} V^{r,\pi^*}(x), \quad \text{for any } x \text{ and any policy } \pi \in G,$$

*where for the fixed population of bandits $G$ consists of all policies, and for the dynamic population of bandits*

- *$G$ is the set of all stable policies $\pi$, or,*
- *$C_k(j,a) > 0$, for all $j, k, a$, and $G$ is the set of all rate-stable and mean rate-stable policies.*

The proof may be found in Appendix D and consists of the following steps: Given a policy $\pi^* \in \Pi(x^*)$, we show that the fluid-scaled steady-state queue length vector converges to $x^*$. Since $x^*$ is an optimal solution of the fluid control problem (LP) with $x(0) = x$ and has cost value $v^*(x)$, this implies that the fluid-scaled cost under policy $\pi^*$ converges to $v^*(x)$. Since $v^*(x)$ serves as a lower bound on the average cost, this allows us to conclude for asymptotic optimality of the priority policy $\pi^*$.

# 5 Whittle's index policy

In Section 4.3 we showed that priority policies inside the set $\Pi^*$ are asymptotically optimal. In this section we will derive that Whittle's index policy is included in this set of policies $\Pi^*$.

In Section 5.1, we first define Whittle's index policy. In Section 5.2 and Section 5.3, we then give sufficient conditions under which Whittle's index policy is asymptotically optimal, both in the case of a fixed population of bandits, and in the case of a dynamic population of bandits, respectively.

## 5.1 Relaxed-constraint optimization problem and Whittle's indices

Whittle's index policy was proposed by Whittle [50] as an efficient heuristic for the multi-class restless bandit problem. Each bandit is assigned a Whittle's index, which is a function of the state the bandit is in. Whittle's index policy activates those bandits having currently the highest indices. In this section, we will describe how these Whittle's indices are derived.

In order to define Whittle's indices, we consider the following optimization problem: find a stationary and Markovian policy that minimizes

$$\mathbb{C}_x^f \left( \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} C_k(j,a) X_{j,k}^{\pi,a}(\cdot) \right), \text{ with } f \in \{av, \beta\}, \tag{10}$$

under the constraint (2), where

$$\mathbb{C}_x^{av}(Y(\cdot)) := \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_x \left( \int_0^T Y(t) \mathrm{d}t \right), \tag{11}$$

represents the average-cost criterion and

$$\mathbb{C}_x^{\beta}(Y(\cdot)) := \mathbb{E}_x \left( \int_0^\infty \mathrm{e}^{-\beta t} Y(t) \mathrm{d}t \right),$$

$\beta > 0$, represents the discounted-cost criterion. The objective as stated in Section 3 is the average-cost criterion. In Section 5.3, it will become clear why we need to introduce here the discounted-cost criterion as well.

### 5.1.1 Relaxed-constraint optimization problem

The restless property of the bandits makes the above-described optimization problem often infeasible to solve. Instead, Whittle [50] proposed to study the so-called *relaxed-constraint optimization problem*, which is defined as follows: find a policy that minimizes (10) under the relaxed constraint

$$\mathbb{C}_x^f \left( \sum_{k=1}^{K} \sum_{j=1}^{J_k} X_{j,k}^{\pi,1}(\cdot) \right) \leq \alpha(f), \tag{12}$$

with $\alpha(av) = \alpha$ and $\alpha(\beta) = \int_0^\infty \alpha \mathrm{e}^{-\beta t} \mathrm{d}t = \alpha/\beta$ for $\beta > 0$. That is, the constraint that at most $\alpha$ bandits can be made active at any moment in time is replaced by its time-average or discounted version, (12). Hence, the cost under the optimal policy of the relaxed-constraint optimization problem provides a lower bound on the cost for any policy that satisfies the original constraint.

In standard restless bandit problems, the constraint (12) needs to be satisfied in the strict sense, that is, with an "=" sign. In this paper we allow however strictly less than $\alpha$ bandits to be active at a time. In order to define Whittle's indices we therefore introduce so-called *dummy bandits*. That is, besides the initial population of bandits, we assume there are $\alpha(f)$ additional bandits that will never change state. We denote the state these bandits are in by $B$ and the cost of having a dummy bandit in state $B$ is $C_B(a) = 0$, $a = 0, 1$. The introduction of these $\alpha(f)$ dummy bandits allows to reformulate the relaxed-constraint problem as follows: minimize (10) under the relaxed constraint

$$\mathbb{C}_x^f \left( X_B^{\pi,1}(\cdot) \right) + \mathbb{C}_x^f \left( \sum_{k=1}^{K} \sum_{j=1}^{J_k} X_{j,k}^{\pi,1}(\cdot) \right) = \alpha(f). \tag{13}$$

This constraint is equivalent to (12) since, for a given set of active bandits, activating additional dummy bandits does not modify the behavior of the system.

Using the Lagrangian approach, we write the relaxed-constraint problem (minimize (10) under constraint (13)) as the problem of finding a policy $\pi$ that minimizes

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k}\mathbb{C}_x^f\left(C_k(j,0)X_{j,k}^{\pi,0}(\cdot)+C_k(j,1)X_{j,k}^{\pi,1}(\cdot)+\nu X_{j,k}^{\pi,1}(\cdot)\right)+\mathbb{C}_x^f\left(\nu X_B^{\pi,1}(\cdot)\right). \tag{14}$$

The Lagrange multiplier $\nu$ can be viewed as the cost to be paid per active bandit. From Lagrangian relaxation theory we have that there exists a value of the Lagrange multiplier $\nu$ such that the constraint (13) is satisfied.

Since there is no longer a common constraint for the bandits, problem (14) can be decomposed into several *subproblems*, one for each bandit: for each class-$k$ bandit the subproblem is to minimize

$$\mathbb{C}^f\left(C_k(J_k(\cdot),A_k^\pi(\cdot))+\nu\mathbf{1}_{(A_k^\pi(\cdot)=1)}\right), \tag{15}$$

where $J_k(t)$ 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 $\pi$. We take as convention that $J_k(t)=0$ and $A_k(t)=0$ if the bandit is not present (or no longer present) in the system at time $t$ and set $C_k(0,0)=0$. For each dummy bandit the problem is to minimize

$$\nu\ \mathbb{C}^f\left(\mathbf{1}_{(A_B^\pi(\cdot)=1)}\right), \tag{16}$$

with $A_B^\pi(t)$ the action chosen for the dummy bandit at time $t$ under policy $\pi$.

We can now define Whittle's index.

**Definition 5.1 (Whittle's index)** *For a given optimization criterion $f$, we define Whittle's index $\nu_k^f(j)$ as the least value of $\nu$ for which it is optimal in* (15) *to make the class-$k$ bandit in state $j$ passive.*

*Similarly, we define the index $\nu_B^f$ as the least value of $\nu$ for which it is optimal in* (16) *to make a dummy bandit passive.*

Indexability is the property that allows to characterize an optimal policy for the relaxed optimization problem.

**Definition 5.2 (Indexability)** *A bandit is* indexable *if the set of states in which passive is an optimal action in* (15)*, denoted by $D(\nu)$, increases in $\nu$. That is, $\nu'<\nu$ implies $D(\nu')\subset D(\nu)$.*

We note that the dynamics of a bandit in state $B$ is independent of the action chosen. Since $\nu$ represents the cost to be paid when active, it will be optimal in (16) to make a bandit in state $B$ passive if and only if $\nu\geq 0$. As a consequence, a dummy bandit is always indexable and $\nu_B^f=0$.

We call the problem indexable if all bandits are indexable. Note that whether or not a problem is indexable can depend on the choice for $f$ (and $\beta$). We refer to [35] for a survey on indexability results. In particular, [35] presents sufficient conditions for a restless bandit to be indexable and provides a method to calculate Whittle's indices. Sufficient conditions for indexability can also be found in [30, 45].

If the bandit problem is indexable, an optimal policy for the subproblem (15) is then such that the class-$k$ bandit in state $j$ is made active if $\nu_k^f(j)>\nu$, is made passive if $\nu_k^f(j)<\nu$, and any action is optimal if $\nu_k^f(j)=\nu$, [50].

An optimal solution to (10) under the relaxed constraint (13) is obtained by setting $\nu$ at the appropriate level $\nu^*$ such that (13) is satisfied. A class-$k$ bandit in state $j$ is then made active if $\nu_k^f(j)>\nu^*$, and kept passive if $\nu_k^f(j)<\nu^*$. When a class-$k$ bandit is in a state $j$ such that $\nu_k^f(j)=\nu^*$, one needs to appropriately randomize the action in this state such that the relaxed constraint (13) is satisfied, [46, 50]. In the case $\nu^*=0$, we take the convention that the randomization is done among the bandits in state $B$ (possible since there are exactly $\alpha(f)$ dummy bandits), while any class-$k$ bandit in a state $j$ with $\nu_k^f(j)=0$ is kept passive.

Since $\nu_B^f=0$, a dummy bandit has higher priority than a class-$k$ bandit in state $j$ with $\nu_k^f(j)\leq 0$. Together with constraint (13) and the fact that there are $\alpha(f)$ dummy bandits, we conclude that any class-$k$ bandit in state $j$ with $\nu_k^f(j)\leq 0$ is kept passive in the relaxed optimization problem. In particular, this implies that $\nu^*\geq 0$.

### 5.1.2 Whittle's index policy as heuristic

The optimal control for the relaxed problem is not feasible for the original optimization problem having as constraint that at most $\alpha$ bandits can be made active *at any moment in time*. Whittle [50] therefore proposed the following heuristic:

**Definition 5.3 (Whittle's index policy)** *For a given optimization criterion $f$, Whittle's index policy activates the $\alpha$ bandits having currently the highest* non-negative *Whittle's index value $\nu_k^f(j)$. In case different states have the same value for the Whittle index, an arbitrary fixed priority rule is used. We denote Whittle's index policy by $\nu^f$.*

If $\nu_k^f(j) < \nu_l^f(i)$, then a class-$l$ bandit in state $i$ is given higher priority than a class-$k$ bandit in state $j$ under Whittle's index policy. Analogously to the optimal solution of the relaxed optimization problem, a class-$k$ bandit in state $j$ with $\nu_k^f(j) \leq 0$ will never be made active under Whittle's index policy. It can therefore happen that strictly less than $\alpha$ bandits are made active, even though there are more than $\alpha$ bandits present.

Whittles indices result from solving (15). Since the latter does not depend on $\alpha$, $\lambda_k$, and $X_k(0)$, we can conclude that Whittle's index policy is a *robust* policy, see Definition 4.5. In the next two sections we will prove that Whittle's index policy is asymptotically optimal, both for the static and dynamic population.

**Remark 5.4 (Multi actions)** *In this remark we define Whittle's index policy in the case of multiple actions. For that we need to assume a stronger form of indexability: there is an index $\nu_k^f(j)$ and an activation mode $a_k^f(j)$ such that an optimal solution of* (15) *is to make a class-k bandit in state $j$ active in mode $a_k^f(j)$ if $\nu < \nu_k^f(j)$ and to keep it passive if $\nu > \nu_k^f(j)$. Whittle's index rule is then defined as in Section 5.1.1, replacing the action $a = 1$ by $a = a_k^f(j)$.*

*If the restless bandit problem satisfies this stronger form of indexability, then one can reduce the multi-action problem to the single-action problem and hence all asymptotic optimality results as obtained in Section 5.2 and Section 5.3 will be valid (replacing action $a = 1$ by $a = a_k(j)$).*

## 5.2 Asymptotic optimality for a fixed population of bandits

In this section we consider a fixed population of indexable bandits and show that Whittle's index policy, defined for the time-average cost criterion $f = av$, is asymptotically optimal.

We will need the following assumption, which was also made in [46].

**Assumption 5.5** *For every $k$, the process describing the state of a class-k bandit is unichain, regardless of the policy employed.*

The next proposition shows that Whittle's index policy is included in the set of priority policies $\Pi^*$. The proof can be found in Appendix E.

**Proposition 5.6** *Consider a fixed population of bandits. If Assumption 5.5 holds and if the restless bandit problem is indexable for the average-cost criterion, then there is an $x^* \in X^*$ such that Whittle's index policy $\nu^{av}$ is included in the set $\Pi(x^*) \subset \Pi^*$.*

We can now conclude that Whittle's index policy is asymptotically optimal.

**Corollary 5.7** *Consider a fixed population of bandits. If the assumptions of Proposition 5.6 are satisfied and if Condition 4.10 holds for Whittle's index policy $\nu^{av}$, then*

$$\lim_{r\to\infty} V^{r,\nu^{av}}(x) \leq \liminf_{r\to\infty} V_-^{r,\pi}(x),$$

*for any $x$ and any policy $\pi$.*

**Proof:** From Proposition 4.14 and Proposition 5.6 we obtain the desired result. □

The above corollary was previously proved by Weber and Weiss in [46] for the case of symmetric bandits,

i.e., $K = 1$. We note that the assumptions made in [46] in order to prove the asymptotic optimality result are the same as the ones in Corollary 5.7.

The proof technique used in Weber and Weiss [46] is different from the one used here. In [46] the cost under an optimal policy is lower bounded by the optimal cost in the relaxed problem and upper bounded by the cost under Whittle's index policy. By showing that both bounds converge to the same value, the fluid approximation, the asymptotic optimality of Whittle's index policy is concluded. Obtaining a lower bound for a dynamic population does not seem straightforward. This is why we undertook in this paper a different proof approach that applies as well for a dynamic population, see Section 5.3.

## 5.3 Asymptotic optimality for a dynamic population of bandits

In this section we will introduce an index policy for the dynamic population of bandits, based on Whittle's indices, and show it to be asymptotically optimal. More precisely, we show the index policy to be included in the set of asymptotically optimal policies $\Pi^*$, as obtained in Section 4.3.

