On the Whittle index of Markov Modulated Restless Bandits

Santiago Guillermo Duran, Urtzi Ayesta, Ina Maria Maaike Verloop

On the Whittle index of Markov Modulated Restless Bandits*

S. Duran*[,a,c] and U. Ayesta[a,b,c,d] and I.M. Verloop[a,c]

[a] CNRS, IRIT, 2 rue C. Camichel, 31071 Toulouse, France.
[b] IKERBASQUE - Basque Foundation for Science, 48011 Bilbao, Spain
[c] Université de Toulouse, INP, 31071 Toulouse, France
[d] UPV/EHU, Univ. of the Basque Country, 20018 Donostia, Spain

February 15, 2022

Abstract

In this paper we study a Multi-Armed Restless Bandit Problem (MARBP) subject to time fluctuations. This model has numerous applications in practice, like in cloud computing systems or in wireless communications networks. Each bandit is formed by two processes: a controllable process and an environment. The transition rates of the controllable process are determined by the state of the environment, which is an exogenous Markov process. The decision maker has full information on the state of every bandit, and the objective is to determine the optimal policy that minimises the long-run average cost.

Given the complexity of the problem, we set out to characterise the Whittle index, which is obtained by solving a relaxed version of the MARBP. As reported in the literature, this heuristic performs extremely well for a wide variety of problems. Assuming that the optimal policy of the relaxed problem is of threshold type, we provide an algorithm that finds Whittle's index. We then consider a multi-class queue with linear cost and impatient customers. For this model, we show threshold optimality, prove indexability, and obtain Whittle's index in closed-form. We also study the limiting regimes in which the environment is relatively slower and faster than the controllable process. By numerical simulations, we assess the suboptimality of Whittle's index policy in a wide variety of scenarios, and the general observation is that, as in the case of standard MARBP, the suboptimality gap of Whittle's index policy is small.

1 Introduction

The Multi-Armed Restless Bandit Problem (MARBP), introduced in [36], represents an important class of Markov Decision Processes (MDPs) with numerous applications spanning resource allocation, machine maintenance, health-care systems, and others. In an MARBP, there are multiple concurrent projects or bandits and the decision maker decides which R bandits to simultaneously activate. The decision maker knows the states of all bandits and the cost in every state, and aims at minimising the average cost. The state of a bandit evolves stochastically according to transition rates that depend on whether the bandit is active. In general, MARBPs cannot be solved

*Corresponding author: sgduran91@gmail.com

analytically, except for some toy examples. An MARBP can be solved numerically via dynamic programming, however, this is a computationally intractable task for realistic model sizes.
Whittle [36] developed a methodology to obtain a heuristic by solving a relaxed version of the MARBP in which R bandits are activated only *on average*. The obtained heuristic, nowadays known as Whittle's index policy, relies on calculating the Whittle index for each of the bandits, and activating in every decision epoch the bandit with the highest Whittle index. It has been reported on numerous instances that Whittle's index policy provides strikingly good performance, and it has been shown to be asymptotically optimal as the number of bandits grows large, see [35, 34]. Fundamental questions regarding Whittle's index policy concern their existence and their complexity in computation. To prove existence, one needs to establish a technical property known as *indexability*. Computing Whittle's index might be involved, and in practice the indices are computed on a problem-to-problem basis (see Section 2), either numerically or analytically.
In this paper we are interested in the case in which the state of the bandit is formed by two distinct processes. The dynamics of the first process depends on whether the bandit is activated, whereas the dynamics of the second process does not. The first process will be referred to as *controllable*, whereas the second will be referred to as the *environment*. The transition rates of the controllable process depend on the state of the environment. These so-called *Markov-Modulated Restless Multi-Armed Bandit Problems* (MM-MARBP) arise naturally when there is an exogenous phenomena that cannot be controlled. For example in the context of wireless communications, the available download rate depends on the weather conditions, and in a cloud computing systems the arrival rates of new jobs fluctuate over time. As considered in the original paper by Whittle, in this paper we focus on the average performance criterion.
Adding an *environment* implies that the state of a bandit is no longer unidimensional. This renders the calculation of Whittle's index much more complex. In fact, as detailed in Section 2, most of the papers in which the Whittle index is calculated assume that the state of a bandit is unidimensional. In our main contribution, we consider an arbitrary bandit and under the assumption that the optimal policy is of threshold type, we provide an algorithm that finds Whittle's index in any state. In the case the *environment* changes at a slower time scale than the *controllable* process, we derive an analytical expression for Whittle's index. We then consider the particular case of queues with abandonments, where impatient customers can leave the system, regardless of whether they are being served or not. The state of the queue is the *controllable* process, and the arrival rates, service rates, and abandonment rates depend on the state of an exogenous *environment* process that visits *two* states. We show that threshold policies are optimal, and in the case of linear holding costs and assumptions on the parameters, we prove indexability and derive Whittle's index in closed form. The expression obtained for the index depends on which of the two states the environment process is in. By numerical simulations we assess the suboptimality of Whittle's index policy in a wide variety of scenarios, and the general observation is that, as in the case of standard MARBPs, the suboptimality gap of Whittle's index policy in an MM-MARBP is small.
The paper is organised as follows. In Section 2 we give an overview of related work and in Section 3 we describe the model. In Section 4 we present the relaxation of the original problem and Whittle's index. In Section 5 we introduce the threshold policies and, assuming they are optimal, we provide the algorithm to determine Whittle's index. In addition, we characterise Whittle's index when the time-scale of the environment is much slower than that of the controllable process. In Section 6 we study the multi-class queue with abandonments. Optimality of threshold policies is proved in Section 6.1. In Section 6.2 we prove indexability and obtain expressions for Whittle's index. The

proofs of the theorems are in Section 6.3. In Section 7 we derive Whittle's index for a multi-class queue without abandonments living in a Markov-modulated environment. Finally, in Section 8 we numerically evaluate the performance of Whittle's index policy. For ease of reading, many proofs are presented in the appendix.

2 Related literature

A classical reference for MDPs is [30], and a comprehensive coverage for MABP and MARBP is given in [16]. Even though Whittle's seminal work introduced Whittle's index within the context of average cost criterion, a large body of work has focused on tackling an MARBP under the total discounted cost criterion. For the discounted cost criterion and finite state space, [26] provides a thorough analysis and efficient algorithms (based on linear programming) to establish indexability and to give an expression for Whittle's index. The same approach was undertaken to obtain the Whittle index for a general MARBP in [25]. By letting the discounting factor tend to one, we can retrieve the index for the average cost criterion, see for instance [18, 4]. Another approach to calculate the index lies in sweeping the state space, by recursively identifying and calculating the states with higher Whittle's indices. This can be done by iterative schemes, see for example [9, 10, 8], and in some particular cases analytically, see [6, 28, 23].

The references above consider unidimensional bandits, which is a critical assumption in order to establish indexability, and in turn to calculate the index. On the other hand, literature on multidimensional MARBPs, as is the case for MM-MARBP, is scarce. The main difficulty lies in establishing indexability, i.e., ordering the states, in a multidimensional space. Important exceptions are [1, 3], in which the authors derive Whittle's index. These papers model the problem of scheduling tasks in a wireless setting: the *controllable* process is the remaining service time of tasks, which is a decreasing process, and the *environment* is given by the capacity of the channel, which fluctuates over time in an uncontrollable manner.

The above references consider independent environments for each of the bandits. In [15] the authors consider an MDP made of independent objects evolving in a common environment. It is shown that as the number of objects tends to infinity, the optimal policy converges to the optimal policy of a deterministic discrete time system.

In [14], MM-MARBP are studied when the environment is unobservable, that is, the decision maker cannot observe the state of the *environment*, and it can take its decision only based on the state of the *controllable* process. It is shown that as the number of bandits grow large and the environment changes state relatively fast, a set of priority policies – that includes an averaged version of Whittle's index – is asymptotically optimal. In the numerical section, we will compare the performance of the averaged Whittle's index policy to that of the index policy in the observable setting as studied in this paper. There are a few other papers that have investigated the optimal control of particular queueing systems – in an asymptotic regime – operating in an unobservable environment. In [12], the authors study a single-server multiclass queueing network in heavy traffic with modulated arrival and service rates. In the main result, it is shown that an "averaged" version of the classical $c\mu$-rule is asymptotically optimal. In [5], the authors investigate optimal control of a many-server system with modulated arrivals, service and abandonment rates in the Halfin-Whitt regime.

In this work, we use the multi-class abandonment queue as a case study where customers may

abandon before having their service finished. Abandonments is an undesirable effect as it induces wasted resources. It has therefore been largely studied in recent literature, see for example the Special Issue in Queuing Systems on queuing systems with abandonments [19] and the survey [13] in the many-server settings. We further mention the work of [22], where the authors formulate the abandonment queue as a MARBP and derive a closed-form expression for Whittle's index. In this paper, we extend this work by considering the abandonment queue with arrival rates, service rates and abandonment rates that fluctuate over time.

3 Model description

We consider a multi-armed restless bandit problem in continuous time. There are N bandits in the system, each bandit is composed by a controllable process and an environment process. The controllable process lives in the state space $\mathcal{X} = \{0, 1, \ldots\}$. A bandit can be kept passive or made active, with the constraint that at most R bandits can be made active at a time, $R \in \mathbb{N}$ and $R \leq N$. The transition rates of the controllable process of a bandit depend on whether it is made active or kept passive and on the current state of the environment, as defined below.
The environments are exogenous Markov processes living in the state space $\mathcal{Z} = \{1, 2, \ldots\}$, whose evolution is independent of the state of the controllable processes or the actions taken. Let $D_k(t) = d \in \mathcal{Z}$ denote the state of the environment of bandit k at time t, $k = 1, \ldots, N$, and $r_k^{(dd')}$ the transition rate of $D_k(t)$ from state d to d'. We assume the environments $D_k(t)$ are positive recurrent. We denote by $\phi_k(d)$ the stationary probability of environment $D_k(t)$ to be in state d.
We note that no further assumptions are made on the correlation between different environments. However, in the numerical examples, we focus on the following two special cases:

- *Independent environments:* The variables $(D_k(t))_{k=1}^N$ are independently distributed. As a consequence, given the action, the evolution of the two-dimensional state of a bandit, $(M_k(t), D_k(t))$, is independent of the others. Hence, this setting falls within the classical MARBP with a 2-dimensional state space.

- *Common environment:* There is only one environment affecting all bandits, that is, $D_1(t) = \ldots = D_N(t) = D(t)$. When environment $D(t)$ changes state from d to d', it changes the transition rates of all bandits at once. In this case there is a correlation between bandits, which does not fit in the standard MARBP model.

Let φ denote the policy that determines the action taken for each bandit. We assume that policies are Markovian, that is, they can base their decisions only on the current state of the bandits. In particular, this implies that decisions can depend on the state the environments are in. That is, the decision maker can observe the environments.
For a given policy φ, let $M_k^\varphi(t) \in \mathcal{X}$ denote the state of the controllable process of bandit k at time t, $k = 1, \ldots, N$. Note that $\left(M_k^\varphi(t), D_k(t)\right)$ describes the two-dimensional state of bandit k. The full description

$$\left(M_1^\varphi(t), D_1(t), \ldots, M_N^\varphi(t), D_N(t)\right),$$

is then a Markov process. Throughout this paper, we assume this process to be ergodic. We note that obtaining ergodicity conditions on the parameters is out of the scope of this paper.

We denote the passive action by $a = 0$ and the active action by $a = 1$. We further denote by $A_k^{\varphi}(t) \in \{0, 1\}$ the action taken for bandit k at time t under policy φ, and $\vec{A}^{\varphi}(t) := (A_1^{\varphi}(t), \ldots, A_N^{\varphi}(t))$ the actions taken for all the bandits. The constraint of making at most R bandits active can be expressed as

$$\sum_{k=1}^{N} A_k^{\varphi}(t) \leq R, \quad \forall t \geq 0. \tag{1}$$

We define the set of feasible policies Φ, as the set of all Markovian policies that satisfy (1).
When action a is applied to bandit k and its environment is in state d, its controllable process makes a transition from state m to state m' at rate $q_k(m'|m, d, a)$. Let $C_k(m, d, a)$ denote the cost per unit of time for bandit k when the controllable process is in state m, the environment is in state d and the action taken is a. For any k, we assume that $C_k(m, d, a)$ is a convex non-decreasing function in m for any d, a.
The objective of the optimisation problem is to find the policy φ that minimises the long-run average holding cost under constraint (1), i.e., solve:

$$\min_{\varphi \in \Phi} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left(\int_0^T \sum_{k=1}^{N} C_k(M_k^{\varphi}(t), D_k(t), A_k^{\varphi}(t)) \mathrm{d}t \right). \tag{2}$$

Finding a solution for the constrained problem (2) is PSPACE-hard, i.e., it can not be solved using a polynomial amount of memory nor a polynomial amount of time, see [16, 24, 37]. However, we refer to Whittle [36] where a Lagrangian relaxation method is introduced allowing to obtain efficient index policies for the original problem. In Section 4 this is applied to the MARBP with uncontrollable environments.

4 Relaxation and Whittle's index policy

In this section, we introduce the relaxed version of the Markov-modulated multi-armed restless bandit problem. The main idea of this methodology, as proposed by Whittle in [36], is to solve an unconstrained problem obtained via a Lagrangian relaxation approach, instead of solving problem (2) under constraint (1). This leads to a significant simplification of the problem: instead of solving the problem for N bandits simultaneously, one solves the problem separately for each of the N bandits. The solution for the relaxed problem can be described by the so-called Whittle index, which forms the basis for a heuristic for the original problem, known as the Whittle index policy, defined later on in this section. the Whittle index policy for the standard MARBP has been proved to be well-performing in many important examples, see Niño-Mora [27], and asymptotically optimal under certain circumstances, see Weber and Weiss [35], Ji et al. [20], Ouyang et al. [29] and Verloop [34].
We now study the relaxed problem, that is, the constraint on the number of active bandits must be satisfied on average, and not in every decision epoch:

$$\limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left(\int_0^T \sum_{k=1}^{N} A_k^{\varphi}(t) \mathrm{d}t \right) \leq R, \tag{3}$$

or equivalently, $\limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left(\int_0^T \sum_{k=1}^{N} \left(1 - A_k^{\varphi}(t) \mathrm{d}t\right) \right) \geq N - R$. We denote the set of policies satisfying constraint (3) by Φ^{REL}, and we note that $\Phi \subseteq \Phi^{REL}$. We now consider the problem

of finding a policy φ that minimises (2) under constraint (3). We use the Lagrangian multipliers approach to rewrite the following unconstrained version of the relaxed problem: find a policy φ that minimises

$$\min_{\varphi} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \left(\sum_{k=1}^{N} C_k(M_k^{\varphi}(t), D_k(t), A_k^{\varphi}(t)) - W(R - N + \sum_{k=0}^{N}(1 - A_k^{\varphi}(t)))\right) \mathrm{d}t\right). \quad (4)$$

The Lagrange multiplier W can be viewed as a subsidy for making a bandit passive. The key observation made by Whittle is that problem (4) can be decomposed into N subproblems, one for each bandit, due to the fact that there is no longer a common constraint. Thus, the solution to (4) is obtained by combining the solution to N separate optimisation problems, that is,

$$\min_{\varphi} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \left(C_k(M_k^{\varphi}(t), D_k(t), A_k^{\varphi}(t)) - W(1 - A_k^{\varphi}(t))\right) \mathrm{d}t\right), \quad (5)$$

for each bandit k. Under ergodicity conditions, we define $g_k^{\varphi}(W)$ as

$$g_k^{\varphi}(W) := \mathbb{E}\left(C_k(M_k^{\varphi}, D_k, A_k^{\varphi})\right) - W\mathbb{E}(\mathbf{1}_{A_k^{\varphi}=0}), \quad (6)$$

where M_k^{φ}, D_k and A_k^{φ} are the respective steady-state variables for a given bandit under policy φ. Then, problem (5) is equivalent to the problem $\min_{\varphi} g_k^{\varphi}(W)$.

We now introduce the indexability notion. Let $P_k(W) \subset \mathcal{X} \times \mathcal{Z}$ denote the set of states (m, d) in which it is optimal to be passive when the subsidy for passivity is W. A bandit is called indexable if $P_k(W)$ increases in W.

Definition 1. *Bandit k is indexable if $W < W'$ implies $P_k(W) \subseteq P_k(W')$.*

In other words, indexability implies that if for $W = W_0$ it is optimal to be passive in state (m, d), then it is also optimal to be passive for any value for the subsidy $W \geq W_0$. In [26] a survey on indexability results can be found. In particular, in [26] an algorithmic method is provided in order to determine if, for a given set of parameters, the problem is indexable. In [35], the authors run statistical tests with random parameters to check indexability, obtaining that around 90% of the cases are indexable. Examples where indexability has been proved for particular model instances can be found in [23, 36]. In this work, we establish indexability for a multi-class queue with linear cost and impatient customers under certain conditions, see Section 6.2.

Indexability guarantees the existence of Whittle's index, which is defined as follows: it is the smallest value for the subsidy such that it is optimal to be passive in that state.

Definition 2. *When bandit k is indexable, Whittle's index in state (m, d) is defined by $W_k(m, d) := \inf\{W : (m, d) \in P_k(W)\}$.*

We can now give an optimal solution to the relaxed control problem (4) for a given W. At any moment in time t, make active all bandits whose current Whittle's index exceeds the subsidy for passivity, i.e., $W_k(M_k(t), D_k(t)) > W$. A standard Lagrangian argument together with the fact that the cost functions C_k are convex non-decreasing, gives that there exists a multiplier W such that constraint (3) is satisfied.

Since the solution to the relaxed optimisation problem will in general not be feasible for the original problem (2) with constraint (1), Whittle proposed a heuristic based on the Whittle index. We will refer to this policy as *Whittle's index policy*.

Definition 3 (Whittle's index policy)**.** *At time t, the Whittle index policy prescribes to make active the R bandits having currently the largest value for their Whittle's index $W_k(M_k(t), D_k(t))$.*

5 Calculation of Whittle's index

In this section, we present our main results for the general model. In Section 5.1, we provide an algorithm to calculate the Whittle index, and in Section 5.2, we obtain Whittle's index when the environment changes very slowly compared to the state of the controllable process.
Since we focus on the relaxed problem for one bandit (5), we omit the dependence on k for ease of notation throughout this section.

5.1 Threshold policies

We assume throughout the paper that there exists a threshold policy (defined below) that is an optimal solution of (5). This is typically the case in many queueing models. For the case study of an abandonment queue, we prove it to hold in Section 6.1. In this section, we further assume the bandit is indexable, as defined in Definition 1.
Threshold policies are defined through a vector $\vec{n} = (n_d : d \in \mathcal{Z})$, where $n_d \in \{-1, 0, 1, \ldots\} \cup \{\infty\}$, for all d. Threshold policy $\vec{n}$ activates the bandit if and only if the controllable process is above the threshold n_d when the state of the environment is d. In other words, $A^{\vec{n}}(m, d) = 1$ if $m > n_d$ and $A^{\vec{n}}(m, d) = 0$ if $m \leq n_d$. We denote by $\pi^{\vec{n}}(m, d)$ the stationary probability of the process $\left(M^{\vec{n}}(t), D(t)\right)$ to be in state (m, d).
Alternatively, for some problems, an appropriate definition of threshold policy is to activate the bandit if and only if the controllable process is *below* a threshold. The analysis of both cases is similar, and we choose the former case in our presentation. See [23, Section 3.2] for a case where both types of threshold policies are considered.
Since we assume that threshold policies are optimal, in order to obtain Whittle's index, one can focus on finding the optimal threshold policies for any value of W. Below, we will present an algorithm that allows us to determine those optimal threshold policies. A similar algorithm was presented in [22], where an expression for Whittle's index is obtained for the classical abandonment queue. In [22], the state space of the bandits is one-dimensional, so that the threshold value is one-dimensional as well. In this paper, we present a generalised version of that algorithm as we will have multiple threshold values, one threshold per environment.
First recall that the average cost (see (6)) under a threshold policy $\vec{n}$ is given by

$$g^{\vec{n}}(W) := \sum_{d \in \mathcal{Z}} \sum_{m=0}^{\infty} C\left(m, d, a\right) \pi^{\vec{n}}(m, d) - W \sum_{d \in \mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m, d). \tag{7}$$

Hence, the optimal average cost can be written as

$$g(W) := \min_{\vec{n}} g^{\vec{n}}(W).$$

In Figure 1, the function $g(W)$ and the functions $g^{\vec{n}}(W)$ are plotted. Note that for any $\vec{n}$, $g^{\vec{n}}(W)$ is a non-increasing linear function in W whose slope equals $-\sum_{d \in \mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m, d)$, and $g(W)$ is a lower envelope of the functions $g^{\vec{n}}(W)$. In particular, one observes that the horizontal axis can be split in intervals, where in each interval a different threshold policy is optimal. To find these intervals, note that for $W = -\infty$ the threshold policy $\vec{n}^{-1} := (-1, -1, \ldots)$ is optimal. This follows because the slope under threshold policy $(-1, -1, \ldots)$ equals $-\sum_{d \in \mathcal{Z}} \sum_{m=0}^{-1} \pi^{\vec{n}}(m, d) = 0$, while the slope of all other threshold policies are strictly negative. Now one can look to the first value

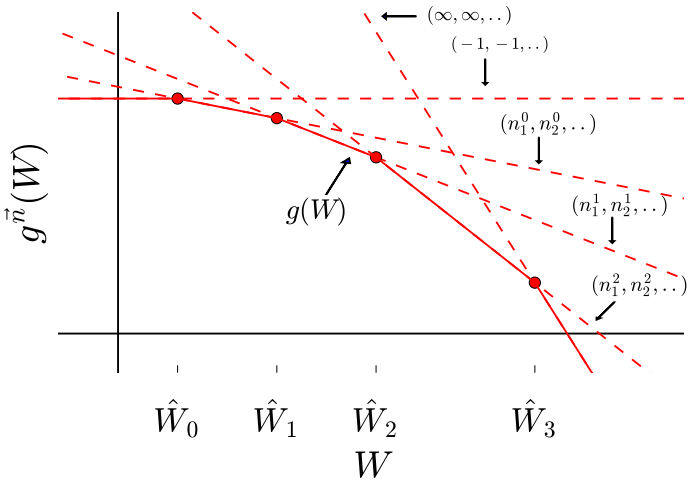


Figure 1: Lower envelop $g(W) := \min_{\vec{n}} g^{\vec{n}}(W)$.

of the subsidy, $\hat{W}_0$, where a linear function $g^{\vec{n}}(W)$ crosses $g^{(-1,-1,\ldots)}(W)$, so that $(-1,-1,\ldots)$ is optimal if and only if $W \in (-\infty, \hat{W}_0]$. Let $\vec{n}^0$ be the corresponding threshold policy of this linear function. Note that in case there are two threshold policies that turn optimal in the crossing point $\hat{W}_0$, then the one having the steepest slope will be the one that minimises when the subsidy increases slightly. Hence, we choose that one in case of a tie. In case there are more than one minimisers with the same slope, one chooses the maximum value of n_d in each component. Then, one knows that Whittle's index is given by $W(m,d) = \hat{W}_0$, for $m = 0, \ldots, n_d^0$ for all d such that $n_d^0 \geq 0$, since $\hat{W}_0$ is the smallest value such that it is optimal to be passive in those states. Similarly, one can now determine the first value of the subsidy $W > \hat{W}_0$, denoted by $\hat{W}_1$, where a linear function $g^{\vec{n}}(W)$ crosses $g^{\vec{n}^0}(W)$. Let $\vec{n}^1$ be the corresponding threshold policy. Now, the Whittle index for states (m,d), with $n_d^0 < m \leq n_d^1$, is given by $W(m,d) = \hat{W}_1$. We can repeat this procedure as long as a new crossing point occurs as W increases. In case in step j no new crossing point occurs, this implies that there exists a $\vec{n}^{j-1}$ such that $g(W) = g^{\vec{n}^{j-1}}(W)$ for $W \geq \hat{W}_{j-1}$. Hence, for states (m,d), with $m > n_d^{j-1}$, it is optimal to be active for any value of W, that is, the Whittle index is not defined for those states.

