Asymptotic optimal control of multi-class restless bandits

Ina Maria Maaike Verloop

Asymptotic optimal control of multi-class restless bandits

I.M. Verloop[1,2]
[1] CNRS-IRIT, 2 rue C. Camichel, 31071 Toulouse Cedex 7, France
[2] Université de Toulouse; INP-ENSEEIHT, IRIT, 31071 Toulouse, France

October 6, 2012

Abstract

We study the asymptotic optimal control of multi-class restless bandits. A restless bandit is a controllable 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 make active in order to minimize the overall average cost associated to the states the bandits are in. Since finding the optimal control is typically intractable, we study an asymptotic regime instead that is obtained by letting the number of bandits that can be simultaneously made active grow proportionally with the population of bandits. We consider both a fixed population of bandits as well as a dynamic population of bandits where bandits can depart and new bandits can arrive over time to the system. We propose a class of priority policies, obtained by solving a linear program, that are proved to be asymptotically optimal under certain technical conditions. Indexability of the bandits is not required for the result to hold. For a fixed population of bandits, the technical conditions reduce to checking that a differential equation has a global attractor. For a dynamic population of bandits additional conditions are needed due to the infinite state space.

In case the bandits are indexable, we prove that Whittle's index policy is included in the class of asymptotically optimal policies. This generalizes the result of Weber and Weiss (1990) who showed asymptotic optimality of Whittle's index policy for a symmetric fixed population of bandits, to the setting of (i) several classes of bandits and (ii) possible arrivals of new bandits. In order to prove the main results we combine fluid-scaling techniques with linear programming results. This is a different proof approach than that taken in Weber and Weiss, and, in contrary to the latter, allows to include arrivals of new bandits to the system.

Finally we present a case study of *impatient* bandits: We show that the technical conditions related to the infinite state space are always satisfied and, hence, asymptotic optimality can be concluded once the global attractor property is proved. For the special case of a multi-class $M/M/S$ queue with impatient bandits the latter is satisfied and henceforth we can derive an asymptotically optimal index policy.

Keywords: Restless bandits, asymptotic optimality, Whittle's index policy, arm-aquiring 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 α bandits can be made active. The aim is to find a control that determines at each decision epoch which bandits to make active in order to minimize the overall cost associated to the states the bandits are in. In the by now classical multi-armed bandit model [16] it is assumed that only active bandits can change from state. In [41] Whittle introduced so-called restless bandits, where a bandit can also change its state while it is passive, possibly according to a different law from that which applies when it is active. The multi-armed restless bandit problem is a general 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 [17, 25, 42] for further references, applications, and possible extensions studied in the literature.

In 1979, Gittins introduced index-based policies for the non-restless bandit problem. He associated to each bandit an index being a function of the state a bandit is in and defined the policy that makes those α bandits active having currently the greatest indices (see [15]). This policy is known as the Gittins index

policy. It was first proved by Gittins that this policy is optimal in case $\alpha = 1$ [15] for the time-average and discounted cost criteria. Extensions of the optimality result of Gittins index policy when new bandits may arrive over time (for example Poisson arrivals or Bernouilli arrivals) were obtained in [39, 40]. Note that for $\alpha > 1$ the optimality result does not necessarily go through. In [31] sufficient conditions on the reward processes are given though in order to guarantee optimality of the Gittins policy for the discounted cost criterion when $\alpha > 1$.

In the presence of restless bandits, finding an optimal control is typically intractable. Whittle [41] proposed therefore to solve a relaxed optimization problem where the constraint of having at most α bandits active at a time is relaxed to a time-average or discounted version of the constraint. He observed that in case the bandits satisfy a so-called indexability property, an optimal solution to the relaxed optimization problem can be described by index values. The latter, in their turn, provide a heuristic for the original restless bandit problem, which is in the literature referred to as Whittle's index policy. In fact, this policy reduces to Gittins index policy when passive bandits are static (the non-restless case). In [18] the authors extend Whittle's index heuristic to the setting where for each bandit from multiple actions may be chosen, 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 e.g. wireless downlink scheduling [5, 23, 30], systems with delayed state observation [12], broadcast systems [33], multi-channel access models [1, 24], stochastic scheduling problems [2, 19, 28] and scheduling in the presence of impatient customers [7, 29].

As opposed to Gittins policy, Whittle's index policy is in general not an optimal solution. Only in some particular cases optimality has been proved. See for example [1, 24] where this has been proved for a symmetric restless bandit problem modeling a multi-channel access system. For a general restless bandit model, in [22] Whittle's index policy has been shown to be optimal for $\alpha = 1$ in case (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 exist for asymptotic regimes: In [41] 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 [37] assuming that the differential equation describing the fluid approximation of the system has a global attractor. Not all bandit problems satisfy this condition though, as was demonstrated in [37] with an example for which Whittle's index policy is not asymptotically optimal. A more recent result on asymptotic optimality is in [30] where the authors consider the model as studied in [24] with two classes of bandits. They prove asymptotic optimality of Whittle's index policy under a recurrence condition. The latter condition replaces the global attractor condition needed in [37] and is 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 here to [5, 6, 7, 23] where this has been studied in the context of wireless downlink channels. In particular, in [5] Whittle's index policy was obtained assuming no future arrivals of bandits and numerically shown to perform very well when including arrivals of bandits (the dynamic setting). In [6] it was shown that this heuristic is in fact maximum stable and asymptotically fluid optimal under the dynamic setting. We note that the asymptotic regime studied in [6] is different than the one as proposed by Whittle [37]. 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 [37] (as well as in this paper) the number of active bandits scales. In fact, in [6] it was shown that any policy that gives strict priority to a bandit being currently in its best possible channel condition (breaking ties according to a size-based rule), is asymptotically fluid optimal. This leaves however unanswered the question of which bandit to serve when there is currently no bandit present in its best possible state. As far as we know, no further results on (asymptotic) optimality in restless bandit models exist for the dynamic setting.

In this paper we study the asymptotic optimal control of a general multi-class restless bandit problem and 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. We derive a set of priority policies that are asymptotically optimal when certain conditions are satisfied. For a fixed population of bandits these conditions reduce to a certain differential equation having a global attractor, which coincides with the condition as needed by Weber and Weiss [37]. For a dynamic population of bandits additional technical conditions are needed due to the infinite state space.

In case the bandits are indexable we prove that Whittle's index policy is contained in the class of asymptotically optimal policies, both for a fixed population as well as for a dynamic population of bandits. This generalizes the result of Weber and Weiss [37] from a symmetric fixed population of bandits to the setting of: (i) several classes of bandits, and (ii) possible arrivals of new bandits. Moreover, as opposed to Whittle's index policy, we do not need the restless bandits to be indexable in order to define asymptotically optimal policies.

To illustrate the applicability of the results, we present a case study of *impatient* bandits. We call a bandit impatient if there is a positive probability that it abandons the system when passive. This can represent practical situations such as impatient customers, companies that go bankrupt, perishable items, etc. We then show that the technical conditions related to the infinite state space are always satisfied and, hence, asymptotic optimality can be concluded once the global attractor property is proved. For the special case of a multi-class *M/M/S* queue with impatient bandits the latter is satisfied and henceforth we derive an asymptotically optimal policy.

In the paper we will consider a small variation of the standard restless bandit formulation: Instead of having at each moment in time exactly α bandits active, we allow strictly less than α bandits to be active at a time. We handle this by introducing so-called dummy bandits. In particular, we derive that it is asymptotically optimal to make those bandits active having currently the greatest, *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 made active.

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 fluid dynamics being an LP problem we can determine its optimal equilibrium point, which we prove to be 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 by the corresponding LP problem whose optimal solution 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 [37] and allows to include arrivals of bandits to the system, whereas the approach of [37] does not. In the context of restless bandits, an LP proof approach was previously used in [8, 26, 27]. In [26, 27] it allowed to characterize and compute indexability of restless bandits. In [8] a set of LP relaxations was presented, providing performance bounds for the restless bandit problem under the discounted-cost criterion. In addition, in [8] the primal-dual index heuristic was proposed (its definition does not require indexability of the system), and proved to have a suboptimality guarantee. 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.

The remainder of the paper is organized as follows. In Section 2 we define the multi-class restless bandit problem. The asymptotic optimality results for a class of priority policies are presented in Section 3. In Section 4 we define Whittle's index policy and prove its asymptotic optimality, both for a fixed population as well as for a dynamic population of bandits. In Section 5 we present the results for impatient bandits.

2 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 $p_k(0, j)$ this bandit starts in state $j \in \{1, \ldots, J_k\}$.

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. At each decision epoch, the controller can choose for each bandit between two actions: action $a = 0$, that is, making the bandit passive, or action $a = 1$, that is, making the bandit active. If control a is performed on a class-k bandit in state i, it goes with rate $q_k^a(i, j)$ to state j, $j = 0, 1, \ldots, J_k$, $j \neq i$, where we interpret $j = 0$ as the fact that the bandit has departed from the system. Note that the evolution of one bandit (given its action) is independent of that of all the other bandits. We write $q_k^a(j, j) = -\sum_{i=0, i\neq j}^{J_k} q_k^a(j, i)$ and define $q_k^a(j) := -q_k^a(j, j)$ as the total transition rate out of state j for a class-k bandit under action a. At any moment in time at most α bandits can be made active. We note that this problem is referred to as the *restless* bandit problem

since even though a bandit is passive, it can still change from state.

For a given policy π 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 see action a. 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.

The cost per unit of time of having a class-k customer in state j under action a is equal to $C^a_{j,k} \in \mathbb{R}$. We note that the cost $C^a_{j,k}$ can be negative, i.e., representing a reward. The objective is to minimize the long-run average holding cost among all Markovian policies (policies that base their decisions on the current state and time), i.e., find a Markovian policy π^* that minimizes

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\Big(\int_{t=0}^{T} C^a_{j,k} X^{\pi,a}_{j,k}(t)\mathrm{d}t\Big), \tag{1}$$

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

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

Throughout this paper we will consider two different scenarios: a fixed population of bandits, and a dynamic population of bandits. Below we further specify the model for the different settings:

- *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., $q^a_k(j, 0) = 0$ for all $k = 1, \ldots, K$, $j = 1, \ldots, J_k$, $a = 0, 1$. Hence, $X^\pi_k(t) = X_k(0)$ for any time t and any policy π.
- *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 $q^a_k(j, 0) > 0$.

For a given policy π, we will call the system stable if the process $\vec{X}^\pi(t)$ has a unique invariant probability distribution with finite first moment. Note that it can be that the system is unstable. This depends strongly on the employed policy. In the case of a fixed population of bandits, the state space is finite, hence the process $X^\pi(t)$ being unichain would be a sufficient condition for stability. In the case of a dynamic population of bandits the stability condition is more involved. We will therefore assume throughout the paper that the maximum stability conditions are satisfied, that is, the traffic parameters are such that there exists a policy that makes the system stable.

Throughout the paper we will need to make additional restrictions on the model and policies considered in order for the results to hold. When doing so, we will make a distinction between (i) assumptions made on the model parameters (referred to as "Assumption") and (ii) conditions that are posed on the policy under investigation (referred to as "Condition").

Remark 2.1 (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 the possibility of multiple actions per bandit, that is, a class-k bandit in state j can chose from any action $a \in \{0, \ldots, A_{j,k}\}$ and at most α bandits can be non-passive at a time, i.e., $\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=1}^{A_{j,k}} X^a_{j,k}(t) \le \alpha$. We note that all results that will be obtained in Section 3 will go through in the multi-action context. For ease of exposition we chose however $A_{j,k} = 1$, this being the scenario for which Whittle's index is defined, see Section 4.*

3 Fluid analysis and asymptotic optimality

In this section we consider a fluid formulation of the multi-class restless bandit problem, which allows to derive a class of priority policies that asymptotically minimize the cost of the original stochastic model (as given in (1)). More precisely, in Section 3.1 we introduce a fluid control problem and show that its optimal fluid cost provides a lower bound on the cost in the original stochastic model. In Section 3.2 we then define a class of priority policies for the original stochastic model based on the optimal fluid solution, which are shown to be asymptotically optimal in Section 3.3.