Recall that our objective is to find a policy that asymptotically minimizes the average-cost criterion (11). We do however not make use of Whittle's index policy $\nu^{av}$ for the following reason: Consider a class-$k$ bandit and the relaxed optimization problem (15), with $f = av$. Any policy that makes sure that the class-$k$ bandit leaves after a finite amount of time has an average cost equal to zero and is hence an optimal solution. In order to derive a non-trivial index rule, the authors of [5, 7] consider instead the Whittle indices corresponding to the discounted-cost criterion ($f = \beta$, $\beta > 0$). An index rule for the average-cost criterion is then obtained by considering the limiting values as $\beta \downarrow 0$. We propose here the same. For a given class $k$, let $\beta_l \downarrow 0$ be some subsequence such that the limit

$$\nu_k^{lim}(j) := \lim_{l\to\infty} \nu_k^{\beta_l}(j)$$

exists, for all $j = 1, \ldots, J_k$. The limit can possibly be equal to $\infty$. The index policy $\nu^{lim}$ activates the $\alpha$ bandits having currently the highest *non-negative* index value $\nu_k^{lim}(j)$. In this section we will show asymptotic optimality of $\nu^{lim}$. In order to do so, we will need that class-$k$ bandits are indexable under the $\beta_l$-discounted cost criterion, for $l$ large enough. In addition, we will need the following assumption on the model parameters.

**Assumption 5.8** *For all $k = 1, \ldots, K$, the set of optimal solutions of the linear program*

$$\begin{aligned} \min_x \sum_{j=1}^{J_k} (C_{j,k}^0 x_{j,k}^0 + C_{j,k}^1 x_{j,k}^1 + \nu x_{j,k}^1) \\ s.t. \;\; 0 = \lambda_k p_k(0,j) + \sum_{a=0}^{1} \sum_{i=1}^{J_k} x_{i,k}^a q_k(j|i,a), \quad \forall\, j, \\ x_{j,k}^a \geq 0, \;\forall\, j, a, \end{aligned}$$

*is bounded when $\nu > 0$.*

We note that this assumption is always satisfied if $C_k(j,0) > 0$ and $C_k(j,1) \geq 0$, for all $j, k$, since $x_{j,k}^{*,1}$ and $x_{j,k}^{*,0}$ are upperbounded by the cost value of a feasible solution divided by $\nu + C_k(j,1) > 0$ and $C_k(j,0) > 0$, respectively.

The proposition below shows that Whittle's index policy $\nu^{lim}$ is included in the set of priority policies $\Pi^*$. The proof can be found in Appendix E.

**Proposition 5.9** *Consider a dynamic population of bandits. For a given class $k$, let $\beta_l \downarrow 0$ be some subsequence such that the limit*

$$\nu_k^{lim}(j) := \lim_{l\to\infty} \nu_k^{\beta_l}(j)$$

*exists, for all $j = 1, \ldots, J_k$. If Assumption 5.8 holds and if the discounted restless bandit problem is indexable for $\beta_l \leq \overline{\beta}$, with $0 < \overline{\beta} < 1$, then there is an $x^* \in X^*$ such that Whittle's index policy $\nu^{lim}$ is included in the set $\Pi(x^*) \subset \Pi^*$.*

We can now conclude for asymptotic optimality of Whittle's index policy $\nu^{lim}$.

**Corollary 5.10** *Consider a dynamic population of bandits. If the assumptions of Proposition 5.9 are satisfied and if Condition 4.10 and Condition 4.12 hold for Whittle's index policy $\nu^{lim}$, then*

$$\lim_{r\to\infty} V^{r,\nu^{lim}}(x) \leq \liminf_{r\to\infty} V_{-}^{r,\pi}(x), \quad \text{for all } x \text{ and any policy } \pi \in G, \tag{17}$$

*where*

- *$G$ consists of all stable policies, or*
- *$C_k(j,a) > 0$, for all $j,k,a$, and $G$ consists of all rate-stable and mean rate-stable policies.*

**Proof:** The result follows directly from Proposition 4.14 and Proposition 5.9. □

The above result for the dynamic population shows that the heuristic $\nu^{lim}$, which is based on a model without arrivals, is in fact nearly optimal in the presence of arrivals. In addition, Whittle's index policy $\nu^{lim}$ is robust, that is, it does not depend on the arrival characteristics of new bandits or on the exact number of bandits that can be made active.

**Remark 5.11 (Multi actions)** *In order to define $\nu^{lim}$ in the case of multiple actions per bandit, we need to assume that, for $\beta_l$ small enough, the stronger form of indexability (defined in Remark 5.4) holds. In addition, the optimal activation mode for a class-k bandit in state $j$, denoted by $a_k^{\beta_l}(j)$, cannot depend on $\beta_l$, i.e., $a_k^{\beta_l}(j) = a_k(j)$.*

# 6 On the global attractor property

In Proposition 4.14, asymptotic optimality of priority policies in the set $\Pi^*$ was proved under the global attractor property (Condition 4.10) and a technical condition (Condition 4.12). In this section we further discuss the global attractor property. The latter is concerned with the process $x^*(t)$, defined by the ODE (9), to have a global attractor. We recall that in [46] the same global attractor property was required in order to prove asymptotic optimality of Whittle's index policy for a fixed population of symmetric bandits ($K = 1$). In addition, the authors of [46] presented an example for which Whittle's index policy is not asymptotically optimal (and hence, does not satisfy the global attractor property.)

For a fixed population of symmetric indexable bandits, the global attractor property was proved to always hold under Whittle's index policy if a bandit can be in at most three states ($J = 3$), see [47]. However, in general no sufficient conditions are available in order for $x^*$ to be a global attractor of $x^{\pi^*}(t)$. A necessary condition was provided in [46, Lemma 2], where for a fixed population of symmetric bandits it was proved that indexability is necessary in order for Whittle's index policy to satisfy the global attractor property, for any value of $\alpha$ and $x(0)$. We emphasize that when the system is non-indexable, there can still exist priority policies in $\Pi^*$ (possibly non-robust) that satisfy the global attractor property.

The asymptotic optimality result of Whittle's index policy for the case $K = 1$, [46], has been cited extensively. The global attractor property is often verified only numerically. Note that in the context of mean field interaction models, convergence of the stationary measure also relies on a global attractor assumption of the corresponding ODE, see for example [8]. In a recent paper, the authors of [37] proved asymptotic optimality of Whittle's index policy for a very specific model with only two classes of bandits (fixed population of bandits) under a recurrence condition. The latter condition replaced the global attractor condition, however, the authors needed as well to resort to numerical experiments in order to verify this recurrence condition.

In the remainder of this section we describe the necessity of the global attractor property and the technical challenges in the case this condition is not satisfied.

Optimal fluid control problems have been widely studied in the literature in order to obtain asymptotically optimal policies for the stochastic model. In the context of this paper, the fluid control problem related to our results would be to find the optimal control $u^*(t)$ that minimizes

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} C_k(j,a) x_{j,k}^{u,a}(t) \mathrm{d}t, \tag{18}$$

where the dynamics of $x_{j,k}^{u,a}(t)$ is described by (3). The optimal control $u^*(t)$ is then to be translated back to the stochastic model in such a way that it is asymptotically optimal. When stating the global

attractor property, the above is exactly what we have in mind. In fact, instead of solving this transient fluid control problem, we directly consider an optimal equilibrium point of the fluid model and propose a priority policy based on this equilibrium point. When the global attractor property is satisfied, this implies that the optimal equilibrium point is indeed reached by the associated strict priority control, and hence this priority control solves (18).

When for any $\pi^* \in \cup_{x^* \in X^*} \Pi(x^*) = \Pi^*$ the global attractor property is not satisfied, this means that there does not exist a priority control $u(t) = \pi^* \in \Pi(x^*)$ such that the fluid process $x^{\pi^*}(t)$ converges to $x^*$. In that case, we can be in either one of the following two situations: 1) There exists a control $u^*(t)$ for which the process $x^{u^*}(t)$ does have as global attractor $x^* \in X^*$, where $X^*$ was defined as the set of optimal equilibrium points. This control $u^*(t)$ might not be of priority type. 2) There does not exist any control that has a global attractor $x^* \in X^*$. In the latter case, the optimal control $u^*(t)$ can be such that the process $x^{u^*}(t)$ behaves cyclically or shows chaotic behavior, or the process converges to a non-optimal equilibrium point. Hence, in the case Condition 4.10 is not satisfied, in both Situation 1) and Situation 2), one needs to determine the exact transient behaviour of the optimal control of (18), $u^*(t)$, which in its turn needs to be translated back to the stochastic model. We leave this as subject for future research.

# 7 Case study: a multi-server queue with abandonments

In this section we study a multi-class multi-server system with impatient customers, the multi-class *M/M/S+M* system. This is an example of a restless bandit problem with a dynamic population. We will derive a robust priority policy that is in the set $\Pi^*$ and show that it satisfies the two conditions needed in order to conclude for asymptotic optimality.

The impact of abandonments has attracted considerable interest from the research community, with a surge in recent years. To illustrate the latter, we can mention the recent Special Issue on abandonments in Queueing Systems [24] and the survey paper [13] on abandonments in a many-server setting

We consider a multi-class system with $S$ servers working in parallel. At any moment in time, each server can serve at most one customer. Class-$k$ customers arrive according to a Poisson process with rate $\lambda_k > 0$ and require an exponentially distributed service with mean $1/\mu_k < \infty$. Server $s$, $s = 1, \ldots, S$ works at speed 1. Customers waiting (being served) abandon the queue after an exponentially distributed amount of time with mean $1/\theta_k$ ($1/\tilde{\theta}_k$), with $\theta_k > 0$, $\tilde{\theta}_k \geq 0$, for all $k$. Having one class-$k$ customers waiting in the queue (in service) costs $c_k$ ($\tilde{c}_k$) per unit of time. Each abandonment of a waiting class-$k$ customer (class-$k$ customer being served) costs $d_k$ ($\tilde{d}_k$). We are interested in finding a policy $\pi$ that minimizes the long-run average cost

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{k=1}^{K} \mathbb{E}_x \left( \int_0^T \left( c_k X_k^{\pi,0}(t) + \tilde{c}_k X_k^{\pi,1}(t) \right) \mathrm{d}t + d_k R_k^{\pi}(T) + \tilde{d}_k \tilde{R}_k^{\pi}(T) \right),$$

where $X_k^{\pi,0}(t)$ ($X_k^{\pi,1}(t)$) denotes the number of class-$k$ customers in the queue (in service) at time $t$ and $R_k^{\pi}(t)$ ($\tilde{R}_k^{\pi}(t)$) denotes the number of abandonments of waiting class-$k$ customers (class-$k$ customers being served) in the interval $[0, t]$.

Representing each customer in the queue (in service) by a passive (active) bandit, the problem can be addressed within the framework of a restless bandit model with the following parameters: $J_k = 1, q_k(0|1,0) = \theta_k > 0, q_k(0|1,1) = \mu_k + \tilde{\theta}_k, C_k(1,0) = c_k + d_k\theta_k, C_k(1,1) = \tilde{c}_k + \tilde{d}_k\tilde{\theta}_k$, $k = 1, \ldots, K$, and $\alpha = S$, where we used that $\mathbb{E}_x(R_k^{\pi}(T)) = \theta_k \mathbb{E}_x(\int_0^T X_k^{\pi,0}(t)\mathrm{d}t)$ and $\mathbb{E}_x(\tilde{R}_k^{\pi}(T)) = \tilde{\theta}_k \mathbb{E}_x(\int_0^T X_k^{\pi,1}(t)\mathrm{d}t)$. A bandit can only be in two states (state 0 or state 1), hence indexability follows directly (for any choice of $\beta$).

We now define an index policy that we will prove to be included in the set $\Pi^*$. For each class $k$ we set:

$$\iota_k := q_k(0|1,1) \left( \frac{C_k(1,0)}{q_k(0|1,0)} - \frac{C_k(1,1)}{q_k(0|1,1)} \right).$$

The index policy $\iota$ is then defined as follows: At any moment in time serve (at most) $S$ customers present in the system that have the highest, strictly positive, index values, $\iota_k$. If a customer belongs to a class that has a negative index value, then this customer will never be served.

Before continuing, we first give an interpretation of the index $\iota_k$. The term $1/q_k(0|1,a)$ is the time it takes until a bandit under action $a$ leaves the system. Hence, $C_k(1,a)/q_k(0|1,a)$ is the cost for applying

action $a$ on a class-$k$ bandit until it leaves the system. The difference $\frac{C_k(1,0)}{q_k(0|1,0)} - \frac{C_k(1,1)}{q_k(0|1,1)}$ is the reduction in cost when making a class-$k$ bandit active (instead of keeping him passive), so that the index $\iota_k$ represents the reduction in cost per time unit when class $k$ is made active. Also note that the index rule $\iota$ does not depend on the arrival rate of the customers or the number of servers present in the system, hence it is a robust rule, see Definition 4.5.

By solving the LP problem corresponding to the multi-server queue with abandonments, we obtain in Proposition 7.1 that the index policy $\iota$ is included in $\Pi^*$.