To formalize the above procedure, we introduce the following notation for the crossing points of the linear functions $g^{\vec{n}}(W)$. We denote by $\overline{W}(\vec{n}, \vec{n}') := \left\{ W \in \mathbb{R} \;\middle|\; g^{\vec{n}}(W) = g^{\vec{n}'}(W) \right\}$, the set of multipliers W such that the expected cost under threshold policies $\vec{n}$ and $\vec{n}'$ is equal. In case the slopes are not equal, i.e., $\sum_{d \in \mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) \neq \sum_{d \in \mathcal{Z}} \sum_{m=0}^{n'_d} \pi^{\vec{n}'}(m,d)$, the set $\overline{W}(\vec{n}, \vec{n}')$ has a unique element, which from (7) is given by:

$$\overline{W}(\vec{n}, \vec{n}') = \frac{\sum_{d \in \mathcal{Z}} \sum_{m=0}^{\infty} C(m,d,a)\, \pi^{\vec{n}}(m,d) - \sum_{d \in \mathcal{Z}} \sum_{m=0}^{\infty} C(m,d,a)\, \pi^{\vec{n}'}(m,d)}{\sum_{d \in \mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) - \sum_{d \in \mathcal{Z}} \sum_{m=0}^{n'_d} \pi^{\vec{n}'}(m,d)}. \tag{8}$$

In case the slopes are equal, the linear functions $g^{\vec{n}}(W)$ and $g^{\vec{n}'}(W)$ are parallel to each other. Then, $\overline{W}(\vec{n}, \vec{n}')$ is either the empty set, or the full state space.

In Algorithm 1 we summarise the above. Under the assumption of threshold optimality and indexability, the output of Algorithm 1 are the threshold policies $\vec{n}$ that optimize (5) for each W and Whittle's index. A mathematical proof of Algorithm 1 would follow according to similar steps as in Glazebrook et al. [17], where an algorithm is developed for finding Whittle's index in the context of admission control and routing of impatient customers.

Algorithm 1. *Define $\vec{n}^{-1}$ the vector equal to -1 in every coordinate, that is, under policy $\vec{n}^{-1}$ the bandit is active in all environments. Then, for $j \geq 0$,*

Step j *Let*

$$E_j = \{\vec{n} : \sum_{d\in\mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) \neq \sum_{d\in\mathcal{Z}} \sum_{m=0}^{n_d^{j-1}} \pi^{\vec{n}^{j-1}}(m,d) \text{ and } n_d \geq n_d^{j-1},\ \forall d\}.$$

If $E_j \neq \emptyset$, compute

$$\hat{W}_j = \inf_{\vec{n}\in E_j} \overline{W}(\vec{n}, \vec{n}^{j-1}). \tag{9}$$

Denote by $\vec{n}^j$ the minimiser of (9) *in case the latter is unique. In case of a tie, choose the minimisers of* (9) *(denoted by $\vec{n}^{j,i}$, $i = 1, \ldots, I$) that have the steepest slope $-\sum_{d\in\mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d)$ and set $\vec{n}^j$ s.t. $n_d^j = \max_i n_d^{j,i}$. Define $W(m,d) := \hat{W}_j$ for $n_d^{j-1} < m \leq n_d^j$, for every d. Go to step j+1.*

If $E_j = \emptyset$, then for all states (m,d) with $m > n_d^{j-1}$ the Whittle index is not defined.

Remark 1. *The numerical computation of this algorithm requires the set over which the minimisation is done to be finite. Hence, in case $|\mathcal{X}| = \infty$ or $|\mathcal{Z}| = \infty$, the results obtained are an approximation. If, instead, $|\mathcal{X}| < \infty$ and $|\mathcal{Z}| < \infty$, and the steady state distribution of the threshold policies is obtained beforehand, the complexity of the algorithm is $\mathcal{O}\left(\left(|\mathcal{X}|^{|\mathcal{Z}|}\right)^2\right)$. This can be seen as follows. (i) In each step, the algorithm calculates the crossing point $\overline{W}(\vec{n}, \vec{n}^{j-1})$ for all threshold policies $\vec{n}$ in E_j. This step has a complexity that corresponds to the number of threshold policies, i.e., $\mathcal{O}\left(|\mathcal{X}|^{|\mathcal{Z}|}\right)$. (ii) Then, in the worst case, the algorithm repeats step (i) for each threshold policy, resulting in a complexity of $\mathcal{O}\left(\left(|\mathcal{X}|^{|\mathcal{Z}|}\right)^2\right)$.*

5.2 Slowly changing environment

In this section, we give an analytical solution of Whittle's index when the environment changes at a much slower time scale than the dynamics of the controllable process of the bandit. We show that in the limit, the Whittle index in state (m,d) coincides with the Whittle index of a bandit that only sees environment d.

Before giving the results, we introduce some notation for a bandit that always sees environment d. That is, the bandit lives in the state space $\mathcal{X} = \{0, 1, \ldots\}$ and its transition rates are $q(m'|m,a) := q(m'|m,d,a)$ for $m, m' \in \mathcal{X}$ and $a = 0, 1$. Let $(p^{n_d,(d)}(m))$ denote the corresponding steady-state probability under threshold policy n_d. We further let $W^{(d)}(m)$ be Whittle's index in state m of a bandit that always sees environment d.

The different time scales are obtained by scaling the transition rates of the environment by β: $\beta r^{(dd')}$, and taking the limit $\beta \to 0$. In order to obtain an analytical expression for Whittle's index in the slow regime, we will assume the environment can be in a finite set of states, i.e., $|\mathcal{Z}| < \infty$. When the environment changes state at a much slower time scale than the controllable process, the conditional (on the environment) steady-state behavior of the bandit is that of a bandit whose environment never changes. This is stated formally below. The proof can be found in Appendix 10.1.

Lemma 1. *Assume $|\mathcal{Z}| = Z < \infty$ and let the transitions of the environment be scaled by β, i.e., $\beta r^{(dd')}$. Then it holds that*

$$\lim_{\beta \to 0} \pi^{\vec{n}}(m,d) = \phi(d) p^{n_d,(d)}(m), \qquad \forall m \in \mathbb{N}_0. \tag{10}$$

The following lemma states that in the slow regime, the index value given by Algorithm 1 can be found by changing the threshold value in only one of the environments. The proof can be found in Appendix 10.1. For a given β, we denote by $\vec{n}^j(\beta) = (n_1^j(\beta), \ldots, n_Z^j(\beta))$, $Z = |\mathcal{Z}|$ the values of $\vec{n}^j$ as defined in Algorithm 1.

Lemma 2. *Let $|\mathcal{Z}| < \infty$ and the transitions of the environment be scaled by β, $\beta r^{(dd')}$. Assume $(\hat{n}_1^j, \ldots, \hat{n}_Z^j) := \lim_{\beta \to 0}(n_1^j(\beta), \ldots, n_Z^j(\beta))$ exists, for all j, and that the family $\{M^{\vec{n}}, \beta\}$ is uniform integrable, for any threshold policy $\vec{n}$. Furthermore, assume $\sum_{m=0}^{n} p^{n,(d)}(m) - \sum_{m=0}^{n'} p^{n',(d)}(m) \geq 0$ for any d and any pair $n > n'$. Then,*

$$\lim_{\beta \to 0} \inf_{\vec{n} \in E_j} \frac{\sum_{d=1}^{Z} \sum_{m=0}^{\infty} C(m,d,a)\, \pi^{\vec{n}}(m,d) - \sum_{d=1}^{Z} \sum_{m=0}^{\infty} C(m,d,a)\, \pi^{\vec{n}^{j-1}(\beta)}(m,d)}{\sum_{d=1}^{Z} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) - \sum_{d=1}^{Z} \sum_{m=0}^{n_d^{j-1}(\beta)} \pi^{\vec{n}^{j-1}(\beta)}(m,d)} \tag{11}$$

$$= \min_{d=1,\ldots,Z} \inf_{n > \hat{n}_d^{j-1}} \frac{\sum_{m=0}^{\infty} C(m,d,a)\, p^{n,(d)}(m) - \sum_{m=0}^{\infty} C(m,d,a)\, p^{\hat{n}_d^{j-1},(d)}(m)}{\sum_{m=0}^{n} p^{n,(d)}(m) - \sum_{m=0}^{\hat{n}_d^{j-1}} p^{\hat{n}_d^{j-1},(d)}(m)} \tag{12}$$

$$= \frac{\sum_{m=0}^{\infty} C(m,d_0,a)\, p^{\hat{n}_{d_0}^{j},(d_0)}(m) - \sum_{m=0}^{\infty} C(m,d_0,a)\, p^{\hat{n}_{d_0}^{j-1},(d_0)}(m)}{\sum_{m=0}^{\hat{n}_{d_0}^{j}} p^{\hat{n}_{d_0}^{j},(d_0)}(m) - \sum_{m=0}^{\hat{n}_{d_0}^{j-1}} p^{\hat{n}_{d_0}^{j-1},(d_0)}(m)}, \tag{13}$$

where d_0 is such that $\hat{n}_{d_0}^{j} \neq \hat{n}_{d_0}^{j-1}$ (there is at least one such d_0).

The following proposition states that in the slow regime the index $W(m,d)$ converges to the index for a fixed environment d, $W^{(d)}(m)$.

Proposition 1. *Let $|\mathcal{Z}| < \infty$ and the transitions of the environment be scaled by β, $\beta r^{(dd')}$. Under the assumptions of Lemma 2, we have that for any m, d,*

$$\lim_{\beta \to 0} W(m,d) = W^{(d)}(m).$$

The uniform integrability property needs to be checked case by case. For the abandonment queue, as studied in Section 6, we give an alternative proof for the above, which does not require to verify uniform integrability, see Proposition 3.

Proof of Proposition 1: The index $W(m_0, d_0)$ is as obtained from the Algorithm 1, through the

values $n_d^j(\beta)$. Since we assume that the limit of $n_d^j(\beta)$ exists and lives on $\mathbb{N}$, we have $n_d^j(\beta) = \hat{n}_d^j$ for β small enough. Hence, we can choose j such that $n_{d_0}^{j-1}(\beta) < m_0 \leq n_{d_0}^j(\beta)$ and we have

$$\begin{aligned}&\lim_{\beta\to 0} W(m_0, d_0)\\ &= \lim_{\beta\to 0} \inf_{\vec{n}\in E_j} \frac{\sum_{d=1}^{Z}\sum_{m=0}^{\infty} C(m,d,a)\,\pi^{\vec{n}}(m,d) - \sum_{d=1}^{Z}\sum_{m=0}^{\infty} C(m,d,a)\,\pi^{\vec{n}^{j-1}(\beta)}(m,d)}{\sum_{d=1}^{Z}\sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) - \sum_{d=1}^{Z}\sum_{m=0}^{n_d^{j-1}(\beta)} \pi^{\vec{n}^{j-1}(\beta)}(m,d)}.\end{aligned}$$

For d_0 it holds that $\hat{n}_{d_0}^j \neq \hat{n}_{d_0}^{j-1}$. Hence, by Lemma 2, the latter is equal to

$$\lim_{\beta\to 0} W(m_0, d_0) = \inf_{n > \hat{n}_{d_0}^{j-1}} \frac{\sum_{m=0}^{\infty} C(m,d_0,a)\,p^{n,(d_0)}(m) - \sum_{m=0}^{\infty} C(m,d_0,a)\,p^{\hat{n}_{d_0}^{j-1},(d_0)}(m)}{\sum_{m=0}^{n} p^{n,(d_0)}(m) - \sum_{m=0}^{\hat{n}_{d_0}^{j-1}} p^{\hat{n}_{d_0}^{j-1},(d_0)}(m)}. \quad (14)$$

The latter is the Whittle index of a bandit that always sees environment d_0, $W^{(d_0)}(m_0)$. This follows by applying Algorithm 1 to the bandit that always sees environment d_0. This concludes the proof. □

5.3 Asymptotic optimality of Whittle's index policy

In [35], the authors considered the standard MARBP model (i.e., no environments) and proved that Whittle's index policy is optimal as the number of bandits and the number of active bandits, R, scale proportionally. This result holds for the standard MARBP setting and requires a so-called global attractor property to be satisfied for the corresponding deterministic set of differential equations.
In the case of independently distributed environments, the state (m, d) is a two-dimensional state of a classical bandit. Hence, the asymptotic optimality result of [35] directly applies. When the environments are correlated on the other hand, no asymptotic result is known. We leave this as future research.

6 Abandonment queue in a Markovian environment

In this section we study a multi-class queue with abandonments living in an observable environment. There are N classes of jobs. Each class is associated an environment process $D_k(t)$ that can be either in state 1 or state 2. We restrict to two states, as this will allow to obtain a closed-form expression for Whittle index. For ease of notation, we define $r_k^{(d)} := r_k^{(d,3-d)}$ for $d = 1, 2$. When the class-k environment is in state d, new class-k jobs arrive according to a Poisson process with rate $\lambda_k^{(d)}$. They require an exponentially distributed amount of service with parameter $\mu_k^{(d)}$. A class-k job abandons the system after an exponentially distributed amount of time with parameter $\theta_k^{(d)}$. Let $C_k(m, d, a)$ denote the cost per unit of time for holding m class-k jobs in the system when the environment is in state d and action a is taken. We assume that the cost function satisfies

$$C_k(m,d,0) - C_k((m-1)^+,d,0) \leq C_k(m+1,d,1) - C_k(m,d,1) \leq C_k(m+1,d,0) - C_k(m,d,0), \quad (15)$$

for all m, d, such that $m \geq 0$. This property is directly implied, for example, when (i) $C_k(m, d, a) = C_k(m, d)$, or when (ii) $C_k(m, d, a) = C_k((m - a)^+, d)$. Here, (i) represents holding cost of jobs in the *system* and case (ii) represents holding costs of jobs waiting for service in the *queue*.

The objective is to minimise the time-average holding cost, see (2).
Similar to [21], we can cast this abandonment model into a Markov-modulated MARBP. There are N bandits, where each bandit represents a class of jobs. The state of bandit k is simply the number of class-k jobs in the system. We then have the following transition rates for bandit k:

$$q_k(m+1|m,d,a) = \lambda_k^{(d)} \quad \text{and} \quad q_k((m-1)^+|m,d,a) = m\theta_k^{(d)} + a\mu^{(d)},$$

for $m \in \mathbb{N}_0, d = 1,2$ and $a = 0,1$. Activating a bandit is equivalent to serving this class. At most $R < N$ classes can be served at a time.
In the remainder of this section, we calculate Whittle's index for one class/bandit. In order to use Algorithm 1, we first show in Section 6.1 that an optimal solution of the relaxed problem is of threshold type. Then, in Section 6.2, we obtain a closed-form expression for Whittle's index in the case of linear holding cost.

6.1 Threshold policies

In this section, we prove that an optimal solution of the relaxed problem (5) for the abandonment model is of threshold type. We henceforth focus on one bandit. For ease of exposition, we remove the subscript k.

Proposition 2. *For each W, there exists an $\vec{n}(W) = (n_1(W), n_2(W))$ such that the threshold policy $\vec{n}(W)$ is an optimal solution of the relaxed problem* (5).

Proof: The value function $V(m,d)$ satisfies the Bellman's optimality equation for average costs [30], which in this case is

$$\begin{aligned}
&(\mu^{(d)} + m\theta^{(d)} + \lambda^{(d)} + r^{(d)})V(m,d) + g = \\
&\lambda^{(d)}V(m+1,d) + m\theta^{(d)}V((m-1)^+,d) + r^{(d)}V(m,3-d) \\
&+ \min\left\{C(m,d,0) - W + \mu^{(d)}V(m,d), C(m,d,1) + \mu^{(d)}V((m-1)^+,d)\right\},
\end{aligned} \tag{16}$$

where g is the averaged cost incurred under the optimal policy. Proving optimality of a threshold policy is equivalent to proving that if in state $m+1$ (with $m \geq 0$) it is optimal to be passive, it is also optimal to be passive in state m. Regarding (16), we have to show that $C(m+1,d,0) - W + \mu^{(d)}V(m+1,d) \leq C(m+1,d,1) + \mu^{(d)}V(m,d)$ implies that $C(m,d,0) - W + \mu^{(d)}V(m,d) \leq C(m,d,1) + \mu^{(d)}V((m-1)^+,d)$. A sufficient condition to prove this is Property (15) together with convexity of the value function for each d, m, so $2V(m,d) \leq V(m+1,d) + V((m-1)^+,d)$, which we prove next.
In order to prove the convexity of $V(m,d)$, we use the value iteration method. However, this technique needs uniformly bounded transition rates. We therefore consider a truncated space at L and smooth the arrival transitions. We define the value function for the truncated system with parameter L as $V^L(m,d)$. In Appendix 10.2.1 we show that $V^L(m,d)$ is a convex function, and in Appendix 10.2.2 that sufficient conditions hold in order to apply [7, Theorem 3.1], to state convergence of $V^L \to V$ as $L \to \infty$. With these two results, convexity of V is concluded. □

Below we prove properties on $\pi^{\vec{n}}(m,d)$, the steady-state probability of having m jobs in the system and being in environment d under threshold policy $\vec{n}$. The proofs can be found in Appendix 10.3 and Appendix 10.4.

The first property can be seen as a rate-conservation law and is based on the fact that, for any given policy, the long-run number of arrivals into a given environment must be the same as the long-run number of departures out of that environment. We state this result in Lemma 3 in the context of threshold policies.

Lemma 3. *Under threshold policy $\vec{n}$ it holds that*

$$\lambda^{(d)}\phi(d) \quad + \quad r^{(3-d)}\mathbb{E}\left(M^{\vec{n}}\mathbf{1}_{(D=3-d)}\right) = \left(\theta^{(d)} + r^{(d)}\right)\mathbb{E}\left(M^{\vec{n}}\mathbf{1}_{(D=d)}\right) + \mu^{(d)} \sum_{m=n_d+1}^{\infty} \pi^{\vec{n}}(m,d) \quad (17)$$

for $d = 1, 2$, where $M^{\vec{n}}$ denotes the random variable with distribution $\pi^{\vec{n}}$.

Lemma 4 proves monotonicity properties on $\sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d)$. This quantity represents the probability of observing environment d and not serving the class.

Lemma 4.

1. *The function $\sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d)$ is non-decreasing in n_d, for $d = 1, 2$.*
2. *The function $\sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d)$ is non-increasing in n_{3-d}, for $d = 1, 2$.*

The first property follows naturally, as increasing the threshold n_d implies that this class is kept passive in more states in environment d. Hence, the probability of being passive in environment d increases. If instead n_{3-d} increases, the number of states where the class is passive in environment d are the same. However, in environment $3-d$ the class is being served in less states, hence this diminishes the death rates in certain states. This allows us to prove that the probability of being passive in environment d decreases, i.e., the second property.

6.2 Whittle's index for linear cost

In this section, we assume linear holding cost, that is, $C(m,d,a) = cm$, with $c \in \mathbb{R}_{\neq 0}$. We first prove that the abandonment problem is indexable, and then give an expression for Whittle's index. For ease of reading, the proofs of the theorems are in Section 6.3.
We introduce the following values:

$$W^{(d)} := c\mu^{(d)} \frac{\theta^{(3-d)} + r^{(1)} + r^{(2)}}{\theta^{(1)}\theta^{(2)} + r^{(1)}\theta^{(2)} + r^{(2)}\theta^{(1)}}, \quad (18)$$

for $d = 1, 2$. The value $W^{(d)}$ has the following interpretation (which will be proved in Lemma 6): If the subsidy W equals $W^{(d)}$, and the class is kept passive in environment $3-d$, then any threshold value for environment d gives the same performance. Throughout Sections 6.2 and 6.3, we assume w.l.o.g. that $W^{(1)} \leq W^{(2)}$, or equivalently $\frac{\mu^{(1)}}{\theta^{(1)}+r^{(1)}+r^{(2)}} \leq \frac{\mu^{(2)}}{\theta^{(2)}+r^{(1)}+r^{(2)}}$. The parameters $W^{(1)}$ and $W^{(2)}$ will play a key role in the characterisation of the optimal solution of the relaxed problem, see Figure 2, and hence in proving indexability and Whittle's index. The results presented in Figure 2 are proved in Section 6.3.
Given $W^{(1)} \leq W^{(2)}$, we make the following two technical conditions, which are needed in order to prove indexability. Numerical simulations however suggest that indexability holds for any set of parameters.

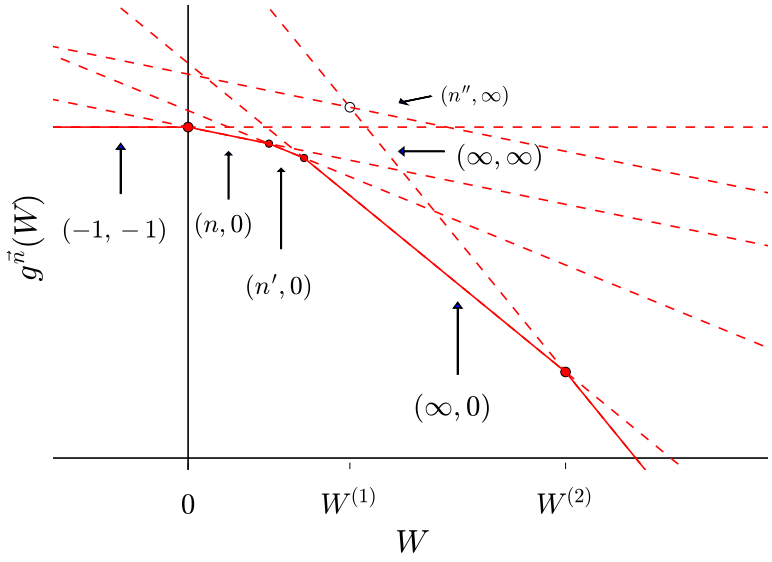


Figure 2: Optimal threshold policies for the abandonment queue.

Condition 1.

$$\mu^{(1)} \leq \mu^{(2)} \text{ and } \theta^{(1)} \geq \theta^{(2)}.$$

Condition 2.

$$W^{(2)} \leq \frac{c\mu^{(2)}}{\theta^{(2)} + r^{(2)}}.$$

Theorem 1. *Assume Conditions 1 and 2 hold. Then all bandits are indexable.*

A closed-form expression for the Whittle index $W(m, d)$ can now be given.

Theorem 2. *Assume the bandit is indexable, and* $\dfrac{\mu^{(1)}}{\theta^{(1)} + r^{(1)} + r^{(2)}} < \dfrac{\mu^{(2)}}{\theta^{(2)} + r^{(1)} + r^{(2)}}$. *We define the subsequence* $(n_j)_{j \geq -1}$ *as follows:* $n_{-1} = -1, n_0 = 0$, *and for* $j \geq 1$ *let*

$$n_j := \underset{n > n_{j-1}}{\arg\min}\, \overline{W}((n_{j-1}, 0), (n, 0)), \tag{19}$$

where $\overline{W}((n_{j-1}, 0), (n, 0))$ *is given by* (8). *In case there is more than one minimiser in* (19), *choose the one that minimises* $-\sum_{d=1}^{2} \sum_{m=0}^{n_d} \pi^{(n,0)}(m, d)$.
Whittle's index $W(m, d)$ *is given by*

$$W(m, d) = \begin{cases} 0 & \text{for } m = 0 \\ \overline{W}((n_{j-1}, 0), (n_j, 0)) & \text{for } n_{j-1} < m \leq n_j, \text{ and } d = 1 \\ W^{(2)} & \text{for } m \geq 1 \text{ and } d = 2, \end{cases} \tag{20}$$

where $\overline{W}((n_{j-1},0),(n,0))$ *is given by* (8). *Moreover,* $W(m,1) \le W^{(1)} \le W^{(2)}$ *for all* m.
If $\dfrac{\mu^{(1)}}{\theta^{(1)}+r^{(1)}+r^{(2)}} = \dfrac{\mu^{(2)}}{\theta^{(2)}+r^{(1)}+r^{(2)}}$, *then* $W(m,1) = W^{(1)} = W^{(2)}$ *for all* m.

Remark 2. *If* $\mu^{(1)}-\mu^{(2)} < \theta^{(2)}$, *then the slope of the linear function* $g^{(n,0)}$, $-\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{(n,0)}(m,d)$, *is a non-increasing sequence in* n, *as it will be stated in Proposition 5. As a consequence, in that case, for each step* j *in Theorem 2, the minimiser* n_j *with the steepest slope is the largest minimiser* n.

In the following corollary we consider the particular case where $n_j = j$ for all $j \ge -1$, or equivalently, the sequence $\overline{W}((n-1,0),(n,0))$ is strictly increasing in n.