3.1 Fluid control problem and lower bound

The fluid control problem arises from the original stochastic model by only taking into account 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. Its dynamics is then described by

$$\frac{\mathrm{d}x^u_{j,k}(t)}{\mathrm{d}t} = \lambda_k p_k(0,j) + \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k} x^{u,a}_{i,k}(t) q^a_k(i,j) - \sum_{a=0}^{1} x^{u,a}_{j,k}(t) q^a_k(j), \tag{3}$$

with the constraint on the total amount of active fluid given by $\sum_{k=1}^{K}\sum_{j=1}^{J_k} x^{u,1}_{j,k}(t) \leq \alpha$, for all $t \geq 0$. We are interested in finding an optimal equilibrium point of the fluid dynamics, as given in (3), with as goal to minimize its holding cost. Hence, we pose the following linear optimization problem:

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

where the constraint (5) follows since by (3) we have $\sum_{j=1}^{J_k} \frac{\mathrm{d}}{\mathrm{d}t} x^u_{j,k}(t) = 0$ if $\lambda_k = 0$, for all $t \geq 0$.

We denote by x^* an optimal solution of the above problem (LP), assuming it exists, and denote the optimal value by

$$V^*(\alpha) := \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C^a_{j,k} x^{*,a}_{j,k}.$$

We now prove some results for the optimization problem (LP).

Lemma 3.1 *The feasible set of (LP) is non-empty and $V^*(\alpha) < \infty$.*

Proof: Let π be a policy for which a unique invariant distribution exists having finite first moment. Let $X_k(0) = x_k(0)$.

For a fixed population, we have that $\lim_{t\to\infty} \frac{X^\pi_{j,k}(t)}{t} \leq \lim_{t\to\infty} \frac{X_k(0)}{t} = 0$, for all j,k. For a dynamic population, stability of policy π implies rate-stability, that is, $\lim_{t\to\infty} \frac{X^\pi_{j,k}(t)}{t} = 0$, for all j,k. Hence, in both cases we conclude that

$$\lim_{t\to\infty} \frac{X^\pi_{j,k}(t)}{t} = 0, \text{ for all } j,k. \tag{6}$$

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

$$X^\pi_{j,k}(t) = X^\pi_{j,k}(0) + N^{\lambda_k p_k(0,j)}(t) + \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k} N^{q^a_k(i,j)}\left(\int_0^t X^{\pi,a}_{i,k}(s)\mathrm{d}s\right) - \sum_{a=0}^{1} N^{q^a_k(j)}\left(\int_0^t X^{\pi,a}_{j,k}(s)\mathrm{d}s\right), \tag{7}$$

where $N^\theta(t)$ is a Poisson process with rate θ. By the ergodic theorem [11], 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 (7) by t, using that $N^\theta(at)/t \to a\theta$ as $t \to \infty$, and together with (6), we obtain that

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

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 $V^*(\alpha) < \infty$. □

The optimal solution of the fluid control problem (LP) serves as a lower bound on the cost of the original stochastic optimization problem (1). This is proved in Lemma 3.2 and Lemma 3.3 for the fixed and dynamic population, respectively. The proofs can be found in Appendix A.

Lemma 3.2 *Assume a fixed population of bandits. For any policy π we have*

$$\sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}\left(\int_0^T C^a_{j,k} X^{\pi,a}_{j,k}(t) \mathrm{d}t \right) \geq V^*(\alpha). \tag{8}$$

Lemma 3.3 *Assume a dynamic population of bandits. For any policy π that is stable we have*

$$\sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}\left(\int_0^T C^a_{j,k} X^{\pi,a}_{j,k}(t) \mathrm{d}t \right) \geq V^*(\alpha). \tag{9}$$

If $C^a_{j,k} > 0$ for all j, k, a, then for any policy π we have

$$\sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\left(\int_0^T C^a_{j,k} X^{\pi,a}_{j,k}(t) \mathrm{d}t \right) \geq V^*(\alpha), \tag{10}$$

and if $C^a_{j,k} > 0$ for all j, k, a, then for any policy π that is either rate-stable (i.e., $\lim_{t\to\infty} \sum_{j,k} \frac{X^\pi_{j,k}(t)}{t} = 0$ almost surely) or mean rate-stable (i.e., $\lim_{t\to\infty} \sum_{j,k} \frac{\mathbb{E}(X^\pi_{j,k}(t))}{t} = 0$), we have

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

3.2 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.

Below we define a set of priority policies that will play a key role in the remaining of the paper.

Definition 3.4 (Set of priority policies) *We define the set of priority policies Π^* as follows:*

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

with

$X^* :=$

$\{x^*$ *an optimal solution of (LP) s.t.* $x^{*,0}_{j,k} x^{*,1}_{j,k} = 0 \ \forall j, k$, *with the exception of at most one pair of indices*$\}$,

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

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

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

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

In Section 3.2.1 and 3.2.2 we comment on the non-emptiness of the class of policies Π^* and in Section 3.3 we will show that a policy in this class is asymptotically optimal if it satisfies certain conditions. We emphasize that we do not require the bandits to be indexable in order to define the set of priority policies Π^*. This as opposed to the definition of Whittle's index policy, which is only well defined in case the system is indexable.

Before continuing we first give an example of Definition 3.4.

Example 3.5 *Assume* $K = 2$ *and* $J_k = 2$. *Let* x^* *be such that for class 1 we have* $x_{1,1}^{*,0} = 0$, $x_{2,1}^{*,0} = 4$, $x_{1,1}^{*,1} = 3$, $x_{2,1}^{*,1} = 1$ *and for class 2 we have* $x_{1,2}^{*,0} = 2$, $x_{2,2}^{*,0} = 0$, $x_{1,2}^{*,1} = 0$, $x_{2,2}^{*,1} = 5$ *and* $\alpha = 10$. *The priority policies associated to* x^* *in the set* $\Pi(x^*)$, *as defined in Definition 3.4, 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* $x_{1,1}^{*,1} + x_{2,1}^{*,1} + x_{1,2}^{*,1} + x_{2,2}^{*,1} = 9 < \alpha$, *class-2 bandits in state 1 are never made active. Hence, 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* $(2,2) \succ (1,1) \succ (2,1)$, *where a state* (j,k) *denotes a class-k bandit in state* j. *In both cases state* $(1,2)$ *is never made active.*

In order to prove asymptotic optimality of a policy $\pi^* \in \Pi^*$, as will be done in Section 3.3, we will need to investigate the ordinary differential equation (ODE), $x^*(t)$, defined by the transition rates of the process $X^{\pi^*}(t)$. The transition rates of the Markov process $X^{\pi^*}(t)$ are as follows:

$$x \to x + e_{j,k} \qquad \text{at rate } \lambda_k p_k(0,j), \quad k = 1,\dots,K,\ j = 1,\dots,J_k, \tag{11}$$

$$x \to x - e_{j,k} \qquad \text{at rate } \sum_{a=0}^{1} x_{j,k}^{a} q_k^a(j,0), \quad k = 1,\dots,K,\ j = 1,\dots,J_k, \tag{12}$$

$$x \to x - e_{j,k} + e_{i,k} \qquad \text{at rate } \sum_{a=0}^{1} x_{j,k}^{a} q_k^a(j,i), \quad k = 1,\dots,K,\ i,j = 1,\dots,J_k, \tag{13}$$

where $x_{j,k}^1 = \min\left((\alpha - \sum_{(i,l)\in S_{j,k}^*} x_{i,l})^+, x_{j,k}\right)$, if $(j,k) \notin I^*$, 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. Further, $S_{j,k}^*$ is defined as the set of all states $(i,l), i = 1,\dots,J_l, l = 1,\dots,K$, such that class-$l$ bandits in state i have higher priority than class-k bandits in state j under policy π^*, and I^* is the set of all states that will never be made active under policy π^*. The process $x^*(t)$ is hence defined by

$$\frac{\mathrm{d}x_{j,k}^*(t)}{\mathrm{d}t} = \lambda_k p_k(0,j) + \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k} x_{i,k}^{*,a}(t) q_k^a(i,j) - \sum_{a=0}^{1} x_{j,k}^{*,a}(t) q_k^a(j), \quad \forall\, j,k, \tag{14}$$

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

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

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

It follows directly that an optimal solution x^* of (LP) is in fact an equilibrium point of the ODE $x^*(t)$ under a policy $\pi^* \in \Pi(x^*)$.

Lemma 3.6 *Let* $\pi^* \in \Pi^*$ *and let* x^* *be a point such that* $\pi^* \in \Pi(x^*)$. *Then* x^* *is an equilibrium point of the ODE* $x^*(t)$ *as defined in* (14) *for the policy* π^*.

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 ODE $x^*(t)$. □

Besides x^* being an equilibrium point, in order to prove asymptotic optimality of π^* it will be needed that x^* is the unique equilibrium point of (14) and in addition is a global attractor, i.e., all trajectories converge to x^*. This is not true in general, which is why we state it as a condition that a policy needs to satisfy. In fact, in Section 5.2 we will show this condition to be satisfied for a specific model with impatient bandits.

Condition 3.7 *Given a policy $\pi^* \in \Pi(x^*) \subset \Pi^*$, for an $x^* \in X^*$. The point x^* is the unique equilibrium point and the global attractor of the ODE $x^*(t)$ as defined in (14) for the policy π^*.*

We conclude this section with a discussion on the existence of policies in the set Π^*. In Subsection 3.2.1 we consider a fixed population of bandits and Subsection 3.2.2 considers the dynamic case.

3.2.1 Fixed population of bandits

The following result states that for a fixed population of bandits the set Π^* is always non-empty.

Lemma 3.8 *Assume a fixed population of bandits. Then the set of priority policies Π^* is non-empty.*

Proof: 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 [32, 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 3.4.

Otherwise, let S denote the set of states (i,l) such that $x^*_{i,l} = 0$. By (4), if $(j,k) \in S$, then $\sum_{a=0}^{1} \sum_{i \neq j} x^{*,a}_{i,k} q^a_k(i,j) = 0$. That is, $x^{*,a}_{i,k} q^a_k(i,j) = 0$ 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

$$0 = \sum_{a=0}^{1} \sum_{i=1, i \neq j, (i,k) \in S^c}^{J_k} x^{*,a}_{i,k} q^a_k(i,j) - \sum_{a=0}^{1} x^{*,a}_{j,k} \tilde{q}^a_k(j), \quad \forall\, j,k,$$

with $\tilde{q}^a_k(j) := \sum_{i=0, i \neq j, (i,k) \in S^c}^{J_k} q^a_k(j,i)$. 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 3.4.

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 3.4. □

Remark 3.9 *In [8] 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 a 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 3.4, and hence is included in the set of asymptotically optimal priority policies Π^*.*

3.2.2 Dynamic population of bandits

Before stating the lemma on the non-emptiness of the set Π^* for the dynamic population, we first introduce two technical assumptions.

Assumption 3.10 *The set of optimal solutions of (LP) is bounded.*

The above assumption is always satisfied if $C^0_{j,k} > 0$ for all j,k, since $x^1_{j,k} \leq \alpha$ and $x^{*,0}_{j,k}$ is upperbounded by the cost value of a feasible solution divided by $C^0_{j,k} > 0$.