**Proposition 7.1** *Policy $\iota$ is contained in the set $\Pi^*$.*

*In addition, when $\iota_1 > \iota_2 > \cdots > \iota_K$, policy $\iota$ coincides with Whittle's index policy $\nu^{lim}$.*

**Proof:** For the multi-class multi-server system with abandonments, the linear program (LP) is given by:

$$\begin{aligned} &\min_x \sum_k (c_k x_k^0 + \tilde{c}_k x_k^1 + d_k \theta_k x_k^0 + \tilde{d}_k \tilde{\theta}_k x_k^1), \\ &\text{s.t. } 0 = \lambda_k - \mu_k x_k^1 - \theta_k x_k^0 - \tilde{\theta}_k x_k^1, \\ &\quad \sum_{k=1}^K x_k^1 \leq S, \text{ and } x_k^0, x_k^1 \geq 0. \end{aligned} \tag{19}$$

Equation (19) implies $x_k^0 = \frac{\lambda_k - (\mu_k + \tilde{\theta}_k) x_k^1}{\theta_k}$. Hence, the above linear program is equivalent to solving

$$\begin{aligned} &\max_x \sum_k \left( (c_k + d_k \theta_k) \frac{\mu_k + \tilde{\theta}_k}{\theta_k} - \tilde{c}_k - \tilde{d}_k \tilde{\theta}_k \right) x_k^1, \\ &\text{s.t. } \sum_{k=1}^K x_k^1 \leq S, \quad \text{and} \quad 0 \leq x_k^1 \leq \frac{\lambda_k}{\mu_k + \tilde{\theta}_k}. \end{aligned}$$

The optimal solution is to assign maximum values to those $x_k^1$ having the highest values for $\iota_k = (c_k + d_k \theta_k) \frac{\mu_k + \tilde{\theta}_k}{\theta_k} - \tilde{c}_k - \tilde{d}_k \tilde{\theta}_k$, with $\iota_k > 0$, until the constraint $\sum_k x_k^1 \leq S$ is saturated. Denote this optimal solution by $x^*$. Assume the classes are ordered such that $\iota_1 \geq \iota_2 \geq \cdots \geq \iota_K$. Hence, one can find an $l$ such that: (1) for all $k < l$ it holds that $x_k^{*,1} = \frac{\lambda_k}{\mu_k + \theta_k}$, and hence $x_k^{*,0} = 0$, (2) for $k = l$ it holds that $0 \leq x_l^{*,1} \leq \frac{\lambda_l}{\mu_l + \theta_l}$ and hence $x_l^{*,0} \geq 0$, (3) and for all $k > l$ it holds that $x_k^{*,1} = 0$. This gives that the index policy $\iota$ is included in the set $\Pi(x^*) \subset \Pi^*$, see Definition 4.4.

When $\iota_1 > \iota_2 > \cdots > \iota_K$ it follows directly that $\iota$ is the unique policy that is in the set $\Pi^*$ for any value of $S$ or $\lambda_k$. Since Whittle's index policy is by definition robust and is in the set $\Pi^*$ (Proposition 5.9), we obtain that $\iota$ and Whittle's index policy have the same priority ordering. □

Note that the *M/M/S+M* system belongs to the class of problems as described in Proposition 4.13. Hence, Condition 4.12 is satisfied. The global attractor property follows from [3], where this property was proved for a slightly different model. We therefore have the following optimality result for the index policy $\iota$.

**Proposition 7.2** *Consider a system with $Sr$ servers working in parallel and arrival rates $\lambda_k r$, $k = 1, \ldots, K$. The index policy $\iota$ is asymptotically optimal as $r \to \infty$, i.e., for any $x$ and any policy $\pi$,*

$$\begin{aligned} &\lim_{r\to\infty} \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_x \left( \int_0^T \sum_{k=1}^K \left( (c_k + d_k \theta_k) \frac{X_k^{r,\iota,0}(t)}{r} + (\tilde{c}_k + \tilde{d}_k \tilde{\theta}_k) \frac{X_k^{r,\iota,1}(t)}{r} \right) \mathrm{d}t \right) \\ &\leq \liminf_{r\to\infty} \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}_x \left( \int_0^T \sum_{k=1}^K \left( (c_k + d_k \theta_k) \frac{X_k^{r,\pi,0}(t)}{r} + (\tilde{c}_k + \tilde{d}_k \tilde{\theta}_k) \frac{X_k^{r,\pi,1}(t)}{r} \right) \mathrm{d}t \right). \end{aligned}$$

**Proof:** In Proposition 7.1 we showed that $\iota$ is included in $\Pi(x^*)$, with $x^*$ as given in the proof of Proposition 7.1. In Appendix H we prove that the process $x^\iota(t)$, as defined in (9), has the point $x^*$ as a unique global attractor, i.e., Condition 4.10 is satisfied. From Proposition 4.13 we obtain that Condition 4.12 is satisfied. Further, note that any policy $\pi$ gives a stable system, since $\theta_k > 0$ for all $k$. Together with Proposition 4.14 we then obtain that the index policy $\iota$ is asymptotically optimal. □

**Remark 7.3 (Existing results in literature)** *In [29], a single-server queue with abandonments has been studied. Whittle's index policy was there derived by modeling the system as a fixed population of restless bandits: each bandit representing a class and the state of a bandit representing the number of customers in the queue. The latter implies that $J_k = \infty$, for all $k$, hence it does not fall inside the framework of this paper. The results obtained in [29] apply to general holding cost functions. In the case of linear holding costs, as considered in this section, the index rule as derived in [29] coincides with policy $\iota$. We further note that even though in [29] the arrival characteristics are taken into account when calculating Whittle's indices, the final result is independent on the arrival characteristics. For non-linear holding cost, this is no longer the case.*

*For the case $\tilde{c}_k = 0, \tilde{\theta}_k = 0$ and $c_k + d_k\theta_k > 0$, the asymptotic optimality of the policy $\iota$ in a multi-server setting has previously been proved in [3, 4]. Note that in this setting the performance criterion is the weighted number of customers present in the queue. If $\sum \lambda_k/\mu_k > S$, that is, the overload situation, the fluid-scaled cost $v^*(S)$ will be non-zero, and hence the optimality result is useful. This is not the case when $\sum \lambda_k/\mu_k < S$, the underload setting, as was also observed in [3, 4]: in underload we have for any non-idling policy $x_k^{*,0} = 0$, $\forall\ k$, see Equation (46), which together with $\tilde{c}_k = 0$ implies $v^*(S) = 0$, that is, in equilibrium the cost is zero for any non-idling policy. In [28] the* transient behavior *of the fluid model has been studied for the underload setting. It was shown that the optimal transient fluid control is in fact a state-dependent strategy and hence no longer a strict priority policy.*

*For a discrete-time model with one server and $\tilde{\theta}_k = 0, \tilde{c}_k = c_k > 0$, Whittle's index $\nu_k^{lim}$ has been derived in [7]. This index $\nu_k^{lim}$ coincides with the Whittle's index $\iota_k$ for the continuous-time model. In this setting, the fluid-scaled cost is always strictly positive: $v^*(S) = 0$ would imply that $x_k^* = 0$, however this contradicts with Equation (19), which would read $0 = \lambda_k$. Hence the asymptotic optimality result applies to both the underload and overload regime.*

# 8 Non-indexable restless bandits

The set of priority policies, $\Pi^*$, consists of more than one policy, and hence, it is not direct which priority policy to choose. For an indexable restless bandit problem, Whittle's index policy is inside the set $\Pi^*$ and is robust, that is, it does not depend on $\alpha, \lambda_k, X_k(0)$, $k = 1, \ldots, K$. This is therefore an obvious choice, and Whittle's index policy has been extensively tested numerically for different applications and shown to perform well, see for example [1, 2, 7, 6, 14, 21, 22, 29, 30, 36, 40] and the examples in the book [19]. In this section we therefore focus our attention on non-indexable restless bandits. In Section 8.1 we describe how to select a priority policy from the (possibly large) set of priority policies $\Pi^*$ and in Section 8.2 their performance is numerically evaluated outside the asymptotic regime.

## 8.1 Policy selection

In this section we describe how to select priority policies from the set $\Pi^*$. In order to do so, we will need the following technical lemma that gives a characterization of an optimal solution of the (LP) problem. Note that this lemma is valid for both indexable and non-indexable examples. We refer to Appendix G for the proof.

**Lemma 8.1** *In the case of a dynamic population of bandits, assume that the set of optimal solutions of (LP) is bounded and either $p_k(j) > 0$, for all $j, k$, or $C_k(j, 0) > 0$, for all $j, k$.*

*For either a fixed or dynamic population of bandits, there exists at least one optimal solution of (LP), $x^*$, such that $x_{j,k}^{*,0} x_{j,k}^{*,1} > 0$ for at most one pair $(j, k)$.*

The assumption that the set of optimal solutions of (LP) is bounded is always satisfied if $C_k(j, 0) > 0$, for all $j, k$. This follows since $x_{j,k}^{*,1} \leq \alpha$ and $x_{j,k}^{*,0} \leq \left(\overline{C} - \sum_{j,k} C_k(j,1) x_{j,k}^{*,1}\right) / C_k(j,0) < \infty$, with $\overline{C} < \infty$ the cost value of a feasible solution.

In the remainder of this section we will write $\Pi^*(\alpha)$ instead of $\Pi^*$ to emphasize the dependence on $\alpha$, that is, the number of bandits that can be simultaneously made active. In the case of indexable bandits, there exists priority policies that are inside $\Pi^*(\alpha)$, for all $\alpha$, for example Whittle's index policy. In general, this is not the case for non-indexable bandits. Below we therefore describe how one can select priority policies from the set $\Pi^*(\alpha)$ as $\alpha$ changes.

From Lemma 8.1, we have that, for a fixed $\alpha$, there exists at least one optimal solution of (LP), $x^*(\alpha)$, such that $x_{j,k}^{*,0}(\alpha) x_{j,k}^{*,1}(\alpha) > 0$, for at most one pair $(j, k)$. In particular, we can define $0 = \alpha_0 < \alpha_1 <$

$\alpha_2 < \ldots < \alpha_M$ and $\alpha_{M+1} = \infty$, such that for a given interval $[\alpha_i, \alpha_{i+1})$ the binding constraints of the (LP) and the basis of an optimal solution do not change. Hence, there are pairs $(j_i, k_i)$ and sets $H_i, L_i$ and $\tilde{L}_i$ such that, for any $\alpha \in [\alpha_i, \alpha_{i+1})$, it holds that

$$\begin{aligned}
&x^{*,0}_{j,k}(\alpha) = 0 \text{ and } x^{*,1}_{j,k}(\alpha) \geq 0, \text{ for all } (j,k) \in H_i,\\
&x^{*,0}_{j_i,k_i}(\alpha) \geq 0 \text{ and } x^{*,1}_{j_i,k_i}(\alpha) \geq 0,\\
&x^{*,0}_{j,k}(\alpha) \geq 0 \text{ and } x^{*,1}_{j,k}(\alpha) = 0, \text{ for all } (j,k) \in L_i,\\
&x^{*,0}_{j,k}(\alpha) = 0 \text{ and } x^{*,1}_{j,k}(\alpha) = 0, \text{ for all } (j,k) \in \tilde{L}_i,
\end{aligned}$$

and either $\sum_{k=1}^{K}\sum_{j=1}^{J_k} x^{*,1}_{j,k}(\alpha) = \alpha$ or $\sum_{k=1}^{K}\sum_{j=1}^{J_k} x^{*,1}_{j,k}(\alpha) < \alpha$.

When choosing a priority policy from the set $\Pi^*(\alpha)$, we propose to choose the same policy for any $\alpha \in [\alpha_i, \alpha_{i+1})$. This policy is chosen in the following way:

- Class-$k$ bandits in state $j$ with $(j,k) \in H_i$ receive highest priority.
- Class-$k_i$ bandits in state $j_i$ receive lower priority than class-$k$ bandits in state $j$ with $(j,k) \in H_i$.
- For class-$k$ bandits in state $j$ with $(j,k) \in L_i$ we have to distinguish between two situations:
  - (i) if $\sum_{k=1}^{K}\sum_{j=1}^{J_k} x^{*,1}_{j,k}(\alpha) < \alpha$, that is, there is capacity left unused, then any class-$k$ bandit in state $j$, with $(j,k) \in L_i$, will never be made active.
  - (ii) if $\sum_{k=1}^{K}\sum_{j=1}^{J_k} x^{*,1}_{j,k}(\alpha) = \alpha$, then the capacity constraint is binding. We will allow bandits in the set $L_i$ to be made active only if this would have happened when there would have been more capacity $\alpha$ available. Hence, a class-$k$ bandit in state $j$, $(j,k) \in L_i$, receives lower priority than a class-$\tilde{k}$ bandit in state $\tilde{j}$, $(\tilde{j},\tilde{k}) \in H_i \cap \{j_i, k_i\}$, if there is an $n > i$ such that $(j,k) \in H_n \cap \{j_n, k_n\}$. If there does not exist such $n$, then such bandits are never made active.
- Class-$k$ bandits in state $j$ with $(j,k) \in \tilde{L}_i$ are never made active.

It is left open how to set the priority ordering within the high priority states $H_i$ and the low priority states $L_i$. One way would be to chose the priorities such that the priority ordering changes minimally as $\alpha$ changes to other intervals.

## 8.2 Performance evaluation

We now turn our attention to a particular non-indexable example and numerically evaluate the selection method as explained in the previous section. We took the continuous-time version of the example given in [35, Section 2.2]. We consider a fixed population of bandits, and each bandit can be in three states. The cost structure is given by

$$(C(1,0), C(2,0), C(3,0)) = (-0.458, -0.5308, -0.6873)$$

and

$$(C(1,1), C(2,1), C(3,1)) = (-0.9631, -0.7963, -0.1057).$$

The transition matrices $Q^0 = (q(j|i,0))_{i,j}$ and $Q^1 = (q(j|i,1))_{i,j}$ are given by

$$Q^0 = \begin{pmatrix} -0.8098 & 0.4156 & 0.3942 \\ 0.5676 & -0.5809 & 0.0133 \\ 0.0191 & 0.1097 & -0.1288 \end{pmatrix} \quad \text{and} \quad Q^1 = \begin{pmatrix} -0.2204 & 0.0903 & 0.1301 \\ 0.1903 & -0.8137 & 0.6234 \\ 0.2901 & 0.3901 & -0.6802 \end{pmatrix}. \tag{20}$$

Our aim in this section is to numerically evaluate the performance of priority policies in $\Pi^*$ outside the asymptotic regime. In particular, we evaluate the performance when $\alpha = 1$, that is, at most one bandit can be made active at a time, and we let the number of bandits, $X(0)$, vary.