Corollary 1. *If* $\dfrac{\mu^{(1)}}{\theta^{(1)}+r^{(1)}+r^{(2)}} < \dfrac{\mu^{(2)}}{\theta^{(2)}+r^{(1)}+r^{(2)}}$ *and the sequence* $(\overline{W}((n-1,0),(n,0)))_{n\in\mathbb{N}}$ *is strictly increasing in* n, *then Whittle's index* $W(m,d)$ *is given by*

$$W(m,d) = \begin{cases} 0 & \text{for } \ m=0 \\ \overline{W}((m-1,0),(m,0)) & \text{for } \ d=1 \\ W^{(2)} & \text{for } \ d=2 \text{ and } m \ge 1. \end{cases} \tag{21}$$

Although we could not prove the strict-increasing property for the crossing points $\overline{W}((n-1,0),(n,0))$, based on numerical observations we believe this holds whenever $W^{(1)} \neq W^{(2)}$.

Remark 3. *The queuing model with abandonments without an environment has been studied in [22]. That is,* $\lambda^{(1)} = \lambda^{(2)} = \lambda$, $\mu^{(1)} = \mu^{(2)} = \mu$ *and* $\theta^{(1)} = \theta^{(2)} = \theta$. *Then, Whittle's index as in Theorem 2 equals* $c\mu/\theta$, *which is in agreement with [22, Section 6.1].*

We now scale the transition rates of the environments, to further characterise Whittle's index in the two extreme cases of very slow changing or very fast changing environments. As in Section 5.2, β is the scaling parameter and $\beta r^{(d)}$ is the transition rate of the environment. The proof can be found in Appendix 10.5.

Proposition 3. *Assume the transition rates of the environment are scaled as* $\beta r^{(d)}$.
As $\beta \to 0$, *it holds that*

$$\lim_{\beta\to 0} W(m,d) = c\frac{\mu^{(d)}}{\theta^{(d)}}, \quad \forall m,d.$$

Assume the limit $\lim_{\beta\to\infty}\overline{W}((n_{j-1},0),(n_j,0))$ *exists. As* $\beta\to\infty$, *it holds that*

$$\lim_{\beta\to\infty} W(m,d) = \begin{cases} 0 & \text{for } \ m=0 \\ \lim_{\beta\to\infty}\overline{W}((n_{j-1},0),(n_j,0)) & \text{for } \ n_{j-1} < m \le n_j, \text{ and } d=1 \\ c\dfrac{\mu^{(2)}}{\overline{\theta}} & \text{for } \ d=2, \end{cases}$$

where $\overline{\theta} := \sum_{d=1}^{2}\phi(d)\theta^{(d)}$.

For the slow regime, we observe that the Whittle Index $W(m,d)$ coincides with that of the Whittle index when the bandit always sees environment d, given by $\frac{c\mu^{(d)}}{\theta^{(d)}}$, see [21].
When the environment changes fast compared to the controllable state of the bandit, the Whittle index remains state dependent. In environment 2, the index simplifies to $c\frac{\mu^{(2)}}{\overline{\theta}}$. This index is very similar to

(i) $c\frac{\mu}{\theta}$, which is Whittle's index when there are *no environments.*

(ii) $c\frac{\overline{\mu}}{\overline{\theta}}$, an index that was proved to be asymptotically optimal in case the *environments are unobservable [14].*

Hence, we see that when the environment is observable and changes very fast, the Whittle index depends on the departure rate of the current environment and the averaged abandonment rate. In case the environment is unobservable, the averaged value is taken both for the departure rate and for the abandonment rate.

6.3 Proof of Theorems 1 and 2

In this section, we present the proofs of Theorems 1 and 2. These proofs are based on the characterisation of the optimal solution of the relaxed optimisation problem. This characterisation can be found in Propositions 4 and 6. For the lemmas and propositions stated in the section, the proofs can be found in Appendix 10.6.
We start by rewriting $\overline{W}(\vec{n},\vec{n}')$, the subsidy such that the expected cost under both threshold policies $\vec{n}$ and $\vec{n}'$ are equal. From (8), we obtain

$$\overline{W}(\vec{n},\vec{n}') = c\cdot\frac{\sum_{d=1}^{2}\sum_{m=0}^{\infty} m\pi^{\vec{n}}(m,d) - \sum_{d=1}^{2}\sum_{m=0}^{\infty} m\pi^{\vec{n}'}(m,d)}{\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{\vec{n}}(m,d) - \sum_{d=1}^{2}\sum_{m=0}^{n'_d}\pi^{\vec{n}'}(m,d)}, \tag{22}$$

given that the denominator in (22) is not equal to 0.
The rate conservation property of Lemma 3 allows us to give an alternative expression for $\overline{W}(\vec{n},\vec{n}')$. For that, we define

$$s_d(\vec{n},\vec{n}') := \sum_{m=0}^{n_d}\pi^{\vec{n}}(m,d) - \sum_{m=0}^{n'_d}\pi^{\vec{n}'}(m,d),$$

as the difference between the probability of being passive under threshold policies $\vec{n}$ and $\vec{n}'$, while the environment is in state d.

Lemma 5. *In case $s_1(\vec{n},\vec{n}') + s_2(\vec{n},\vec{n}') \neq 0$, it holds that*

$$\overline{W}(\vec{n},\vec{n}') \;=\; tW^{(1)} + (1-t)W^{(2)}, \tag{23}$$

where $t := \dfrac{s_1(\vec{n},\vec{n}')}{s_1(\vec{n},\vec{n}') + s_2(\vec{n},\vec{n}')}$.

Consider two policies that are always passive in environment $d=1$, i.e., $\vec{n}=(\infty,n_2)$ and $\vec{n}'=(\infty,n_2')$, for any pair n_2, n_2'. We have that $\sum_{m=0}^{\infty}\pi^{(\infty,n_2)}(m,1)=\phi(1)$, where $\phi(1)$ is the stationary measure of the environment for being in state $d=1$. Hence, $s_1(\vec{n},\vec{n}') = \sum_{m=0}^{\infty}\pi^{(\infty,n_2)}(m,1) -$

$\sum_{m=0}^{\infty} \pi^{(\infty,n_2')}(m,1) = 0$, and if $s_2(\vec{n},\vec{n}') = \sum_{m=0}^{n_2} \pi^{(\infty,n_2)}(m,2) - \sum_{m=0}^{n_2'} \pi^{(\infty,n_2')}(m,2) \neq 0$, then from Lemma 5 we have

$$\overline{W}((\infty,n_2),(\infty,n_2')) = W^{(2)}. \tag{24}$$

Similarly, we have $\overline{W}((n_1,\infty),(n_1',\infty)) = W^{(1)}$. Following similar reasonings, we derive properties for the policy that never serves the bandit, i.e., (∞,∞).

Lemma 6. *We consider the threshold policy that never serves the bandit, (∞,∞). For any policy $\vec{n} = (n_1,n_2)$ it holds that $\overline{W}((\infty,\infty),\vec{n}) \in [W^{(1)},W^{(2)}]$. Furthermore, $\overline{W}((\infty,\infty),\vec{n}) = W^{(2)}$ if and only if $n_1 = \infty$ and $\overline{W}((\infty,\infty),\vec{n}) = W^{(1)}$ if and only if $n_2 = \infty$.*

In the following proposition, we characterise optimal threshold policies of the relaxed optimisation problem (7) as a function of the subsidy W. For a visual representation of the optimal policies, we refer to Figure 2, where the optimal threshold policies are indicated as W varies.

Proposition 4.
1. *The threshold policy $(-1,-1)$ is an optimal solution of (7) when $W \leq 0$. When $W < 0$, it is the unique optimal threshold policy. When $W = 0$, the optimal threshold policies are $(-1,-1),(-1,0)$, $(0,-1)$ and $(0,0)$.*
2. *Assume Conditions 1 and 2 hold. For $W \in \left[0, W^{(2)}\right)$ the optimal threshold solutions of (7) are of the form $(n,0)$ with $n \geq -1$.*
3. *The threshold policy (∞,∞) is an optimal solution of (7) when $W \geq W^{(2)}$. When $W > W^{(2)}$, it is the unique optimal threshold policy. When $W = W^{(2)}$, the optimal threshold policies are of the form (∞,n) with $n \geq 0$.*

Proposition 5. *Assume Condition 1 holds. Then the sequence $\left(-\sum_{d=1}^{2}\sum_{m=0}^{n_d} \pi^{(n,0)}(m,d)\right)_{n\in\mathbb{N}}$, which represents the slope of the linear functions $(g^{(n,0)})_{n\in\mathbb{N}}$, is a non-increasing sequence.*

We can now prove Theorem 1.

Proof of Theorem 1: By Proposition 2, an optimal solution of (7) is of threshold form. For a given subsidy W, let $n_d(W)$ denote the minimum value for a threshold in environment d such that the threshold policy $(n_d(W),\tilde{n}_{3-d})$, for some $\tilde{n}_{3-d}$, is optimal.
If

$$n_1(W) \leq n_1(W+\Delta) \text{ and } n_2(W) \leq n_2(W+\Delta), \tag{25}$$

with $\Delta > 0$, then, the bandit is indexable. Equation (25) will be proved below.
First assume $W < 0$. Then, by Proposition 4, the optimal threshold policy is $(-1,-1)$. In $W = 0$ it turns to $(0,0)$, and for $W > 0$ it is of the form $(n,0)$. Hence, for $W \leq 0$, (25) holds.
Now assume $0 \leq W < W^{(2)}$. If $W + \Delta > W^{(2)}$, then the inequality (25) is trivially true. Now assume $W+\Delta < W^{(2)}$. By Proposition 4, it follows that there are n and n' such that $\vec{n}(W) = (n,0)$ and $\vec{n}(W+\Delta) = (n',0)$. Since the function $g(W)$ is a lower envelope, the linear function $g^{(n',0)}(W)$ has a steeper slope than the linear function $g^{(n,0)}(W)$, as it can be seen in Figure 1. Furthermore, in Proposition 5 we proved that the slope of the linear functions $g^{(n,0)}(W)$ is non-increasing in n. Therefore, $n \leq n'$, and (25) holds.
Finally, for $W \geq W^{(2)}$, the optimal threshold policy is (∞,∞), hence $n_d(W) = \infty$ for both $d = 1,2$, and (25) is trivially true. □

In order to prove Theorem 2, we first show that threshold policy $(\infty, 0)$ is an optimal solution of (7) when $W \in \left[W^{(1)}, W^{(2)}\right]$. Again, we refer to Figure 2 for a visual representation.

Proposition 6. *The threshold policy* $(\infty, 0)$ *is an optimal solution of* (7) *when* $W \in \left[W^{(1)}, W^{(2)}\right]$. *If* $W \in (W^{(1)}, W^{(2)})$, *then it is the unique optimal threshold policy. If* $W = W^{(2)}$, *then any threshold policy* (∞, n) *with* $n \geq 0$ *is optimal, and no other threshold policy is.*

Proof of Theorem 2: Recall that the Whittle index of a state (m, d) is the smallest value for the subsidy W such that the optimal threshold policy makes that state passive.
First consider states $(0, d)$. From Proposition 4, we have that in state 0 it is optimal to be active when $W < 0$ and passive if $W = 0$. Hence, we have that $W(0, d) = 0$, for both $d = 1, 2$.
Now consider states $(m, 2)$, with $m > 0$. From Proposition 6, we have that for $W \in (W^{(1)}, W^{(2)})$ the unique optimal threshold policy is $(\infty, 0)$. For $W \geq W^{(2)}$, the optimal threshold policy is (∞, ∞), since it is the steepest one. Hence, $W^{(2)}$ is the smallest value for the subsidy W such that in state $(m, 2)$ the optimal action is to be passive, that is, $W(m, 2) = W^{(2)}$ for every $m \geq 0$.
Now consider states of the form $(m, 1)$, with $m > 0$. Note that for $W = 0$, in state 0 it is optimal to be passive, and for $W \in (W^{(1)}, W^{(2)})$ the optimal threshold policy is $(\infty, 0)$. As a consequence, and since the bandit is indexable, when $0 \leq W < W^{(2)}$, an optimal threshold policy is of the form $(n, 0)$. Algorithm 1 characterises Whittle's index by iteratively defining $\hat{W}_j$. From Proposition 4, we have that for $W < 0$ the unique optimal threshold policy is $\vec{n}^{-1} = (-1, -1)$, and for $W = 0$ the optimal threshold policies are $(-1, -1), (-1, 0), (0, -1)$ and $(0, 0)$. Hence, $\hat{W}_0 = 0$. Furthermore, since the bandit is indexable, for $W \geq 0$ it is optimal to be passive in state 0. Therefore, among those four threshold policies, the optimal threshold policy in the interval $[0, \hat{W}_1]$ is $\vec{n}^0 = (0, 0)$. Then, in $W = \hat{W}_1$ the linear function of the following optimal threshold policy $(n_1, 0)$ crosses the linear function $g^{(0,0)}$. Inductively, the increasing sequence $(n_j)_{j \geq 0}$, provides the set of minimising policies inside the set $\{(n, 0)\}_{n \geq 0}$. As a consequence, in order to determine the smallest value of W such that the optimal threshold policy makes a state $(m, 1)$ passive, it is enough to determine the value j such that $n_{j-1} < m \leq n_j$. In other words, $W(m, 1) = \overline{W}((n_{j-1}, 0), (n_j, 0))$ for any j and $n_{j-1} < m \leq n_j$.
We are left to prove that $\overline{W}\left((n_j, 0), (n_{j-1}, 0)\right) \leq W^{(1)}$ for every j. To do so, note that from the definition of the crossing points n_j, the linear functions $g^{(n_{j-1},0)}$ and $g^{(n_j,0)}$ are not parallel, hence $s_1((n_j, 0), (n_{j-1}, 0)) + s_2((n_j, 0), (n_{j-1}, 0)) \neq 0$. Then, from (23) we have that

$$\overline{W}\left((n_j, 0), (n_{j-1}, 0)\right) = W^{(1)} + (1 - t)\left(W^{(2)} - W^{(1)}\right),$$

with $1 - t := \dfrac{s_2((n_j, 0), (n_{j-1}, 0))}{s_1((n_j, 0), (n_{j-1}, 0)) + s_2((n_j, 0), (n_{j-1}, 0))}$. From Property 2) in Lemma 4 it follows that $s_2((n_j, 0), (n_{j-1}, 0)) < 0$. We state the following property between arbitrary linear functions in $\mathbb{R}$: if two linear functions cross each other in a given $\overline{W}$, the one that minimises for $W \geq \overline{W}$ is steeper than the one that minmises for $W \leq \overline{W}$. In this case, since $g^{(n_{j-1},0)}(W)$ minimises for $W \leq \overline{W}\left((n_j, 0), (n_{j-1}, 0)\right)$ and $g^{(n_j,0)}(W)$ minimises for $W \geq \overline{W}\left((n_j, 0), (n_{j-1}, 0)\right)$, $g^{(n_j,0)}$ is steeper than $g^{(n_{j-1},0)}$, i.e., $s_1((n_j, 0), (n_{j-1}, 0)) + s_2((n_j, 0), (n_{j-1}, 0)) > 0$. Then $1 - t \leq 0$, and as a consequence, $\overline{W}\left((n_j, 0), (n_{j-1}, 0)\right) \leq W^{(1)}$ for every j. □

7 Multi-class queue in a Markovian environment

In this section we study a queueing model without abandonments with an observable random environment and general holding cost. That is, we consider the model as described in Section 6 where now the abandonment rates are set equal to zero. First, we present the maximum stability conditions. Secondly, we derive an index policy, based on the results obtained in the previous section for an abandonment queue (by letting $\theta \to 0$).

7.1 Maximum stability conditions

We first provide the maximum stability conditions, that is the conditions on the parameters such that there exists a policy that makes the system stable. This stability result follows from [31].

Proposition 7 (Proposition 1, Section 4, in [31])**.** *Let $\phi(\vec{d})$ be the probability that bandit k sees environment d_k, $k = 1, \ldots, N$. Recall that $\phi_k(d_k)$ denotes the marginal distribution.*
If there exists a policy such that the multi-class queue with environments is stable, then there exists a vector $\vec{\gamma} = (\gamma_{k\vec{d}} : 1 \leq k \leq N, \vec{d} \in \mathcal{Z}^N)$, such that

$$\gamma_{k\vec{d}} \geq 0 \text{ and } \sum_{k=1}^{N} \gamma_{k\vec{d}} \leq \phi(\vec{d}), \text{ for all } k, \vec{d}, \tag{26}$$

and

$$\sum_{d \in \mathcal{Z}} \lambda_k^{(d)} \phi_k(d) \leq \sum_{\vec{d} \in \mathcal{Z}^N} \mu_k^{(d_k)} \gamma_{k\vec{d}}, \quad \forall 1 \leq k \leq N. \tag{27}$$

If there exists a vector $\vec{\gamma}$ such that (26) *and*

$$\sum_{d \in \mathcal{Z}} \lambda_k^{(d)} \phi_k(d) < \sum_{\vec{d} \in \mathcal{Z}^N} \mu_k^{(d_k)} \gamma_{k\vec{d}}, \quad \forall 1 \leq k \leq N, \tag{28}$$

then there exists a policy that makes the system stable.

Proof: For the sake of readability we present a sketch of the proof.
If the system is stable under some policy φ, then we consider the process under this policy in a stationary regime, and let $\gamma_{k\vec{d}}^{\varphi}$ be the average fraction of time that the environment is $\vec{d}$ and bandit k is active. The obtained vector $\vec{\gamma}^{\varphi}$ satisfies (26) by definition. It also satisfies Equation (27), which can be seen by contradiction. Assume it did not hold, then at least one bandit would grow indeterminately towards infinite with probability 1.
If (26) and (28) hold for a certain $\vec{\gamma}$, we define the policy φ as the policy that, under environment $\vec{d}$, makes bandit k active with probability $\gamma_{k\vec{d}}/\phi(\vec{d})$ and with probability $1 - \dfrac{\sum_{k=1}^{N} \gamma_{k\vec{d}}}{\phi(\vec{d})}$ does not make active any bandit. Then policy φ allocates to bandit k on average a service rate equal to $\sum_{\vec{d} \in \mathcal{Z}^N} \mu_k^{(d_k)} \gamma_{k\vec{d}}$, which by (28) is larger than its arrival rate. This implies stability. □

m θ	1	0.8	0.6	0.4	0.2	0.1	0.08	0.06	0.04	0.02	0.01	0.005	0.001
1	1.63	1.7	1.78	1.87	1.97	2.02	2.03	2.04	2.06	2.07	2.07	2.08	2.08
5	3.67	4.06	4.55	5.19	6.03	6.57	6.69	6.81	6.94	7.08	7.15	7.18	7.21
10	4.93	5.64	6.61	7.99	10.10	11.65	12.02	12.42	12.84	13.29	13.53	13.65	13.75
15	5.62	6.57	7.89	9.90	13.31	16.08	16.78	17.55	18.39	19.32	19.82	20.08	20.29
20	5.99	7.17	8.76	11.29	15.89	19.97	21.06	22.27	23.62	25.16	26	26.45	26.82

Table 1: Whittle's index convergence as θ tends to 0.

7.2 Whittle's index policy

In this section we assume that the environment of each bandit can be either in state 1 or state 2. For the multi-class queue without abandonments, one cannot directly apply Algorithm 1 in order to get Whittle's index. The reason for this is that the value of the threshold does not always impact the average obtained subsidy for passivity, and hence does not provide a Whittle index. For example, assume $\mu^{(1)} < \mu^{(2)}$. Then, as the subsidy grows from $-\infty$ to ∞, threshold policies of the form $(n, 0)$, $n = 1, 2, \ldots$, will be optimal, and for large enough W, the threshold policy $(\infty, 0)$ is optimal. However, once in environment 1 it is optimal to be passive in all states, when now comparing different threshold values for environment 2, each such threshold will have the same steady-state probability of being passive. Hence, there is no difference in the average obtained subsidy for passivity between threshold policies of the form (∞, n) and $(\infty, n + n')$. This means that no index can be defined for states $(m, 2)$. A similar observation was made for the classical multi-class queue without environments, see [22, Section 7]. In order to obtain Whittle's index for the multi-class queue with environments, we therefore assume there are abandonments, and then let the abandonment rate scale to zero.

We assume linear holding cost. Let $W^\theta(m, d)$ be the Whittle index in the presence of abandonments (as derived in Section 6.2), with $\theta_1 = \theta_2 = \theta$, with $\theta > 0$. It is direct from Theorem 2 that if $\mu^{(1)} < \mu^{(2)}$, then

$$\lim_{\theta \to 0} \theta W^\theta(m, 2) = c\mu^{(2)}, \text{ for all } m \geq 1. \tag{29}$$

That is, in environment 2, one needs to consider the scaled index. For environment 1, no scaling is needed. In fact, we believe the following to be true:

Conjecture 1.

$$\lim_{\theta \to 0} W^\theta(m, 1) < \infty.$$

This conjecture is based on our numerical observations. In Table 1 we show the Whittle index as the abandonment rate approaches zero. We consider a multi-class queue with the following parameters: $\lambda^{(d)} = 4$, for $d = 1, 2$, and $\mu_1^{(1)} = 5, \mu_1^{(2)} = 8$, hence $\mu^{(1)} < \mu^{(2)}$. The rates for the environment are $r^{(1)} = 17$ and $r^{(2)} = 15$. We set the abandonment rate as $\theta^{(d)} = \theta$ for $d = 1, 2$. In Table 1, we show the index for different states m. It can be seen that, as θ tends to 0, the index $W^\theta(m, 1)$ seems to converge.

In order to prove this conjecture, one needs to study $\overline{W}((n_{j-1}, 0), (n_j, 0))$, which by (23) depends on the steady-state distributions. A perturbation approach as presented in [2, Theorem 2], would allow to write the steady state as an expansion in θ. However, the results of [2] do not directly apply since there the transition rates are assumed to be uniformly bounded.

We can now define a heuristic for the multi-class queue with a random environment. Let us assume there are N bandits and R can be made active. Define for each bandit k, $1 \leq k \leq N$, d_k such that $\mu^{(3-d_k)} \leq \mu^{(d_k)}$. In every decision epoch make active the bandits that are currently in their state d_k. If there are more than R, choose the R ones having the largest value for $c_k \mu^{(d_k)}$. If there are less than R, make also active the bandits currently in state $3 - d_k$ having the largest value for $\lim_{\theta \to 0} W_k^{\theta}(m, 3 - d_k)$.
We extend the previous example to numerically illustrate the performance of the heuristics. We consider a model with two classes of users $k = 1, 2$, and two states for the environment, $\mathcal{Z} = \{1, 2\}$. At each decision epoch the decision maker chooses which user to serve, that is, $R = 1$. We simulate our heuristic and compare its average cost to that of an optimal policy obtained via value iteration. More details on the models used for the simulations can be found in Section 8.2.
We choose the following parameters: the arrival rates are $\lambda_k^{(d)} = 4$, for $k = 1, 2, d = 1, 2$. The departure rates are $\mu_1^{(1)} = 5, \mu_1^{(2)} = 8, \mu_2^{(1)} = 21, \mu_2^{(2)} = 27$, hence $\mu_k^{(1)} < \mu_k^{(2)}$ for $k = 1, 2$. We consider three models for the environment: a model with independent and identically distributed environments, a model with independent environments with different distributions, and a model with common environments for both classes of users, i.e., they see the same state of the environment at each moment in time. The parameters for the environments are as follows:

- In the first case, both environments $D_1(t), D_2(t)$ are independent and identically distributed and their transition rates are given by $r_k^{(1)} = 15$ and $r_k^{(2)} = 17$, $k = 1, 2$.
- In the second case, the environments are independent and their transition rates are given by $r_1^{(1)} = 15, r_1^{(2)} = 17$, $r_2^{(1)} = 10$ and $r_2^{(2)} = 2$.
- In the third case, the environments are common with $r^{(1)} = 15$ and $r^{(2)} = 17$, $k = 1, 2$.

The obtained performances are, for each case:

- In the first case, the performance of our heuristic is $g^{HEUR} = 1.13$ and for the optimal policy is $g^{OPT} = 1.11$. In other words, our heuristic presents a suboptimality gap of 1.82%.
- In the second case, the performance of our heuristic is $g^{HEUR} = 0.8042$ and for the optimal policy is $g^{OPT} = 0.8044$. In other words, our heuristic presents a suboptimality gap of 0.02%.
- In the third case, the performance of our heuristic is $g^{HEUR} = 1.41$ and for the optimal policy is $g^{OPT} = 1.36$. In other words, our heuristic presents a suboptimality gap of 4.3%.