Assumption 3.11 *One of the following holds:*

- $p_k(0,j) > 0$ *for all* j,k, *or,*

- *for all $\epsilon > 0$ small enough, the set of optimal solutions of (LP(ϵ)) is bounded and non-empty, with $LP(\epsilon)$ defined by*

$$
\begin{aligned}
(LP(\epsilon)) \qquad & \min_x \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{A_{j,k}} C_{j,k}^a x_{j,k}^a \\
& s.t. \;\; 0 = \lambda_k (p_k(0,j) + \epsilon) + \sum_{a=0}^{1} \sum_{i=1, i \neq j}^{J_k} x_{i,k}^a q_k^a(i,j) - \sum_{a=0}^{1} x_{j,k}^a q_k^a(j), \quad \forall\, j,k, \qquad (15) \\
& \sum_{k=1}^{K} \sum_{j=1}^{J_k} x_{j,k}^1 \leq \alpha, \\
& x_{j,k}^a \geq 0, \quad \forall\, j,k,a.
\end{aligned}
$$

As before, the boundedness of optimal solutions of $(LP(\epsilon))$ is always satisfied in case $C_{j,k}^0 > 0$ for all j,k.

The following lemma shows that Assumption 3.10 and Assumption 3.11 are sufficient conditions in order for the set of priority policies Π^* to be non-empty. We note that in case the bandit problem is indexable, as defined in Section 4.3, the non-emptiness of the set Π^* follows directly from the fact that Whittle index policy is included in Π^*, as will be proved in Section 4.3.

Lemma 3.12 *Assume a dynamic population of bandits. If Assumptions 3.10 and 3.11 are satisfied, then the set of priority policies Π^* is non-empty.*

Proof: First assume $p_k(0,j) > 0$ 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 $x_{j,k}^{*,a} > 0$. Since the set of optimal solutions of (LP) is non-empty and bounded (Assumption 3.10), from standard LP theory, see [32, 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 $x_{j,k}^{*,a} = 0$, and in at most one combination (j,k) the components $x_{j,k}^{*,a}$ can be positive in both actions $a = 0$ and $a = 1$.

Now assume that for all $\epsilon > 0$ small enough the set of optimal solutions of (LP(ϵ)) is bounded. By sensitivity results of linear programming theory we have that the same basis provides an optimal solution for (LP(ϵ)), for all $0 \leq \epsilon < \bar{\epsilon}$ and $\bar{\epsilon} \geq 0$ small enough. We denote the corresponding optimal solution by $x^*(\epsilon)$. By (15) 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 ϵ) such that $x_{j,k}^{*,a}(\epsilon) = 0$ and for at most one state (j,k) (independent of ϵ) the components $x_{j,k}^{*,a}(\epsilon)$ can be strictly positive for both actions $a = 0$ and $a = 1$.

Note that $(LP(0)) = (LP)$. Hence, by Assumption 3.10 and using [10, Corollary 1], we obtain that the correspondence that gives for each ϵ the set of optimal solutions of (LP(ϵ)) is upper semicontinuous in the point $\epsilon = 0$. Being a compact-valued correspondence, it follows that there exists a sequence ϵ_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. □

3.3 Asymptotic optimality of priority policies

In this section we present results showing the asymptotic optimality of priority policies in the set Π^*. In particular, we obtain that, after a *fluid scaling*, the long-run average holding cost is optimized under a priority policy in Π^* (satisfying certain additional conditions).

We will consider the multi-class 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 per time unit. 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 $X_{j,k}^r(0) = r \cdot x_{j,k}(0)$, with $x_{j,k}(0) \geq 0$.

For a given policy π, 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 will be interested in the process under the fluid scaling, i.e.,

space is scaled linearly with the parameter r:

$$x_{j,k}^{r,\pi,a}(t) := X_{j,k}^{r,\pi,a}(t)/r. \tag{16}$$

We consider the cost of the stochastic model after fluid scaling, that is, we are interested in

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \mathbb{E}\left(\int_0^T C_{j,k}^a x_{j,k}^{r,\pi,a}(t) \mathrm{d}t \right). \tag{17}$$

Our goal is to find policies that minimize the above as $r \to \infty$. We directly obtain from Lemma 3.2 and Lemma 3.3 that $V^*(\alpha)$ provides a lower bound on this cost. We will therefore call a policy *asymptotically optimal* when the fluid-scaled cost (17) converges to this lower bound.

In the remainder of this section we prove asymptotic optimality of priority policies in the set Π^* for both the setting of a fixed and a dynamic population of bandits. The proofs consist in 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) and has cost value $V^*(\alpha)$, this implies that the fluid-scaled cost (17) under policy π^* converges to $V^*(\alpha)$. Since $V^*(\alpha)$ serves as a lower bound on the average cost, this allows us to conclude for asymptotic optimality of the priority policy π^*.

3.3.1 Fixed population of bandits

For a fixed population of bandits the asymptotic optimality result is presented in the following proposition.

Proposition 3.13 *Assume a fixed population of bandits. For a given policy $\pi^* \in \Pi(x^*) \subset \Pi^*$, assume Condition 3.7 is satisfied and the Markov process $x^{r,\pi^*}(t)$ is unichain (there is a state x such that there is a path from any state to x, i.e., state x is recurrent), for any r. Then*

$$\lim_{r\to\infty} \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \mathbb{E}\left(\int_0^T C_{j,k}^a x_{j,k}^{r,\pi^*,a}(t) \mathrm{d}t \right) = V^*(\alpha).$$

In particular, for any policy π we have

$$\begin{aligned} &\lim_{r\to\infty} \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \mathbb{E}\left(\int_0^T C_{j,k}^a x_{j,k}^{r,\pi^*,a}(t) \mathrm{d}t \right) \\ &\leq \liminf_{r\to\infty} \liminf_{T\to\infty} \frac{1}{T} \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \mathbb{E}\left(\int_0^T C_{j,k}^a x^{r,\pi,a}(t) \mathrm{d}t \right). \end{aligned} \tag{18}$$

Proof: The transition rates of the process $x^{r,\pi^*}(t)$ are as follows:

$$x \to x - \tfrac{e_{j,k}}{r} + \tfrac{e_{i,k}}{r} \qquad \text{at rate } \; r \sum_{a=0}^{1} x_{j,k}^a q_k^a(j,i), \quad k = 1,\dots,K, \; i,j = 1,\dots,J_k, \tag{19}$$

where $x_{j,k}^1 = \min\left((\alpha - \sum_{(i,l)\in S_{j,k}^*} x_{i,l})^+, x_{j,k} \right)$, if $(j,k) \notin I^*$, 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 only zeros except for component (j,k) which is equal to one. Here $S_{j,k}^*$ is defined as the set of all states $(i,l), i = 1,\dots,J_l, l = 1,\dots,K$, such that class-$l$ bandits in state i have higher priority than class-k bandits in state j under policy π^*, and I^* is the set of all states that will never be made active under policy π^*.

The transition rate from x to $x + l/r$ has the form $rb_l(x)$, see (19), with $l \in \mathcal{L}$ and $\mathcal{L}$ composed of a finite number of vectors in $\mathbb{N}^{\sum_k J_k}$. Hence, the process $x_{j,k}^{r,\pi^*}(t)$ belongs to the family of density dependent population processes as defined in [13, Chapter 11].

Note that the ODE $x^*(t)$ as defined in (14) can equivalently be written as $\frac{\mathrm{d}x^*(t)}{\mathrm{d}t} = F(x^*(t))$, with $F(x^*) = \sum_{l\in\mathcal{L}} l b_l(x^*)$. We note that $F(\cdot)$ is Lipschitz continuous. From Condition 3.7 we have that x^* is the unique global attractor of $x^*(t)$.

The state space of $x^{r,\pi^*}(t)$ is finite. Hence, the unichain assumption guarantees that for each r there is a unique equilibrium distribution p^{r,π^*}, [35], and the family $\{p^{r,\pi^*}\}$ is tight (i.e., for every $\epsilon > 0$

there is some compact subset K of the state space such that $p^{r,\pi^*}(K) \geq 1-\epsilon$ for all r). Together with Condition 3.7 and [14, Corollary 5] it follows then that $p^{r,\pi^*}(x)$ converges to the Dirac measure in x^*, the global attractor of $x^*(t)$, defined in (14). Hence, we can write

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\lim_{r\to\infty}\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\left(\int_0^T C_{j,k}^a x_{j,k}^{r,\pi^*,a}(t)\mathrm{d}t\right)\\
&=\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\lim_{r\to\infty}\sum_x p^{r,\pi^*}(x)C_{j,k}^a x_{j,k}^a=\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}C_{j,k}^a x_{j,k}^{*,a}=V^*(\alpha),
\end{aligned}$$

where the first step follows from the ergodicity theorem [11, 35], the second step (interchange of limit and summation) follows since $p^{r,\pi^*}(x)=0$ when $x_{j,k}^a > x_k(0)$ and since p^{r,π^*} 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 $V^*(\alpha)$ is a lower bound on the steady-state cost, as shown in Lemma 3.2. □

3.3.2 Dynamic population of bandits

Due to the infinite state space in case of a dynamic population setting we need to pose further conditions on the policy π^* in order to obtain asymptotic optimality results. In Section 5.1 we will show these conditions to be always satisfied in case of impatient bandits.

Condition 3.14 *Given a policy $\pi^* \in \Pi^*$.*

- *The process $x^{r,\pi^*}(t)$ is irreducible and has an invariant probability distribution p^{r,π^*} with a finite first moment, for all r.*
- *The family $\{p^{r,\pi^*}, r\}$ is tight.*
- *The family $\{p^{r,\pi^*}, r\}$ is uniform integrable.*

We can now state the asymptotic optimality result for a dynamic population of bandits.

Proposition 3.15 *Assume a dynamic population of bandits. For a given policy $\pi^* \in \Pi(x^*) \subset \Pi^*$, assume Condition 3.7 and Condition 3.14 are satisfied. Then,*

$$\lim_{r\to\infty}\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\mathbb{E}\left(\int_0^T C_{j,k}^a x_{j,k}^{r,\pi^*,a}(t)\mathrm{d}t\right)=V^*(\alpha).$$

In particular, for any policy π that is stable we have

$$\begin{aligned}
&\lim_{r\to\infty}\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\mathbb{E}\left(\int_0^T C_{j,k}^a x_{j,k}^{r,\pi^*,a}(t)\mathrm{d}t\right)\\
&\leq\lim_{r\to\infty}\liminf_{T\to\infty}\frac{1}{T}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\mathbb{E}\left(\int_0^T C_{j,k}^a x^{r,\pi,a}(t)\mathrm{d}t\right).
\end{aligned}\tag{20}$$

If $C_{j,k}^a > 0$, for all j,k,a, then for any policy π we have

$$\begin{aligned}
&\lim_{r\to\infty}\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\mathbb{E}\left(\int_0^T C_{j,k}^a x_{j,k}^{r,\pi^*,a}(t)\mathrm{d}t\right)\\
&\leq\lim_{r\to\infty}\liminf_{T\to\infty}\limsup_{T\to\infty}\frac{1}{T}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\mathbb{E}\left(\int_0^T C_{j,k}^a x^{r,\pi,a}(t)\mathrm{d}t\right),
\end{aligned}\tag{21}$$

and if $C^a_{j,k} > 0$, for all j, k, a, then for any policy π that is either rate-stable or mean rate-stable we have

$$\lim_{r\to\infty}\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\mathbb{E}\left(\int_0^T C^a_{j,k}x^{r,\pi^*,a}_{j,k}(t)\mathrm{d}t\right)$$
$$\leq \liminf_{r\to\infty}\liminf_{T\to\infty}\frac{1}{T}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\mathbb{E}\left(\int_0^T C^a_{j,k}x^{r,\pi,a}(t)\mathrm{d}t\right). \tag{22}$$

Proof: The transition rates of the process $x^{r,\pi^*}(t)$ are defined as follows:

$$x \to x + \tfrac{e_{j,k}}{r} \qquad \text{at rate } \; r\lambda_k p_k(0,j), \quad k = 1,\dots,K,\ j = 1,\dots,J_k, \tag{23}$$

$$x \to x - \tfrac{e_{j,k}}{r} \qquad \text{at rate } \; r\sum_{a=0}^{1} x^a_{j,k}q^a_k(j,0), \quad k = 1,\dots,K,\ j = 1,\dots,J_k, \tag{24}$$

$$x \to x - \tfrac{e_{j,k}}{r} + \tfrac{e_{i,k}}{r} \qquad \text{at rate } \; r\sum_{a=0}^{1} x^a_{j,k}q^a_k(j,i), \quad k = 1,\dots,K,\ i,j = 1,\dots,J_k, \tag{25}$$

where $x^1_{j,k} = \min\left((\alpha - \sum_{(i,l)\in S^*_{j,k}} x_{i,l})^+, x_{j,k}\right)$, if $(j,k) \notin I^*$, and $x^1_{j,k} = 0$ otherwise, $x^0_{j,k} = x_{j,k} - x^1_{j,k}$, and $e_{j,k}$ is a vector composed of all zeros except for component (j,k) which is one. Here $S^*_{j,k}$ is defined as the set of all states $(i,l), i = 1,\dots,J_l, l = 1,\dots,K$, such that class-$l$ bandits in state i have higher priority than class-k bandits in state j under policy π^*, and I^* is the set of all states that will never be made active under policy π^*.