For a given value of $\alpha$ and $X(0)$, the set $\Pi^*$ consists of more than one policy. Before presenting the numerical results, we will therefore first describe the priority policies we considered using the selection method as given in the previous section. In Table 1, one can find the structure of an optimal basic solution of (LP), obtained numerically, when fixing $\alpha = 1$ and letting the number of bandits present

| $\alpha = 1$ | Optimal basic solution | | | |
|---|---|---|---|---|
| $1 \leq x(0) \leq 2.4$ | $x_1^{*,0} = 0, x_1^{*,1} > 0$ | $x_2^{*,0} > 0, x_2^{*,1} = 0$ | $x_3^{*,0} > 0, x_3^{*,1} = 0$ | (5) not binding |
| $2.4 \leq x(0) \leq 3.6$ | $x_1^{*,0} = 0, x_1^{*,1} > 0$ | $x_2^{*,0} > 0, x_2^{*,1} > 0$ | $x_3^{*,0} > 0, x_3^{*,1} = 0$ | (5) binding |
| $3.6 \leq x(0) \leq 7.36$ | $x_1^{*,0} > 0, x_1^{*,1} > 0$ | $x_2^{*,0} = 0, x_2^{*,1} > 0$ | $x_3^{*,0} > 0, x_3^{*,1} = 0$ | (5) binding |
| $7.36 \leq x(0)$ | $x_1^{*,0} > 0, x_1^{*,1} = 0$ | $x_2^{*,0} = 0, x_2^{*,1} > 0$ | $x_3^{*,0} > 0, x_3^{*,1} = 0$ | (5) binding |

Table 1: Optimal basic solutions for the (LP) problem

| | priority ordering | always passive | name of policy |
|---|---|---|---|
| $x(0) = 1, 2$ | 1 | 2,3 | prio1 |
| $x(0) = 3$ | $1 \succ 2$ | 3 | prio12 |
| $x(0) \geq 4$ | $2 \succ 1$ | 3 | prio21 |

Table 2: Selected priority policies

in the system, $x(0)$, increase. We note that equivalently we could have taken $x(0) = \bar{x}$ fixed and let $\alpha$ decrease, simply by a change of variable in the (LP) problem.

We can now characterize the priority policies, see also Table 2. Consider $x(0) = 1$ or $x(0) = 2$. In that case we derive from Table 1 that a bandit in state 1 receives priority (by Definition 4.4). Since the constraint (5) is not binding, a bandit in state 2 or 3 will never be made active (by Definition 4.4). Hence, $\Pi(x^*)$ consists of the policy that only makes bandits active in state 1. This policy is referred to as "prio1". Now consider $x(0) = 3$. Then, Definition 4.4 prescribes that, for any policy in $\Pi(x^*)$, state 1 has strict priority over state 2, and state 3 is either never made active, or has lowest priority. Note that for smaller values of $x(0)$ (equivalent to considering higher values of $\alpha$), state 3 is not made active either. Hence, as explained in the previous section we choose to keep bandits in state 3 passive, that is, we focus on the policy "prio12". Now consider $4 \leq x(0) \leq 7$. Then, Definition 4.4 prescribes that, for any policy in $\Pi(x^*)$, state 2 has strict priority over state 1, and state 3 is either never made active, or has lowest priority. Note that state 3 is never made active for $x(0) < 4$. Hence, as explained in the previous section, we chose to do the same for $4 \leq x(0) \leq 7$, that is, we focus on policy "prio21". Now consider $x(0) \geq 8$. Then Definition 4.4 prescribes that, for any policy in $\Pi(x^*)$, state 2 has strict priority and that states 1 and 3 are either never made active or have lowest priority. For smaller values of $x(0)$, class 1 is made active, while class 3 is never made active. Hence, as explained in the previous section, we chose to do the same for $x(0) \geq 8$, that is, we focus on policy "prio21".

Any policy gives a unichain Markov chain, hence Condition 4.12 is satisfied. We therefore have that any priority policy in $\Pi^*$ that satisfies the global attractor property, as in Condition 4.10, is asymptotically optimal. Numerically we evaluated the global attractor property and found the following: for $x(0) = 1$, the policy prio1 has $x^*$ as global attractor, while policies prio12 and prio123, which also belong to $\Pi^*$, converge to a non-optimal equilibrium point. For $x(0) = 3$, there does not exist a priority policy that converges to $x^*$. For example, the fluid dynamics under prio12 converges to an equilibrium where state 1 is sometimes passive (and state 2 and 3 are never active), while the optimal point $x^*$ never makes state 1 passive. For $4 \leq x(0) \leq 7$, the set $\Pi^*$ consists of the policies prio21 and prio213, both of them have $x^*$ as global attractor. For $x(0) \geq 8$, the set $\Pi^*$ consists of prio2, prio21 and prio213, all of them have $x^*$ as global attractor.

We have numerically evaluated the performance of the priority policies as described in Table 2 against both the optimal policy (obtained numerically by value iteration) and against other priority policies. In Figure 1 we plot the relative suboptimality gap (in %) for the different policies when $\alpha = 1$ and let the number of bandits, $X(0) = x(0)$, vary on the horizontal axis. The line referred to as "Selected" plots for each given $x(0)$ the selected priority policy as given in Table 2. We observe that these selected policies always have the smallest sub-optimality gap.

Prio123 and prio213 are inside the class of asymptotically optimal policies, $\Pi^*$, for $X(0) = 3$ and $X(0) \geq 4$, respectively, however, the selection process, as described in Section 8.1, does not select these policies. In fact, we observe that prio123 and prio213 are outperformed by our selected priority policies. Below we will see that this suboptimality gap can be made arbitrarily large.

The difference in performance between different priority policies is not that large in this example. For other instances of non-indexable bandits, including dynamic populations, the differences can be larger though. For this particular example, we note however that the suboptimality gap can be made arbitrarily

large by adequately changing the values for $C(3,1), q(1|3,1)$ and $q(2|3,1)$. These parameters do not affect the performance of policies that never activate state 3 (including the priority policies in Table 2), but do influence the performance of prio123 and prio213. By making the cost of being active in state 3, $C(3,1)$, larger, and the transition rates when being active in state 3 smaller, the performance of these policies degrades. As an example, in Figure 2 we plot the suboptimality gaps when taking $C(3,1) = 5$ and $q(1|3,1) = q(2|3,1) = 0.001$ and we observe larger optimality gaps. Furthermore, note that for a fixed $X(0)$, the gap will grow linearly in $C(3,1)$ and hence can be made arbitrarily large.

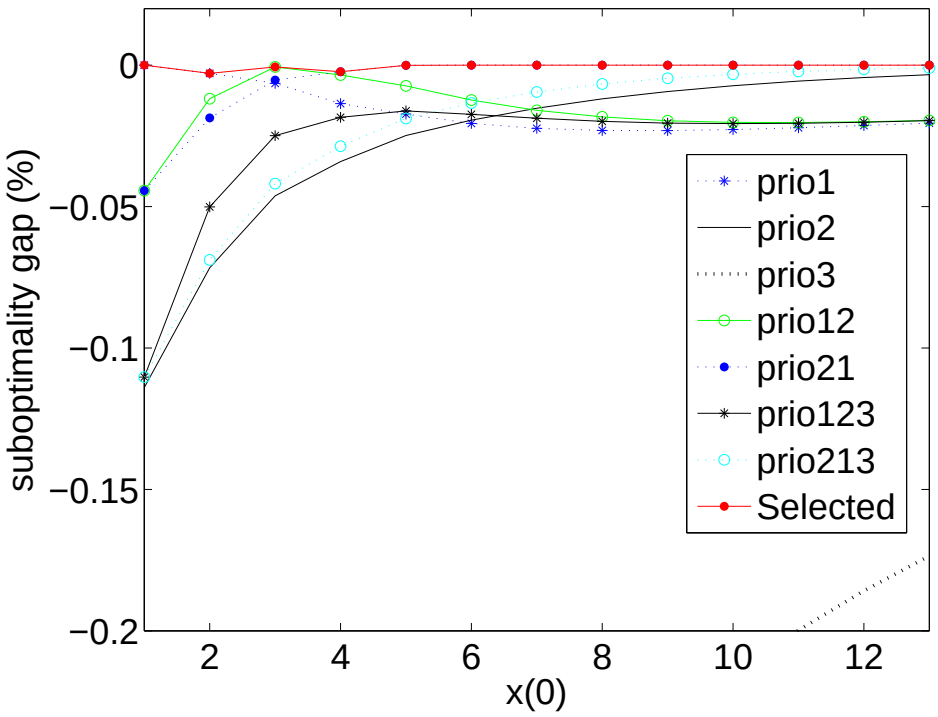


Figure 1: Suboptimality gap of priority policies for non-indexable example

## 9 Conclusion and further research

In this paper we studied the general multi-class restless-bandit problem for both the setting of a fixed population of bandits as well as a dynamic population of bandits. Using linear-programming techniques, the paper provided a unified approach to derive a set of asymptotically optimal priority policies, $\Pi^*$, which does not rely on indexability of the system. Under the indexability assumption, Whittle's index policy was shown to be inside this class. This is one of the first works that proposes heuristics for *non-indexable* settings. As future work it would therefore be interesting to further understand their performance outside the asymptotic regime.

The global attractor property is crucial in order to prove asymptotic optimality of the priority policies $\Pi^*$, as explained in Section 6. Finding sufficient conditions under which the global attractor property holds for policies in $\Pi^*$ is therefore important on its own. Another interesting research thread is to characterize asymptotic optimal policies for models that do not satisfy the global attractor property, as discussed in Section 6.

In addition, it would be interesting to investigate whether Condition 4.12 holds in greater generality for restless bandit problems. For example, Condition 4.12 a) concerns stability of the system under a strict priority policy resulting from the fluid analysis. In general, care has to be taken when applying a fluid optimal control directly to the stochastic system, as they might not succeed in making the system stable, see for example [42, 44]. We believe though that the set $\Pi^*$ contains policies that do provide a stable system, however, this is a subject for future research. As an example, we refer to [6] where

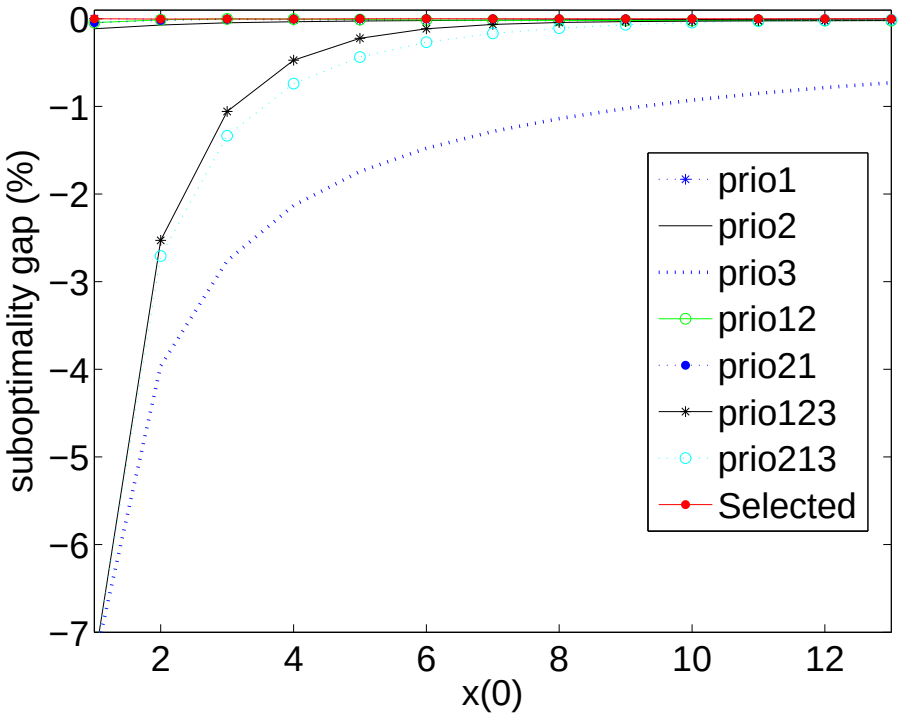


Figure 2: Suboptimality gap of priority policies for non-indexable example

a restless bandit problem was studied that modeled a system with state-dependent capacity. In that problem, certain priority policies (e.g. the myopic $c\mu$ rule, which is not in $\Pi^*$) yield an unstable system, while other priority policies, including Whittle's index policy, keep the system stable.

Another interesting research avenue would be to extend this paper to the general setting of multi-actions. That is, in each state one can choose from $A_k(j)$ different actions, given the constraint that $\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=1}^{A_k(j)} w_k^a(j)X_{j,k}^a(t)$ is less than or equal to $\alpha$, with $w_k^a(j) \geq 0$ the weight of action $a$. This paper discussed the case of $w_k^a(j) = 1$, while in [26] asymptotic optimality of an index policy has been investigated for $w_k^a(j) = a$.

## Acknowledgements

The author is grateful to Urtzi Ayesta, Balakrishna J. Prabhu, Nicolas Gast, Peter Jacko, José Niño-Mora, and Philippe Robert for valuable discussions and comments, and to the referees for their constructive reviews.

## References

[1] 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.

[2] 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.

[3] R. Atar, C. Giat, and N. Shimkin. The $c\mu/\theta$ rule for many-server queues with abandonment. *Operations Research*, 58(5):1427–1439, 2010.

[4] R. Atar, C. Giat, and N. Shimkin. On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost. *Queueing Systems*, 67(2):127–144, 2011.

[5] 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.

[6] U. Ayesta, M. Erausquin, M. Jonckheere, and I.M. Verloop. Scheduling in a random environment: stability and asymptotic optimality. *IEEE/ACM Transactions on Networking*, 21(1):258–271, 2013.

[7] 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.

[8] 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.

[9] D. Bertsimas and J. Niño-Mora. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research*, 48(1):80–90, 2000.

[10] P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1999.

[11] M.J. Cánovas, M.A. López, and J. Parra. On the continuity of the optimal value in parametric linear optimization: Stable discretization of the Lagrangian dual of nonlinear problems. *Set-Valued Analysis*, 13:69–84, 2005.

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

[13] J.G. Dai and S. He. Many-server queues with customer abandonment: A survey of diffusion and fluid approximations. *Journal of Systems Science and Systems Engineering*, 21:1–36, 2012.

[14] 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.

[15] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, New York, 1986.