8 Numerical evaluation

In this section the performance under Whittle's index policy is compared to the performance of an optimal policy obtained via value iteration. We consider the abandonment model with two classes of users and two states for the environment(s), $\mathcal{Z} = \{1, 2\}$. The decision epochs occur when there is a change of state, either in the queue length or in the environment. At each decision epoch the decision maker chooses which user to serve, that is, $R = 1$. We assume linear holding costs that do not depend on the environment or action taken, i.e., $C_k(m, d, k) = m$.

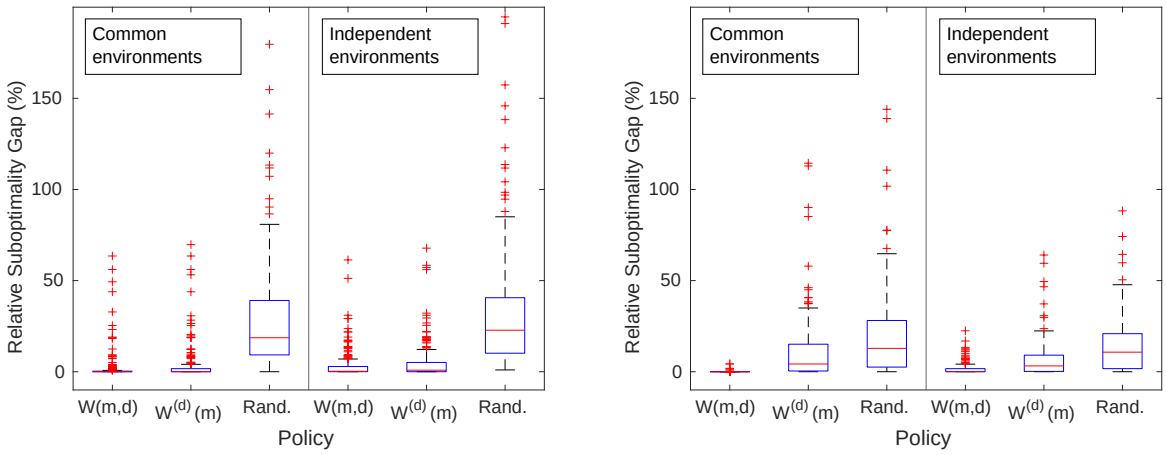


Figure 3: Suboptimality gap with (left) random parameters and (right) constrained parameters.

We plot the relative suboptimality gap in percentage for the different policies. We denote by g^{φ} the performance under policy φ. Then, the suboptimality gap for policy φ is given by $100 * (g^{\varphi} - g^{OPT})/g^{OPT}$, where g^{OPT} is the average cost under an optimal policy.
In Section 8.1, we generate a large number of parameters and using boxplots we show the suboptimality gaps under different policies. In Section 8.2 we focus on one particular set of parameters. In Section 8.3, we discuss how much one can gain by observing the state of the environments.

8.1 Boxplots

In this section we consider two different models for the environments. The first model is that of one common environment, and the second model is that of independent identically distributed environments. Numerically we calculate the suboptimality gap under Whittle's index policy $W(m,d)$, Whittle's index policy for a fixed environment $W^{(d)}(m)$, and the policy that for each set of parameters chooses uniformly at random which bandit to serve in each state (1/2 of probability of serving each class, in each state). We do this for 200 sets of randomly generated parameters in order to obtain a boxplot. This allows us to plot for each policy the 25th and 75th percentiles, the median value with an horizontal line and the outliers with "+".
Figure 3 (left) considers random sets of parameters chosen as follows: The parameters are chosen uniformly at random such that $\lambda_k^{(d)} = \lambda \in [1,\ 10]$, $\mu_k^{(d)} \in [1,\ 20]$, $r_k^{(d)} = r^{(d)} \in [1,\ 20]$, and $\theta_k^{(d)} = \theta \in [0.1,\ 0.5]$, for $k = 1, 2$ and $d = 1, 2$. We observe that the suboptimality gap of policies $W(m,d)$ and $W^{(d)}(m)$ is very small, both for independent environments as well as for a common environment. The random policy shows worse performance having a median of the suboptimality gap in 18%.
The fact that the policy $W^{(d)}(m)$ performs similar to Whittle's index policy $W(m,d)$ is surprising. The former does not take into account that the environment changes dynamically over time. To show that $W^{(d)}(m)$ is not an efficient policy, we narrow our set of randomly generated parameters. As before, we chose the parameters $\lambda_k^{(d)}, r^{(d)}$, $r_k^{(d)}$, and $\theta_k^{(d)}$. However, the departure rates are chosen differently. In environment $d = 2$, the departure rates are equal for both classes, i.e.,

$\mu_1^{(2)} = \mu_2^{(2)} = \mu$, where the value μ is chosen uniformly at random in the interval [1, 20]. In environment $d = 1$, we take $\mu_1^{(1)} = \mu - a$ and $\mu_2^{(1)} = \mu + a$, where $a \in [0, \mu]$ is chosen uniformly at random. Now, the fixed policy $W^{(d)}(m)$ will give equal priority to classes 1 and 2 in environment 2. However, since the environment will change later to state $d = 1$, it could have been better to prioritize class 1 in environment 2, because in environment 1 it will have a lower departure rate, while class 2 has a higher departure rate. This effect of changing environments is taken into account in the Whittle index policy $W(m, d)$.
Figure 3 (right) plots the result for 200 samples. As expected, Whittle's index policy has an suboptimality gap close to 0, as in the previous example, while this is not the case for Whittle's index policy for a fixed environment, $W_k^{(d)}(m)$. For the latter policy, the median is 4% and the 75th percentile is 15% for the common environment setting and the median is 3% and the 75th percentile is 9% for the independent environments setting.

8.2 Particular example

From the above boxplots, we conclude that even in a common environment, Whittle's index performs rather well. Below we show that this is not always the case. We obtain a set of parameters such that Whittle's index policy has a good performance when the environments are independent, but it performs bad when the environments are common for both bandits.
We choose the following parameters: the arrival rates are $\lambda_k^{(d)} = 4\gamma$, $\gamma > 0$, for $k = 1, 2, d = 1, 2$, and hence independent of the environment. The departure rates are $\mu_1^{(1)} = 8, \mu_1^{(2)} = 5, \mu_2^{(1)} = 27, \mu_2^{(2)} = 21$, hence $\mu_1^{(d)} < \mu_2^{(d)}$, for each environment d. The abandonment rates are $\theta_1^{(1)} = 0.1, \theta_1^{(2)} = 0.1, \theta_2^{(1)} = 0.4, \theta_2^{(2)} = 0.3$, hence $\theta_1^{(d)} < \theta_2^{(d)}$, for each environment d. These parameters satisfy the following inequalities:

$$1) \;\; W_2^{(1)} << W_1^{(1)} \;\text{ and }\; 2) \;\; W_1^{(2)} < W_2^{(1)}.$$

As such, when the environment is in state $(D_1(t), D_2(t)) = (1, 1)$, the indices (relation 1) indicate that preference is leaning towards serving class 1. This is surprising, since the departure rate for class 1, $\mu_1^{(1)} = 8$ is much smaller than the departure rate for class 2 in environment 1, $\mu_2^{(1)} = 27$. However, when $(D_1(t), D_2(t)) = (2, 1)$, then from relation 2, one sees that preference leans towards serving class 2 at its high departure rate 27. When the two environments are independent, one visits this state a positive fraction of time and hence profits from the highest departure rate for class 2. If instead the environment for both classes is common, one is never in state $(D_1(t), D_2(t)) = (2, 1)$, explaining why Whittle's index policy can have a large suboptimality gap.
We consider three cases for the environment parameters:

- in the first set, both environments $D_1(t), D_2(t)$ are identically distributed and their transition rates are given by $r_k^{(1)} = 15\beta$ and $r_k^{(2)} = 17\beta$, $k = 1, 2$, with $\beta > 0$. Thus, $\phi_k^{(1)} = 17/32$ and $\phi_k^{(2)} = 15/32$. Indicated in the plot by "i.i.d.".
- In the second set, the environments are non-identical and their transition rates are given by $r_1^{(1)} = 15\beta, r_1^{(2)} = 17\beta$, $r_2^{(1)} = 10\beta$ and $r_2^{(2)} = 2\beta$, where $\beta > 0$. Thus, $\phi_1^{(1)} = 17/32$, $\phi_1^{(2)} = 15/32$, $\phi_2^{(1)} = 1/6$ and $\phi_2^{(2)} = 5/6$. Indicated in the plot by "non-identical".

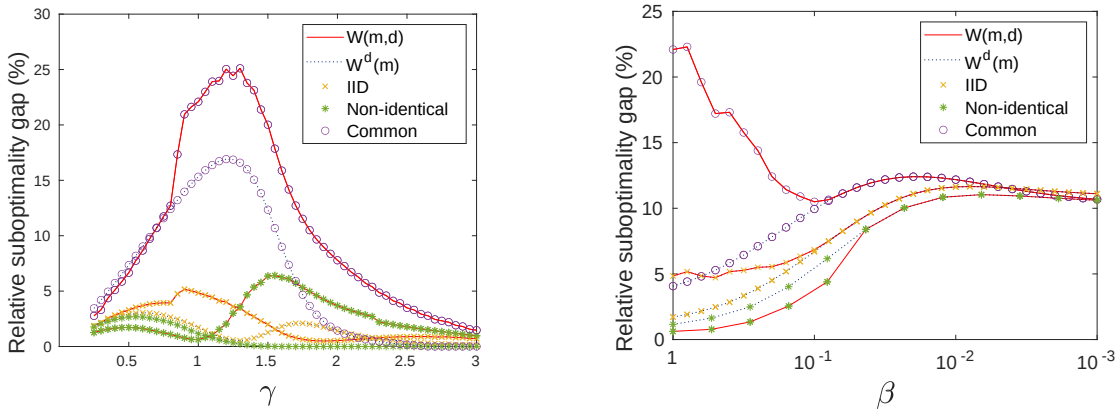


Figure 4: Suboptimality gap of Whittle's index policy and the index $W^{(d)}(m)$ as a function of (left) the scaling parameter of arrival rates (γ) and (right) the scaling parameter of the transition rates for the environment (β).

- In the third set, the environments are common with $r^{(1)} = 15\beta$ and $r^{(2)} = 17\beta$, $k = 1, 2$, with $\beta > 0$. Thus, $\phi^{(1)} = 17/32$ and $\phi^{(2)} = 15/32$. Indicated in the plot by "common".

Note that Condition 1 and Condition 2, which were needed in order to prove indexability, are not satisfied for the above parameters. However, numerically we observed that for this parameter setting, the system is indexable.

8.2.1 Scaling arrivals

We first study the suboptimality gap as the load in the system changes. We set $\beta = 1$. In Figure 4 (left) the relative suboptimality gap under Whittle's index policy (denoted by $W(m, d)$) is plotted as a function of γ. For the independent environment settings, the gap is not larger than 7%, and is 1% when in overload. However, in a common environment, the gap can be around 25%.

8.2.2 Scaling speed of the environments

We now study the suboptimality gap as the speed of the environments changes.
In Figure 4, (right) the suboptimality gap for both policies $W(m, d)$ and $W^{(d)}(m)$ is plotted as a function of the speed of the transitions of the environment (and $\gamma = 1$). For the two independent environment settings, we observe that for β around 10^{-2}, the performance under Whittle's index policy $W(m, d)$ and the policy $W^{(d)}(m)$ are very similar. Their suboptimality gap is less than 12%. On the other hand, for the common environment, the suboptimality gap under Whittle's index policy is around 23% when the transition rates of the environments are not scaled. Surprisingly, the fixed Whittle index performs rather well for any choice of β.
Recall from Proposition 1 that Whittle's index $W(m, d)$ converges as $\beta \to 0$ to $W^{(d)}(m)$. This explains why in Figure 4, (right) the performance of Whittle's index policy converges towards the fixed Whittle index policy.

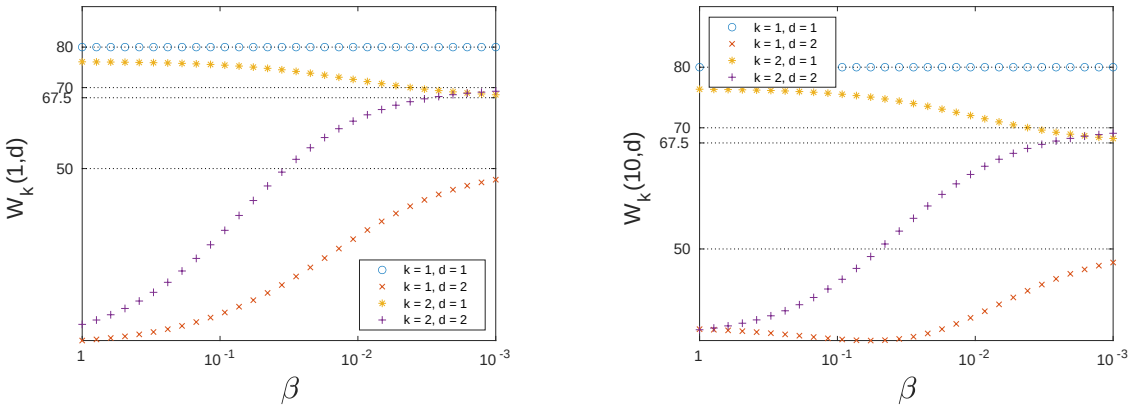


Figure 5: Whittle's index as a function of the scaling parameter of the transition rates for the environment. (left) state $m = 1$, (right) state $m = 10$. The horizontal lines represent the values $c\mu_k^{(d)}/\theta_k^{(d)}$, $d = 1, 2$ and $k = 1, 2$.

Under linear cost, $W^{(d)}(m)$ is equal to $c\mu_k^{(d)}/\theta_k^{(d)}$, for $k = 1, 2$ and $d = 1, 2$. In Figure 5, we plot Whittle's index in state $m = 1$ and $m = 10$, respectively, as well as the values $c\mu_k^{(d)}/\theta_k^{(d)}$. We see in both cases the convergence as $\beta \to 0$.

8.3 Unobservable environments

In practice, the environment might not be observable due to technical constraints. In this section, we assess the performance degradation in case information on the environment is not available.
In [14] it was shown that so-called *averaged Whittle index policy* are asymptotically optimal when the state of environments is unobservable, as the number of bandits grows large together with the speed of the environment. In particular, from [14] together with [22], one obtains that for the abandonment multi-class queue with linear cost, the averaged Whittle index $\overline{W}_k(m)$, is given by

$$\overline{W}_k(m) = c_k \frac{\overline{\theta}_k}{\overline{\mu}_k},$$

with $\overline{\theta}_k := \sum_{d \in \mathcal{Z}} \phi_k(d)\theta_k^{(d)}$ and $\overline{\mu}_k := \sum_{d \in \mathcal{Z}} \phi_k(d)\mu_k^{(d)}$.
We present here boxplots that compare the performance of the averaged Whittle index policy, $\overline{W}(m)$, to the Whittle index policy, $W(m, d)$, obtained for the observable model. We further include the policy that for each set of parameters chooses at random which bandit to serve in each state. We consider two models for the environment: (i) one common environment for both classes, (ii) independent identically distributed environments.
We first consider 200 sets of randomly generated parameters. In Figure 6 (left) the parameters are chosen uniformly at random such that $\lambda_k^{(d)} = \lambda \in [1,\ 10]$, $\mu_k^{(d)} \in [1,\ 20]$, $r_k^{(d)} = r^{(d)} \in [1,\ 20]$, and $\theta_k^{(d)} \in [0.1,\ 0.5]$, for $k = 1, 2$ and $d = 1, 2$. We observe that the suboptimality gap of policy $W(m, d)$ is very small, with a median of 0% for both the common environment and the independent environments setting. The suboptimality gaps of policy $\overline{W}(m)$ are larger, with a median of 7% for

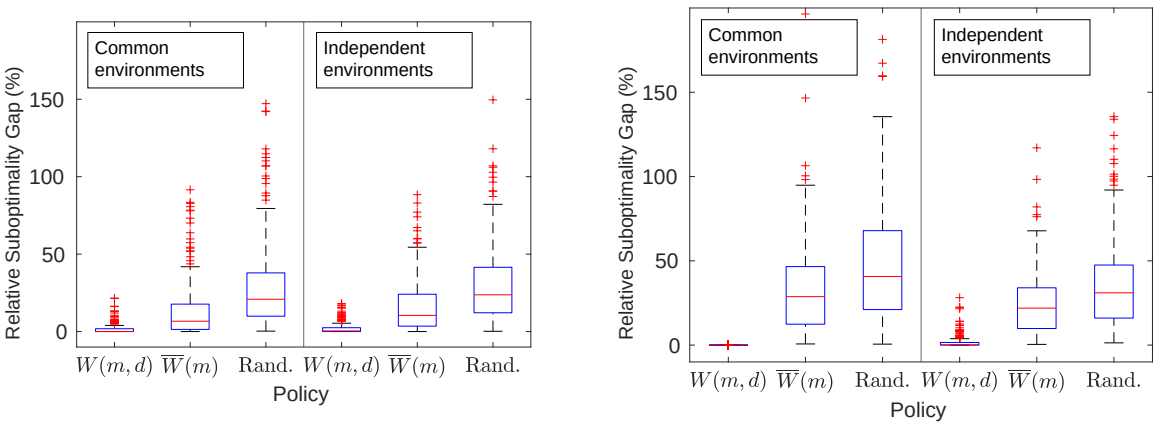


Figure 6: Suboptimality gap for unobservable policies with (left) random parameters and (right) constrained parameters.

the common environment setting and 10% for the independent environments setting. The random policy shows worse performance, where the median of the suboptimality gap is above 20% in both cases.

For Figure 6 (right), we narrow our set of randomly generated parameters. In particular, we choose parameters such that the optimal action depends on the state of the environment. The averaged index policy will not be able to mimic this, since it is a static policy. We fix the parameters such that $W_1^{(1)} >> W_2^{(1)}$ and $W_1^{(2)} << W_2^{(2)}$. Hence, in environment 1 it will be optimal to serve bandit 1 with high probability, and in environment 2 it will be optimal to serve bandit 2 with high probability. In the boxplot of Figure 6 (right), we still observe that the median of the suboptimality gap of policy $W(m,d)$ is 0% in both cases, while the median of the suboptimality gap of policy $\overline{W}(m)$ is 29% for the common environment setting and 22% for the independent environments setting.

In Figure 6, we considered that the environment has transition rates of the same order as the main process. In a separate analysis we consider environments whose transition rates are 100 times larger than the transitions rates of the controllable process. We recall that the averaged Whittle's index policy $\overline{W}(m)$ was proved to be optimal in a rapidly varying unobservable environment [14]. In the boxplots of Figure 7 the parameters are chosen as in Figure 6, except for the rates of the environment, where we take $r_k^{(d)} = r^{(d)} \in [100,\ 2000]$. We consider 200 sets of parameters for each boxplot. In this case we observe that the suboptimality gap of policy $W(m,d)$ is still small, below 1% in both boxplots and in both settings. The median of the suboptimality gap of policy $\overline{W}(m)$ for the random parameters (left) is 9% for the common environment setting, and 12% for the independent environments setting, and for the constrained parameters (right) is 24% for the common environment setting, and 20% for the independent environments setting.

We conclude that in both regimes, with normal speed and with fast speed, there is an important loss in performance if we consider an unobservable policy.

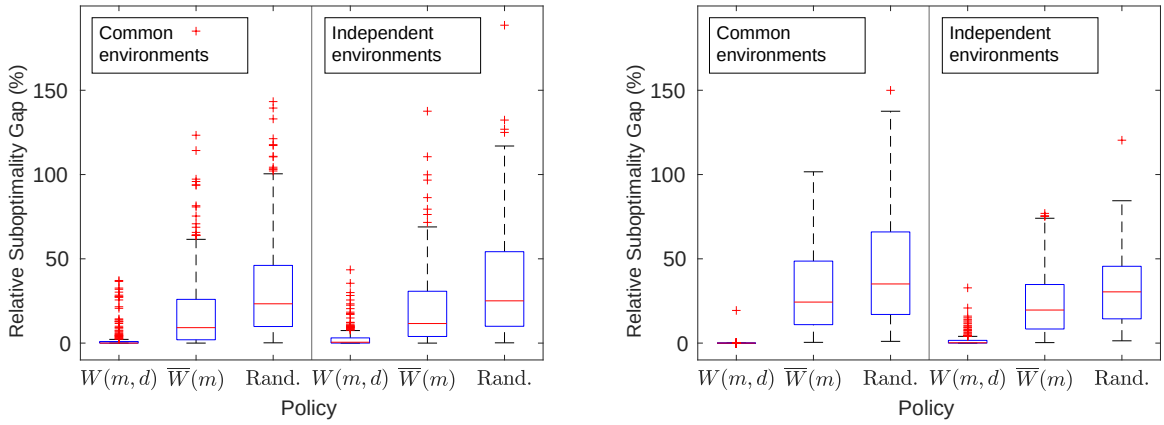


Figure 7: Suboptimality gap for unobservable policies with a fast environment, with (left) random parameters and (right) constrained parameters.

9 Conclusions and further work

In this paper we have introduced and studied a stochastic control problem with environments. The transition rates of the controllable processes depended on the state of the environments, which were assumed to be exogenous processes. Given the complexity of the problem, we focused on the approximate approach pioneered by Whittle, which is known to yield extremely good performing heuristics.

We took as a case study a multi-class queue with abandonments, for which we obtained a closed-form expression for Whittle's index. These results are restricted to having only two environment states. They crucially rely on the rate-conservation law stated in Lemma 3, which allowed us to simplify the formula for Whittle's index, see Lemma 5. We did not succeed in generalizing this approach to an arbitrary number of environment states. As a first step in future research, it would be worthwhile to consider more than two environment states, but with a specific structure for the transition of the environment state such as birth-and-death, cyclic, etc.

Acknowledgments

Research partially supported by the French Agence Nationale de la Recherche (ANR) through the project ANR-15-CE25-0004 (ANR JCJC RACON) and by the Department of Education of the Basque Government through the Consolidated Research Group MATHMODE (IT1294-19).

References

[1] S. Aalto, P. Lassila, and P. Osti. Whittle index approach to size-aware scheduling with time-varying channels. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 57–69, 2015.

[2] E. Altman, K.E. Avrachenkov, and R. Núnez-Queija. Perturbation analysis for denumerable markov chains with application to queueing models. *Advances in Applied Probability*, 36(3):839–853, 2004.

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

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

[5] A. Arapostathis, A. Das, G. Pang, and Y. Zheng. Optimal control of markov-modulated multiclass many-server queues. *Stochastic Systems*, 9(2):83–181, 2019.

[6] N. T. Argon, L. Ding, K. D. Glazebrook, and S. Ziya. Dynamic routing of customers with general delay costs in a multiserver queuing system. *Probability in the Engineering and Informational Sciences*, 23(2):175–203, 2009.

[7] S. Bhulai, A.C. Brooms, and F.M. Spieksma. On structural properties of the value function for an unbounded jump markov process with an application to a processor sharing retrial queue. *Queueing Systems*, 76(4):425–446, 2014.

[8] V.S. Borkar, G.S. Kasbekar, S. Pattathil, and P. Shetty. Opportunistic scheduling as restless bandits. *IEEE Transactions on Control of Network Systems*, 2017.

[9] V.S. Borkar and S. Pattathil. Whittle indexability in egalitarian processor sharing systems. *Annals of Operations Research*, pages 1–21, 2017.

[10] V.S. Borkar, K. Ravikumar, and K. Saboo. An index policy for dynamic pricing in cloud computing under price commitments. *Applicationes Mathematicae*, 44:215–245, 2017.

[11] R.J. Boucherie and N.M. Van Dijk. *Queueing networks: a fundamental approach*, volume 154. Springer Science & Business Media, 2010.

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

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

[14] S. Duran and I.M. Verloop. Asymptotic optimal control of markov-modulated restless bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1):7, 2018.

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

[16] J. Gittins, K. Glazebrook, and R. Weber. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, Chichester, 1989.

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

[18] K.D. Glazebrook, H.M. Mitchell, and P.S. Ansell. Index policies for the maintenance of a collection of machines by a set of repairmen. *European Journal of Operational Research*, 165(1):267–284, 2005.