As in the proof of Proposition 3.13, the transition rates have the form $rb_l(x)$, see (23)–(25), and hence the process $x^{r,\pi^*}_{j,k}(t)$ belongs to the family of density dependent population processes.

Note that the ODE $x^*(t)$ as defined in (14) can equivalently be written as $\frac{x^*(t)}{\mathrm{d}t} = F(x^*(t))$, with $F(x^*) = \sum_{l\in\mathcal{L}} lb_l(x^*)$. We note that $F(\cdot)$ is Lipschitz continuous. From Condition 3.7 we have that x^* is the unique global attractor of $x^*(t)$.

From the irreducibility assumption and the existence of an invariant distribution, we obtain that there is a unique invariant distribution (ergodicity theorem [20, Section 6.9]). From Condition 3.7, Condition 3.14 and [14, Corollary 5] we obtain that $p^{r,\pi^*}(x)$ converges to the Dirac measure in x^*, the global attractor of $x^*(t)$, defined in (14), and we can write

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\lim_{r\to\infty}\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\left(\int_0^T C^a_{j,k}x^{r,\pi^*,a}_{j,k}(t)\mathrm{d}t\right)$$
$$= \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\lim_{r\to\infty}\sum_x p^r(x)C^a_{j,k}x^a_{j,k} = \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}C^a_{j,k}x^{*,a}_{j,k} = V^*(\alpha),$$

where the first step follows from the ergodicity theorem [35, 11] (applicable since the first moment of p^{r,π^*} is finite), the second step follows from uniform integrability of $\{p^{r,\pi^*}\}$ and the fact that p^{r,π^*} converges to the Dirac meausre in x^*, and the last step follows since x^* is an optimal solution of (LP).

We conclude the proof by noting that $V^*(\alpha)$ is a lower bound on the steady-state cost, as shown in Lemma 3.3. □

4 Whittle's index policy

In Section 3.3 we found that priority policies inside the set Π^* are asymptotically optimal. In this section we will show that, under certain conditions, Whittle's index policy is in fact included in Π^* and is hence asymptotically optimal.

In Section 4.1 we first define Whittle's index policy. In Section 4.2 and Section 4.3 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.

4.1 Relaxed constraint optimization problem and Whittle's indices

Whittle's index policy has been proposed by Whittle [41] as an efficient heuristic for the multi-class restless bandit problem ([17, Section 6] and [42, Section II.14]). Under this policy each bandit is associated a Whittle's index being a function of the state a bandit is in, making those bandits active having currently the greatest indices. In this section we will describe how these Whittle's indices are derived. In order to do so, we will need to introduce a new optimization problem. We note that this problem is different than that originally posed in Section 2. In fact, in Section 4.2 and Section 4.3 we make the connection with the original problem as we show that Whittle's index policy is asymptotically optimal for the original problem as posed in Section 2.

The optimization problem: In order to define the Whittle's indices, we consider the following optimization problem: Assume we have a population of bandits (at least one bandit from each class) and assume there are no further arrivals in the future. At any moment in time at most α bandits can be made active and we restrict ourselves to the class of *stationary* and Markovian policies. As cost criterion we will consider one of the following: the average-cost criterion

$$\mathbb{C}^{av}(Y(\cdot)) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T Y(t)\mathrm{d}t\right), \tag{26}$$

or the discounted-cost criterion

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

for $\beta \in (0,1)$. The objective is to find a policy that minimizes

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\mathbb{C}^f(C^a_{j,k}X^{\pi,a}_{j,k}(\cdot)), \tag{27}$$

with $f \in \{av, \beta\}$. We note that the objective posed in (1) was that of the average-cost criterion. In Section 4.3 it will become clear why in this section we also need to introduce the discounted-cost criterion.

Relaxed constraint optimization problem: The restless property of the bandits makes the above-described optimization problem often infeasible to solve. Instead, Whittle [41] proposed to study the so-called *relaxed constraint optimization problem*, which is defined as follows: find a policy that minimizes (27) under the relaxed constraint

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k}\mathbb{C}^f(X^{\pi,1}_{j,k}(\cdot)) \le \alpha(f), \tag{28}$$

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 α bandits can be made active at any moment in time is replaced by its time-average or discounted version, (28). 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.

Dummy bandits: In standard restless bandit problems the constraint (28) needs to be satisfied in the strict sense, that is, with an "=" sign. In this paper we allowed however strictly less than α bandits to be active at a time. In order to be able to deal with this within the original framework of Whittle, we 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, so $q^a(B,B) = 1$ for all a. The introduction of these $\alpha(f)$ dummy bandits allows to reformulate the relaxed constraint problem as follows: Minimize (27) under the relaxed constraint

$$\mathbb{C}^f(X^{\pi,1}_B(\cdot)) + \sum_{k=1}^{K}\sum_{j=1}^{J_k}\mathbb{C}^f(X^{\pi,1}_{j,k}(\cdot)) = \alpha(f). \tag{29}$$

This constraint is equivalent to (28) since, for a given set of active bandits, activating additional dummy bandits does not modify the behavior of the system.

Lagrangian relaxation: Using the Lagrangian approach, we write the relaxed constraint problem (minimize (27) under constraint (29)) as the problem of finding a policy π that minimizes

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k}\mathbb{C}^f(C^0_{j,k}X^{\pi,0}_{j,k}(\cdot)+C^1_{j,k}X^{\pi,1}_{j,k}(\cdot)+\nu X^{\pi,1}_{j,k}(\cdot))+\mathbb{C}^f(\nu X^{\pi,1}_B(\cdot)). \tag{30}$$

The Lagrange multiplier ν 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 ν such that the constraint (29) is satisfied.

Optimization problem per bandit: Since there is no more a common constraint for the bandits, problem (30) can be decomposed into several subproblems, one for each bandit. So, for each class-k bandit the subproblem is to minimize

$$\mathbb{C}^f(C^{A_k(\cdot)}_{J_k(\cdot),k}+\nu\mathbf{1}_{(A_k(\cdot)=1)}), \tag{31}$$

where $J_k(t)$ and $A_k(t)$ denote the state and action chosen for the class-k bandit at time t, respectively, and for each dummy bandit the problem is to minimize

$$\nu\ \mathbb{C}^f(\mathbf{1}_{(A_B(\cdot)=1)}), \tag{32}$$

with $A_B(t)$ the action chosen for the dummy bandit at time t.

Whittle's index: For a given optimization criterion f, Whittle defines the index $\nu^f_{j,k}$ as the least value of ν for which it could be optimal in (31) to make the class-k bandit in state j passive. We refer to $\nu^f_{j,k}$ as *Whittle's index* for the optimization criterion f. Similarly, we define the index ν^f_B as the least value of ν for which it could be optimal in (32) to make a dummy bandit passive.

A bandit is called *indexable* if the set of states where it could be optimal in (31) to make a class-k bandit passive forms an increasing sequence of sets (from the empty set to the set $\{1,\ldots,J_k\}$), as ν increases. We refer to the problem as being indexable if all bandits are indexable. Note that whether a problem is indexable or not can depend on the choice for f (and β). We refer to [28] for a survey on indexation results. In particular, [28] 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 [24, 36].

We note that the dynamics of a bandit in state B is independent of the action chosen. Since ν represents the cost to be paid when active, it will be optimal in (32) 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 $\nu^f_B = 0$.

Optimal solution per bandit: If the bandit problem is indexable, an optimal policy for the subproblem (31) is then such that the class-k bandit in state j is made active if $\nu^f_{j,k} > \nu$, is made passive if $\nu^f_{j,k} < \nu$, and any action is optimal if $\nu^f_{j,k} = \nu$, [41].

Optimal solution of relaxed constraint problem: An optimal solution to (27) under the relaxed constraint (29) is now obtained by setting ν at the appropriate level ν^* such that (29) is satisfied and to apply to each bandit the optimal solution, i.e., make a class-k bandit in state j active if $\nu^f_{j,k} > \nu^*$, and passive if $\nu^f_{j,k} < \nu^*$. Further, the value for ν^* will be such that $\nu^* = \nu^f_{j^0,k^0}$ for an (j^0,k^0) with $\nu^f_{j^0,k^0} \geq 0$. States (j^i,k^i) with $\nu^f_{j^i,k^i} = \nu^*$, $i = 1,\ldots,I$, will be ordered arbitrarily, for example, $(j^1,k^1) \succ \ldots \succ (j^I,k^I)$). Then there is an $\tilde{i} \in \{1,\ldots,I\}$ and $\gamma \in [0,1]$ such that the relaxed constraint (29) is satisfied when making bandits active in states (j^i,k^i) with $i < \tilde{i}$, making bandits passive in states (j^i,k^i) with $i > \tilde{i}$, and making bandits in state $(j^{\tilde{i}},k^{\tilde{i}})$ active with probability $\gamma \in [0,1]$, [37, 41]. In case $\nu^* = 0$, we take the convention that the randomization is done among the bandits in state B, while any class-k bandit in a state j with $\nu^f_{j,k} = 0$ is kept passive.

Since there are always $\alpha(f)$ bandits in state B, we necessarily have that an optimal action for class-k bandits in state j with $\nu^f_{j,k} < 0 = \nu^f_B$ (having lower priority than bandits in state B) is to be passive. In particular, this implies that we can assume $\nu^* \geq 0$ (when $\nu^* \geq 0$, any class-k bandit in state j with $\nu^f_{j,k} < 0$ will be made passive since $\nu^f_{j,k} < \nu^*$).

Heuristic: Whittle's index policy: Obviously the above optimal control for the relaxed problem is not feasible for the original optimization problem (as posed in the beginning of Section 4.1) having as constraint that at most α bandits can be made active *at any moment in time*. Whittle [41] therefore proposed the following heuristic: At any moment in time, make α bandits active having currently the

greatest Whittle's indices. Hence, if $\nu^f_{j,k} < \nu^f_{i,l}$, then a class-$l$ bandit in state i is given higher priority than a class-k bandit in state j. Recall that a class-k bandit in state j with $\nu^f_{j,k} \leq 0$ will never be made active in the optimal solution of the relaxed problem. Hence, analogously, we define that under Whittle's index policy a class-k bandit in state j will never be made active if $\nu^f_{j,k} \leq 0$ (even though it would be possible to activate this bandit). Further, we note that in case different states have the same Whittle index, an arbitrary fixed priority rule is used.

The above described heuristic is referred to in the literature as *Whittle's index policy* for the average cost ($f = av$) or discounted cost ($f = \beta$) criterion. In the next two sections we will prove that this heuristic is asymptotically optimal, both for the static and dynamic population.