[16] N. Gast and B. Gaujal. A mean field model of work stealing in large-scale systems. In *Proceedings of ACM SIGMETRICS*, pages 13–24, New York NY, USA, 2010.

[17] J.C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B*, 41(2):148–177, 1979.

[18] J.C. Gittins. *Multi-Armed Bandit Allocation Indices*. Wiley, Chichester, 1989.

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

[20] K.D. Glazebrook, D.J. Hodge, and C. Kirkbride. General notions of indexability for queueing control and asset management. *Annals of Applied Probability*, 21:876–907, 2011.

[21] 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.

[22] 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.

[23] X. Guo, O. Hernández-Lerma, and T. Prieto-Rumeau. A survey of recent results on continuous-time Markov decision processes. *TOP*, 14:177–261, 2006.

[24] J. Hasenbein and D. Perry (Eds). Special issue on queueing systems with abandonments. *Queueing Systems*, 75:111–384, 2013.

[25] D. J. Hodge and K. D. Glazebrook. Dynamic resource allocation in a multi-product make-to-stock production system. *Queueing Systems*, 67:333–364, 2011.

[26] D. J. Hodge and K. D. Glazebrook. On the asymptotic optimality of greedy index heuristics for multi-action restless bandits. *Advances in Applied Probability*, 2015. To appear.

[27] P. Jacko. Optimal index rules for single resource allocation to stochastic dynamic competitors. In *Proceedings of ValueTools*, 2011.

[28] M. Larrañaga, U. Ayesta, and I.M. Verloop. Dynamic fluid-based scheduling in a multi-class abandonment queue. *Performance Evaluation*, 70(10):841–858, 2013.

[29] 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.

[30] 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.

[31] 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.

[32] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2009.

[33] J. Niño-Mora. Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, 33(1):76–98, 2001.

[34] J. Niño-Mora. Characterization and computation of restless bandit marginal productivity indices. In *ACM SMCtools*, New York, USA, 2007.

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

[36] 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.

[37] 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.

[38] D.G. Pandelis and D. Teneketzis. On the optimality of the Gittins index rule for multi-armed bandits with multiple plays. *Mathematical Methods of Operations Research*, 50:449–461, 1999.

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

[40] 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.

[41] P. Robert. *Stochastic Networks and Queues*. Springer-Verlag, New York, 2003.

[42] A.N. Rybko and A.L. Stolyar. Ergodicity of stochastic processes describing the operation of open queuing networks. *Problems of Information Transmission*, 28:199–220, 1992.

[43] H.C. Tijms. *A First Course in Stochastic Models*. Wiley, England, 2003.

[44] I.M. Verloop and R. Núñez-Queija. Assessing the efficiency of resource allocations in bandwidth-sharing networks. *Performance Evaluation*, 66:59–77, 2009.

[45] R.R. Weber. Comments on: Dynamic priority allocation via restless bandit marginal productivity indices. *TOP*, 15:211–216, 2007.

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

[47] R.R. Weber and G. Weiss. Addendum to "On an index policy for restless bandits". *Journal of Applied Probability*, 23:429–430, 1991.

[48] G. Weiss. Branching bandit processes. *Probability in the Engineering and Informational Sciences*, 2:269–278, 1988.

[49] P. Whittle. Arm-acquiring bandits. *The Annals of Probability*, 9(2):284–292, 1981.

[50] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298, 1988.

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

# Appendix A: Proof of Lemma 4.1

Set $X_k(0) = x_k(0)$. Let $\pi$ be a policy for which a unique invariant distribution exists having finite first moment. Stability of policy $\pi$ implies rate-stability, that is,

$$\lim_{t\to\infty} \frac{X^{\pi}_{j,k}(t)}{t} = 0, \text{ for all } j,k. \tag{21}$$

Note that $\int_0^t X^{\pi,a}_{j,k}(s)\mathrm{d}s$ 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 $X^{\pi}_{j,k}(t)$:

$$\begin{aligned} X^{\pi}_{j,k}(t) &= X^{\pi}_{j,k}(0) + N^{\lambda_k p_k(j)}(t) + \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k} N^{q_k(j|i,a)}\left(\int_0^t X^{\pi,a}_{i,k}(s)\mathrm{d}s\right) \\ &\quad - \sum_{a=0}^{1}\sum_{i=0,i\neq j}^{J_k} N^{q_k(i|j,a)}\left(\int_0^t X^{\pi,a}_{j,k}(s)\mathrm{d}s\right), \end{aligned} \tag{22}$$

where $N^{\lambda_k p_k(j)}(t)$ and $N^{q_k(j|i,a)}(t)$ are independent Poisson processes having as rates $\lambda_k p_k(j)$ and $q_k(j|i,a)$, respectively, $i,j = 1,\ldots,J_k$, $k = 1,\ldots,K$, $a = 0,1$. By the ergodic theorem [12], we obtain that $\frac{1}{t}\int_0^t X^{\pi,a}_{j,k}(s)\mathrm{d}s$ converges to the mean, denoted by $\overline{X}^{\pi,a}_{j,k} < \infty$, for all $j,k,a$. Hence, when dividing both sides in (22) by $t$, using that $N^{\theta}(at)/t \to a\theta$ as $t\to\infty$, and together with (21), we obtain that

$$0 = \lambda_k p_k(j) + \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k} q_k(j|i,a)\overline{X}^{\pi,a}_{i,k} - \sum_{a=0}^{1}\sum_{i=0,i\neq j}^{J_k} \overline{X}^{\pi,a}_{j,k} q_k(i|j,a), \text{ a.s.},$$

that is, $\overline{X}^{\pi}$ satisfies Equation (4). By definition, $\overline{X}^{\pi}$ satisfies $\sum_{k,j}\overline{X}^{\pi,1}_{j,k} \leq \alpha$, $\overline{X}^{\pi,a}_{j,k} \geq 0$ and if $\lambda_k = 0$, then $\sum_{j=1}^{J_k}\sum_{a=0}^{1}\overline{X}^{\pi,a}_{j,k} = x_k(0)$. Hence, $\overline{X}^{\pi}$ is a feasible solution of (LP).

Since the feasible set is non-empty and the objective is to minimize the cost, the optimal value satisfies $v^*(x(0)) < \infty$. □

# Appendix B: Proof of Lemma 4.3

By Fatou's lemma we have

$$V^{\pi}_{-}(x) \geq \mathbb{E}_x\left(\liminf_{T\to\infty}\frac{1}{T}\int_0^T \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k(j,a)X^{\pi,a}_{j,k}(t)\mathrm{d}t\right).$$

Hence, it is sufficient to prove that

$$\liminf_{T\to\infty}\frac{1}{T}\int_0^T \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k(j,a)X^{\pi,a}_{j,k}(t)\mathrm{d}t \geq v^*(x), \text{ almost surely}, \tag{23}$$

with $X(0) = x$.

Consider a fixed realization $\omega$ of the process. We note that Equation (23) is trivially true if it holds that $\liminf_{T\to\infty} \frac{1}{T}\int_0^T \sum_{k=1}^K \sum_{j=1}^{J_k} \sum_{a=0}^1 C_k(j,a) X_{j,k}^{\pi,a}(t)\mathrm{d}t = \infty$, since $v^*(x) < \infty$ (see Lemma 4.1). Hence, it remains to be verified that (23) holds when the LHS of (23) is finite.

First assume either a fixed population of bandits, or a dynamic population of bandits under a stable policy $\pi$. Since the LHS of (23) is finite, we can consider the subsequence $t_n$ corresponding to the liminf sequence. For a fixed population of bandits, we have $\frac{1}{T}\int_0^T X_{j,k}^{\pi,a}(t)\mathrm{d}t \le X_k(0) = x_k$. Hence, there is a subsequence $t_{n_l}$ of $t_n$ such that $\frac{1}{t_{n_l}}\int_0^{t_{n_l}} X_{j,k}^{\pi,a}(t)\mathrm{d}t$ converges to a constant $\overline{X}_{j,k}^{\pi,a}$, for all $j,k,a$. In the case of a dynamic population, given the policy $\pi$ is stable, we have by the ergodicity theorem [12] that $\frac{1}{T}\int_0^T X_{j,k}^{\pi,a}(t)\mathrm{d}t$ converges to the mean, here denoted by $\overline{X}_{j,k}^{\pi,a}$.

In addition, it holds that $\lim_{t\to\infty} X_{j,k}^{\pi,a}(t)/t = 0$, for all $j,k,a$. For the fixed population this follows since $\lim_{t\to\infty} X_{j,k}^{\pi,a}(t)/t \le \lim_{t\to\infty} X_k(0)/t = 0$, and for the dynamic population this follows since any stable policy is rate stable.

When studying (22) in the point $t_{n_l}$, dividing both sides by $t_{n_l}$ and using that $N^\theta(t)/t \to \theta$ as $t\to\infty$, we can now conclude that $0 = \lambda_k p_k(j) + \sum_{a=0}^1 \sum_{i=1}^{J_k} q_k(j|i,a)\overline{X}_{i,k}^{\pi,a}$. By (2) we also have that $\sum_{k=1}^K \sum_{j=1}^{J_k} \overline{X}_{j,k}^{\pi,1} \le \alpha$. In addition, if $\lambda_k = 0$, then $\sum_{j,a} \overline{X}_{j,k}^{\pi,a} = X_k(0) = x_k$. Hence $\overline{X}^\pi$ is a feasible solution of (LP) with $x(0) = x$. We conclude that

$$\begin{aligned}\liminf_{T\to\infty} \frac{1}{T}\int_0^T \sum_{k=1}^K \sum_{j=1}^{J_k} \sum_{a=0}^1 C_k(j,a) X_{j,k}^{\pi,a}(t)\mathrm{d}t &= \lim_{l\to\infty} \frac{1}{t_{n_l}}\int_0^{t_{n_l}} \sum_{k=1}^K \sum_{j=1}^{J_k} \sum_{a=0}^1 C_k(j,a) X_{j,k}^{\pi,a}(t)\mathrm{d}t \\ &= \sum_{k=1}^K \sum_{j=1}^{J_k} \sum_{a=0}^1 C_k(j,a)\overline{X}_{j,k}^{\pi,a} \ge v^*(x),\end{aligned}$$

which proves $V_-^\pi(x) \ge v^*(x)$.

We now consider a dynamic population of bandits and take $\pi$ to be rate-stable. In addition, assume $C_k(j,a) > 0$, for all $j,k,a$. Again we consider the subsequence $t_n$ corresponding to the liminf sequence of (23). So

$$\lim_{n\to\infty} \frac{1}{t_n}\int_0^{t_n} \sum_{k=1}^K \sum_{j=1}^{J_k} \sum_{a=0}^1 C_k(j,a) X_{j,k}^{\pi,a}(t)\mathrm{d}t < \infty. \tag{24}$$

Since $C_k(j,a) > 0$, this implies that the sequence $\frac{1}{t_n}\int_0^{t_n} X_{j,k}^{\pi,a}(t)\mathrm{d}t$ is bounded, for all $j,k,a$. By the Bolzano-Weierstrass theorem, there exists a subsubsequence $t_{n_l}$ of $t_n$ and values $\overline{X}_{j,k}^{\pi,a}$'s such that $\lim_{l\to\infty} \frac{1}{t_{n_l}}\int_0^{t_{n_l}} X_{j,k}^{\pi,a}(t)\mathrm{d}t = \overline{X}_{j,k}^{\pi,a}$, for all $j,k,a$. In addition, by rate stability we have that $\lim_{t\to\infty} X_{j,k}^{\pi,a}(t)/t = 0$, a.s., for all $j,k,a$. The proof follows now in the same way as above.

The proof in the case of mean-rate stability goes along similar lines as that for rate stability and is therefore not included here. □

# Appendix C: Proof of Proposition 4.13

Consider an arbitrary priority policy $\pi$ for which $X^{r,\pi}(t)$ is irreducible. We first prove stability and then show the tightness and uniform integrability.

**Stability:** The Markov process $X^{r,\pi}(t)$ has unbounded transition rates, however, it does not die in finite time (upward jumps are of the order 1). Hence, once we prove the multi-step drift criterion [32, 41], we can conclude that there is a unique invariant distribution measure. The multi-step drift criterion will consist here in proving that there are $\delta > 0$, $T < \infty$, $d > 0$ and a stopping time $\tau$, such that $\mathbb{E}_x(\tau) \le T$ for all $x$ and

$$\mathbb{E}_x\Big(\sum_{k=1}^K \sum_{j=1}^{J_k} X_k^{r,\pi,0}(\tau)\Big) - \sum_{k=1}^K \sum_{j=1}^{J_k} x_{j,k}^0 \le -\delta,$$

for any $x \in D^c$, with $D := \{x : \sum_{k=1}^K \sum_{j=1}^{J_k} x_{j,k}^0 \le d\}$. In other words, for any initial state $x$ outside the compact set $D$, there is a negative drift (lower bounded by $-\delta$) towards the set $D$.

We define the stopping time $\tau$ as the first moment that an active bandit is made passive. Hence, during the interval $[0, \tau]$ the collection of passive bandits does not change.

First assume there exists an $x$ such that $\mathbb{E}_x(\tau) = \infty$. This implies that when starting in state $x$, the collection of passive and active bandits remains fixed. Hence, each passive class-$k$ bandit evolves according to the transition rates $q_k(j|i, 0)$. The number of passive class-$k$ bandits is therefore equivalent to that in an $M/G/\infty$ queue with arrival rate $\lambda_k r$ and phase-type distributed service requirements as described by the transitions of a passive class-$k$ bandit. We note that the $M/G/\infty$ queue is stable. By irreducibility, for any starting point, the process will be in state $x$ after a finite expected amount of time, hence, stability follows.