[19] J. Hasenbein and D. Perry (Eds.). Special issue on queueing systems with abandonments, 2013.

[20] B. Ji, GG. R Gupta, M. Sharma, X. Lin, and N.B. Shroff. Achieving optimal throughput and near-optimal asymptotic delay performance in multichannel wireless networks with low complexity: a practical greedy scheduling policy. *IEEE/ACM Transactions on Networking*, 23(3):880–893, 2014.

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

[22] M. Larrañaga, U. Ayesta, and I.M. Verloop. Asymptotically optimal index policies for an abandonment queue with convex holding cost. *Queueing Systems*, 81(2-3):99–169, 2015.

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

[24] 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.

[25] J. Niño-Mora. Restless bandit marginal productivity indices, diminishing returns, and optimal control of make-to-order/make-to-stock M/G/1 queues. *Mathematics of Operations Research*, 31(1):50–84, 2006.

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

[27] J. Niño-Mora and S.S. Villar. Sensor scheduling for hunting elusive hiding targets via whittle's restless bandit index policy. In *International Conference on NETwork Games, Control and Optimization (NetGCooP 2011)*, pages 1–8. IEEE, 2011.

[28] M. Opp, K. Glazebrook, and V.G. Kulkarni. Outsourcing warranty repairs: Dynamic allocation. *Naval Research Logistics (NRL)*, 52(5):381–398, 2005.

[29] W. Ouyang, A. Eryilmaz, and N.B. Shroff. Asymptotically optimal downlink scheduling over markovian fading channels. In *2012 Proceedings IEEE INFOCOM*, pages 1224–1232. IEEE, 2012.

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

[31] A.L. Stolyar. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability*, 14(1):1–53, 2004.

[32] H.C. Tijms. *Stochastic modelling and analysis: a computational approach*. John Wiley & Sons, Inc., 1986.

[33] N.M. van Dijk. Approximate uniformization for continuous-time markov chains with an application to performability analysis. *Stochastic processes and their applications*, 40(2):339–357, 1992.

[34] I.M. Verloop. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *The Annals of Applied Probability*, 26(4):1947–1995, 2016.

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

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

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

10 Appendix

10.1 Proof of Lemmas of Section 5.2.

10.1.1 Proof of Lemma 1:

The measure $p^{n_d,(d)}(m)$ is the stationary distribution of the one-dimensional process with the following rates for $m \geq 0$: $q(m+1|m) = \lambda^{(d)}$, and $q((m-1)^+|m) = m\theta^{(d)} + \mathbf{1}_{(m>n_d)}\mu^{(d)}$, Let $p^{n_d,(d)}(m)$ be the stationary measure of this process, which satisfies the following balance equations:

$$\begin{aligned} &\left(\lambda^{(d)} + m\theta^{(d)} + \mathbf{1}_{(m>n_d)}\mu^{(d)}\right) p^{n_d,(d)}(m) \\ &= \mathbf{1}_{(m>0)}\lambda^{(d)} p^{n_d,(d)}(m-1) + \left[(m+1)\theta^{(d)} + \mathbf{1}_{(m+1>n_d)}\mu^{(d)}\right] p^{n_d,(d)}(m+1), \end{aligned} \tag{30}$$

for all $m \geq 0$, d.
The balance equations for $\pi^{\vec{n}}(m,d)$ are

$$\begin{aligned} &(\lambda^{(d)} + m\theta^{(d)} + \mathbf{1}_{(m>n_d)}\mu^{(d)} + \beta r^{(d)})\pi^{\vec{n}}(m,d) \\ &= \mathbf{1}_{(m>0)}\lambda^{(d)}\pi^{\vec{n}}(m-1,1) + \left[(m+1)\theta^{(d)} + \mathbf{1}_{(m+1>n_d)}\mu^{(d)}\right]\pi^{\vec{n}}(m+1,d) \\ &\quad + \sum_{d' \neq d} \beta r^{(dd')}\pi^{\vec{n}}(m,d'), \end{aligned} \tag{31}$$

for all $m \geq 0$, d. The stationary probability measure that satisfies the balance equations is unique. As $\beta \to 0$, (31) is equal to (30) with $p^{n_d,(d)}(m)$ replaced by $\lim_{\beta\to 0}\pi^{\vec{n}}(m,d)$. After normalisation, we hence have that $\lim_{\beta\to 0}\pi^{\vec{n}}(m,d) = \phi(d)p^{n_d,(d)}(m)$, for all $m \geq 0$. □

10.1.2 Proof of Lemma 2:

Recall that for a given β, $\vec{n}^j(\beta)$, $j = 0, \ldots,$ is the minimisation vector obtained in (9), and $\vec{\hat{n}}^j = \lim_{\beta\to 0}\vec{n}^j(\beta)$.
For $n \geq \hat{n}_d^j$, define

$$\begin{aligned} \hat{A}_d^j(n) &:= \sum_{m=0}^{\infty} C\,(m,d,a)\,p^{n,(d)}(m) - \sum_{m=0}^{\infty} C\,(m,d,a)\,p^{\hat{n}_d^j,(d)}(m) \\ \hat{B}_d^j(n) &:= \sum_{m=0}^{n} p^{n,(d)}(m) - \sum_{m=0}^{\hat{n}_d^j} p^{\hat{n}_d^j,(d)}(m). \end{aligned}$$

We define the function

$$f^j(\vec{n}) := \lim_{\beta\to 0} \frac{\sum_{d=1}^{Z}\sum_{m=0}^{\infty} C\,(m,d,a)\,\pi^{\vec{n}}(m,d) - \sum_{d=1}^{Z}\sum_{m=0}^{\infty} C\,(m,d,a)\,\pi^{\vec{n}^j(\beta)}(m,d)}{\sum_{d=1}^{Z}\sum_{m=0}^{n_d}\pi^{\vec{n}}(m,d) - \sum_{d=1}^{Z}\sum_{m=0}^{n_d^j(\beta)}\pi^{\vec{n}^j(\beta)}(m,d)}, \tag{32}$$

for $\vec{n} \in E_j$. By Lemma 1 and since $\vec{M}^{\vec{n}(\beta)}$ is uniform integrable, we have that $f^j(\vec{n})$ is equal to

$$\frac{\sum_{d=1}^{Z}\sum_{m=0}^{\infty} C\,(m,d,a)\,\phi(d)p^{n_d,(d)}(m) - \sum_{d=1}^{Z}\sum_{m=0}^{\infty} C\,(m,d,a)\,\phi(d)p^{\hat{n}_d^j,(d)}}{\sum_{d=1}^{Z}\sum_{m=0}^{n_d}\phi(d)p^{n_d,(d)} - \sum_{d=1}^{Z}\sum_{m=0}^{\hat{n}_d^j}\phi(d)p^{\hat{n}_d^j,(d)}}.$$

Hence, equivalently, we can write $f^j(\vec{n}) = \dfrac{\phi(1)\hat{A}_1^j(n_1) + \ldots + \phi(Z)\hat{A}_Z^j(n_Z)}{\phi(1)\hat{B}_1^j(n_1) + \ldots + \phi(Z)\hat{B}_Z^j(n_Z)}$.
By definition of $\vec{\hat{n}}^{(j+1)}$ and since $\vec{n}^{j+1}(\beta) = \vec{\hat{n}}^{j+1}$ for β small enough, $f^j(\vec{n})$ is minimised in $\vec{\hat{n}}^{j+1}$, that is,

$$f^j(\vec{\hat{n}}^{j+1}) = \inf_{\vec{n}\in E_j} f^j(\vec{n}). \tag{33}$$

In particular, this implies, $f^j(\vec{\hat{n}}^{j+1}) \leq \min(f^j(\hat{n}_1^{j+1}, \hat{n}_2^j, \ldots, \hat{n}_Z^j), \ldots, f^j(\hat{n}_1^j, \hat{n}_2^j \ldots, \hat{n}_Z^{j+1}))$, that is,

$$\frac{\phi(1)\hat{A}_1^j(\hat{n}_1^{j+1}) + \ldots + \phi(Z)\hat{A}_Z^j(\hat{n}_Z^{j+1})}{\phi(1)\hat{B}_1^j(\hat{n}_1^{j+1}) + \ldots + \phi(Z)\hat{B}_Z^j(\hat{n}_Z^{j+1})} \leq \min\left(\frac{\hat{A}_1^j(\hat{n}_1^{j+1})}{\hat{B}_1^j(\hat{n}_1^{j+1})}, \ldots, \frac{\hat{A}_Z^j(\hat{n}_Z^{j+1})}{\hat{B}_Z^j(\hat{n}_Z^{j+1})}\right). \tag{34}$$

Now, assume there is a strict inequality in (34), and assume $\dfrac{\hat{A}_1^j(\hat{n}_1^{j+1})}{\hat{B}_1^j(\hat{n}_1^{j+1})} \leq \dfrac{\hat{A}_2^j(\hat{n}_2^{j+1})}{\hat{B}_2^j(\hat{n}_2^{j+1})}, \ldots, \dfrac{\hat{A}_Z^j(\hat{n}_Z^{j+1})}{\hat{B}_Z^j(\hat{n}_Z^{j+1})}$.
The strict inequality in (34) implies

$$\begin{aligned} \hat{B}_1^j(\hat{n}_1^{j+1})\left(\phi(1)\hat{A}_1^j(\hat{n}_1^{j+1}) + \ldots + \phi(Z)\hat{A}_Z^j(\hat{n}_Z^{j+1})\right) &< \hat{A}_1^j(\hat{n}_1^{j+1})\left(\phi(1)\hat{B}_1^j(\hat{n}_1^{j+1}) + \ldots + \phi(Z)\hat{B}_Z^j(\hat{n}_Z^{j+1})\right), \\ \text{that is,} \quad \frac{\phi(2)\hat{A}_2^j(\hat{n}_2^{j+1}) + \ldots + \phi(Z)\hat{A}_Z^j(\hat{n}_Z^{j+1})}{\phi(2)\hat{B}_2^j(\hat{n}_2^{j+1}) + \ldots + \phi(Z)\hat{B}_Z^j(\hat{n}_Z^{j+1})} &< \frac{\hat{A}_1^j(\hat{n}_1^{j+1})}{\hat{B}_1^j(\hat{n}_1^{j+1})}. \end{aligned} \tag{35}$$

The LHS in (35) can be rewritten as the following convex combination,

$$\sum_{i=2}^{Z} \alpha_i \frac{\hat{A}_i^j(\hat{n}_i^{j+1})}{\hat{B}_i^j(\hat{n}_i^{j+1})},$$

with $\alpha_i = \dfrac{\phi(i)\hat{B}_i^j(\hat{n}_i^{j+1})}{\phi(2)\hat{B}_2^j(\hat{n}_2^{j+1}) + \ldots + \phi(Z)\hat{B}_Z^j(\hat{n}_Z^{j+1})}$ for $i = 2, \ldots, N$, $\alpha_i \geq 0$ (since $\vec{n} \in E_j$ and by assumption in Lemma 2), and $\sum_{i=2}^{Z} \alpha_i = 1$. Together with (35), this gives that $\sum_{i=2}^{Z} \alpha_i \dfrac{\hat{A}_i^j(\hat{n}_i^{j+1})}{\hat{B}_i^j(\hat{n}_i^{j+1})} < \dfrac{\hat{A}_1^j(\hat{n}_1^{j+1})}{\hat{B}_1^j(\hat{n}_1^{j+1})}$. The latter gives contradiction with the assumption that $\dfrac{\hat{A}_1^j(\hat{n}_1^{j+1})}{\hat{B}_1^j(\hat{n}_1^{j+1})} \leq \dfrac{\hat{A}_2^j(\hat{n}_2^{j+1})}{\hat{B}_2^j(\hat{n}_2^{j+1})}, \ldots, \dfrac{\hat{A}_Z^j(\hat{n}_Z^{j+1})}{\hat{B}_Z^j(\hat{n}_Z^{j+1})}$.
Hence, by contradiction we proved that the inequality in (34) is an equality, that is,

$$f^j(\vec{\hat{n}}^{j+1}) = \min(f^j(\hat{n}_1^{j+1}, \hat{n}_2^j, \ldots, \hat{n}_Z^j), \ldots, f^j(\hat{n}_1^j, \hat{n}_2^j \ldots, \hat{n}_Z^{j+1})). \tag{36}$$

Assume without loss of generality that $f^j(\hat{n}_1^{j+1}, \hat{n}_2^j, \ldots, \hat{n}_Z^j) \leq f^j(\hat{n}_1^j, \hat{n}_2^{j+1}, \ldots, \hat{n}_Z^j), \ldots, f^j(\hat{n}_1^j, \hat{n}_2^j \ldots, \hat{n}_Z^{j+1})$.
We are left to prove that

$$f^j(\hat{n}_1^{j+1}, \hat{n}_2^j, \ldots, \hat{n}_Z^j) = \inf_{n > \hat{n}_1^j} \frac{\sum_{m=0}^{\infty} C(m,1,a)\, p^{n,(1)}(m) - \sum_{m=0}^{\infty} C(m,1,a)\, p^{\hat{n}_1^j,(1)}(m)}{\sum_{m=0}^{n} p^{n,(1)}(m) - \sum_{m=0}^{\hat{n}_1^j} p^{\hat{n}_1^j,(1)}(m)}. \tag{37}$$

Let n_1^* be the n such that the infimum is taken on the RHS. Hence, the RHS can equivalently be written as $f^j(n_1^*, \hat{n}_2^j, \ldots, \hat{n}_Z^j)$. We prove (37) by contradiction. That is, assume $f^j(\hat{n}_1^{j+1}, \hat{n}_2^j, \ldots, \hat{n}_Z^j) > f^j(n_1^*, \hat{n}_2^j, \ldots, \hat{n}_Z^j)$. Since $f^j(\vec{\hat{n}}^{j+1}) = f^j(\hat{n}_1^{j+1}, \hat{n}_2^j, \ldots, \hat{n}_Z^j) > f^j(n_1^*, \hat{n}_2^j, \ldots, \hat{n}_Z^j)$, we have contradiction with the fact that f^j is minimised in $\vec{\hat{n}}^{j+1}$.
Combining (33), (36), and (37), we conclude the proof. □

10.2 Proof of Proposition 2:

For a given $L \gg 1$, we truncate the state space at L and we smooth the arrival transitions by

$$q^L(m+1|m,d,a) := \lambda^{(d)} \left(1 - \frac{m}{L}\right)^+, \tag{38}$$

for $m = 0, 1, \ldots, L$.

10.2.1 Convexity of V^L

We assume $1 = \lambda^{(1)} + \lambda^{(2)} + \mu^{(1)} + \mu^{(2)} + r^{(1)} + r^{(2)} + L\theta^{(1)} + L\theta^{(2)}$ for the uniformization constant, without loss of generality. For any d and $m = 0, \ldots, L$, we initialize by defining $V_0^L(m,d) = 0$ and

$$\begin{aligned}
V_{t+1}^L(m,d) = &\left(1 - \frac{m}{L}\right) \lambda^{(d)} V_t^L(\min\{m+1, L\}, d) \\
&+ r^{(d)} V_t^L(m, 3-d) + m\theta^{(d)} V_t^L((m-1)^+, d) \\
&+ \min\{-W + C(m,d,0) + \mu^{(2)} V_t^L(m,d), C(m,d,1) + \mu^{(2)} V_t^L((m-1)^+, d)\} \\
&+ \frac{m}{L} \lambda^{(d)} V_t^L(m,d) + (L-m)\theta^{(d)} V_t^L(m,d) \\
&+ \left(\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}\right) V_t^L(m,d).
\end{aligned}$$

We will prove that

$$2V_t^L(m,d) \leq V_t^L((m-1)^+, d) + V_t^L(m+1, d), \text{ for } 0 \leq m \leq L-1, \tag{39}$$

for any t, i.e. the convexity of V^L.
We first prove by induction in t that $V_t^L(m,d)$ is non-decreasing in m. Note that $V_0^L(m,d) = 0$ is non-decreasing by definition, so we assume $V_t^L(m,d)$ is non-decreasing and we prove that

$$V_{t+1}^L(m+1,d) - V_{t+1}^L(m,d) \geq 0 \text{ for } 0 \leq m \leq L-1. \tag{40}$$

We study the inequality splitting in terms according to the parameter that is multiplying. Firstly, we look to the terms multiplied by $\lambda^{(d)}$ in $V_{t+1}^L(m+1,d) - V_{t+1}^L(m,d)$, so we have

$$\begin{aligned}
&\lambda^{(d)} \left(1 - \frac{m+1}{L}\right) V_t^L(\min\{m+2, L\}, d) + \lambda^{(d)} \frac{m+1}{L} V_t^L(\min\{m+1, L\}, d) \\
&\quad - \lambda^{(d)} \left(1 - \frac{m}{L}\right) V_t^L(\min\{m+1, L\}, d) - \frac{m}{L} \lambda^{(d)} V_t^L(m,d) \\
&= \lambda^{(d)} \left(1 - \frac{m+1}{L}\right) V_t^L(\min\{m+2, L\}, d) + \frac{m}{L} \lambda^{(d)} V_t^L(\min\{m+1, L\}, d) \\
&\quad - \lambda^{(d)} \left(1 - \frac{m+1}{L}\right) V_t^L(\min\{m+1, L\}, d) - \frac{m}{L} \lambda^{(d)} V_t^L(m,d) \\
&= \lambda^{(d)} \left(1 - \frac{m+1}{L}\right) (V_t^L(\min\{m+2, L\}, d) - V_t^L(\min\{m+1, L\}, d)) \\
&\quad + \frac{m}{L} \lambda^{(d)} (V_t^L(\min\{m+1, L\}, d) - V_t^L(m,d)) \geq 0.
\end{aligned}$$

The last inequality is due to the inductive hypothesis for $V_t^L(m,d)$. Let us consider now the terms multiplied by $\theta^{(d)}$:

$$\begin{aligned}
&(m+1)\theta^{(d)}V_t^L(m,d) + (L-m-1)\theta^{(d)}V_t^L(\min\{m+1,L\},d)\\
&\quad - m\theta^{(d)}V_t^L((m-1)^+,d) - (L-m)\theta^{(d)}V_t^L(m,d)\\
&= m\theta^{(d)}V_t^L(m,d) + (L-m-1)\theta^{(d)}V_t^L(\min\{m+1,L\},d)\\
&\quad - m\theta^{(d)}V_t^L((m-1)^+,d) - (L-m-1)\theta^{(d)}V_t^L(m,d)\\
&\geq\ m\theta^{(d)}(V_t^L(m,d) - V_t^L((m-1)^+,d))\\
&\quad + (L-m-1)\theta^{(d)}(V_t^L(\min\{m+1,L\},d) - V_t^L(m,d)) \geq 0.
\end{aligned}$$

The last inequality holds because of the non-decreasing hypothesis for $V_t^L(m,d)$. For the terms multiplied by $r^{(d)}$ it is straightforward that

$$r^{(d)}V_t^L(m+1,3-d) - r^{(d)}V_t^L(m,3-d) \geq 0,$$

because of the inductive hypothesis as well. Equivalently for the dummy transition terms,

$$(\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)})\ (V_t^L(m+1,d) - V_t^L(m,d)) \geq 0.$$

Finally, we consider the minimisation terms in $V_{t+1}^L(m+1,d) - V_{t+1}^L(m,d)$, where the inequality is a consequence of the non-decreasing property of $V_t^L(m,d)$ and C, and the fact that if $A \geq B$ and $A' \geq B'$, then $\min\{A,A'\} \geq \min\{B,B'\}$:

$$\begin{aligned}
&\min\{-W + C(\min\{m+1,L\},d,0) + \mu^{(2)}V_t^L(\min\{m+1,L\},d),\\
&\qquad C(\min\{m+1,L\},d,1) + \mu^{(2)}V_t^L(m,d)\}\\
&\geq \min\{-W + C(m,d,0) + \mu^{(2)}V_t^L(m,d),\\
&\qquad C(m,d,1) + \mu^{(2)}V_t^L((m-1)^+,d)\}.
\end{aligned}$$

This concludes the proof of (40), i.e., V_t^L is non-decreasing.
We consider now equation (39). Note that for $m=0$ the equation reduces to $V_t^L(0,d) \leq V_t^L(1,d)$, which is true for every t and every d, because V_t^L is non-decreasing in m. Then for $1 \leq m \leq L-1$, we make an analogous reasoning: we prove it by induction in t and we split the inequalities according to the multiplying parameters. For the initial step $V_0^L(m) = 0$ the inequality holds, so we assume it holds for $V_t^L(m,d)$ and we study the inequality (39) for $t+1$. Note that

$$\begin{aligned}
2V_{t+1}^L(m,d) =& 2\left(1 - \frac{m}{L}\right)\lambda^{(d)}V_t^L(m+1,d) + 2\frac{m}{L}\lambda^{(d)}V_t^L(m,d)\\
&+ 2r^{(d)}V_t^L(m,3-d)\\
&+ 2m\theta^{(d)}V_t^L(m-1,d) + 2(L-m)\theta^{(d)}V_t^L(m,d)\\
&+ 2\min\{-W + C(m,d,0) + \mu^{(2)}V_t^L(m,d), C(m,d,1) + \mu^{(2)}V_t^L(m-1,d)\}\\
&+ 2\left(\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}\right)V_t^L(m,d)
\end{aligned} \tag{41}$$

The term $V_{t+1}^L(m-1,d) + V_{t+1}^L(m+1,d)$ equals

$$
\begin{aligned}
&\lambda^{(d)}\left(1-\frac{m-1}{L}\right)V_t^L(m,d) + \lambda^{(d)}\left(1-\frac{m+1}{L}\right)V_t^L(\min\{m+2,L\},d)\\
&+\lambda^{(d)}\frac{m-1}{L}V_t^L(m-1,d) + \lambda^{(d)}\frac{m+1}{L}V_t^L(m+1,d)\\
&+r^{(d)}V_t^L(m-1,3-d) + r^{(3-d)}V_t^L(m-1,d) + r^{(d)}V_t^L(m+1,3-d) + r^{(3-d)}V_t^L(m+1,d)\\
&+(m-1)\theta^{(d)}V_t^L((m-2)^+,d) + (m+1)\theta^{(d)}V_t^L(m,d)\\
&+(L-m+1)\theta^{(d)}V_t^L(m-1,d) + (L-m-1)\theta^{(d)}V_t^L(m+1,d)\\
&+\min\{-W + C(m-1,d,0) + \mu^{(2)}V_t^L(m-1,d), C(m-1,d,1) + \mu^{(2)}V_t^L((m-2)^+,d)\}\\
&+\min\{-W + C(m+1,d,0) + \mu^{(2)}V_t^L(m+1,d), C(m+1,d,1) + \mu^{(2)}V_t^L(m,d)\} \qquad (42)\\
&+\left(\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}\right)V_t^L(m-1,d)\\
&+\left(\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}\right)V_t^L(m+1,d).
\end{aligned}
$$

We first compare the two terms multiplied by $\lambda^{(d)}$ in (41) to check they are smaller than or equal to the ones multiplied by $\lambda^{(d)}$ in (42). First assume $1 \le m \le L-2$. The terms multiplied by $\lambda^{(d)}$ in (41) are

$$
\begin{aligned}
&2\left(1-\frac{m}{L}\right)V_t^L(m+1,d) + 2\frac{m}{L}V_t^L(m,d)\\
&\quad= 2\left(1-\frac{m+1}{L}\right)V_t^L(m+1,d) + 2\frac{m}{L}V_t^L(m,d) + \frac{2}{L}V_t^L(m,d)\\
&\quad\le \left(1-\frac{m+1}{L}\right)V_t^L(m,d) + \left(1-\frac{m+1}{L}\right)V_t^L(m+2,d) + 2\frac{m}{L}V_t^L(m,d) + \frac{2}{L}V_t^L(m+1,d)\\
&\quad= \left(1-\frac{m-1}{L}\right)V_t^L(m,d) - \frac{2}{L}V_t^L(m,d) + \left(1-\frac{m+1}{L}\right)V_t^L(m+2,d)\\
&\qquad\quad + 2\frac{m}{L}V_t^L(m,d) + \frac{2}{L}V_t^L(m+1,d)\\
&\quad= \left(1-\frac{m-1}{L}\right)V_t^L(m,d) + \left(1-\frac{m+1}{L}\right)V_t^L(m+2,d)\\
&\qquad\quad + 2\frac{m-1}{L}V_t^L(m,d) + \frac{2}{L}V_t^L(m+1,d) \qquad (43)
\end{aligned}
$$