Remark 4.1 (Robustness) *We note that Whittle's indices are robust in the sense that they do not depend on the arrival characteristics of new bandits or on the exact number of bandits that can be made active, α. The Whittle's indices do depend on f (and β) though.*

4.2 Asymptotic optimality for a fixed population of bandits

In this section we will prove asymptotic optimality of Whittle's index policy in the setting of a fixed population of bandits. More precisely, we show that Whittle's index policy (defined for the time-average cost criterion $f = av$) is included in the set of asymptotically optimal policies Π^*, as obtained in Section 3.3, in case the bandit problem is indexable.

We will need the following assumption. (The same assumption was made in [37].)

Assumption 4.2 *The transition rates of the state of one class-k bandit are such that they form a unichain (there is a state $j \in \{1, \ldots, J_k\}$ such that there is a path from any state $i \in \{1, \ldots, J_k\}$ to state j), for all k, regardless of the policy employed.*

The next proposition shows that Whittle's index policy is included in the set of asymptotically optimal policies. The proof can be found in Appendix B.

Proposition 4.3 *Consider a fixed population of bandits. If Assumption 4.2 holds and if the restless bandit problem is indexable for the average-cost criterion, then there is an $x^{av} \in X^*$ such that Whittle's index policy $(\nu^{av}_{j,k})$ is included in the set $\Pi(x^{av}) \subset \Pi^*$.*

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

Corollary 4.4 *Consider a fixed population of bandits. If the assumptions of Proposition 4.3 are satisfied and if Condition 3.7 holds for Whittle's index policy $(\nu^{av}_{j,k})$, then policy $(\nu^{av}_{j,k})$ is asymptotically optimal. That is, for any policy π,*

$$\lim_{r\to\infty} \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \mathbb{E}\Big(\int_0^T C^a_{j,k} x^{r,\nu^{av},a}_{j,k}(t)\mathrm{d}t\Big)$$
$$\leq \liminf_{r\to\infty} \liminf_{T\to\infty} \frac{1}{T} \sum_{k=1}^{K} \sum_{j=1}^{J_k} \sum_{a=0}^{1} \mathbb{E}\Big(\int_0^T C^a_{j,k} x^{r,\pi,a}_{j,k}(t)\mathrm{d}t\Big).$$

Proof: From Proposition 3.13 and Proposition 4.3 we obtain the desired result. □

Remark 4.5 (Weber and Weiss, 1990) *The above corollary was previously proved by Weber and Weiss in [37] for the case of symmetric bandits, i.e., $K = 1$. We note that the assumptions made in [37] in order to prove the asymptotic optimality result are the same as the ones in Corollary 4.4.*

The proof technique used in Weber and Weiss is different from the one used here. In [37] 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. In this paper we have used another approach in order to prove the result. In particular, our approach allowed to include the case of a dynamic population of bandits, see Section 4.3, whereas the approach of [37] did not (note that for a dynamic population, the optimal cost cannot be lower bounded by any relaxed optimization problem).

The asymptotic optimality result under Whittle's index policy for the case $K = 1$ [37] has been cited extensively. However, in almost all cases, the global attractor property (Condition 3.7) needed for the asymptotic optimality result to hold could only be checked numerically. An exception is the case of symmetric bandits ($K = 1$) with $J = 2$ or $J = 3$, for which the global attractor property was proved in [38].

4.3 Asymptotic optimality for a dynamic population of bandits

In this section we will introduce an index policy based on Whittle's indices and show it to be asymptotically optimal in the setting of a dynamic population of bandits. More precisely, we show the index policy to be included in the set of asymptotically optimal policies Π^*, as obtained in Section 3.3.

Recall that our objective is to find a policy that asymptotically minimizes the average-cost criterion (26). However, we cannot make use of Whittle's index policy with indices $\nu^{av}_{j,k}$ as we did in the previous section: any policy that makes sure that the class-k bandit leaves after a finite amount of time is an optimal solution of the class-k subproblem (31) for $f = av$ (the average cost will be equal to zero), and hence, no useful heuristic for a priority structure can be derived. In order to derive a non-trivial index rule, the authors of [5, 7] therefore consider instead the Whittle's indices corresponding to the subproblem (31) for the discounted-cost criterion, that is, $f = \beta$, $\beta > 0$, and studied their limiting values as $\beta \downarrow 0$. We will do the same: for each k, let $\beta_l \to 0$ be some subsequence such that

$$\nu^{lim}_{j,k} := \lim_{l\to\infty} \nu^{\beta_l}_{j,k}$$

is well-defined for all $j = 1, \ldots, J_k$. We will refer to the index policy $(\nu^{lim}_{j,k})$ as Whittle's index policy for the average-cost criterion in the dynamic population setting.

We will need the following assumptions.

Assumption 4.6 *For all $k = 1, \ldots, K$, the set of optimal solutions of the linear program*

$$\begin{aligned} \min_x \; & \sum_{j=1}^{J_k} (C^0_{j,k} x^0_{j,k} + C^1_{j,k} x^1_{j,k} + \nu x^1_{j,k}) \\ s.t. \; & 0 = \lambda_k p_k(0,j) + \sum_{a=0}^{1} \sum_{i=1, i\neq j}^{J_k} x^a_{i,k} q^a_k(i,j) - \sum_{a=0}^{1} x^a_{j,k} q^a_k(j), \quad \forall\, j, \\ & x^a_{j,k} \geq 0, \; \forall\, j, a. \end{aligned}$$

is bounded when $\nu > 0$.

We note that Assumption 4.6 is always satisfied in case $C^0_{j,k} > 0$ for all j, k. The next assumption concerns ties in the index values.

Assumption 4.7 *In case there are two different states (j,k) and (i,l) such that $\nu^{lim}_{j,k} = \nu^{lim}_{i,l} = \nu$, then $\nu = \infty$. That is, all index values are unique, besides the index value ∞ which can be obtained in several states.*

The next proposition shows that Whittle's index policy $(\nu^{lim}_{j,k})$ is included in the set of asymptotically optimal policies. The proof can be found in Appendix C.

Proposition 4.8 *Consider a dynamic population of bandits. If Assumption 4.6 and Assumption 4.7 hold and if the discounted restless bandit problem is indexable for $\beta \in (0, \overline{\beta}]$, with $\overline{\beta}$ such that $0 < \overline{\beta} < 1$, then there is an $x^{lim} \in X^*$ such that Whittle's index policy $(\nu^{lim}_{j,k})$ is included in the set $\Pi(x^{lim}) \subset \Pi^*$.*

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

Corollary 4.9 *Consider a dynamic population of bandits. If the assumptions of Proposition 4.8 are satisfied and if Condition 3.7 and Condition 3.14 hold for Whittle's index policy $(\nu^{lim}_{j,k})$, then policy $(\nu^{lim}_{j,k})$ is asymptotically optimal, that is, the statements in (20), (21) and (22) hold with π^* replaced by the policy $(\nu^{lim}_{j,k})$.*

Proof: The result follows directly from Proposition 3.15 and Proposition 4.8. □

Remark 4.10 (No arrivals versus arrivals) *The above result shows that the heuristic $(\nu^{lim}_{j,k})$ which is based on a model without arrivals is in fact nearly optimal in the presence of arrivals.*

5 Case study: Impatient restless bandits

In the paper we obtained results on the asymptotic optimality of priority policies in general restless bandit problems. In order for the results to go through, a policy needs to satisfy certain technical conditions. In this section we will verify these conditions for a particular restless bandit problem in the setting of a *dynamic* population of bandits. We note that we did not attempt to include an example for the fixed population of bandits, knowing the difficulty in proving its corresponding condition (the global attractor property) already for the case $K = 1$, see the comment at the end of Section 4.2.

In the context of a dynamic population of bandits, a sufficient condition for a policy $\pi^* \in \Pi^*$ to be asymptotically optimal as given in Proposition 3.15 is that: (i) the ODE $x^*(t)$ has a global attractor (same condition as needed for a fixed population of bandits), *plus* (ii) Condition 3.14 is satisfied. In the remainder of this section we will further discuss the latter condition as it is different from that for the fixed population. More precisely, we will describe a class of restless bandit problems for which Condition 3.14 is always satisfied. This class of problems, which are characterized by the presence of impatient bandits, can be found in Subsection 5.1. In Subsection 5.2 we further specify the model to a multi-class *M/M/S* queue with impatient bandits and derive an index policy that is included in the set Π^*, satisfies the global attractor property and, hence, is asymptotically optimal.

5.1 Condition 3.14 and asymptotic optimality

In this section we discuss a general restless bandit problem in the presence of impatient bandits. More precisely, we assume that any *passive* bandit has a strictly positive probability of abandoning the system, that is, $q_k^0(j,0) > 0$ for all $j = 1, \ldots, J_k$ and $K = 1, \ldots, k$[1]. The assumption of impatient bandits might seem rather strong. However, for many real-life situations it is reasonable to assume that a bandit might abandon the system before "being served", think for example of customers that become impatient and abandon the queue/system, companies that go bankrupt, perishable items, etc.

In the following proposition we will prove that, in the presence of impatient bandints, any policy satisfies Condition 3.14. In particular, this implies that if one proves that the ODE corresponding to a priority policy $\pi^* \in \Pi^*$ has a global attractor, then asymptotic optimality can be concluded from Proposition 3.15.

Proposition 5.1 *Assume $q_k^0(j,0) > 0$ for all $j = 1, \ldots, J_k$, $k = 1, \ldots, K$. For any policy, Condition 3.14 is satisfied.*

In addition, any policy $\pi^* \in \Pi(x^*) \subset \Pi^*$ that satisfies Condition 3.7 is asymptotically optimal, that is, the statements in (20)*,* (21) *and* (22) *hold.*

Proof: Consider an arbitrary policy π. The Markov process $X^{r,\pi}(t)$ has unbounded transition rates, however, it follows that it does not die in finite time (upward jumps are of the order 1). Hence, once we prove the ergodicity criterion, as for example given in [34, Proposition 8.14], we can conclude that there is a unique invariant distribution measure.

The mean drift of the process $\sum_k X_k^{r,\pi}(t)$, given it starts at time 0 in state x, is smaller than or equal to

$$\lambda r - \hat{q} \cdot \sum_{k=1}^{K} (x_k - r\alpha), \tag{33}$$

with $\lambda = \sum_{k=1}^{K} \lambda_k$ and $\hat{q} := \min_{j,k} q_k^0(j,0) > 0$. Hence, there exists a $\Delta := (\lambda r + \alpha \hat{q} r + \delta)/\hat{q}$, with $\delta > 0$, such that if $\sum_k x_k > \Delta$, then the drift is smaller than or equal to $\lambda r - (\hat{q} \sum_k x_k - \alpha \hat{q} r) < -\delta < 0$. The ergodicity criterion is hence satisfied and from [34, Proposition 8.14] we obtain that there is a unique invariant probability distribution for the process $X^r(t)$, for any r. Recall that we denote the invariant probability distribution of $X^{r,\pi}(t)/r$ by $p^{r,\pi}$.

The process $\{\sum_{k=1}^{K} X_k^r(t)\}$ can be stochastically upper bounded by the queue length process of an $M/M/\infty$ queue with arrival rate $\lambda r = \sum_{k=1}^{K} \lambda_k r$ and departure rate $\hat{q}$, having αr permanent customers in the queue. We denote this queue length process by $Y^r(t)$. The stationary distribution of the process $\{Y^r(t)\}$, is distributed as $Y^r \stackrel{d}{=} \alpha r + N^r$, where N^r is Poisson with parameter $\lambda r/\hat{q}$ [34]. It can be checked that N^r/r converges to the Dirac measure in the point $\lambda/\hat{q}$, as $r \to \infty$. By Prohorov's theorem it then

[1] We believe this condition can be weakened to having one state $j_k \in \{1, \ldots, J_k\}$, for each k, such that $q_k^0(j_k, 0) > 0$ and state j_k is positive recurrent under the policy that always keeps the class-k bandit passive.

follows that the family $\{Y^r/r\}$ is tight [34]. Furthermore, since $\mathbb{E}(Y^r/r) = \alpha+\lambda/\hat{q}$ and $\mathbb{E}(\lim_{r\to\infty} Y^r/r) = \alpha + \lambda/\hat{q}$, a.s., we obtain from [9, Theorem 3.6] that the family $\{Y^r/r\}$ is uniform integrable.