We now assume $\mathbb{E}_x(\tau) < \infty$, for all $x$. Since there is a finite number of states $J_k < \infty$ and the state transitions are exponential, it follows directly that there exists a $T < \infty$ such that $\mathbb{E}_x(\tau) < T$, for all $x$. Note that the passive bandits behave independently during the interval $[0, \tau]$. The probability that a passive bandit departs in the interval $[0, \tau]$ can be lower bounded by $p_0$ with $p_0 > 0$. This follows from the assumption that state 0 is positive recurrent under the policy that always keeps the class-$k$ bandit passive. Hence, the mean number of passive bandits that leave during the interval $[0, \tau]$ is larger than or equal to $p_0 \sum_{k=1}^{K} \sum_{j=1}^{J_k} x_{j,k}^0$. We therefore have as mean drift

$$\begin{aligned}
&\mathbb{E}_x(\sum_{k=1}^{K} \sum_{j=1}^{J_k} X_k^{r,\pi,0}(\tau)) - \sum_{k=1}^{K} \sum_{j=1}^{J_k} x_{j,k}^0 \\
&\leq \lambda r \mathbb{E}_x(\tau) + 1 - p_0 \sum_{k}^{K} \sum_{j=1}^{J_k} x_{j,k}^0 < \lambda r T + 1 - p_0 d,
\end{aligned}$$

for all $x \in D^c$. The $+1$ in the mean drift is due to the active bandit that becomes passive at time $\tau$. Choosing $d = (\lambda r T + 1 + \delta)/p_0$, we conclude that $\mathbb{E}_x(\sum_{k=1}^{K} \sum_{j=1}^{J_k} X_k^{r,\pi,0}(\tau)) - \sum_{k=1}^{K} \sum_{j=1}^{J_k} x_{j,k}^0 \leq -\delta$. Hence, by the multistep drift criterion we obtain that there is a unique invariant probability distribution for the process $X^{r,\pi}(t)$, for any $r$. Recall that we denote this distribution by $p^{r,\pi}$.

**Tightness and uniform integrability:** In order to prove tightness and uniform integrability, we will define a process that serves as a stochastic upper bound on $X_k^{r,\pi,0}(t)$. First note that $\max_i q_k(j|i, 1)\alpha$ is the maximum rate at which active bandits go to state $j$. Hence, $\overline{\lambda}_k := \lambda_k + \sum_{j=1}^{J_k} \max_i q_k(j|i, 1)\alpha$ is an upper bound on the arrival rate of new passive class-$k$ bandits. For the upper bound process, we assume that once a bandit is passive, it will never be made active again. Hence, the time such a passive bandit stays in the system can be described by the state transitions of a passive class-$k$ bandit. We define $B_k$ as the distribution described by the state transition rates $q_k(j|i, 0)$, with a certain initial probability $\tilde{p}_0$. Choosing an appropriate value for $\tilde{p}_0$, the $B_k$ describes the time a passive class-$k$ bandit stays in the system. Let $Y_k^r(t)$ be the number of customers in a $M/G/\infty$ queue with arrival rate $\overline{\lambda}_k$ and service requirement $B_k$. This process is an upper bound on $X_k^{r,\pi,0}(t)$.

The stationary distribution of the process $\{Y_k^r(t)\}$ is given by a Poisson distribution with parameter $\overline{\lambda}_k r \mathbb{E}(B_k)$ [41]. It can be checked that this distribution converges to the Dirac measure in the point $\overline{\lambda}_k \mathbb{E}(B_k)$, as $r \to \infty$. By Prohorov's theorem it then follows that the family $\{Y_k^r/r\}$ is tight [41]. Furthermore, since $\mathbb{E}(Y_k^r/r) = \overline{\lambda}_k \mathbb{E}(B_k)$ and $\mathbb{E}(\lim_{r\to\infty} Y_k^r/r) = \overline{\lambda}_k \mathbb{E}(B_k)$, a.s., we obtain from [10, Theorem 3.6] that the family $\{Y_k^r/r\}$ is uniform integrable.

At most $\alpha$ bandits are active, hence $\sum_k Y_k^r(t)/r + \alpha$ represents a stochastic upper bound on the queue length process $\sum_{k=1}^{K} X_k^{r,\pi}(t)/r$. This implies that the family $\{p^{r,\pi}\}$ is tight and uniform integrable as well. □

# Appendix D: Proof of Proposition 4.14

We denote by $S_k^{\pi^*}(j)$ the set of all combinations $(i, l), i = 1, \ldots, J_l, l = 1, \ldots, K$, such that class-$l$ bandits in state $i$ have higher priority than class-$k$ bandits in state $j$ under policy $\pi^*$, and $I^{\pi^*}$ is the set of all states that will never be made active under policy $\pi^*$. The transition rates of the process $X^{r,\pi^*}(t)/r$ are

then defined as follows:

$$x \to x + \tfrac{e_{j,k}}{r} \qquad \text{at rate } \; r\lambda_k p_k(j), \quad k = 1, \ldots, K, \; j = 1, \ldots, J_k, \tag{25}$$

$$x \to x - \tfrac{e_{j,k}}{r} \qquad \text{at rate } \; r \sum_{a=0}^{1} x_{j,k}^{a} q_k(0|j,a), \quad k = 1, \ldots, K, \; j = 1, \ldots, J_k, \tag{26}$$

$$x \to x - \tfrac{e_{j,k}}{r} + \tfrac{e_{i,k}}{r} \qquad \text{at rate } \; r \sum_{a=0}^{1} x_{j,k}^{a} q_k(i|j,a), \quad k = 1, \ldots, K, \; i,j = 1, \ldots, J_k, \quad i \neq j, \tag{27}$$

where $x_{j,k}^{1} = \min\left((\alpha - \sum_{(i,l) \in S_k^{\pi^*}(j)} x_{i,l})^+, x_{j,k}\right)$, if $(j,k) \notin I^{\pi^*}$, and $x_{j,k}^{1} = 0$ otherwise, $x_{j,k}^{0} = x_{j,k} - x_{j,k}^{1}$, and $e_{j,k}$ is a vector composed of all zeros except for component $(j,k)$ which is one.

From (25)–(27) it follows that there exists a continuous function $b_l(x)$, with $l \in \mathcal{L}$ and $\mathcal{L}$ composed of a finite number of vectors in $\mathbb{N}^{\sum_k J_k}$, such that the transition rates of the process $x^{r,\pi^*}(t)$ from $x$ to $x + l/r$ have the form $r b_l(x)$. Hence, the process $X_{j,k}^{r,\pi^*}(t)/r$ belongs to the family of density dependent population processes as defined in [15, Chapter 11].

Note that the process $x^{\pi^*}(t)$ as defined in (9) can equivalently be written as $\frac{\mathrm{d}x^{\pi^*}(t)}{\mathrm{d}t} = F(x^{\pi^*}(t))$, with $F(x^*) = \sum_{l \in \mathcal{L}} l b_l(x^*)$, where $F(\cdot)$ is Lipschitz continuous. From Condition 4.10 we have that $x^*$ is the unique global attractor of $x^{\pi^*}(t)$.

Together with the fact that the family $\{p^{r,\pi^*}\}$ is tight, we then obtain from [16, Theorem 4] that $p^{r,\pi^*}(x)$ converges to the Dirac measure in $x^*$, the global attractor of $x^{\pi^*}(t)$. Hence, we can write

$$\lim_{r\to\infty} V_+^{r,\pi^*}(x) = \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \lim_{r\to\infty} \sum_{x} p^{r,\pi^*}(x) C_k(j,a) x_{j,k}^{a} = \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} C_k(j,a) x_{j,k}^{*,a} = v^*(x),$$

where the first step follows from the ergodicity theorem [43, 12] (applicable since the first moment of $p^{r,\pi^*}$ is finite), the second step (interchange of limit and summation) follows from uniform integrability of $\{p^{r,\pi^*}\}$ and the fact that $p^{r,\pi^*}$ converges to the Dirac measure in $x^*$, and the last step follows since $x^*$ is an optimal solution of (LP).

We conclude the proof by noting that $v^*(x)$ is a lower bound on the steady-state cost, as shown in Lemma 4.3. □

# Appendix E: Proof of Proposition 5.6

Recall that the relaxed optimization problem for $f = av$ consists in finding a stationary and Markovian policy that minimizes

$$\lim_{T\to\infty} \frac{1}{T} \mathbb{E}_x \left( \int_0^T \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} C_k(j,a) X_{j,k}^{\pi,a}(t) \mathrm{d}t \right), \tag{28}$$

under the relaxed constraint

$$\lim_{T\to\infty} \frac{1}{T} \mathbb{E}_x \left( \int_0^T \sum_{k=1}^{K} \sum_{j=1}^{J_k} X_{j,k}^{\pi,1}(t) \mathrm{d}t \right) \leq \alpha. \tag{29}$$

For a given policy $\pi$, we denote by $x_{j,k}^{\pi,a}$ the (stationary) state-action frequencies, that is, the average fraction of time the class-$k$ bandit is in state $j$ and action $a$ is chosen. Assumption 5.5 implies that these frequencies exist and satisfy the balance equations, that is, they satisfy

$$0 = \sum_{a=0}^{1} \sum_{i=0, i\neq j}^{J_k} q_k(i|j,a) x_{j,k}^{\pi,a} - \sum_{a=0}^{1} \sum_{i=1, i\neq j}^{J_k} q_k(j|i,a) x_{i,k}^{\pi,a}, \; \forall j,$$

or, by definition of $q_k(j|j,a) = -\sum_{i=0, i\neq j}^{J_k} q_k(i|j,a)$, this can be written as,

$$0 = \sum_{a=0}^{1} \sum_{i=1}^{J_k} q_k(j|i,a) x_{i,k}^{\pi,a}, \; \forall j.$$

We will restrict ourselves to the class of policies that are symmetric for bandits in the same class. We can do this without loss of generality, since an optimal solution of the relaxed problem, given by Whittle's indices, is symmetric. Having $X_k(0)$ bandits in class $k$, Equations (28) and (29) can now equivalently be written as

$$\sum_{k=1}^{K} X_k(0) \sum_{j=1}^{J_k} \left( C_k(j,0) x_{j,k}^{\pi,0} + C_k(j,1) x_{j,k}^{\pi,1} \right) \quad \text{and} \quad \sum_{k=1}^{K} X_k(0) \sum_{j=1}^{J_k} x_{j,k}^{\pi,1} \leq \alpha,$$

respectively.

The relaxed optimization problem can now be formulated as the following linear program (D):

$$\begin{aligned} (D) \qquad \min_{x} & \sum_{k=1}^{K} X_k(0) \sum_{j=1}^{J_k} \left( C_k(j,0) x_{j,k}^{0} + C_k(j,1) x_{j,k}^{1} \right) \\ \text{s.t.} \;\; & 0 = \sum_{a=0}^{1} \sum_{i=1}^{J_k} q_k(j|i,a) x_{i,k}^{a}, \;\; \forall \; j,k, \\ & \sum_{k=1}^{K} X_k(0) \sum_{j=1}^{J_k} x_{j,k}^{1} \leq \alpha, \\ & \sum_{j=1}^{J_k} \sum_{a=0}^{1} x_{j,k}^{a} = 1, \;\; \forall \; k, \quad x_{j,k}^{a} \geq 0, \;\; \forall \; k,j,a. \end{aligned} \tag{30}$$

We have that for any feasible solution $(x_{j,k}^{a})$ of (D) there is a stationary policy $\pi$ such that the state-action frequencies $x_{j,k}^{\pi,a}$ coincide with the value of the feasible solution $x_{j,k}^{a}$ [39, Theorem 8.8.2 b)]. Hence, for any optimal (symmetric) policy $\pi^*$ of the relaxed optimization problem, the state-action frequencies $x_{j,k}^{\pi^*,a}$ provide an optimal solution of (D). We further note that $(x_{j,k}^{\pi^*} X_k(0))$ is an optimal solution of (LP) with $x(0) = X(0)$.

We assume the restless bandit problem is indexable. Hence, an optimal policy of the relaxed optimization problem is described in Section 5.1, and will be denoted here by $\tilde{\pi}^*$. We recall that policy $\tilde{\pi}^*$ is described by a value $\nu^* \geq 0$ and is such that a class-$k$ bandit in state $j$ is made active if $\nu_k^{av}(j) > \nu^*$ and is kept passive if $\nu_k^{av}(j) < \nu^*$. Hence, the state-action frequencies under $\tilde{\pi}^*$ satisfy

$$\begin{aligned} x_{j,k}^{\tilde{\pi}^*,0} &= 0 \;\; \text{when } \nu_k^{av}(j) > \nu^*, \\ x_{j,k}^{\tilde{\pi}^*,1} &= 0 \;\; \text{when } \nu_k^{av}(j) < \nu^*. \end{aligned} \tag{31}$$

By definition of policy $\tilde{\pi}^*$, for states $(\hat{j},\hat{k})$ with $\nu_{\hat{k}}^{av}(\hat{j},) = \nu^*$ a class-$\hat{k}$ bandit in state $\hat{j}$ is made active with a certain probability, hence $x_{\hat{j},\hat{k}}^{\tilde{\pi}^*,0} \geq 0$ and $x_{\hat{j},\hat{k}}^{\tilde{\pi}^*,1} \geq 0$.