Because of convexity, we know that $2\frac{m-1}{L}V_t^L(m,d) \le \frac{m-1}{L}(V_t^L(m-1,d) + V_t^L(m+1,d))$, hence

the third term in (43) can be bounded, and we obtain:

$$
\begin{aligned}
&2\left(1-\frac{m}{L}\right)V_t^L(m+1,d)+2\frac{m}{L}V_t^L(m,d)\\
&\quad\le\left(1-\frac{m-1}{L}\right)V_t^L(m,d)+\left(1-\frac{m+1}{L}\right)V_t^L(m+2,d)\\
&\qquad+\frac{m-1}{L}\left(V_t^L(m-1,d)+V_t^L(m+1,d)\right)+\frac{2}{L}V_t^L(m+1,d)\\
&\quad\le\left(1-\frac{m-1}{L}\right)V_t^L(m,d)+\left(1-\frac{m+1}{L}\right)V_t^L(m+2,d)\\
&\qquad+\frac{m-1}{L}V_t^L(m-1,d)+\frac{m+1}{L}V_t^L(m+1,d),
\end{aligned}
\tag{44}
$$

which is the same as the terms multiplied by $\lambda^{(d)}$ in (42). Now assume $m=L-1$, then inequality (39) reduces to $2(1-2/L)V_t^L(L-1,d)\le(1-2/L)(V_t^L(L-2,d)+V_t^L(L,d))$, which holds because of convexity of V_t^L.
We consider now the terms multiplied by $\theta^{(d)}$. We need to prove

$$
\begin{aligned}
&2mV_t^L(m-1,d)+2(L-m)V_t^L(m,d)\\
&\le(m-1)V_t^L((m-2)^+,d)+(m+1)V_t^L(m,d)+(L-m+1)V_t^L(m-1,d)\\
&\quad+2V_t^L(m-1,d)+(L-m-1)V_t^L(m+1,d),
\end{aligned}
$$

or, equivalently,

$$
\begin{aligned}
&2(m-1)V_t^L(m-1,d)+2(L-m-1)V_t^L(m,d)\\
&\le(m-1)V_t^L((m-2)^+,d)+(m-1)V_t^L(m,d)+(L-m-1)V_t^L(m-1,d)\\
&\quad+(L-m-1)V_t^L(m+1,d).
\end{aligned}
$$

This last inequality is obtained from the convexity property for $2V_t^L(m-1,d)$ and $2V_t^L(m,d)$ on the lhs.
For the terms multiplied by $r^{(d)}$, and the dummy transitions $(\lambda^{(3-d)}+\mu^{(3-d)}+r^{(3-d)}+L\theta^{(3-d)})$, the inequalities to prove are

$$
\begin{aligned}
&2r^{(d)}V_t^L(m,3-d)\le r^{(d)}V_t^L(m-1,3-d)+r^{(d)}V_t^L(m+1,3-d)\quad\text{ and}\\
&2(\lambda^{(3-d)}+\mu^{(3-d)}+r^{(3-d)}+L\theta^{(3-d)})V_t^L(m,d)\\
&\qquad\le(\lambda^{(3-d)}+\mu^{(3-d)}+r^{(3-d)}+L\theta^{(3-d)})\ (V_t^L(m-1,d)+V_t^L(m+1,d)),
\end{aligned}
$$

which are a direct consequence of convexity of V_t^L.
We lastly consider the minimisation terms, for which we analyse each possible combination of optimal actions in states $m-1$ and $m+1$. Since at time t, V_t^L is convex, the optimal actions satisfy the optimality of threshold policies. Denote by $a_m^*\in\{0,1\}$ the optimal action in state m, where action 0 (1) is passive (active). Then, since the threshold policy is optimal at time t, (a_{m-1}^*,a_{m+1}^*) equals $(0,0)$, $(0,1)$ or $(1,1)$. We also use Property (15), regarding the cost function. For $a^*=(0,1)$ and $1\le m\le L-1$ we have

$$
\begin{aligned}
&2\min\{-W + C(m,d,0) + \mu^{(2)}V_t^L(m,d), C(m,d,1) + \mu^{(2)}V_t^L(m-1,d)\} \\
&\quad \le -W + C(m,d,0) + \mu^{(2)}V_t^L(m,d) + C(m,d,1) + \mu^{(2)}V_t^L(m-1,d) \\
&\quad \le -W + C(m-1,d,0) + \mu^{(2)}V_t^L(m,d) + C(m+1,d,1) + \mu^{(2)}V_t^L(m-1,d) \\
&\quad = \min\{-W + C(m-1,d,0) + \mu^{(2)}V_t^L(m-1,d), C(m-1,d,1) + \mu^{(2)}V_t^L((m-2)^+,d)\} \\
&\qquad + \min\{-W + C(m+1,d,0) + \mu^{(2)}V_t^L(m+1,d), C(m+1,d,1) + \mu^{(2)}V_t^L(m,d)\},
\end{aligned}
$$

where in the last equality the value for the minimums are given by the optimal action $a^* = (0,1)$. For $a^* = (0,0)$, we use convexity of C and V_t^L:

$$
\begin{aligned}
&2\min\{-W + C(m,d,0) + \mu^{(2)}V_t^L(m,d), C(m,d,1) + \mu^{(2)}V_t^L(m-1,d)\} \\
&\quad = -2W + 2C(m,d,0) + 2\mu^{(2)}V_t^L(m,d) \\
&\quad \le -2W + C(m-1,d,0) + C(m+1,d,0) + \mu^{(2)}V_t^L(m-1,d) + \mu^{(2)}V_t^L(m+1,d) \\
&\quad = \min\{-W + C(m-1,d,0) + \mu^{(2)}V_t^L(m-1,d), C(m-1,d,1) + \mu^{(2)}V_t^L((m-2)^+,d)\} \\
&\qquad \min\{-W + C(m+1,d,0) + \mu^{(2)}V_t^L(m+1,d), C(m+1,d,1) + \mu^{(2)}V_t^L(m,d)\}.
\end{aligned}
$$

Equivalently for $a^* = (1,1)$, we make use of convexity properties:

$$
\begin{aligned}
&2\min\{-W + C(m,d,0) + \mu^{(2)}V_t^L(m,d), C(m,d,1) + \mu^{(2)}V_t^L(m-1,d)\} \\
&\quad = 2C(m,d,1) + 2\mu^{(2)}V_t^L(m-1,d) \\
&\quad \le C(m-1,d,1) + C(m+1,d,1) + \mu^{(2)}V_t^L((m-2)^+,d) + \mu^{(2)}V_t^L(m,d) \\
&\quad = \min\{-W + C(m-1,d,0) + \mu^{(2)}V_t^L(m-1,d), C(m-1,d,1) + \mu^{(2)}V_t^L((m-2)^+,d)\} \\
&\qquad \min\{-W + C(m+1,d,0) + \mu^{(2)}V_t^L(m+1,d), C(m+1,d,1) + \mu^{(2)}V_t^L(m,d)\}.
\end{aligned}
$$

This finishes the proof of (39) for $t+1$, hence function V_t^L is convex. By [30, Chapter 9.4] it follows that $V_t^L(\cdot) - tg \to V^L$ as $t \to \infty$, where g is the averaged cost incurred under the optimal policy. Hence, convexity of V_t^L implies convexity of V^L. □

10.2.2 Hypothesis needed for [7, Theorem 3.1]

In this section, we verify that $V^L \to V$ as $L \to \infty$. In particular, we verify that the sufficient conditions as stated in [7, Theorem 3.1] hold.

For ease of notation, we introduce $q^L((m',d')|(m,d),a)$ as the transition rate of the truncated process from state (m,d) to state (m',d'), when action a is applied. That is, $q^L((m+1,d)|(m,d),a) = \lambda^{(d)}\left(1 - \frac{m}{L}\right)^+$, $q^L(((m-1)^+,d)|(m,d),a) = m\theta^{(d)} + a\mu^{(d)}$, and $q^L((m,3-d)|(m,d),a) = r^{(d)}$, for $m \in \mathbb{N}_0, d = 1,2$ and $a = 0,1$.

In order to state the sufficient conditions of [7, Theorem 3.1], we need the following definition.

Definition 4. *A function $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}_+$ is a moment function if there exists an increasing sequence of finite sets $(E_n)_{n\in\mathbb{N}} \subset \mathcal{X}\times\mathcal{Z}$ such that $\lim_{n\to\infty} E_n = \mathcal{X}\times\mathcal{Z}$ and $\inf\{f(m,d) : (m,d) \notin E_n\} \to \infty$ as $n \to \infty$.*

The conditions, as stated in [7, Theorem 3.1], are:

1. There exists a function $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}_+$, constants $\alpha, \beta > 0$ and $M > 0$ such that

$$\sum_{(m',d') \in \mathcal{X} \times \mathcal{Z}} q^L((m',d')|(m,d),a)\ f(m',d') \leq -\alpha f(m,d) + \beta \mathbf{1}_{\{m<M\}}(m,d), \text{ for all } m,d,\varphi,L.$$

2. $(a,L) \to q^L((m',d')|(m,d),a)$ and $(a,L) \to \sum_{(m',d') \in \mathcal{X} \times \mathcal{Z}} q^L((m',d')|(m,d),a) f(m',d')$ are continuous functions in a and L for all (m,d) and (m',d').

We take the function $f(m,d) = e^{\epsilon m}$ with $\epsilon > 0$. We define the sets $E_n = \{(0,1),(0,2),\ldots,(n,1),(n,2)\}$ for each n, which are finite, $\lim_{n\to\infty} E_n = \mathcal{X} \times \mathcal{Z}$ and $\inf\{f(m,d) : (m,d) \notin E_n\} \to \infty$ as $n \to \infty$. Condition 1 can be reduced to show that there exists $\epsilon > 0$, $M > 0$ and a constant $\alpha > 0$ such that

$$\sum_{(m',d') \in \mathcal{X} \times \mathcal{Z}} q^L((m',d')|(m,d),a)\ f(m',d') \leq -\alpha f(m,d),$$

for $d = 1,2$ and $m > M$, that is,

$$\lambda^{(d)} \left(1 - \frac{m}{L}\right)^+ e^{\epsilon(m+1)} + (m\theta^{(d)} + a\mu^{(d)}) e^{\epsilon(m-1)} + r^{(d)} e^{\epsilon m}$$
$$\left(\lambda^{(d)} \left(1 - \frac{m}{L}\right)^+ + m\theta^{(d)} + a\mu^{(d)} + r^{(d)}\right) e^{\epsilon m} \leq -\alpha e^{\epsilon m},$$

for $d = 1,2$ and $m > M$, where a is the action taken in state (m,d) under policy φ. The inequality can be rewritten as

$$\lambda^{(d)} \left(1 - \frac{m}{L}\right)^+ (e^{\epsilon} - 1) + (m\theta^{(d)} + a\mu^{(d)})(e^{-\epsilon} - 1) \leq -\alpha, \tag{45}$$

Note that, since $\left(1 - \frac{m}{L}\right)^+$ is decreasing in m, there exists a constant C such that $\lambda^{(d)} \left(1 - \frac{m}{L}\right)^+ (e^{\epsilon} - 1) < C$. For the other term, since $(e^{-\epsilon} - 1) < 0$, there exists an M such that for $m > M$, $(m\theta^{(d)} + a\mu^{(d)})(e^{-\epsilon} - 1) < -C$. Then there exists $\alpha > 0$ such that inequality (45) holds for $d = 1,2$ and $m > M$ and Condition 1 is proved. For Condition 2, the continuity of the functions $(a,L) \to q^L((m',d')|(m,d),a)$ and $(a,L) \to \sum_{(m',d') \in \mathcal{X} \times \mathcal{Z}} q^L((m',d')|(m,d),a) f(m',d')$ in a and L holds by definition of the transition rates. □

10.3 Proof of Lemma 3

Since each class-k job abandons with a rate $\theta_k^{(d)} > 0$ when in environment d, it is certain that the system is stable and existence of the steady-state distribution is guaranteed. Stability implies that the long-run number of job arrivals will be the same as the long-run number of departures. The same applies to a subsystem corresponding to one state of the environment. That is, for environment d, we consider the process $M^{\vec{n}}(t)\mathbf{1}_{(D(t)=d)}$ and refer to it as the process of the subsystem d. Hence, when the environment is in state $D(t) = d$, the subsystem d has all the jobs, and the subsystem in environment $3 - d$ is empty. Accordingly, when the environment changes its state, all jobs "move" from one subsystem to the other subsystem.

Since the Markov chain is unichain, stability also implies that the long-run number of arrivals and departures in each of these subsystems will be equal. This is precisely what Equation (17) captures. To see this, we note that the long-run number of arrivals of new customers to subsystem d is $\lambda^{(d)}\phi(d)$. Arrivals to subsystem d might also come from subsystem $3-d$ when the environment changes from state $3-d$ to state d. This happens with rate $r^{(3-d)}$ and gives an average increase in the number of jobs in subsystem d by $\mathbb{E}\left(M^{\vec{n}}\mathbf{1}_{(D=3-d)}\right)$.
Regarding the departures, the long-run number of departures due to service completions in subsystem d equals $\mu^{(d)}\sum_{m=n_d+1}^{\infty}\pi^{\vec{n}}(m,d)$. The long-run number of departures due to abandonments in subsystem d equals $\theta^{(d)}\mathbb{E}\left(M^{\vec{n}}\mathbf{1}_{(D=d)}\right)$. All customers leave subsystem d in a batch in case the environment changes from state d to state $3-d$. This happens with rate $r^{(d)}$ and gives an average decrease of $\mathbb{E}\left(M^{\vec{n}}\mathbf{1}_{(D=d)}\right)$. Equation (17) now follows by equating the long-run number of arrivals and the long-run number of departures. □

10.4 Proof of Lemma 4

10.4.1 Proof of Property 1) in Lemma 4:

Without loss of generality, we assume that the increasing term is n_1. Then, for a given (n_1, n_2), we will show that

$$\sum_{m=0}^{n_1}\pi^{(n_1,n_2)}(m,1) \leq \sum_{m=0}^{n_1+1}\pi^{(n_1+1,n_2)}(m,1).$$

The idea of the proof relies on using the comparison result 9.3.2 in [11, Chapter 9], for both processes given by the threshold policies (n_1, n_2) and (n_1+1, n_2). For that, we define the following cost function:

$$C^{\vec{n}}(m,d,a) := \begin{cases} 1 & \text{if } d=1 \text{ and } m \leq n_1 \\ 0 & \text{otherwise.} \end{cases}$$

In other words, $C^{\vec{n}}(m,d,a)=1$ if and only if the state of the environment is 1 and the process is in a passive state for policy $\vec{n}$. We also define the resulting expected reward per unit time,

$$G^{\vec{n}} := \sum_{m=0}^{n_1}\pi^{\vec{n}}(m,1).$$

The remainder of the proof consists in showing

$$G^{(n_1,n_2)} \leq G^{(n_1+1,n_2)}. \tag{46}$$

Since the result 9.3.2 in [11, Chapter 9] requires a uniformizable process and our abandonment rates grow linearly in n in an infinite state space, we consider a truncated version. Let L be the limited capacity for the truncated version of the process, with $L > \max\{n_1+1, n_2\}$. We denote by $q^{\vec{n},L}((m',d')|(m,d),a)$ the transition rate of the process from state (m,d) to state (m',d'), when action a is applied, where action a is determined by the threshold policy $\vec{n}$.
We introduce the uniformization constant $H(L) := \max_d\left\{\lambda^{(d)}+\mu^{(d)}+r^{(d)}+L\theta^{(d)}\right\}$ and the transition probabilities $P^{\vec{n},L}\left((m',d')|(m,d),a\right)$ obtained after the standard uniformization approach

[32, P.110] for each step of length $H(L)$. Let $V_t^{\vec{n},L}(m,d)$ denote the expected cumulative cost over t steps under threshold policy $\vec{n}$ when starting in state (m,d). Then $V_t^{\vec{n},L}$ satisfies the relation

$$V_{t+1}^{\vec{n},L}(m,d) = \frac{C^{\vec{n}}(m,d,\mathbf{1}_{m>n_d})}{H(L)} + \sum_{(m',d')} P^{\vec{n},L}\left((m',d')|(m,d),\mathbf{1}_{m>n_d}\right) V_t^{\vec{n},L}(m',d').$$

Let $G_L^{\vec{n}}$ denote the expected reward for the truncated processes, and, using result 9.3.2 in [11, Chapter 9], we will prove that

$$G^{(n_1,n_2),L} \leq G^{(n_1+1,n_2),L}. \tag{47}$$

In order to apply result 9.3.2 in [11, Chapter 9], we need to prove that for all states (m,d) and $t \geq 0$,

$$\begin{aligned} & C^{(n_1+1,n_2)}(m,d,a) - C^{(n_1,n_2)}(m,d,a) \\ & + \sum_{(m',d')} \left[q^{(n_1+1,n_2),L}((m',d')|(m,d),a) - q^{(n_1,n_2),L}((m',d')|(m,d),a)\right] \\ & \qquad \cdot \left[V_t^{(n_1,n_2),L}(m',d') - V_t^{(n_1,n_2),L}(m,d)\right] \\ & \geq 0. \end{aligned} \tag{48}$$

Note that for $(m,d) \neq (n_1+1,1)$, $C^{(n_1+1,n_2)}(m,d,a) = C^{(n_1,n_2)}(m,d,a)$, and $q^{(n_1+1,n_2),L}((m',d')|(m,d),a) = q^{(n_1,n_2),L}((m',d')|(m,d),a)$, for any (m',d'). Thus the inequality holds directly. It remains to check the state $(m,d) = (n_1+1,1)$. The only difference in rates between $q^{(n_1+1,n_2),L}$ and $q^{(n_1,n_2),L}$ is the transition to state $(n_1,1)$. Hence, inequality (48) simplifies to

$$1 - \mu^{(1)}\left[V_t^{(n_1,n_2),L}(n_1,1) - V_t^{(n_1,n_2),L}(n_1+1,1)\right] \geq 0,$$

or equivalently,

$$V_t^{(n_1,n_2),L}(n_1,1) - V_t^{(n_1,n_2),L}(n_1+1,1) \leq \frac{1}{\mu^{(1)}}. \tag{49}$$

By induction, we can prove the following more general result. For ease of reading, its proof appears in Appendix 10.4.2.

Lemma 7.

$$V_t^{(n_1,n_2),L}(m,d) - V_t^{(n_1,n_2),L}(m+1,d) \leq \frac{1}{\mu^{(1)}}, \qquad \forall\ 0 \leq m \leq L-1, \quad d = 1,2.$$

With this Lemma, the proof for the truncated processes is done and (47) holds. To generalize this for the original processes with unbounded rates we use the following result from [33, Theorem 3.1]. There exist constants K_1, K_2 such that

$$\begin{aligned} \left|G^{(n_1,n_2),L} - G^{(n_1,n_2)}\right| &\leq \frac{K_1}{H(L)}, \\ \left|G^{(n_1+1,n_2),L} - G^{(n_1+1,n_2)}\right| &\leq \frac{K_2}{H(L)}. \end{aligned}$$

As a consequence and after (47), we get the following relation:

$$G^{(n_1,n_2)} \leq G^{(n_1,n_2),L} + \frac{K_1}{H(L)} \leq G^{(n_1+1,n_2),L} + \frac{K_1}{H(L)} \leq G^{(n_1+1,n_2)} + \frac{K_2}{H(L)} + \frac{K_1}{H(L)}.$$

Since this holds for every L, and $H(L) \to \infty$ when $L \to \infty$, we conclude (46) and the proof is done. □

10.4.2 Proof of Lemma 7.

To simplify notation, since (n_1, n_2) and L are fixed in the lemma, we will write V_t for $V_t^{(n_1,n_2),L}$ and H for $H(L)$. The inequality to prove is

$$V_t(m,d) - V_t(m+1,d) \leq \frac{1}{\mu^{(1)}}, \qquad \forall\ 0 \leq m \leq L-1, \quad d = 1,2.$$

We initialize with $k = 0$, $V_0(m,d) = 0$ for every (m,d), and for $k = 1$,

$$V_1(m,d) = \begin{cases} 1/H & \text{if } d = 1 \text{ and } m \leq n_1 \\ 0 & \text{otherwise.} \end{cases}$$

As a consequence, $\sup_{(m,d)} |V_1(m,d) - V_1(m+1,d)| = \dfrac{1}{H} \leq \dfrac{1}{\mu^{(1)}}$.

We assume now $V_t(m,d) - V_t(m+1,d) \leq \dfrac{1}{\mu^{(1)}}$ for every (m,d), and we prove it for $V_{t+1}(m,d) - V_{t+1}(m+1,d)$. We begin with the state $(n_1, 1)$, where we have

$$\begin{aligned} V_{t+1}(n_1,1) &= \frac{1}{H} + \frac{1}{H}\Big[n_1\theta^{(1)} V_t(n_1-1,1) + r^{(1)} V_t(n_1,2) + \lambda^{(1)} V_t(n_1+1,1) \\ &\qquad + \left(H - n_1\theta^{(1)} - r^{(1)} - \lambda^{(1)}\right) V_t(n_1,1)\Big], \\ V_{t+1}(n_1+1,1) &= \frac{1}{H}\Big[\left((n_1+1)\theta^{(1)} + \mu^{(1)}\right) V_t(n_1,1) + r^{(1)} V_t(n_1+1,2) + \lambda^{(1)} V_t(n_1+2,1) \\ &\qquad + \left(H - (n_1+1)\theta^{(1)} - \mu^{(1)} - r^{(1)} - \lambda^{(1)}\right) V_t(n_1+1,1)\Big]. \end{aligned}$$

Then, the following equation holds, and we apply the inductive hypothesis

$$
\begin{aligned}
&V_{t+1}(n_1,1) - V_{t+1}(n_1+1,1) \\
&= \frac{1}{H} + \frac{1}{H}\Bigg[n_1\theta^{(1)}\left(V_t(n_1-1,1) - V_t(n_1,1)\right) + r^{(1)}\left(V_t(n_1,2) - V_t(n_1+1,2)\right) \\
&\quad + \lambda^{(1)}\left(V_t(n_1+1,1) - V_t(n_1+2,1)\right) - \left(\theta^{(1)} + \mu^{(1)}\right)\left(V_t(n_1,1) - V_t(n_1+1,1)\right) \\
&\quad + \left(H - n_1\theta^{(1)} - r^{(1)} - \lambda^{(1)}\right)\left(V_t(n_1,1) - V_t(n_1+1,1)\right)\Bigg] \\
&= \frac{1}{H} + \frac{1}{H}\Bigg[n_1\theta^{(1)}\left(V_t(n_1-1,1) - V_t(n_1,1)\right) + r^{(1)}\left(V_t(n_1,2) - V_t(n_1+1,2)\right) \\
&\quad + \lambda^{(1)}\left(V_t(n_1+1,1) - V_t(n_1+2,1)\right) \\
&\quad + \left(H - n_1\theta^{(1)} - \theta^{(1)} - \mu^{(1)} - r^{(1)} - \lambda^{(1)}\right)\left(V_t(n_1,1) - V_t(n_1+1,1)\right)\Bigg] \\
&\leq \frac{1}{H} + \frac{1}{H}\Bigg[n_1\theta^{(1)}\frac{1}{\mu^{(1)}} + r^{(1)}\frac{1}{\mu^{(1)}} + \lambda^{(1)}\frac{1}{\mu^{(1)}} \\
&\quad + \left(H - n_1\theta^{(1)} - \theta^{(1)} - \mu^{(1)} - r^{(1)} - \lambda^{(1)}\right)\frac{1}{\mu^{(1)}}\Bigg] \\
&= \frac{H-\theta^{(1)}}{H\mu^{(1)}} \leq \frac{1}{\mu^{(1)}}.
\end{aligned}
$$

For states $(m,d) \neq (n_1,1)$, $C^{\vec{n}}(m,d,a) = C^{\vec{n}}(m+1,d,a)$, which makes the inequality easier to check. Again, using similar algebraic calculations, we can conclude that $V_{t+1}(m,d) - V_{t+1}(m+1,d) \leq 1/\mu^{(1)}$ and the Lemma is proved. □

10.4.3 Proof of Property 2) in Lemma 4:

We compare again the processes (n_1,n_2) with (n_1+1,n_2). In this case we have to prove

$$\sum_{m=0}^{n_2} \pi^{(n_1,n_2)}(m,2) \geq \sum_{m=0}^{n_2} \pi^{(n_1+1,n_2)}(m,2).$$

We denote by $M^{(n_1,n_2)}(t)$ and $M^{(n_1+1,n_2)}(t)$ the controllable processes of the bandit, under policies (n_1,n_2) and (n_1+1,n_2) respectively, with initial state given by the stationary measure. Note that their distribution is given by $\pi^{\vec{n}}$: for a given $m \in \mathcal{X}$,

$$\mathbb{P}(M^{\vec{n}}(t) = m) = \pi^{\vec{n}}(m,1) + \pi^{\vec{n}}(m,2).$$

A simple coupling argument shows that $M^{(n_1,n_2)}(t) \leq_{st} M^{(n_1+1,n_2)}(t)$, since they have the same rates in every state except for $(n_1+1,1)$, where $M^{(n_1+1,n_2)}(t)$ does not serve and, as a consequence, it has a lower death rate. Furthermore, the previous statement is true when looking to the second

environment, i.e., $M^{(n_1,n_2)}(t)\mathbf{1}_{(D(t)=2)} \leq_{st} M^{(n_1+1,n_2)}(t)\mathbf{1}_{(D(t)=2)}$. This inequality implies that for any $n \in \mathcal{X}$,

$$\begin{aligned}\sum_{m=0}^{n} \pi^{(n_1,n_2)}(m,2) &= \mathbb{P}\left(M^{(n_1,n_2)}(t) \leq n, D(t)=2\right) \\ &\geq \mathbb{P}\left(M^{(n_1+1,n_2)}(t) \leq n, D(t)=2\right) \\ &= \sum_{m=0}^{n} \pi^{(n_1+1,n_2)}(m,2).\end{aligned}$$

If we take $n = n_2$, the Proposition is proved. □

10.5 Proof of Proposition 3:

Slow regime. We assume without loss of generality $\frac{\mu^{(1)}}{\theta^{(1)}+r^{(1)}+r^{(2)}} \leq \frac{\mu^{(2)}}{\theta^{(2)}+r^{(1)}+r^{(2)}}$. Note that $\lim_{\beta\to 0} W^{(d)} = c\frac{\mu^{(d)}}{\theta^{(d)}}$ for both $d = 1, 2$. Together with Theorem 2, it is direct that $\lim_{\beta\to 0} W(m,2) = c\frac{\mu^{(2)}}{\theta^{(2)}}$ for $m \geq 0$.