Since $Y^r(t)/r$ represents an upper bound on the queue length process $\sum_{k=1}^K X_k^{r,\pi}(t)/r$, we obtain that the family $\{p^{r,\pi}\}$ is tight and uniform integrable as well. □

5.2 Asymptotically optimal index policy for an *M/M/S* queue

In this section we study a multi-class multi-server system with impatient customers, which is a special case of the restless bandit problem as investigated in the previous section. For this model we are able to prove Condition 3.7 (the ODE having a global attractor), which is needed in order to conclude for asymptotic optimality. In fact, we will derive a closed-form expression for an asymptotically optimal policy.

We consider a multi-server 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 that minimizes the total long-run average cost

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

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 $[0,t]$, under policy π.

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: $J_k = 1, q_k^0(1,0) = \theta_k > 0, q_k^1(1,0) = \mu_k + \tilde{\theta}_k, C_{1,k}^0 = c_k + d_k\theta_k, C_{1,k}^1 = \tilde{c}_k + \tilde{d}_k\tilde{\theta}_k$, $k = 1, \ldots, K$, and $\alpha = S$, where we used that $\mathbb{E}(R_k^\pi(T)) = \theta_k \mathbb{E}(\int_0^T X_k^{\pi,0}(t)\mathrm{d}t)$ and $\mathbb{E}(\tilde{R}_k^\pi(T)) = \tilde{\theta}_k \mathbb{E}(\int_0^T X_k^{\pi,1}(t)\mathrm{d}t)$.

We can now define an index policy that we prove to be included in the set Π^*, see the proposition below. For each class k we define the following index:

$$\iota_k := q_k^1(1,0)\left(\frac{C_{1,k}^0}{q_k^0(1,0)} - \frac{C_{1,k}^1}{q_k^1(1,0)} \right).$$

The index policy (ι_k) 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. In case a customer belongs to a class that has a negative index value, this customer will never be served.

Proposition 5.2 *Policy (ι_k) is contained in the set Π^*.*

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

$$\begin{aligned} &\min_x \sum_k (c_k x_k^0 + \tilde{c}_k x_k^1 + d_k\theta_k x_k^0 + \tilde{d}_k\tilde{\theta}_k x_k^1), \\ &\text{s.t. } 0 = \lambda_k - \mu_k x_k^1 - \theta_k x_k^0 - \tilde{\theta}_k x_k^1, \\ &\quad \sum_{k=1}^K x_k^1 \leq S, \\ &\quad x_k^0, x_k^1 \geq 0. \end{aligned} \tag{34}$$

Equation (34) 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

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

An optimal solution is to assign maximum values to x_k^1 for those classes having the highest values for $(c_k + d_k\theta_k)\frac{\mu_k+\tilde{\theta}_k}{\theta_k} - \tilde{c}_k - \tilde{d}_k\tilde{\theta}_k = \iota_k$, 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 $\iota_1 \geq \iota_2 \geq \ldots \geq \iota_K$. Hence, there is an l such that $x_k^{*,1} = \frac{\lambda_k}{\mu_k+\tilde{\theta}_k}$, and hence $x_k^{*,0} = 0$, for all $k < l$, $0 \leq x_l^{*,1} \leq \frac{\lambda_l}{\mu_l+\tilde{\theta}_l}$ and hence $x_l^{*,0} \geq 0$, and $x_k^{*,1} = 0$, for all $k > l$. Hence, the index policy (ι_k) is included in the set $\Pi(x^*) \subset \Pi^*$, see Definition 3.4. □

We have the following optimality result for the index policy (ι_k).

Proposition 5.3 *Consider a system with Sr servers working in parallel and arrival rates $\lambda_k r$, $k = 1, \ldots, K$. The index policy (ι_k) is asymptotically optimal as $r \to \infty$, i.e., for any policy π,*

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

Proof: In Proposition 5.2 we showed that (ι_k) is included in $\Pi(x^*)$, with x^* as given in the proof of Proposition 5.2. In Appendix D we prove that the ODE $x^*(t)$ corresponding to policy (ι_k), as defined in (14), has the point x^* as a unique global attractor, i.e., Condition 3.7 is satisfied. Together with Proposition 5.1 we then obtain that the index policy (ι_k) is asymptotically optimal. □

We note that the asymptotically optimal index rule (ι_k) for the abandonment problem is robust, i.e., it does not depend on the arrival rate of the customers or the number of servers present in the system. In addition, depending on the parameters, it can happen that it is asymptotically optimal *not* to serve certain classes of customers having a strictly negative index, as was also observed in [7].

Remark 5.4 (Existing results in literature) *In [3, 4] a special case of the multi-class multi-server system has been studied, which is obtained by setting $\tilde{c}_k = 0, \tilde{\theta}_k = 0$ and $c_k + d_k\theta_k > 0$. The authors showed that the index rule with index $(c_k + d_k\theta_k)\mu_k/\theta_k$ asymptotically minimizes the number of customers present in the queue. We note that 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 underloaded system, as was also observed in [3, 4]: In that case $x_k^{*,0} = 0, \ \forall\ k$, and hence $V^*(S) = 0$. However, any non-idling policy will keep the scaled number of waiting customers equal to zero, i.e., has scaled cost equal to zero, and is hence asymptotically optimal.*

The authors of [7] studied a special case of the model (in discrete time) obtained by setting $\tilde{\theta}_k = 0, \tilde{c}_k = c_k > 0$. Hence, their objective is to minimize the number of customers in the system (and not just in the queue as was the case in [3, 4]). The authors proved indexability (for any discount factor) and derived that Whittle's index ν_k^{lim} is given by $\frac{c_k(\mu_k-\theta_k)+d_k\mu_k\theta_k}{\theta_k}$. They refer to the corresponding index policy as the AJN-rule. The latter coincides with the index ι_k as defined in Proposition 5.2. Hence, it follows directly from Proposition 5.3 that the AJN-rule is asymptotically optimal in the continuous-time setting. Finally, note that for the model of [7], the fluid scaled cost is always strictly positive ($V^*(S) = 0$ would imply that $x_k = 0$. However this contradicts with Equation (34) which would read $0 = \lambda_k$). Hence the asymptotic optimality result is useful for both an underloaded as well as for an overloaded regime.

6 Conclusion and further research

We have characterized a class of priority policies that are asymptotically optimal for the general multi-class restless-bandit problem. We studied both the setting of a fixed population of bandits as well as

a dynamic population of bandits. In both cases we showed that Whittle's index policy is included in the class of asymptotically optimal policies. This extends the result of Weber and Weiss [37] in several directions: (i) we allow various classes of restless bandits, (ii) we allow arrivals of new restless bandits to the system, and (iii) we do not need the restless bandits to be indexable in order to define asymptotically optimal policies. In order to obtain the asymptotic optimality results we combined fluid-scaling techniques with linear programming results. This is a different approach than that taken in [37] and in particular allows us to include arrivals of bandits to the system.

We considered a slight variation of the standard restless bandit formulation: Instead of having at each moment in time exactly α bandits active, we allowed strictly less than α bandits to be active at a time. By introducing so-called dummy bandits we showed that restless bandits having a negative Whittle's index are never made active under an optimal policy of the relaxed optimization problem or under Whittle's index policy.

As future work it would be interesting to investigate whether Condition 3.14 would hold in greater generality for restless bandit problems. So far, we showed it to hold in the presence of impatient bandits. A further interesting thread would be to estimate the suboptimality gap of the proposed priority policies outside the fluid-scaling regime.

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*, 2012. To appear.

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

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

[10] 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.

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

[12] 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.

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

[14] 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.

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

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

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

[18] 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.

[19] 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.

[20] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, New York, 2001.

[21] 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.

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

[23] P. Jacko. Value of information in optimal flow-level scheduling of users with Markovian time-varying channels. *Performance Evaluation*, 68:1022–1036, 2011.

[24] 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.

[25] 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.

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

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

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

[29] 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.

[30] 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.

[31] 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.

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

[33] 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 Conf. Comput. Commun.*, pages 1570–1578, 2008.

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

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

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

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

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

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

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

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

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

Appendix A: Proof of Lemma 3.2 and Lemma 3.3:

We first proof Lemma 3.2. By Fatou's lemma we have

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\liminf_{T\to\infty}\frac{1}{T}\mathbb{E}(\int_0^T C_{j,k}^a X_{j,k}^{\pi,a}(t)\mathrm{d}t) \geq \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\mathbb{E}(\liminf_{T\to\infty}\frac{1}{T}\int_0^T C_{j,k}^a X_{j,k}^{\pi,a}(t)\mathrm{d}t).$$

Hence, it is sufficient to prove that

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\liminf_{T\to\infty}\frac{1}{T}\int_0^T C_{j,k}^a X_{j,k}^{\pi,a}(t)\mathrm{d}t \geq V^*(\alpha), \text{ almost surely.} \tag{35}$$

We consider a fixed realization ω of the process. If $\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\liminf_{T\to\infty}\frac{1}{T}\int_0^T C_{j,k}^a X_{j,k}^{\pi,a}(t)\mathrm{d}t = \infty$, then (35) is trivially true, since $V^*(\alpha) < \infty$ (see Lemma 3.1). Hence, assume this is not the case, and consider the subsequence t_n corresponding to the liminf sequence in (35). Since $X_{j,k}^{\pi,a}(t) \leq X_k(0)$, for all t, $\frac{1}{T}\int_0^T X_{j,k}^{\pi,a}(t)\mathrm{d}t$ are bounded sequences. Hence, there is a subsequence t_{n_l} of t_n such that $\frac{1}{t_{n_l}}\int_0^{t_{n_l}} C_{j,k}^a X_{j,k}^{\pi,a}(t)\mathrm{d}t$ converges to a constant $\overline{X}_{j,k}^{\pi,a}$, for all j,k,a. In addition, it holds that $\lim_{t\to\infty} X_{j,k}^{\pi,a}(t)/t \leq \lim_{t\to\infty} X_k(0)/t = 0$, for all j,k,a. When studying (7) 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 obtain

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

By (2) we also have that $\sum_{k=1}^{K}\sum_{j=1}^{J_k}\overline{X}_{j,k}^{\pi,1} \leq \alpha$, hence $\overline{X}^\pi$ is a feasible solution of (LP). We can now conclude that

$$\begin{aligned}\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\liminf_{T\to\infty}\frac{1}{T}\int_0^T C_{j,k}^a X_{j,k}^{\pi,a}(t)\mathrm{d}t &= \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\lim_{l\to\infty}\frac{1}{t_{n_l}}\int_0^{t_{n_l}} C_{j,k}^a X_{j,k}^{\pi,a}(t)\mathrm{d}t &(36)\\ &= \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_{j,k}^a \overline{X}_{j,k}^{\pi,a} \geq V^*(\alpha), &(37)\end{aligned}$$

which proves (8).

We now focus on Lemma 3.3. Let us first assume that policy π is stable. We consider a fixed realization ω of the process. We have by the ergodicity theorem [11] 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 $\lim_{t\to\infty} X_{j,k}^{\pi,a}(t)/t = 0$, a.s., for all j,k,a, since any stable policy is rate stable. The proof of (9) now follows as above for the fixed population.