Since Whittle's index policy gives priority to bandits having highest index value, we directly obtain that Whittle's index policy $\nu^{av}$ satisfies points 1 and 2 of Definition 4.4 when setting $x^* = (x_{j,k}^{\tilde{\pi}^*} X_k(0))$. We now treat point 3 of Definition 4.4: Assume $\sum_{k=1}^{K} \sum_{j=1}^{J_k} x_{j,k}^{\tilde{\pi}^*,1} X_k(0) < \alpha$. Hence, under the optimal policy, on average, strictly less than $\alpha$ bandits are made active. This implies that the remaining fraction of the time the policy makes dummy bandits in state $B$ active. Hence, $\nu_B^{av} \geq \nu^*$. Since $\nu^* \geq 0$ and $\nu_B^{av} = 0$ we necessarily have $\nu^* = 0$. A policy satisfies point 3 of Definition 4.4 if it never makes a class-$k$ bandit in state $j$ active that satisfies

$$x_{j,k}^{\tilde{\pi}^*,1} = 0 \text{ and } x_{j,k}^{\tilde{\pi}^*,0} > 0. \tag{32}$$

From (31) (with $\nu^* = 0$), we obtain that (32) implies $\nu_k^{av}(j) \leq 0$. By definition of Whittle's index policy, a bandit in a state such that $\nu_k^{av}(j) \leq 0$ will never be made active, hence point 3 is satisfied. We therefore conclude that Whittle's index policy $\nu^{av}$ is included in the set of priority policies $\Pi(x^*) \subset \Pi^*$, with $x^* = (x_{j,k}^{\tilde{\pi}^*} X_k(0))$. $\square$

# Appendix F: Proof of Proposition 5.9

Let $\beta \leq \overline{\beta}$ and $\beta > 0$. Whittle's index $\nu_k^\beta(j)$ results from solving the following problem for a class-$k$ bandit:

$$\min_\pi \mathbb{E}_x \left( \int_0^\infty \mathrm{e}^{-\beta t} (C_k(J_k(t), A_k^\pi(t)) + \nu \mathbf{1}_{(A_k^\pi(t)=1)}) \mathrm{d}t \right), \tag{33}$$

see (15), where $A_k^\pi(t) \in \{0,1\}$ and $J_k(t)$ denotes the state of the class-$k$ bandit. This is a continuous-time discounted Markov decision problem in a finite state space. After uniformization ([23, Remark 3.1], [39, Section 11.5.2]) this is equivalent to a *discrete-time* discounted Markov decision problem with discount factor $\tilde{\beta} = \frac{\overline{q}}{\beta+\overline{q}}$, cost function $\tilde{C}_k(j,a) = \frac{C_k(j,a)+\nu \mathbf{1}_{(a=1)}}{\beta+\overline{q}}$, and transition probabilities $\tilde{p}_k^a(i,j) = \frac{q_k(j|i,a)}{\overline{q}} + \mathbf{1}_{(i=j)}$ (recall that $q_k(i|i,a) = -\sum_{j=0, i\neq j}^{J_k} q_k^a(i,j)$), where $\overline{q} := \max_{i,k,a} -q_k(i|i,a) < \infty$. In LP formulation the discrete-time MDP for the class-$k$ bandit is then as follows (see [39, Section 6.9]):

$$\begin{aligned} &\max_v \sum_{j=1}^{J_k} \gamma_{j,k} v(j) \\ &\text{s.t. } \; v(i) - \tilde{\beta} \sum_{j=0}^{J_k} \tilde{p}_k^a(i,j) v(j) \leq \tilde{C}_k(i,a), \quad \forall\ i = 1, \ldots, J_k,\ a = 0, 1, \end{aligned}$$

with $\gamma_{j,k} > 0$ arbitrary. In fact, we will make the choice $\gamma_{j,k} = \lambda_k(p_{0j}^k + \epsilon)$, with $\epsilon > 0$. The dual of the above LP is

$$\begin{aligned} (D_k(\beta,\epsilon)) \qquad &\min_x \sum_{j=1}^{J_k} \frac{C_k(j,0)x_{j,k}^0 + C_k(j,1)x_{j,k}^1 + \nu x_{j,k}^1}{\beta+\overline{q}} \\ &\text{s.t. } \; 0 = \lambda_k(p_k(0,j)+\epsilon) + \sum_{a=0}^{1} \sum_{i=1}^{J_k} \frac{q_k(j|i,a)}{\beta+\overline{q}} x_{i,k}^a - \frac{\beta}{\beta+\overline{q}} \sum_{a=0}^{1} x_{j,k}^a,\ \forall\ j, \\ &\qquad\ x_{j,k}^a \geq 0,\ \forall j, a. \end{aligned} \tag{34}$$

As stated in Section 5.1.1, indexability implies that an optimal policy for the subproblem (33) is described by a priority ordering according to the indices $\nu_k^\beta(j)$: an optimal action in state $j$ is $a = 1$ if $\nu_k^\beta(j) > \nu$ and $a = 0$ if $\nu_k^\beta(j) < \nu$. Recall that a class-$k$ bandit is indexable for each $\beta_l$ (subsequence can depend on the class). Hence, by [39, Theorem 6.9.4], this implies that there exists an optimal solution to $(D_k(\beta_l,\epsilon))$, denoted by $x_k^*(\beta_l,\epsilon)$, such that

$$\begin{aligned} x_{j,k}^{*,0}(\beta_l,\epsilon) &= 0 \quad \text{when } \nu_k^{\beta_l}(j) > \nu, \\ x_{j,k}^{*,1}(\beta_l,\epsilon) &= 0 \quad \text{when } \nu_k^{\beta_l}(j) < \nu. \end{aligned}$$

Since $\lim_{l\to\infty} \nu_k^{\beta_l}(j) = \nu_k^{lim}(j)$, we obtain that there exists an $L(\nu)$ such that for all $l > L(\nu)$ it holds that

$$x_{j,k}^{*,0}(\beta_l,\epsilon) = 0 \quad \text{when } \nu_k^{lim}(j) > \nu, \tag{35}$$

$$x_{j,k}^{*,1}(\beta_l,\epsilon) = 0 \quad \text{when } \nu_k^{lim}(j) < \nu. \tag{36}$$

By change of variable $\tilde{x}_{j,k}^a = x_{j,k}^a/(\beta+\overline{q})$ we obtain that $\tilde{x}_k^*(\beta_l,\epsilon)$ satisfies (35) and (36) and is an optimal solution of $(\tilde{D}_k(\beta_l,\epsilon))$ defined as:

$$\begin{aligned} (\tilde{D}_k(\beta,\epsilon)) \qquad &\min_{\tilde{x}} \sum_{j=1}^{J_k} (C_k(j,0)\tilde{x}_{j,k}^0 + C_k(j,1)\tilde{x}_{j,k}^1 + \nu \tilde{x}_{j,k}^1) \\ &\text{s.t. } \; 0 = \lambda_k(p_k(j)+\epsilon) + \sum_{a=0}^{1} \sum_{i=1, i\neq j}^{J_k} q_k(j|i,a)\tilde{x}_{i,k}^a - \beta \sum_{a=0}^{1} \tilde{x}_{j,k}^a, \quad \forall\ j, \\ &\qquad\ \tilde{x}_{j,k}^a \geq 0,\ \forall j, a. \end{aligned} \tag{37}$$

By Assumption 5.8 we have that the set of optimal solutions of $(\tilde{D}_k(0,0))$ is bounded and non-empty when $\nu > 0$. Hence, from [11, Corollary 1] we obtain that the correspondence that gives for each $(\beta, \epsilon)$ the set of optimal solutions of $(\tilde{D}_k(\beta,\epsilon))$ is upper semicontinuous in the point $(\beta,\epsilon) = (0,0)$. It is a compact-valued correspondence (after summing (37) over all $j$, we have that $\tilde{x}_k = \lambda_k(1+\epsilon J_k)/\beta$, $\beta > 0$). Hence, it follows that there exists a sequence $(\beta_{l_n}, \epsilon_n)$ (with $\beta_{l_n}$ a subsequence of $\beta_l$ and $\epsilon_n \to 0$) such that $\tilde{x}^{*,a}_{j,k}(\beta_{l_n},\epsilon_n) \to \tilde{x}^{*,a}_{j,k}$, as $n \to \infty$, and with $\tilde{x}^*_k$ an optimal solution of $(\tilde{D}_k(0,0))$. For a fixed $\nu$, the components of $\tilde{x}^*_k(\beta_l,\epsilon)$ that are zero are independent of the exact values for $\epsilon > 0$, and $l > L(\nu)$, see (35) and (36). Hence, the limit $\tilde{x}^*_k$, which is an optimal solution of $(\tilde{D}_k(0,0))$, has the same components equal to zero, i.e., (35) and (36) are satisfied for $\tilde{x}^*_k$.

Below we will show that there exists a value $\nu^*$ such that there is a vector $\tilde{y}^*$ that satisfies the following: (i) $\tilde{y}^*_k$ is an optimal solution of $(\tilde{D}_k(0,0))$, for all $k$, with $\nu = \nu^*$, (ii) $\tilde{y}^*$ is an optimal solution of (LP), and (iii) the Whittle index policy $\nu^{lim}$ is included in the set $\Pi(\tilde{y}^*) \in \Pi^*$. The latter then concludes the proof.

In the remainder of the proof we denote by $\tilde{x}^*_k(\nu)$ the above-described optimal solution $\tilde{x}^*_k$ of $(\tilde{D}_k(0,0))$ for a given value $\nu$. We have the following properties:

- **Property 1:**
$$\sum_{k=1}^{K}\sum_{j=1}^{J_k} \tilde{x}^{*,1}_{j,k}(\infty) \leq \alpha. \tag{38}$$
This can be seen as follows. As $\nu \to \infty$, the objective of $(\tilde{D}_k(0,0))$ is to minimize $\sum_{j=1}^{J_k} \tilde{x}^1_{j,k}$. For any feasible solution $x$ of (LP), $x_k$ is in the feasible set of $\tilde{D}_k(0,0)$. Hence, $\sum_{j=1}^{J_k} \tilde{x}^{*,1}_{j,k}(\infty) \leq \sum_{j=1}^{J_k} x^1_{j,k}$ with $x$ a feasible solution of (LP). In addition, we have that $\sum_{k=1}^{K}\sum_{j=1}^{J_k} x^1_{j,k} \leq \alpha$ with $x$ a feasible solution of (LP). This proves (38).

- **Property 2:**
$$\sum_{j=1}^{J_k} \tilde{x}^{*,1}_{j,k}(\nu) \geq \sum_{j=1}^{J_k} \tilde{x}^{*,1}_{j,k}(\tilde{\nu}), \quad \text{for} \quad \nu < \tilde{\nu}. \tag{39}$$
This can be seen as follows: By definition we have $\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k(j,a)\tilde{x}^{*,a}_{j,k}(\nu) + \nu\sum_{j=1}^{J_k}\tilde{x}^{*,1}_{j,k}(\nu) \leq \sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k(j,a)\tilde{x}^{*,a}_{j,k}(\tilde{\nu}) + \nu\sum_{j=1}^{J_k}\tilde{x}^{*,1}_{j,k}(\tilde{\nu})$ and $\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k(j,a)\tilde{x}^{*,a}_{j,k}(\tilde{\nu}) + \tilde{\nu}\sum_{j=1}^{J_k}\tilde{x}^{*,1}_{j,k}(\tilde{\nu}) \leq \sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k(j,a)\tilde{x}^{*,a}_{j,k}(\nu) + \tilde{\nu}\sum_{j=1}^{J_k}\tilde{x}^{*,1}_{j,k}(\nu)$. Substracting the latter inequality from the first, we obtain Equation (39).

- **Property 3:**
$$\sum_{j=1}^{J_k} \tilde{x}^{*,1}_{j,k}(\nu) < \infty \quad \text{for} \quad \nu > 0. \tag{40}$$
This follows since by Assumption 5.8 the set of optimal solutions of $(\tilde{D}_k(0,0))$ is bounded for $\nu > 0$.

We define $\overline{\alpha} := \sum_{k=1}^{K}\sum_{j=1}^{J_k} \tilde{x}^{*,1}_{j,k}(0)$. Equations (38)–(40) imply that there exists a $\nu^* \geq 0$ such that

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k} \tilde{x}^{*,1}_{j,k}((\nu^*)^-) \geq \min(\alpha,\overline{\alpha}) \text{ and } \sum_{k=1}^{K}\sum_{j=1}^{J_k} \tilde{x}^{*,1}_{j,k}((\nu^*)^+) \leq \min(\alpha,\overline{\alpha}). \tag{41}$$

From standard LP theory we know that there exists a $\overline{\nu} < \infty$ such that $\tilde{x}^*_k(\overline{\nu})$ is an optimal solution of $(D_k(0,0))$ for all $\nu \geq \overline{\nu}$, that is $\tilde{x}^*_k(\nu) = \tilde{x}^*_k(\overline{\nu})$ for $\nu \geq \overline{\nu}$. Hence, we can take $\nu^* < \infty$.

From (39) and (41) we obtain that there exists a $\tilde{y}^* = (\tilde{y}^{*,a}_{j,k})$ with $y^*_{\tilde{k}}$ being a convex combination of $\tilde{x}^*_{\tilde{k}}((\nu^*)^-)$ and $\tilde{x}^*_{\tilde{k}}((\nu^*)^+)$ and for $k \neq \tilde{k}$, $\tilde{y}^*_k$ being equal to either $\tilde{x}^*_k((\nu^*)^-)$ or $\tilde{x}^*_k((\nu^*)^+)$, such that $\sum_{k=1}^{K}\sum_{j=1}^{J_k} \tilde{y}^{*,1}_{j,k} = \min(\alpha,\overline{\alpha})$. Note that $\tilde{y}^*_k$ is still a solution of $(\tilde{D}_k(0,0))$, for all $k$. Now, if $\alpha = \min(\overline{\alpha},\alpha)$, it follows directly that $\tilde{y}^*$ is also an optimal solution of (LP). If instead $\overline{\alpha} = \min(\overline{\alpha},\alpha)$, then $\nu^* = 0$ and hence $\tilde{y}^*_k$ is an optimal solution of $(\tilde{D}_k(0,0))$ with $\nu = 0$. After summing over $k$, the latter has the same objective function as (LP). Together with $\sum_{k=1}^{K}\sum_{j=1}^{J_k} \tilde{y}^{*,1}_{j,k} = \overline{\alpha} \leq \alpha$, it follows that $\tilde{y}^*$ is also an optimal solution of (LP).