For $d = 1$, recall from the definition of the crossing points n_j that the linear functions $g^{(n_{j-1},0)}$ and $g^{(n_j,0)}$ are not parallel. Thus, Lemma 5 can be applied, and we obtain that $\overline{W}((n_{j-1},0),(n_j,0))$ is written as a function of $s_1((n_{j-1},0),(n_j,0))$ and $s_2((n_{j-1},0),(n_j,0))$. For any $j \geq 0$, $s_2((n_{j-1},0),(n_j,0)) = \pi^{(n_{j-1},0)}(0,2) - \pi^{(n_j,0)}(0,2)$. From Lemma 1 we can deduce that for any $d \in \mathcal{Z}$ and any pair of policies $\vec{n}, \vec{n}'$ such that $n_d = n'_d$,

$$\lim_{\beta\to 0} \pi^{\vec{n}}(m,d) = \lim_{\beta\to 0} \pi^{\vec{n}'}(m,d) \quad \forall m \geq 0. \tag{50}$$

In particular,

$$\lim_{\beta\to 0} s_2((n_{j-1},0),(n_j,0)) = 0.$$

Then, following from (23), $\lim_{\beta\to 0} \overline{W}((n_{j-1},0),(n_j,0)) = W^{(1)} = c\frac{\mu^{(1)}}{\theta^{(1)}}$, which concludes the proof for the slow regime.

Fast regime. Since $\phi(d) = \frac{r^{(3-d)}}{r^{(1)}+r^{(2)}}$ and $\overline{\theta} = \sum_{d=1}^{2} \phi(d)\theta^{(d)}$, we can write

$$\begin{aligned}\lim_{\beta\to\infty} W^{(d)} &= \lim_{\beta\to\infty} c\mu^{(d)} \frac{\beta(r^{(1)}+r^{(2)}) + \theta^{(3-d)}}{\beta(r^{(1)}\theta^{(2)} + r^{(2)}\theta^{(1)}) + \theta^{(1)}\theta^{(2)}} \\ &= \lim_{\beta\to\infty} c\mu^{(d)} \frac{1 + \frac{\theta^{(3-d)}}{\beta(r^{(1)}+r^{(2)})}}{\frac{\beta(r^{(1)}\theta^{(2)}+r^{(2)}\theta^{(1)})}{\beta(r^{(1)}+r^{(2)})} + \frac{\theta^{(1)}\theta^{(2)}}{\beta(r^{(1)}+r^{(2)})}} \\ &= \lim_{\beta\to\infty} c\mu^{(d)} \frac{1 + \frac{\theta^{(3-d)}}{\beta(r^{(1)}+r^{(2)})}}{(\phi(2)\theta^{(2)} + \phi(1)\theta^{(1)}) + \frac{\theta^{(1)}\theta^{(2)}}{\beta(r^{(1)}+r^{(2)})}} \\ &= c\frac{\mu^{(d)}}{\overline{\theta}},\end{aligned}$$

which gives the value of the index for $d = 2$. □

10.6 Proof of results in Section 6.3.

10.6.1 Proof of Lemma 5: expression for $\overline{W}(\vec{n}, \vec{n}')$.

Since $\sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) + \sum_{m=n_d+1}^{\infty} \pi^{\vec{n}}(m,d) = \phi(d)$, Lemma 3 for $\vec{n}$ states

$$\lambda^{(d)}\phi(d) + r^{(3-d)} \sum_{m=0}^{\infty} m\pi^{\vec{n}}(m,3-d) \quad = \quad (\theta^{(d)} + r^{(d)}) \sum_{m=0}^{\infty} m\pi^{\vec{n}}(m,d) + \mu^{(d)}\phi(d) - \mu^{(d)} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d),$$

or, equivalently,

$$\sum_{m=0}^{\infty} m\pi^{\vec{n}}(m,d) \quad = \quad \left(\lambda^{(d)}\phi(d) + r^{(3-d)} \sum_{m=0}^{\infty} m\pi^{\vec{n}}(m,3-d) - \mu^{(d)}\phi(d) + \mu^{(d)} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) \right) \frac{1}{\theta^{(d)} + r^{(d)}},$$

for $d = 1, 2$. After some algebra, we can solve these two equations in $\sum_{m=0}^{\infty} m\pi^{\vec{n}}(m,1)$ and $\sum_{m=0}^{\infty} m\pi^{\vec{n}}(m,2)$, and then we obtain

$$\begin{aligned} \sum_{m=0}^{\infty} m\pi^{\vec{n}}(m,d) \quad = \quad & \left(\left(\lambda^{(d)} - \mu^{(d)}\right) \phi(d) \left(\theta^{(3-d)} + r^{(3-d)}\right) + \left(\lambda_{3-d} - \mu^{(3-d)}\right) \phi(3-d) r^{(3-d)} \right. \\ & \left. + \mu^{(3-d)} r^{(3-d)} \sum_{m=0}^{n_{3-d}} \pi^{\vec{n}}(m,3-d) + \mu^{(d)} \left(\theta^{(3-d)} + r^{(3-d)}\right) \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) \right) \\ & \cdot \frac{1}{\theta^{(1)}\theta^{(2)} + \theta^{(2)} r^{(1)} + \theta^{(1)} r^{(2)}}, \end{aligned} \tag{51}$$

for $d = 1, 2$. In the numerator in (22), the terms can be regrouped per environment, i.e.

$$\sum_{d=1}^{2} c \left(\sum_{m=0}^{\infty} m\pi^{\vec{n}}(m,d) - \sum_{m=0}^{\infty} m\pi^{\vec{n}'}(m,d) \right). \tag{52}$$

From (51), and together with the notation $s_i(\vec{n}, \vec{n}') := \sum_{m=0}^{n_i} \pi^{\vec{n}}(m,i) - \sum_{m=0}^{n_i'} \pi^{\vec{n}'}(m,i)$, we obtain

$$\sum_{m=0}^{\infty} m\pi^{\vec{n}}(m,d) - \sum_{m=0}^{\infty} m\pi^{\vec{n}'}(m,d) = \frac{\mu^{(d)} \left(\theta^{(3-d)} + r^{(3-d)}\right) s_d(\vec{n}, \vec{n}') + \mu^{(3-d)} r^{(3-d)} s_{3-d}(\vec{n}, \vec{n}')}{\theta^{(1)}\theta^{(2)} + \theta^{(2)} r^{(1)} + \theta^{(1)} r^{(2)}}. \tag{53}$$

Summing we obtain the following expression for (52):

$$\begin{aligned} & \sum_{d=1}^{2} c \left(\sum_{m=0}^{\infty} m\pi^{\vec{n}}(m,d) - \sum_{m=0}^{\infty} m\pi^{\vec{n}'}(m,d) \right) \\ & \qquad = c \cdot \frac{\mu^{(1)} \left(\theta^{(2)} + r^{(1)} + r^{(2)}\right) s_1(\vec{n}, \vec{n}') + \mu^{(2)} \left(\theta^{(1)} + r^{(1)} + r^{(2)}\right) s_2(\vec{n}, \vec{n}')}{\theta^{(1)}\theta^{(2)} + r^{(1)}\theta^{(2)} + r^{(2)}\theta^{(1)}} \\ & \qquad = s_1(\vec{n}, \vec{n}') W^{(1)} + s_2(\vec{n}, \vec{n}') W^{(2)}. \end{aligned}$$

Since the denominator in (22) is $s_1(\vec{n}, \vec{n}') + s_2(\vec{n}, \vec{n}')$, Lemma 5 is proved. □

10.6.2 Proof of Lemma 6: comparison to policy (∞, ∞).

Note that $\sum_{m=0}^{n_i} \pi^{\vec{n}}(m,i) + \sum_{m=n_i+1}^{\infty} \pi^{\vec{n}}(m,i) = \phi(i)$. Hence, if $n_i < \infty$, then $\sum_{m=n_i+1}^{\infty} \pi^{\vec{n}}(m,i) > 0$ and hence $\sum_{m=0}^{n_i} \pi^{\vec{n}}(m,i) < \phi(i)$. From this, it follows that $s_i((\infty,\infty),\vec{n}) = \sum_{m=0}^{\infty} \pi^{(\infty,\infty)}(m,i) - \sum_{m=0}^{n_i} \pi^{\vec{n}}(m,i) = \phi(i) - \sum_{m=0}^{n_i} \pi^{\vec{n}}(m,i) > 0$. As a consequence, if $\vec{n} \neq (\infty,\infty)$, $s_1((\infty,\infty),\vec{n}) + s_2((\infty,\infty),\vec{n}) > 0$, and Lemma 5 can be applied with $0 \leq t \leq 1$. From (23), It follows that $\overline{W}((\infty,\infty),\vec{n}) \in [W^{(1)}, W^{(2)}]$.
In case $n_1 < \infty$, $\sum_{m=0}^{n_1} \pi^{\vec{n}}(m,1) < \phi(1)$, so $s_1((\infty,\infty),\vec{n}) = \phi(1) - \sum_{m=0}^{n_1} \pi^{\vec{n}}(m,1) > 0$ and $t > 0$. Equation (23) can be rewritten as $W^{(2)} - t(W^{(2)} - W^{(1)})$, hence $\overline{W}((\infty,\infty),\vec{n}) < W^{(2)}$ if $n_1 < \infty$. If $n_1 = \infty$, $s_1((\infty,\infty),\vec{n}) = \sum_{m=0}^{\infty} \pi^{(\infty,\infty)}(m,1) - \sum_{m=0}^{\infty} \pi^{(\infty,n_2)}(m,1) = \phi(1) - \phi(1) = 0$, and $t = 0$. In this case, $\overline{W}((\infty,\infty),\vec{n}) = W^{(2)}$.
The reasoning for $n_2 = \infty$ is analogous. □

10.6.3 Proof of Proposition 4

In order to prove Point 1, we start with $W = 0$. The difference between policies $(-1,-1), (-1,0), (0,-1)$ and $(0,0)$ relies on serving or not a queue when it's empty. As a consequence, they have no difference in their dynamics (thus, in their invariant distribution $\pi^{\vec{n}}$), neither in their expected cost $\sum_{d=1}^{2} \sum_{m=0}^{\infty} cm\pi^{\vec{n}}(m,d)$. From the definition of $g^{\vec{n}}(0)$ in Equation (7) we obtain that $g^{(-1,-1)}(0) = g^{(-1,0)}(0) = g^{(0,-1)}(0) = g^{(0,0)}(0)$. Furthermore, if either $n_1 > 0$, or $n_2 > 0$, or both of them are larger than 0, then the expected costs satisfy $\sum_{d=1}^{2} \sum_{m=0}^{\infty} cm\pi^{(0,0)}(m,d) < \sum_{d=1}^{2} \sum_{m=0}^{\infty} cm\pi^{\vec{n}}(m,d)$, because the former is active in every non-zero state. Hence, $g^{(0,0)}(0) < g^{\vec{n}}(0)$ for any $\vec{n} \notin \{(-1,-1),(-1,0),(0,-1),(0,0)\}$.
For $W < 0$, the proof follows considering the slope of the linear functions $g^{\vec{n}}(W)$, which is given by $-\sum_{d=1}^{2} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d)$. For policy $(-1,-1)$ the slope is 0, while $-\sum_{d=1}^{2} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) < 0$ for any $\vec{n} \neq (-1,-1)$. Together with the fact that in $W = 0$ policy $(-1,-1)$ minimises, we get that $(-1,-1)$ is the unique optimal threshold policy for $W < 0$, and the proof of Point 1 is finished.
For the proof of Point 2 we state the following lemma. We present its proof in Appendix 10.6.4.

Lemma 8. *Assume Conditions 1 and 2 hold. When $d = 2$ and $m \geq 0$, being active is an optimal action for $0 \leq W < W^{(2)}$. When $d = 2$ and $m = 0$, being passive is an optimal action for $0 \leq W$.*

From Lemma 8, for $0 \leq W < W^{(2)}$ in environment 2 the optimal action is passive in $m = 0$ and active in $m > 0$, then the optimal threshold policy in environment 2 has to be 0. Thus, the optimal solutions are in the set of policies $\{(n,0)\}_{-1 \leq n \leq \infty}$.
Point 3 follows from Lemma 6 and a comparison of the linear functions. For any policy $\vec{n} = (n_1, n_2)$, $\overline{W}((\infty,\infty),\vec{n}) \leq W^{(2)}$ and the slope of its linear function is $-\sum_{d=1}^{2} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) \geq -1$, since $\pi^{\vec{n}}$ is a probability distribution. For the policy (∞,∞), the slope is $-\sum_{m=0}^{\infty} \pi^{(\infty,\infty)}(m,d) = -1$, thus the linear function $g^{(\infty,\infty)}(W)$ is the steepest one. It follows that for $W \geq W^{(2)}$, (∞,∞) is an optimal solution of (7) and it is the only one for $W > W^{(2)}$. In addition, when $W = W^{(2)}$, only policies of the form (∞, n) have the same cost $g^{\vec{n}}(W)$ as (∞,∞) (see Lemma 6). Hence, when $W = W^{(2)}$, optimal threshold policies are of the form (∞, n). □

10.6.4 Proof of Lemma 8.

We want to show the optimal policies in environment $d = 2$ for $0 \leq W < W^{(2)}$. Recall from (16) that Bellman's optimality equation is given by:

$$\begin{aligned}&(\mu^{(d)} + m\theta^{(d)} + \lambda^{(d)} + r^{(d)})V(m,d) + g = \\ &cm + \lambda^{(d)}V(m+1,d) + m\theta^{(d)}V((m-1)^+,d) + r^{(d)}V(m,3-d) \\ &+ \min\left\{-W + \mu^{(d)}V(m,d), \mu^{(d)}V((m-1)^+,d)\right\}.\end{aligned} \tag{54}$$

Hence, in state (m,d) passive is an optimal action if and only if $-W + \mu^{(d)}V(m,d) \leq \mu^{(d)}V((m-1)^+,d)$. Similarly, active is an optimal action if and only if $-W+\mu^{(d)}V(m,d) \geq \mu^{(d)}V((m-1)^+,d)$. In case $m = 0$, $-W + \mu^{(d)}V(0,d) \leq \mu^{(d)}V(0,d)$ if and only if $W \geq 0$. In other words, in state 0 the optimal action is passive if and only if $W \geq 0$, and both actions are optimal if $W = 0$.
Let $W < W^{(2)}$. We will show that

$$-W + \mu^{(2)}V(m,2) \geq \mu^{(2)}V(m-1,2), \quad \forall m \geq 1,$$

or equivalently,

$$V(m,2) - V(m-1,2) \geq W/\mu^{(2)}, \quad \forall m \geq 1. \tag{55}$$

Since we will use the value iteration method, we formulate an analogous property for the truncated value function V^L, see Lemma 9. We truncate the state space at L and smooth the transition rates from state m to state $m+1$ as in (38). That is, we replace the arrival rates $\lambda^{(d)}$ by

$$\lambda^{(d)}\left(1 - \frac{m}{L}\right)^+,$$

for $m = 0, 1, \ldots, L$. The uniformization constant is taken as

$$\gamma := \lambda^{(1)} + \lambda^{(2)} + \mu^{(1)} + \mu^{(2)} + r^{(1)} + r^{(2)} + L\theta^{(1)} + L\theta^{(2)}, \tag{56}$$

We define the value function $V_t^L(m,d)$ as follows. For any d and $m = 0, \ldots, L$ we initialize by defining $V_0^L(m,d) = 0$ and

$$\begin{aligned}V_{t+1}^L(m,d) =& \frac{cm}{\gamma} + \left(1 - \frac{m}{L}\right)^+ \frac{\lambda^{(d)}}{\gamma}V_t^L(\min\{m+1, L\}, d) \\ &+ \frac{r^{(d)}}{\gamma}V_t^L(m, 3-d) + \frac{m\theta^{(d)}}{\gamma}V_t^L((m-1)^+, d) \\ &+ \frac{1}{\gamma}\min\{-W + \mu^{(d)}V_t^L(m,d), \mu^{(d)}V_t^L((m-1)^+,d)\} \\ &+ \min\left\{\frac{m}{L}, 1\right\}\frac{\lambda^{(d)}}{\gamma}V_t^L(m,d) + \frac{(L-m)\theta^{(d)}}{\gamma}V_t^L(m,d) \\ &+ \frac{1}{\gamma}\left(\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}\right)V_t^L(m,d).\end{aligned}$$

The following lemma states the property needed for V^L.

Lemma 9. *Assume Conditions 1 and 2 hold. For $0 \leq W < W^{(2)}$ there exists an L_0 large enough such that*

$$V^L(m,2) - V^L(m-1,2) \geq W/\mu^{(2)} \quad \forall L \geq L_0 \text{ and } 1 \leq m \leq L. \tag{57}$$

We can conclude now the proof of Lemma 8. From Lemma 9 there exists an L_0 such that inequality (57) holds for all $L \geq L_0$. Since $V^L \to V$, as stated in Section 10.2, inequality (55) is proved, which concludes the proof. □

10.6.5 Proof of Lemma 9:

We define the parameters $W^{L,(d)}$:

$$W^{L,(d)} := c\mu^{(d)} \frac{\theta^{(3-d)} + r^{(1)} + r^{(2)} + \lambda^{(3-d)}/L}{\left(\theta^{(1)} + r^{(1)} + \frac{\lambda^{(1)}}{L}\right)\left(\theta^{(2)} + r^{(2)} + \frac{\lambda^{(2)}}{L}\right) - r^{(1)}r^{(2)}}.$$

Note that $W^{L,(d)} \to W^{(d)}$ as $L \to \infty$. We further define:

$$C_t^{L,(d)}(m) := V_t^L(m,d) - V_t^L(m-1,d), \quad d = 1,2,$$

and we will show that there exists an L_0 and t_0 such that for $L \geq L_0$ and $t \geq t_0$,

$$C_t^{L,(2)}(m) \geq W/\mu^{(2)} \quad \text{for } 1 \leq m \leq L, \tag{58}$$

We write $C_{t+1}^{L,(d)}(m)$ as follows:

$$\begin{aligned}
C_{t+1}^{L,(d)}(m) &= V_{t+1}^L(m,d) - V_{t+1}^L(m-1,d) \qquad (59)\\
&= \frac{c}{\gamma} + \left(1 - \frac{m}{L}\right)^+ \frac{\lambda^{(d)}}{\gamma} C_t^{L,(d)}(m+1) + \frac{r^{(d)}}{\gamma} C_t^{L,(3-d)}(m) + \frac{(m-1)\theta^{(d)}}{\gamma} C_t^{L,(d)}(m-1)\\
&\quad + \frac{1}{\gamma}\min\left\{-W + \mu^{(d)} V_t(m,d), \mu^{(d)} V_t(m-1,d)\right\}\\
&\quad - \frac{1}{\gamma}\min\left\{-W + \mu^{(d)} V_t(m-1,d), \mu^{(d)} V_t((m-2)^+,d)\right\}\\
&\quad + \frac{m-1}{L}\frac{\lambda^{(d)}}{\gamma} C_t^{L,(d)}(m) + \frac{(L-m)\theta^{(d)}}{\gamma} C_t^{L,(d)}(m)\\
&\quad + \frac{\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}}{\gamma} C_t^{L,(d)}(m).
\end{aligned}$$

We have the following properties for functions $C_t^{L,(1)}(m)$ and $C_t^{L,(2)}(m)$.

Lemma 10. *Assume $0 \leq W < W^{(2)}$. Define*

$$t_0 := \min_t \left\{ t \text{ s.t. } \mu^{(d)} C_t^{L,(d)}(m) \geq W \text{ for a pair } (m,d) \right\}.$$

Then there exists an L_1 large enough such that for $L \geq L_1$, $t_0 < \infty$ and $C_t^{L,(d)}(m) = C_t^{L,(d)}$ is constant in m for all $t \leq t_0$.

From the previous lemma, we have that t_0 is the minimum t such that either $\mu^{(1)}C_t^{L,(1)} \geq W$ or $\mu^{(2)}C_t^{L,(2)} \geq W$ for $L \geq L_1$. Under Condition 1, we will prove that at time t_0, it holds that $\mu^{(2)}C_{t_0}^{L,(2)} \geq \mu^{(1)}C_{t_0}^{L,(1)}$, and hence $\mu^{(2)}C_{t_0}^{L,(2)} \geq W$.

Lemma 11. *Assume Condition 1 holds and let* $0 \leq W < W^{(2)}$ *and* $L > 0$. *Then* $\mu^{(2)}C_t^{L,(2)} \geq \mu^{(1)}C_t^{L,(1)}$ *for* $t \leq t_0$.

The previous property allows us to state conditions under which inequality (58) holds in $t = t_0$ for all $L \geq L_1$. Note that for $t > t_0$, Condition 2 will be sufficient to prove that (58) holds for $t \geq t_0$ and for all m.

Lemma 12. *Assume Condition 2 holds and* $0 \leq W < W^{(2)}$. *Then there exists an* L_0 *large enough such that for all* $L \geq L_0$, $C_t^{L,(2)}(m) \geq W/\mu^{(2)}$, $\forall 1 \leq m \leq L$ *and* $t \geq t_0$.

Lemma 12 proves that (58) holds for $0 \leq W < W^{(2)}$ and $L \geq L_0$. Equivalently, (57) holds and the proof of Lemma 9 is finished. □

We provide now the proofs of Lemmas 10, 11 and 12.