We now assume that π is rate-stable and that $C_{j,k}^a > 0$, for all j,k,a. Again we consider a fixed realization ω of the process. If $\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\liminf_{T\to\infty}\frac{1}{T}\int_0^T C_{j,k}^a X_{j,k}^{\pi,a}(t)\mathrm{d}t = \infty$, then the result is

trivially true, since $V^*(\alpha) < \infty$ (see Lemma 3.1). Hence, assume this is not the case, and consider the subsequence t_n corresponding to the liminf sequence. So

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\lim_{n\to\infty}\frac{1}{t_n}\int_0^{t_n} C_{j,k}^a X_{j,k}^{\pi,a}(t)\mathrm{d}t < \infty. \tag{38}$$

Since $C_{j,k}^a > 0$, 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 again as in the fixed population case.

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

We now prove relation (10). In case policy π is mean-rate stable, then it follows directly from relation (9). Now assume policy π is not mean-rate stable. Hence, there exist a k such that there is a subsequence t_n, $t_n \to \infty$, with $\lim_{n\to\infty}\mathbb{E}(X_k^\pi(t_n))/t_n = a$, $a \in (0,\infty]$. So for each $\epsilon > 0$, $\mathbb{E}(X_k^\pi(t_n)) \geq (a-\epsilon)t_n$, with n large enough. Define $\tilde{a} = a - \epsilon > 0$ (with $\epsilon > 0$ small enough). For $t \leq t_n$ we have that

$$\mathbb{E}(X_k^\pi(t_n)) \leq \mathbb{E}(X_k^\pi(t)) + \lambda_k(t_n - t).$$

Hence, for $t \in [t_n(1-\delta), t_n]$,

$$\mathbb{E}(X_k^\pi(t)) \geq \mathbb{E}(X_k^\pi(t_n)) - \lambda_k(t_n - t) \geq \tilde{a}t_n - \lambda_k\delta t_n = \tilde{\delta}t_n,$$

with $\tilde{\delta} := \tilde{a} - \lambda_k\delta$. We take $\delta > 0$ small enough such that $\tilde{\delta} > 0$. We can now conclude that

$$\frac{1}{t_n}\int_0^{t_n}\mathbb{E}(X_k^\pi(t))\mathrm{d}t \geq \frac{1}{t_n}\int_{t_n(1-\delta)}^{t_n}\mathbb{E}(X_k^\pi(t))\mathrm{d}t \geq \frac{1}{t_n}\int_{t_n(1-\delta)}^{t_n}\tilde{\delta}t_n\mathrm{d}t = \tilde{\delta}\delta t_n.$$

The latter goes to infinity as $n \to \infty$. Since $C_{j,k}^a > 0$ and $V^*(\alpha) < \infty$, this proves Relation (10). □

Appendix B: Proof of Proposition 4.3

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

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\lim_{T\to\infty}\frac{1}{T}\mathbb{E}(\int_0^T C_{j,k}^a X_{j,k}^{\pi,a}(t)\mathrm{d}t), \tag{39}$$

under the relaxed constraint

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k}\lim_{T\to\infty}\frac{1}{T}\mathbb{E}(\int_0^T X_{j,k}^{\pi,1}(t)\mathrm{d}t) \leq \alpha. \tag{40}$$

For a given policy π, 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 4.2 implies that these frequencies exist and satisfy the balance equations, that is, they satisfy

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

We restrict ourselves to the class of policies that are symmetric for bandits in the same class[2]. Having $x_k(0)$ bandits in class k, Equations (39) and (40) can then equivalently be written as

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

[2]We can do this without loss of generality, since this is the case for the optimal solution of the relaxed problem as given by Whittle.

respectively. We can now formulate the relaxed optimization problem as the following linear program (D):

$$
\begin{aligned}
(D) \qquad & \min_{x} \sum_{k=1}^{K} x_k(0) \sum_{j=1}^{J_k} \left(C^0_{j,k} x^0_{j,k} + C^1_{j,k} x^1_{j,k} \right) \\
& \text{s.t. } \; 0 = \sum_{a=0}^{1} q^a_k(j) x^a_{j,k} - \sum_{a=0}^{1} \sum_{i=1, i\neq j}^{J_k} q^a_k(i,j) x^a_{i,k}, \; \forall \; j,k, \\
& \qquad \sum_{k=1}^{K} x_k(0) \sum_{j=1}^{J_k} x^1_{j,k} \leq \alpha, \\
& \qquad \sum_{j=1}^{J_k} \sum_{a=0}^{1} x^a_{j,k} = 1, \; \forall \; k, \quad x^a_{j,k} \geq 0, \; \forall \; k,j,a.
\end{aligned} \tag{41}
$$

We have that for any feasible solution $(x^a_{j,k})$ of (D) there is a stationary policy π such that the state-action frequencies $x^{\pi,a}_{j,k}$ coincide with the value of the feasible solution $x^a_{j,k}$ [32, Theorem 8.8.2 b)]. Hence, for any optimal (symmetric) policy π^* of the relaxed optimization problem, the state-action frequencies $x^{\pi^*,a}_{j,k}$ provide an optimal solution of (D). We further note that $(x^{\pi^*}_{j,k} x_k(0))$ is an optimal solution of (LP).

We assume the restless bandit problem is indexable. Hence, an optimal policy of the relaxed optimization problem is described in Section 4.1, and will be denoted here by π^*. We recall that policy π^* is described by a value $\nu^* \geq 0$ and is such that a class-k bandit in state j is served if $\nu^{av}_{j,k} > \nu^*$ and is not served if $\nu^{av}_{j,k} < \nu^*$. Hence, the state-action frequencies under π^* satisfy

$$
\begin{aligned}
x^{\pi^*,0}_{j,k} &= 0 \;\; \text{when } \nu^{av}_{j,k} > \nu^*, \\
x^{\pi^*,1}_{j,k} &= 0 \;\; \text{when } \nu^{av}_{j,k} < \nu^*.
\end{aligned} \tag{42}
$$

By definition of policy π^*, there is at most one state $(\hat{j}, \hat{k})$ with $\nu^{av}_{\hat{j},\hat{k}} = \nu^*$ such that a class-k bandit in state j is made active with a certain probability, hence $x^{\pi^*,0}_{\hat{j},\hat{k}} \geq 0$ and $x^{\pi^*,1}_{\hat{j},\hat{k}} \geq 0$. Any other class-$\tilde{k}$ bandit in state $\tilde{j}$ with $\nu^{av}_{\tilde{j},\tilde{k}} = \nu^*$ is either always active, so $x^{\pi^*,0}_{\tilde{j},\tilde{k}} = 0$, or always passive, so $x^{\pi^*,1}_{\tilde{j},\tilde{k}} = 0$, depending on whether or not it has strict priority over a class-$\hat{k}$ bandit in state $\hat{j}$. Hence, $x^{\pi^*} \in X^*$ (with X^* as defined in Defition 3.4).

Since Whittle's index policy gives priority to bandits having highest index value, we directly obtain that Whittle's index policy $(\nu^{av}_{j,k})$ satisfies points 1 and 2 of Definition 3.4 when setting $x^* = x^{\pi^*}$. We now treat point 3 of Definition 3.4: Assume $\sum_{k=1}^{K} \sum_{j=1}^{J_k} x^{\pi^*,1}_{j,k} < \alpha$. Hence, under the optimal policy, on average, strictly less than α bandits are made active. This implies that the remaining fraction of the time the policy makes dummy bandits in state B active. Hence, $\nu^{av}_B \geq \nu^*$. Since $\nu^* \geq 0$ and $\nu^{av}_B = 0$ we necessarily have $\nu^* = 0$. A policy satisfies point 3 of Definition 3.4 if it never makes a class-k bandit in state j active that satisfies

$$
x^{\pi^*,1}_{j,k} = 0 \text{ and } x^{\pi^*,0}_{j,k} > 0. \tag{43}
$$

From (42) (with $\nu^* = 0$), we obtain that (43) implies $\nu^{av}_{j,k} \leq 0$. By definition of Whittle's index policy, a bandit in a state such that $\nu^{av}_{j,k} \leq 0$ will never be made active, hence point 3 is satisfied. Hence, we conclude that Whittle's index policy $(\nu^{av}_{j,k})$ is included in the set of priority policies $\Pi(x^{\pi^*}) \subset \Pi^*$.

Appendix C: Proof of Proposition 4.8

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

$$
\min_{A_k(\cdot)} \mathbb{E} \left(\int_0^{\infty} \mathrm{e}^{-\beta t} (C^{A_k(t)}_{J_k(t),k} + \nu \mathbf{1}_{(A_k(t)=1)}) \mathrm{d}t \right), \tag{44}
$$

see (31), where $A_k(t) \in \{0,1\}$ 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 ([21, Remark 3.1], [32,

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 $\tilde{C}^a_{j,k} = \frac{C^a_{i,k}+\nu \mathbf{1}_{(a=1)}}{\beta+\overline{q}}$, and transition probabilities $\tilde{p}^a_k(i,j) = \frac{q^a_k(i,j)}{\overline{q}} + \mathbf{1}_{(i=j)}$ (recall that $q^a_k(i,i) = -q^a_k(i) = -\sum_{j=0,i\neq j}^{J_k} q^a_k(i,j)$), where $\overline{q} := \max_{i,k,a} q^a_k(i) < \infty$. In LP formulation the discrete-time MDP for the class-k bandit is then as follows (see [32, Section 6.9]):

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

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

$$\begin{aligned}
(D_k(\beta,\epsilon)) \qquad &\min_x \sum_{j=1}^{J_k} \frac{C^0_{j,k} x^0_{j,k} + C^1_{j,k} x^1_{j,k} + \nu x^1_{j,k}}{\beta+\overline{q}} \\
&\text{s.t. } \; 0 = \lambda_k(p_k(0,j)+\epsilon) + \sum_{a=0}^{1} \sum_{i=1,i\neq j}^{J_k} \frac{q^a_k(i,j)}{\beta+\overline{q}} x^a_{i,k} - \frac{\beta}{\beta+\overline{q}} \sum_{a=0}^{1} x^a_{j,k} - \sum_{a=0}^{1} \frac{q^a_k(j)}{\beta+\overline{q}} x^a_{j,k}, \;\; \forall j, \quad (45) \\
&\qquad\; x^a_{j,k} \geq 0, \;\; \forall j, a.
\end{aligned}$$

As stated in Section 4.1 ("Optimal solution per bandit"), indexability implies that an optimal policy for the subproblem (44) is described by a priority ordering according to the indices $\nu^\beta_{j,k}$: an optimal action in state j is $a = 1$ if $\nu^\beta_{j,k} > \nu$ and $a = 0$ if $\nu^\beta_{j,k} < \nu$. By [32, Theorem 6.9.4 e)], this implies that there exists an optimal solution to $(D_k(\beta,\epsilon))$, denoted by $x^*_k(\beta,\epsilon)$, such that

$$\begin{aligned}
x^{*,0}_{j,k}(\beta,\epsilon) &= 0 \;\; \text{when } \nu^\beta_{j,k} > \nu, \\
x^{*,1}_{j,k}(\beta,\epsilon) &= 0 \;\; \text{when } \nu^\beta_{j,k} < \nu.
\end{aligned}$$

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

$$x^{*,0}_{j,k}(\beta_l,\epsilon) = 0 \;\; \text{when } \nu^{lim}_{j,k} > \nu, \quad (46)$$

$$x^{*,1}_{j,k}(\beta_l,\epsilon) = 0 \;\; \text{when } \nu^{lim}_{j,k} < \nu. \quad (47)$$

By change of variable $\tilde{x}^a_{j,k} = x^a_{j,k}/(\beta+\overline{q})$ we obtain that $\tilde{x}^*_k(\beta,\epsilon)$ satisfies (46) and (47) and is an optimal solution of $(\tilde{D}_k(\beta,\epsilon))$ defined as:

$$\begin{aligned}
(\tilde{D}_k(\beta,\epsilon)) \qquad &\min_{\tilde{x}} \sum_{j=1}^{J_k} (C^0_{j,k}\tilde{x}^0_{j,k} + C^1_{j,k}\tilde{x}^1_{j,k} + \nu \tilde{x}^1_{j,k}) \\
&\text{s.t. } \; 0 = \lambda_k(p_k(0,j)+\epsilon) + \sum_{a=0}^{1} \sum_{i=1,i\neq j}^{J_k} q^a_k(i,j)\tilde{x}^a_{i,k} - \beta \sum_{a=0}^{1} \tilde{x}^a_{j,k} - \sum_{a=0}^{1} q^a_k(j)\tilde{x}^a_{j,k}, \;\; \forall\, j, \quad (48) \\
&\qquad\; \tilde{x}^a_{j,k} \geq 0, \;\; \forall j, a.
\end{aligned}$$

By Assumption 4.6 we have that the set of optimal solutions of $(\tilde{D}_k(0,0))$ is bounded and non-empty when $\nu > 0$. Hence, from [10, Corollary 1] we obtain that the correspondence that gives for each (β,ϵ) 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 (48) 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 (β_{l_n},ϵ_n) (with β_{l_n} a subsequence of β_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 ν, the components of $\tilde{x}^*(\beta_l,\epsilon)$ that are zero are independent of the exact values for $\epsilon > 0$, and $l > L(\nu)$, see (46) and (47). Hence, the limit $\tilde{x}^*$, which is an optimal solution of $(\tilde{D}_k(0,0))$, has the same components equal to zero, i.e., (46) and (47) are satisfied for $\tilde{x}^*$.