It remains to be proved that the Whittle index policy is included in the set $\Pi(\tilde{y}^*) \subset \Pi^*$. Assume for class $\tilde{k}$ the states are ordered such that $\nu^{lim}_{\tilde{k}}(j_1) \leq \nu^{lim}_{\tilde{k}}(j_2) < \ldots \leq \ldots \leq \nu^{lim}_{\tilde{k}}(j_{J_{\tilde{k}}})$. From $\nu^* < \infty$ and Properties (35)–(36) (which hold for $\tilde{x}^*(\nu)$) we have that there are $n^*$ and $\tilde{n}^*$, $n^* \leq \tilde{n}^*$, such that $\nu_{\tilde{k}}(j_{n^*}) = \ldots = \nu_{\tilde{k}}(j_{\tilde{n}^*}) = \nu^*$ and

$$\tilde{x}^{*,1}_{j_m,\tilde{k}}((\nu^*)^-) = 0, \text{ for all } m = 1, \ldots, n^*,$$
$$\tilde{x}^{*,0}_{j_m,\tilde{k}}((\nu^*)^-) = 0, \text{ for all } m = n^*+1, \ldots, J,$$

and

$$\tilde{x}^{*,1}_{j_m,\tilde{k}}((\nu^*)^+) = 0, \text{ for all } m = 1, \ldots, \tilde{n}^*,$$
$$\tilde{x}^{*,0}_{j_m,\tilde{k}}((\nu^*)^+) = 0, \text{ for all } m = \tilde{n}^*+1, \ldots, J.$$

The vector $\tilde{y}^*_{\tilde{k}}$ is a convex combination of $\tilde{x}^*_{\tilde{k}}((\nu^*)^-)$ and $\tilde{x}^*_{\tilde{k}}((\nu^*)^+)$, hence $\tilde{y}^{*,1}_{j_m,\tilde{k}} = 0$ for all $m \leq n^*$ and $\tilde{y}^{*,0}_{j_m,\tilde{k}} = 0$ for all $m \geq \tilde{n}^* + 1$. Hence, Whittle's index policy $\nu^{lim}$ satisfies items 1 and 2 of Definition 4.4 with $x^* = \tilde{y}^*$.

If $\sum_{k=1}^K \sum_{j=1}^{J_k} \tilde{y}^{*,1}_{j,k} < \alpha$, then since $\sum_{k=1}^K \sum_{j=1}^{J_k} \tilde{y}^{*,1}_{j,k} = \min(\alpha, \overline{\alpha})$ we have $\bar{\alpha} < \alpha$, so $\nu^* = 0$. This implies that for any state $(j,k)$ with $\tilde{y}^{*,1}_{j,k} = 0$ and $\tilde{y}^{*,0}_{j,k} > 0$ it follows from Property (35) that $\nu^{lim}_k(j) < (\nu^*)^+ = 0^+$. Hence, by definition of Whittle's index policy $\nu^{lim}$, a bandit in this state will never be made active, which implies that item 3 in Definition 4.4 is satisfied for $x^* = \tilde{y}^*$. It hence follows that Whittle's index policy $\nu^{lim}$ is included in the set of priority policies $\Pi(\tilde{y}^*) \subset \Pi^*$. □

# Appendix G: Proof of Lemma 8.1

For the fixed population, the total number of constraints in (LP) is $\sum_{k=1}^K J_k + 1 + K$. However, since $\sum_{k=1}^K \lambda_k = 0$, one of the constraints in (4) is redundant for each $k$. Hence, the number of independent constraints in (LP) is $\sum_{k=1}^K J_k + 1$.

Since the feasible set of (LP) is bounded, from standard LP theory, see [39, Theorem D.1a], we obtain that there exists an optimal basic feasible solution $x^*$ to (LP). Hence, $x^*$ has $\sum_{k=1}^K J_k + 1$ basic terms and all other terms are equal to zero. If $x^*_{j,k} > 0$ for all $j,k$, then for any $j,k$ there is an action $a$ such that $x^{*,a}_{j,k} = 0$, and in at most one combination $(j,k)$ the components $x^{*,a}_{j,k}$ can be positive in both actions. Hence, $x^*$ satisfies the property in Definition 4.4.

Otherwise, let $S$ denote the set of pairs $(i,l)$ such that $x^*_{i,l} = 0$. By (4), if $(j,k) \in S$, then $\sum_{a=0}^1 \sum_{i \neq j} x^{*,a}_{i,k} q_k(j|i,a) = 0$. That is, $x^{*,a}_{i,k} q_k(j|i,a) = 0$ for all $i = 1, \ldots, J_k$, $a = 0,1$, if $(j,k) \in S$. Hence, for $(j,k) \notin S$, Equation (4) in the point $x^*$ can be rewritten as

$$0 = \sum_{a=0}^{1} \sum_{i=1,(i,k)\in S^c}^{J_k} x^{*,a}_{i,k} q_k(j|i,a), \quad \forall\, j,k,$$

where $q_k(j|j,a) = \sum_{i=0, i\neq j, (i,k)\in S^c}^{J_k} q_k(i|j,a)$. Hence, $x^*$ (restricted to the states $(j,k) \in S^c$) is an optimal solution of (LP) restricted to the set of states $S^c$. Similar as above, the latter has an optimal basic solution with $|S^c| + 1$ basic terms (and all other terms equal to zero). Let $y^*$ denote such an optimal basic solution. Note that $y^*$ is also an optimal solution of (LP) when setting $y^*_{j,k} = 0$ for all states $(j,k) \in S$.

If $y^*_{j,k} > 0$ for all $(j,k) \notin S$, then, since it has $|S^c| + 1$ basic terms, it satisfies that for any $(j,k)$ there is an action $a$ such that $y^{*,a}_{j,k} = 0$, and in at most one combination $(j,k)$ the components $y^{*,a}_{j,k}$ can be positive in both actions. Hence, $y^*$ satisfies the property in Definition 4.4.

If $y^*_{j,k} = 0$ for some $(j,k) \notin S$, the above procedure can be repeated until one ends up with an optimal basic solution that satisfies the properties as given in Definition 4.4. LP-

Now assume a dynamic population of bandits. First assume $p_k(j) > 0$ for all $k,j$. By (4) we have that any feasible solution of (LP) has $x_{j,k} > 0$. Hence, for each $(j,k)$ there exists at least one action $a$ such that $x^a_{j,k} > 0$. Since the set of optimal solutions of (LP) is non-empty and bounded, from standard LP theory, see [39, Theorem D.1a], we obtain that there exists a bounded optimal basic feasible solution

$x^*$ to (LP). We know that $x^*$ has $\sum_{k=1}^K J_k + 1$ basic terms (the number of constraints), and all other terms are equal to zero. Since $x^*_{j,k} > 0$ for all $j,k$, this implies that for any $(j,k)$ there is one action $a$ such that $x^{*,a}_{j,k} = 0$, and in at most one combination $(j,k)$ the components $x^{*,a}_{j,k}$ can be positive in both actions $a = 0$ and $a = 1$.

Now assume $C_k(j,0) > 0$ for all $j,k$. This implies that for all $\epsilon > 0$ small enough the set of optimal solutions of the (LP($\epsilon$)) problem is bounded and non-empty, where $(LP(\epsilon))$ is defined by

$$
\begin{aligned}
(LP(\epsilon)) \qquad & \min_x \sum_{k=1}^K \sum_{j=1}^{J_k} \sum_{a=0}^{A_k(j)} C_k(j,a) x^a_{j,k} \\
& \text{s.t. } 0 = \lambda_k(p_k(j) + \epsilon) + \sum_{a=0}^{1} \sum_{i=1}^{J_k} x^a_{i,k} q_k(j|i,a), \quad \forall\, j,k, \qquad (42) \\
& \sum_{k=1}^K \sum_{j=1}^{J_k} x^1_{j,k} \le \alpha, \\
& x^a_{j,k} \ge 0, \quad \forall\, j,k,a.
\end{aligned}
$$

We note that the assumption $C_k(j,0) > 0$ for all $j,k$ as stated in Lemma 8.1 could have been replaced by the weaker assumption that the set of optimal solutions of (LP($\epsilon$) is bounded and non-empty. By sensitivity results of linear programming theory we have that for $\bar{\epsilon} > 0$ small enough, the same basis provides an optimal solution for (LP($\epsilon$)) for all $0 \le \epsilon < \bar{\epsilon}$. We denote the corresponding optimal solution by $x^*(\epsilon)$. By (42) we have that $x^*_{j,k}(\epsilon) > 0$ for all $\epsilon > 0$. Since for any $0 < \epsilon < \bar{\epsilon}$ the basis of $x^*(\epsilon)$ is the same, we conclude that for any state $(j,k)$ there is one action $a$ (independent on $\epsilon$) such that $x^{*,a}_{j,k}(\epsilon) = 0$ and for at most one state $(j,k)$ (independent of $\epsilon$) the components $x^{*,a}_{j,k}(\epsilon)$ can be strictly positive for both actions $a = 0$ and $a = 1$.

Note that $(LP(0)) = (LP)$. Hence, using [11, Corollary 1] we obtain that the correspondence that gives for each $\epsilon$ the set of optimal solutions of (LP($\epsilon$)) is upper semicontinuous in the point $\epsilon = 0$. Being a compact-valued correspondence, it follows that there exists a sequence $\epsilon_l$ such that $\epsilon_l \to 0$ and $x^*(\epsilon_l) \to x^*$, with $x^*$ being an optimal solution of (LP). Being the limit, $x^*$ has the same components equal to zero (and maybe even more) as $x^*(\epsilon)$ (with $\epsilon < \bar{\epsilon}$). Hence, $x^*$ has the property as stated in the lemma. □

# Appendix H: Condition 4.10 for an *M/M/S+M* queue

Assume the classes are reordered such that $\iota_1 \ge \iota_2 \ge \cdots \ge \iota_K$. We further define $\hat{l} := \arg\min\{l : \iota_l \le 0\}$, so that $\{\hat{l}, \ldots, K\}$ is the set of classes that will never be served. Under policy $\iota$, the ODE as defined in (9) is given by

$$
\begin{aligned}
& \frac{\mathrm{d}x^\iota_k(t)}{\mathrm{d}t} = \lambda_k - x^{\iota,0}_k(t)\theta_k - x^{\iota,1}_k(t)(\mu_k + \tilde{\theta}_k), \quad \forall\, k, && (43) \\
& \text{with } x^{\iota,1}_k(t) = \min\left( \left( S - \sum_{l=1}^{k-1} x^\iota_l(t) \right)^+ , \; x^\iota_k(t) \right), \quad \text{if } k < \hat{l}, \ \forall\, k, && (44) \\
& \qquad x^{\iota,1}_k(t) = 0, \quad \text{if } k \ge \hat{l}, \ \forall\, k, && (45) \\
& \qquad x^{\iota,0}_k(t) = x^\iota_k(t) - x^{\iota,1}_k(t), \quad \forall\, k.
\end{aligned}
$$

This ODE has a unique equilibrium point, which is given by

$$
\begin{aligned}
& x^{*,0}_k = 0, \quad x^{*,1}_k = \frac{\lambda_k}{\mu_k + \tilde{\theta}_k}, \text{ for } k = 1, \ldots, \hat{k}, && (46) \\
& x^{*,0}_{\hat{k}+1} = \frac{\lambda_k - (\mu_k + \tilde{\theta}_k)(S - \sum_{l=1}^{\hat{k}} \frac{\lambda_l}{\mu_l + \tilde{\theta}_l})}{\theta_k}, \quad x^{*,1}_{\hat{k}+1} = S - \sum_{l=1}^{\hat{k}} \frac{\lambda_l}{\mu_l + \tilde{\theta}_l}, \text{ if } \hat{k} + 1 < \hat{l}, && (47) \\
& x^{*,0}_k = \frac{\lambda_k}{\theta_k}, \quad x^{*,1}_k = 0, \text{ for } k \ge \min(\hat{k} + 2, \hat{l}), && (48)
\end{aligned}
$$

where $\hat{k} = \arg\max\{k = 0, 1, \ldots, \hat{l} - 1 : \sum_{l=1}^{k} \frac{\lambda_l}{\mu_l + \tilde{\theta}_l} \leq S\}$. This can be seen as follows. If $x^*$ is an equilibrium point, it follows from (43) that

$$\frac{\lambda_k}{\mu_k + \tilde{\theta}_k} = x_k^{*,1} + x_k^{*,0} \frac{\theta_k}{\mu_k + \tilde{\theta}_k}. \tag{49}$$

We first prove (46). Let $k = 1$ and assume $1 \leq \hat{k}$. Hence, we have $\frac{\lambda_1}{\mu_1 + \tilde{\theta}_1} < S$. By (49) we obtain $x_1^{*,1} < S$. Together with (44), that is, $x_1^{*,1} = \min(S, x_1^*)$, we obtain $x_1^{*,1} = x_1^*$ and hence $x_1^{*,0} = 0$. From (49) we obtain that $x_1^{*,1} = \frac{\lambda_1}{\mu_1 + \tilde{\theta}_1}$. The proof of (46) continues by induction. Assume (46) holds for $k \leq l - 1$, and let $l \leq \hat{k}$. For $k \leq l - 1$ we have that $x_k^{*,1} = \frac{\lambda_k}{\mu_k + \tilde{\theta}_k}$. Since $\sum_{k=1}^{l} \frac{\lambda_k}{\mu_k + \tilde{\theta}_k} \leq S$, by (44) we obtain that $x_l^{*,1} = x_l^*$ and hence $x_l^{*,0} = 0$. From (49) we then obtain that (46) holds for $k = l$ as well.

We now prove (47). Let $\hat{k} + 1 < \hat{l}$. From (46) and (47) we obtain that $S - \sum_{l=1}^{\hat{k}} x_l^* < x_{\hat{k}+1}^*$. So by (44) we obtain $x_{\hat{k}+1}^{*,1} = S - \sum_{l=1}^{\hat{k}} \frac{\lambda_l}{\mu_l + \tilde{\theta}_l}$ as stated in (47).

We now prove (48). From (46) and (47) we obtain that $S \leq \sum_{l=1}^{\hat{k}+1} x_l^*$, hence $x_k^{*,1} = 0$ for $k$ such that $\hat{k} + 1 < k < \hat{l}$. Equation (48) for $k \geq \hat{l}$ follows directly from (45).

In addition, $x^*$ is a global attractor, as was shown in [3, Appendix]. This can be seen by replacing the $\mu_i$ in [3] by $\mu_i + \tilde{\theta}_i$, making the ODE in [3] coincide with our ODE (43).