Below we will show that there exists a value ν^* 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 $(\nu_{j,k}^{lim})$ 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 ν. We have the following properties:

- **Property 1:**

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k} \tilde{x}_{j,k}^{*,1}(\infty) \leq \alpha. \tag{49}$$

  This can be seen as follows. As $\nu \to \infty$, the objective of $(\tilde{D}_k(0,0))$ is to minimize $\sum_{j=1}^{J_k} \tilde{x}_{j,k}^1$. For any feasible solution x of (LP), x_k is in the feasible set of $\tilde{D}_k(0,0)$. Hence, $\sum_{j=1}^{J_k} \tilde{x}_{j,k}^{*,1}(\infty) \leq \sum_{j=1}^{J_k} x_{j,k}^1$ with x a feasible solution of (LP). In addition, we have that $\sum_{k=1}^{K}\sum_{j=1}^{J_k} x_{j,k}^1 \leq \alpha$ with x a feasible solution of (LP). This proves (49).

- **Property 2:**

$$\sum_{j=1}^{J_k} \tilde{x}_{j,k}^{*,1}(\nu) \geq \sum_{j=1}^{J_k} \tilde{x}_{j,k}^{*,1}(\tilde{\nu}), \quad \text{for} \quad \nu < \tilde{\nu}. \tag{50}$$

  This can be seen as follows: By definition we have $\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_{j,k}^a \tilde{x}_{j,k}^{*,a}(\nu) + \nu \sum_{j=1}^{J_k} \tilde{x}_{j,k}^{*,1}(\nu) \leq \sum_{j=1}^{J_k}\sum_{a=0}^{1} C_{j,k}^a \tilde{x}_{j,k}^{*,a}(\tilde{\nu}) + \nu \sum_{j=1}^{J_k} \tilde{x}_{j,k}^{*,1}(\tilde{\nu})$ and $\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_{j,k}^a \tilde{x}_{j,k}^{*,a}(\tilde{\nu}) + \tilde{\nu} \sum_{j=1}^{J_k} \tilde{x}_{j,k}^{*,1}(\tilde{\nu}) \leq \sum_{j=1}^{J_k}\sum_{a=0}^{1} C_{j,k}^a \tilde{x}_{j,k}^{*,a}(\nu) + \tilde{\nu} \sum_{j=1}^{J_k} \tilde{x}_{j,k}^{*,1}(\nu)$. Substracting the latter inequality from the first, we obtain Equation (50).

- **Property 3:**

$$\sum_{j=1}^{J_k} \tilde{x}_{j,k}^{*,1}(\nu) < \infty \quad \text{for} \quad \nu > 0. \tag{51}$$

  This follows since by Assumption 4.6 the set of optimal solutions of $(\tilde{D}_k(0,0))$ is bounded for $\nu > 0$.

We define $\overline{\alpha} := \sum_{k=1}^{K}\sum_{j=1}^{J_k} \tilde{x}_{j,k}^{*,1}(0)$. Equations (49)–(51) imply that there exists a $\nu^* \geq 0$ such that

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k} \tilde{x}_{j,k}^{*,1}((\nu^*)^-) \geq \min(\alpha, \overline{\alpha}) \text{ and } \sum_{k=1}^{K}\sum_{j=1}^{J_k} \tilde{x}_{j,k}^{*,1}((\nu^*)^+) \leq \min(\alpha, \overline{\alpha}). \tag{52}$$

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 (52) we obtain that there exists a convex combination of the two solutions $\tilde{x}^*((\nu^*)^-)$ and $\tilde{x}^*((\nu^*)^+)$, denoted by $\tilde{y}^*$, such that $\sum_{k=1}^{K}\sum_{j=1}^{J_k} \tilde{y}_{j,k}^{*,1} = \min(\alpha, \overline{\alpha})$. Note that $\tilde{y}_k^*$ is still a solution of $(\tilde{D}_k(0,0))$. 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 $\sum_{k=1}^{K}\sum_{j=1}^{J_k} \tilde{y}_{j,k}^{*,1} = \overline{\alpha} \leq \alpha$, 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 $\Pi(\tilde{y}^*) \subset \Pi^*$. Define $J = \sum_{k=1}^{K} J_k$. By Assumption 4.7 we can number the states such that $\nu_{j_1,k_1}^{lim} < \nu_{j_2,k_2}^{lim} < \ldots < \nu_{j_n,k_n}^{lim} = \ldots = \nu_{j_J,k_J}^{lim} = \infty$. From $\nu^* < \infty$ and Properties (46)–(47) (which hold for $\tilde{x}^*(\nu)$) we have that there are n^* and $\tilde{n}^*$, with $0 \leq \tilde{n}^* - n^* \leq 1$, such that

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

and

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

The vector $\tilde{y}^*$ is a convex combination of $\tilde{x}^*((\nu^*)^-)$ and $\tilde{x}^*((\nu^*)^+)$, hence it follows that (j_{n^*+1}, k_{n^*+1}) is the only state in which it can happen that $\tilde{y}^{*,a}_{j_{n^*+1},k_{n^*+1}} > 0$ for both $a = 0$ and $a = 1$. For $m = n^*+2, \ldots, J$ one has $\tilde{y}^{*,0}_{j_m,k_m} = 0$ and for $m = 1, \ldots, n^*$ one has $\tilde{y}^{*,1}_{j_m,k_m} = 0$. Hence $\tilde{y}^* \in X^*$ (with X^* as defined in Definition 3.4) and it follows that Whittle's index policy $(\nu^{lim}_{j,k})$ satisfies the conditions of the set $\Pi(\tilde{y}^*)$ as posed in items 1 and 2 of Definition 3.4.

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 (46) that $\nu^{lim}_{j,k} < (\nu^*)^+ = 0^+$. Hence, by definition of Whittle's index policy $(\nu^{lim}_{j,k})$, a bandit in this state will never be made active, which implies that item 3 in Definition 3.4 is satisfied. It hence follows that Whittle's index policy $(\nu^{lim}_{j,k})$ is included in the set of priority policies $\Pi(\tilde{y}^*) \subset \Pi^*$. □

Appendix D: Condition 3.7 for an *M/M/S* queue with impatient customers

Assume the classes are reordered such that $\iota_1 \geq \iota_2 \geq \ldots \geq \iota_K$. We further denote by $I^* := \{l : \iota_l \leq 0\}$, the set of classes that will never be served. This set is of the form $I^* = \{\hat{l}, \ldots, K\}$. Under policy (ι_k), the ODE as defined in (14) is given by

$$\frac{\mathrm{d}x^*_k(t)}{\mathrm{d}t} = \lambda_k - x^{*,0}_k(t)\theta_k - x^{*,1}_k(t)(\mu_k + \tilde{\theta}_k), \quad \forall\ k, \tag{53}$$

$$\text{with } x^{*,1}_k(t) = \min\left((S - \sum_{l=1}^{k-1} x^*_l(t))^+,\ x^*_k(t)\right), \quad \text{if } k < \hat{l},\ \forall\ k, \tag{54}$$

$$\begin{aligned} x^{*,1}_k(t) &= 0, \quad \text{if } k \geq \hat{l},\ \forall\ k, \\ x^{*,0}_k(t) &= x^*_k(t) - x^{*,1}_k(t), \quad \forall\ k. \end{aligned} \tag{55}$$

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

$$x^{*,0}_k = 0, \quad x^{*,1}_k = \frac{\lambda_k}{\mu_k + \tilde{\theta}_k}, \text{ for } k = 1, \ldots, \hat{k}, \tag{56}$$

$$x^{*,0}_{\hat{k}+1} = \frac{\lambda_k - (\mu_k + \tilde{\theta}_k)(S - \sum_{l=1}^{\hat{k}} \frac{\lambda_l}{\mu_l + \tilde{\theta}_l})}{\theta_k}, \quad x^{*,1}_{\hat{k}+1} = S - \sum_{l=1}^{\hat{k}} \frac{\lambda_l}{\mu_l + \tilde{\theta}_l}, \text{ if } \hat{k}+1 < \hat{l}, \tag{57}$$

$$x^{*,0}_k = \frac{\lambda_k}{\theta_k}, \quad x^{*,1}_k = 0, \text{ for } k \geq \min(\hat{k}+2, \hat{l}), \tag{58}$$

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 (53) that

$$\frac{\lambda_k}{\mu_k + \tilde{\theta}_k} = x^{*,1}_k + x^{*,0}_k \frac{\theta_k}{\mu_k + \tilde{\theta}_k}. \tag{59}$$

We first prove (56). Let $k = 1$ and assume $1 \leq \hat{k}$. Hence, we have $\frac{\lambda_1}{\mu_1 + \tilde{\theta}_1} < S$. By (59) we obtain $x^{*,1}_1 < S$. Together with (54), that is, $x^{*,1}_1 = \min(S, x^*_1)$, we obtain $x^{*,1}_1 = x^*_1$ and hence $x^{*,0}_1 = 0$. From (59) we obtain that $x^{*,1}_1 = \frac{\lambda_1}{\mu_1 + \tilde{\theta}_1}$. The proof of (56) continues by induction. Assume (56) holds for $k \leq l - 1$, and let $l \leq \hat{k}$. For $k \leq l - 1$ we have that $x^{*,1}_k = \frac{\lambda_k}{\mu_k + \tilde{\theta}_k}$. Since $\sum_{k=1}^{l} \frac{\lambda_k}{\mu_k + \tilde{\theta}_k} \leq S$, by (54) we obtain that $x^{*,1}_l = x^*_l$ and hence $x^{*,0}_l = 0$. From (59) we then obtain that (56) holds for $k = l$ as well.

We now prove (57). Let $\hat{k}+1 < \hat{l}$. From (56) and (57) we obtain that $S - \sum_{l=1}^{\hat{k}} x^*_l < x^*_{\hat{k}+1}$. So by (54) we obtain $x^{*,1}_{\hat{k}+1} = S - \sum_{l=1}^{\hat{k}} \frac{\lambda_l}{\mu_l + \tilde{\theta}_l}$ as stated in (57).

We now prove (58). From (56) and (57) we obtain that $S \leq \sum_{l=1}^{\hat{k}+1} x^*_l$, hence $x^{*,1}_k = 0$ for k such that $\hat{k}+1 < k < \hat{l}$. Equation (58) for $k \geq \hat{l}$ follows directly from (55).

In addition, x^* is a global attractor, as was shown in [3, Appendix] (this can be seen by replacing the μ_i in [3] by $\mu_i + \tilde{\theta}_i$, making the ODE in [3] coincide with our ODE (53)).