Proof of Lemma 10: Since the sequence $W^{L,(2)} \to W^{(2)}$ as $L \to \infty$ and $W < W^{(2)}$, we can take L_1 such that for all $L \geq L_1$ $W < W^{L,(2)}$. The property of $C_t^{L,(d)}(m)$ being constant in m is trivial for $t = 0$, since $V_0^L(m) = 0$ for all m, and thus $C_0^{L,(d)}(m) = 0$ as well. For a given $t < t_0$ we assume $C_t^{L,(d)}(m)$ is constant, and we then prove this property for $t+1$. Since $\mu^{(d)}C_t^{L,(d)}(m) < W$, at time t passive is the optimal action in (59) for $m \geq 1$ and for $W \geq 0$ passive is also optimal for $m = 0$. Hence,

$$\begin{aligned} C_{t+1}^{L,(d)}(m) &= \frac{c}{\gamma} + \left(1 - \frac{m}{L}\right)\frac{\lambda^{(d)}}{\gamma}C_t^{L,(d)}(m+1) + \frac{r^{(d)}}{\gamma}C_t^{L,(3-d)}(m) + \frac{(m-1)\theta^{(d)}}{\gamma}C_t^{L,(d)}(m-1) \\ &+ \frac{\mu^{(d)}}{\gamma}C_t^{L,(d)}(m) + \frac{m-1}{L}\frac{\lambda^{(d)}}{\gamma}C_t^{L,(d)}(m) + \frac{(L-m)\theta^{(d)}}{\gamma}C_t^{L,(d)}(m) \\ &+ \frac{\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}}{\gamma}C_t^{L,(d)}(m). \end{aligned}$$

At time t, we have that $C_t^{L,(d)}(m) = C_t^{L,(d)}$ for every d and $m \in \{1, \ldots, L\}$, hence

$$
\begin{aligned}
C_{t+1}^{L,(d)}(m) &= \frac{c}{\gamma} + \left(1 - \frac{m}{L}\right)\frac{\lambda^{(d)}}{\gamma} C_t^{L,(d)} + \frac{r^{(d)}}{\gamma} C_t^{L,(3-d)} + \frac{(m-1)\theta^{(d)}}{\gamma} C_t^{L,(d)} \\
&\quad + \frac{\mu^{(d)}}{\gamma} C_t^{L,(d)} + \frac{m-1}{L}\frac{\lambda^{(d)}}{\gamma} C_t^{L,(d)} + \frac{(L-m)\theta^{(d)}}{\gamma} C_t^{L,(d)} \\
&\quad + \frac{\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}}{\gamma} C_t^{L,(d)} \\
&= \frac{c}{\gamma} + \left(1 - \frac{1}{L}\right)\frac{\lambda^{(d)}}{\gamma} C_t^{L,(d)} + \frac{r^{(d)}}{\gamma} C_t^{L,(3-d)} + \frac{\mu^{(d)}}{\gamma} C_t^{L,(d)} \\
&\quad + \frac{(L-1)\theta^{(d)}}{\gamma} C_t^{L,(d)} + \frac{\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}}{\gamma} C_t^{L,(d)} \\
&= \frac{c}{\gamma} + \frac{\gamma - \theta^{(d)} - r^{(d)} - \lambda^{(d)}/L}{\gamma} C_t^{L,(d)} + \frac{r^{(d)}}{\gamma} C_t^{L,(3-d)}, \qquad (60)
\end{aligned}
$$

where the last inequality follows from the definition of γ, (56). Hence, $C_{t+1}^{L,(d)}(m)$ does not depend on m. To conclude, we show that $t_0 \neq \infty$. Since for $d = 1, 2$, $C_t^{L,(d)}$ is increasing in t, let $C^{L,(d)} := \lim_{t\to\infty} C_t^{L,(d)}$. Following from (60), $C^{L,(1)}$ and $C^{L,(2)}$ are the solutions of the following system of equations.

$$
\begin{cases}
C^{L,(1)} = \dfrac{c}{\gamma} + \dfrac{\gamma - \theta^{(1)} - r^{(1)} - \frac{\lambda^{(1)}}{L}}{\gamma} C^{L,(1)} + \dfrac{r^{(1)}}{\gamma} C^{L,(2)} \\
C^{L,(2)} = \dfrac{c}{\gamma} + \dfrac{\gamma - \theta^{(2)} - r^{(2)} - \frac{\lambda^{(2)}}{L}}{\gamma} C^{L,(2)} + \dfrac{r^{(2)}}{\gamma} C^{L,(1)}
\end{cases} \qquad (61)
$$

After some algebraic computation, we find that the solutions of this set of equations are $C^{L,(1)} = W^{L,(1)}/\mu^{(1)}$ and $C^{L,(2)} = W^{L,(2)}/\mu^{(2)}$. If $t_0 = \infty$, then $\lim_{t\to\infty} \mu^{(2)} C_t^{L,(2)} \leq W < W^{L,(2)}$, which gives a contradiction. Hence, $t_0 < \infty$. □

Proof of Lemma 11: Let $\Delta_t = \mu^{(2)} C_t^{L,(2)} - \mu^{(1)} C_t^{L,(1)}$. Using (60) and after some computation we get

$$
\begin{aligned}
C_{t+1}^{L,(1)} &= C_t^{L,(1)} \left[\frac{\gamma - \theta^{(1)} - r^{(1)} - \frac{\lambda^{(1)}}{L}}{\gamma} + \frac{\mu^{(1)} r^{(1)}}{\gamma \mu^{(2)}}\right] + \Delta_t \frac{r^{(1)}}{\gamma} + c \\
\Delta_{t+1} &= \frac{\mu^{(1)} C_t^{L,(1)}}{\gamma}\left[\theta^{(1)} + \frac{\mu^{(2)} r^{(2)}}{\mu^{(1)}} + r^{(1)} - \theta^{(2)} - r^{(2)} - \frac{\mu^{(1)} r^{(1)}}{\mu^{(2)}}\right] \\
&\quad + \Delta_t \left[\frac{\mu^{(2)}(\gamma - \theta^{(2)} - r^{(2)} - \frac{\lambda^{(2)}}{L}) - \mu^{(1)} r^{(1)}}{\gamma \mu^{(2)}}\right] + \frac{c(\mu^{(2)} - \mu^{(1)})}{\gamma}.
\end{aligned}
$$

We show by induction that both $C_t^{L,(1)} \geq 0$ and $\Delta_t \geq 0$. For $t = 0$ it is trivial. Assume it holds for t. Since $\gamma - \theta^{(1)} - r^{(1)} - \frac{\lambda^{(1)}}{L} > 0$, it follows that for $C_{t+1}^{L,(1)} \geq 0$. For Δ_{t+1} we have that

$\frac{c(\mu^{(2)}-\mu^{(1)})}{\gamma} \geq 0$, because $\mu^{(2)}-\mu^{(1)} \geq 0$ by Condition 1. The term multiplying Δ_t is:

$$\frac{\mu^{(2)}(\gamma-\theta^{(2)}-r^{(2)}-\lambda^{(2)}/L)-\mu^{(1)}r^{(1)}}{\gamma\mu^{(2)}} \geq 0,$$

which is non-negative if and only if

$$\mu^{(2)}(\gamma-\theta^{(2)}-r^{(2)}-\lambda^{(2)}/L) \geq \mu^{(1)}r^{(1)}.$$

This holds because $\mu^{(2)} \geq \mu^{(1)}$ and $(\gamma-\theta^{(2)}-r^{(2)}-\lambda^{(2)}/L) \geq r^{(1)}$, by definition of γ. Finally, since $\mu^{(1)} \leq \mu^{(2)}$ and $\theta^{(2)} \leq \theta^{(1)}$, we have

$$\theta^{(2)}+r^{(2)}+\frac{\mu^{(1)}}{\mu^{(2)}}r^{(1)} \leq \theta^{(1)}+\frac{\mu^{(2)}}{\mu^{(1)}}r^{(2)}+r^{(1)}.$$

Hence $\Delta_{t+1} \geq 0$ and the proof is finished. □

Proof of Lemma 12: We show that there is an L_0 such that the following property holds for every $t \geq t_0$, for all $L \geq L_0$ and all $m \geq 1$:

$$\begin{cases} C_t^{L,(2)}(m) \geq \dfrac{W}{\mu^{(2)}} \\ C_t^{L,(1)}(m) \geq \dfrac{(\theta^{(2)}+r^{(2)}+\lambda^{(2)}/L)W-c\mu^{(2)}}{r^{(2)}\mu^{(2)}}. \end{cases} \tag{62}$$

Since from Condition 2, $W < W^{(2)} \leq \frac{c\mu^{(2)}}{\theta^{(2)}+r^{(2)}}$, there exists an L_2 such that

$$W \leq \frac{c\mu^{(2)}}{\theta^{(2)}+r^{(2)}+\lambda^{(2)}/L},$$

for all $L \geq L_2$, or equivalently, $(\theta^{(2)}+r^{(2)}+\lambda^{(2)}/L)W-c\mu^{(2)} \leq 0$. On the other hand, it is easy to see by induction from (59) that $C_t^{L,(1)}(m) \geq 0$ for all m, t. Hence, $C_t^{L,(1)}(m) \geq \frac{(\theta^{(2)}+r^{(2)}+\lambda^{(2)}/L)W-c\mu^{(2)}}{r^{(2)}\mu^{(2)}}$ for all $L \geq L_2$ and all t.

In Lemmas 10 and 11 we proved that $C_{t_0}^{L,(2)}(m) \geq \frac{W}{\mu^{(2)}}$, holds for $L \geq L_1$ and $t = t_0$. Then we take $L_0 = \max(L_1, L_2)$, we assume inequalities in (62) hold for $t > t_0$, and we prove they are valid for $t+1$. For all states $(m, 2)$ we have that active is an optimal action because $C_{t+1}^{L,(2)}(m) \geq W/\mu^{(2)}$. Hence, from (59):

$$\begin{aligned} C_{t+1}^{L,(2)}(m) &= \frac{c}{\gamma}+\left(1-\frac{m}{L}\right)\frac{\lambda^{(2)}}{\gamma}C_t^{L,(2)}(m+1)+\frac{r^{(2)}}{\gamma}C_t^{L,(1)}(m)+\frac{(m-1)\theta^{(2)}}{\gamma}C_t^{L,(2)}(m-1) \\ &\quad +\frac{\mu^{(2)}}{\gamma}C_t^{L,(2)}(m-1)+\frac{m-1}{L}\frac{\lambda^{(d)}}{\gamma}C_t^{L,(2)}(m)+\frac{(L-m)\theta^{(2)}}{\gamma}C_t^{L,(2)}(m) \\ &\quad +\frac{\lambda^{(1)}+\mu^{(1)}+r^{(1)}+L\theta^{(1)}}{\gamma}C_t^{L,(2)}(m) \\ &\geq \frac{c}{\gamma}+\frac{\gamma-\theta^{(2)}-r^{(2)}-\lambda^{(2)}/L}{\gamma}\frac{W}{\mu^{(2)}}+\frac{r^{(2)}}{\gamma}\frac{(\theta^{(2)}+r^{(2)}+\lambda^{(2)}/L)W-c\mu^{(2)}}{r^{(2)}\mu^{(2)}} \\ &= \frac{W}{\mu^{(2)}}, \end{aligned}$$

where in the inequality we used that $C_t^{L,(1)}(m) \geq \frac{(\theta^{(2)}+r^{(2)}+\lambda^{(2)}/L)W-c\mu^{(2)}}{r^{(2)}\mu^{(2)}}$. On the other hand, $C_t^{L,(1)}(m) \geq \frac{(\theta^{(2)}+r^{(2)}+\lambda^{(2)}/L)W-c\mu^{(2)}}{r^{(2)}\mu^{(2)}}$ holds for all t as it was proved before. This concludes the proof of both inequalities in (62) for $t \geq t_0$, and then Lemma 12 is proved as well. □

10.6.6 Proof of Proposition 5.

The proof here is similar to the one of Property 1) in Lemma 4 in Section 10.4.1, in the sense that we will compare the truncated processes with limited capacity $L > n$, using the comparison result 9.3.2 in [11, Chapter 9]. Here we compare the truncated process under threshold policy $(n, 0)$ with the one under threshold policy $(n+1, 0)$. For the uniformization, as we repeat the same structure, $H(L)$, $P^{\vec{n},L}((m', d')|(m, d), a)$ and $V_t^{\vec{n},L}$ have the same definitions. However, we take as cost function $C^{\vec{n}}(m, d)$, for a given threshold policy $\vec{n} = (n_1, n_2)$,

$$C^{\vec{n}}(m,d) = \begin{cases} 1 & \text{if } m \leq n_d \quad d = 1,2 \\ 0 & \text{otherwise.} \end{cases}$$

So $C^{\vec{n}}(m, d) = 1$ in the passive states, regardless the environment. As a result, the reward per unit time is $G^{\vec{n},L} = \sum_{d=1}^{2} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m, d)$, which is the term to compare between the truncated processes.
We consider the costs $C^{(n,0)}, C^{(n+1,0)}$ and the transition rates $q^{(n,0),L}, q^{(n+1,0),L}$ for the truncated processes under policies $(n, 0)$ and $(n + 1, 0)$, respectively. Then Inequality (48) has to be proved for all the states where $C^{(n+1,0)}$ and $C^{(n,0)}$ or $q^{(n+1,0),L}$ and $q^{(n,0),L}$ are not the same. This happens in the states that are active under one policy $(n, 0)$ and passive under the other one $(n+1, 0)$, i.e., in state $(n + 1, 1)$. The inequality to prove reduces to

$$V_t^{(n,0),L}(n, 1) - V_t^{(n,0),L}(n + 1, 1) \leq \frac{1}{\mu^{(1)}}, \tag{63}$$

for $t \geq 0$. As before, we prove a more general result, as stated in Lemma 13.

Lemma 13. *Under Condition 1,*

$$V_t^{(n,0),L}(m, d) - V_t^{(n,0),L}(m + 1, d) \leq \frac{1}{\mu^{(1)}} \qquad \forall\ 0 \leq m \leq L - 1, \quad d = 1, 2.$$

Now, result 9.3.2 in [11, Chapter 9] can be applied, and we obtain

$$\sum_{d=1}^{2} \sum_{m=0}^{n_d} \pi^{(n,0)}(m, d) \leq \sum_{d=1}^{2} \sum_{m=0}^{n_d} \pi^{(n+1,0)}(m, d)$$

for the truncated processes. Using the same reasoning as in the proof of Property 1) in Lemma 4, we obtain that this property holds as well for the original (untruncated) process.

□

Proof of Lemma 13.
The proof is similar to the one in Section 10.4.2. To simplify notation, since $(n, 0)$ and L are fixed in the lemma, we will write V_t for $V_t^{(n,0),L}$ and H for $H(L)$. The proof goes by induction, and the

cases of $t = 0$ and $t = 1$ are the same as before. We assume $V_t(m,d) - V_t(m+1,d) \leq \frac{1}{\mu^{(1)}}$ for every (m,d). We prove it now for $t+1$. We first consider the state $(0,2)$.

$$\begin{aligned} V_{t+1}(0,2) &= \frac{1}{H} + \frac{1}{H}\left[r^{(2)}V_t(0,1) + \lambda^{(2)}V_t(1,2) + \left(H - r^{(2)} - \lambda^{(2)}\right)V_t(0,2)\right], \\ V_{t+1}(1,2) &= \frac{1}{H}\left[\left(\theta^{(2)} + \mu^{(2)}\right)V_t(0,2) + r^{(2)}V_t(1,1) + \lambda^{(2)}V_t(2,2)\right. \\ &\qquad \left. + \left(H - \theta^{(2)} - \mu^{(2)} - r^{(2)} - \lambda^{(2)}\right)V_t(1,2)\right]. \end{aligned}$$

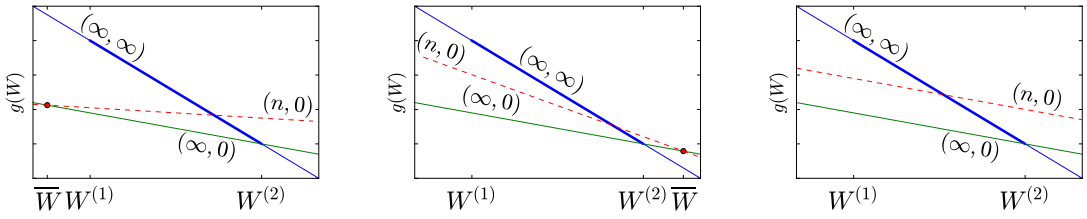
By applying the inductive hypothesis, we obtain

$$\begin{aligned} &V_{t+1}(0,2) - V_{t+1}(1,2) \\ &= \frac{1}{H} + \frac{1}{H}\left[r^{(2)}\left(V_t(0,1) - V_t(1,1)\right) + \lambda^{(2)}\left(V_t(1,2) - V_t(2,2)\right)\right. \\ &\quad \left. - \left(\theta^{(2)} + \mu^{(2)}\right)\left(V_t(0,2) - V_t(1,2)\right) + \left(H - r^{(2)} - \lambda^{(2)}\right)\left(V_t(0,2) - V_t(1,2)\right)\right] \\ &\leq \frac{1}{H} + \frac{1}{H}\left[r^{(2)}\frac{1}{\mu^{(1)}} + \lambda^{(2)}\frac{1}{\mu^{(1)}} + \left(H - \theta^{(2)} - \mu^{(2)} - r^{(2)} - \lambda^{(2)}\right)\frac{1}{\mu^{(1)}}\right] \\ &= \frac{H + \mu^{(1)} - \theta^{(2)} - \mu^{(2)}}{H\mu^{(1)}}. \end{aligned}$$

The inequality $\frac{H + \mu^{(1)} - \theta^{(2)} - \mu^{(2)}}{H\mu^{(1)}} \leq \frac{1}{\mu^{(1)}}$ holds if and only if $\mu^{(1)} - \mu^{(2)} < \theta^{(2)}$. The latter holds from Condition 1 and $\theta^{(2)} > 0$.
Similarly, one obtains the same conclusion for states $(m,d) \neq (0,2)$. This finishes the inductive step and the Lemma is proved. □

10.6.7 Proof of Proposition 6.

The proof relies on comparing $g^{(n,m)}(W)$ for an arbitrary policy (n,m) with $g^{(\infty,0)}(W)$ for $W \in \left[W^{(1)}, W^{(2)}\right]$. In particular, it will be shown that *i)* $g^{(\infty,0)}(W) \leq g^{(n,m)}(W)$ for $W \in [W^{(1)}, W^{(2)}]$, *ii)* $g^{(\infty,0)}(W) < g^{(n,m)}(W)$ for $W \in (W^{(1)}, W^{(2)})$, and *iii)* for $W = W^{(2)}$ the equality $g^{(\infty,0)}(W) = g^{(n,m)}(W)$ holds if and only if $n = \infty$.
To prove the above properties, we distinguish between the two cases $n = \infty$ and $n < \infty$.
We start by assuming $n = \infty$. From Equation (24), it is direct to see that $\overline{W}\left((\infty,m),(\infty,0)\right) = W^{(2)}$, hence $g^{(\infty,m)}(W^{(2)}) = g^{(\infty,0)}(W^{(2)})$. Then, we compare the slopes, recall that $-\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{\vec{n}}(m,d)$ is the slope of $g^{\vec{n}}(W)$ for any threshold policy $\vec{n}$. In this case, the slope of $g^{(\infty,m)}(W)$ is $-(\phi(1) + \sum_{k=0}^{m}\pi^{(\infty,m)}(k,2))$ and the slope of $g^{(\infty,0)}(W)$ is $-(\phi(1) + \pi^{(\infty,0)}(0,2))$. From Property 2) in Lemma 4, we know that $\sum_{k=0}^{m}\pi^{(\infty,m)}(k,2) > \pi^{(\infty,0)}(0,2)$. As a consequence, $g^{(\infty,m)}(W)$ is steeper than $g^{(\infty,0)}(W)$, and therefore, for $W < W^{(2)}$, $g^{(\infty,0)}(W) < g^{(\infty,m)}(W)$ and *i)* and *ii)*



Case a) *Case b)* *Case c)*

Figure 8: Proof of Proposition 6. Comparing slopes of functions $g^{(n,0)}$ and $g^{(\infty,0)}$.

are proved. For the "if and only if" of *iii)* , in case $n \neq \infty$, $g^{(n,m)}(W^{(2)}) \neq g^{(\infty,0)}(W^{(2)})$, because if $g^{(n,m)}(W^{(2)}) = g^{(\infty,0)}(W^{(2)})$, then $g^{(n,m)}(W^{(2)}) = g^{(\infty,\infty)}(W^{(2)})$, and this contradicts Lemma 6.

We now assume $n < \infty$. We split the proof in two steps. In the first step, we compare policies $(\infty, 0)$ and $(n, 0)$. Then, in the second step we compare policies $(n, 0)$ and (n, m). This will be sufficient to conclude the comparison between policies $(\infty, 0)$ and (n, m).

Step 1. In this step we prove that $g^{(\infty,0)}(W) < g^{(n,0)}(W)$, when $W \in \left(W^{(1)}, W^{(2)}\right]$, and $g^{(\infty,0)}(W^{(1)}) \leq g^{(n,0)}(W^{(1)})$ when $W = W^{(1)}$.

To prove Step 1, we distinguish between three cases, as depicted in Figure 8. The cases represent the possible relations between the slopes of the functions $g^{(\infty,0)}(W)$ and $g^{(n,0)}(W)$.

Case a) $-\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{(\infty,0)}(m,d) < -\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{(n,0)}(m,d)$.

In this case $g^{(\infty,0)}(W)$ is steeper than $g^{(n,0)}(W)$. In terms of s_1 and s_2, $-\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{(\infty,0)}(m,d) < -\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{(n,0)}(m,d)$ is equivalent to $s_1\left((\infty,0),(n,0)\right) + s_2\left((\infty,0),(n,0)\right) > 0$. Recall from (23) that

$$\overline{W}\left((\infty,0),(n,0)\right) = W^{(1)} + (1-t)\left(W^{(2)} - W^{(1)}\right),$$

with $t := \dfrac{s_1((\infty,0),(n,0))}{s_1((\infty,0),(n,0)) + s_2((\infty,0),(n,0))}$.

Since $s_1\left((\infty,0),(n,0)\right)+s_2\left((\infty,0),(n,0)\right) > 0$ and $s_2\left((\infty,0),(n,0)\right) = \pi^{(\infty,0)}(0,2)-\pi^{(n,0)}(0,2) \leq 0$ (because of Property 2) in Lemma 4), we have that $t \geq 1$. Since $W^{(2)}-W^{(1)} \geq 0$, $\overline{W}\left((\infty,0),(n,0)\right) \leq W^{(1)}$. This, together with the fact that $g^{(\infty,0)}$ is steeper, implies that for $W > W^{(1)}$, $g^{(\infty,0)}(W) < g^{(n,0)}(W)$ and for $W = W^{(1)}$, $g^{(\infty,0)}(W^{(1)}) \leq g^{(n,0)}(W^{(1)})$.

Case b) $-\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{(\infty,0)}(m,d) > -\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{(n,0)}(m,d)$.

When $g^{(n,0)}(W)$ is steeper the opposite situation occurs. Note that $-\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{(\infty,\infty)}(m,d) = -1$, hence $g^{(\infty,\infty)(W)}$ is the steepest linear function. Besides $\overline{W}((\infty,\infty),(n,0)) < W^{(2)}$ and $\overline{W}((\infty,\infty),(\infty,0)) = W^{(2)}$ because of Lemma 6. Hence, $\overline{W}\left((\infty,0),(n,0)\right) > W^{(2)}$, and as a consequence, for $W \leq W^{(2)}$, $g^{(\infty,0)}(W) < g^{(n,0)}(W)$.

Case c) $-\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{(\infty,0)}(m,d) = -\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{(n,0)}(m,d)$.

Hence $g^{(n,0)}(W)$ and $g^{(\infty,0)}(W)$ are parallel, which means that $\overline{W}\left((\infty,0),(n,0)\right)$ is not defined. Since $\overline{W}((\infty,\infty),(n,0)) < W^{(2)}$ and $\overline{W}((\infty,\infty),(\infty,0)) = W^{(2)}$, we can conclude that $g^{(n,0)}(W)$ and $g^{(\infty,0)}(W)$ are not the same line, and in particular $g^{(n,0)}(W) > g^{(\infty,0)}(W)$ for all W.

Step 2. Assume $n \neq \infty$, and fix any m. For $W \in \left[W^{(1)}, W^{(2)}\right]$, we will prove that $g^{(n,0)}(W) \leq g^{(n,m)}(W)$.
If $g^{(n,0)}(W)$ and $g^{(n,m)}(W)$ are parallel, in order to compare them we can consider the expected cost as $W = 0$, i.e., $\mathbb{E}\left(C(M^{\vec{n}})\right)$. Since policy $(n, 0)$ has larger departure rates than policy (n, m), a simple coupling argument shows that $M^{(n,0)}(t) \leq_{st} M^{(n,m)}(t)$, and hence, $g^{(n,0)}(W) \leq g^{(n,m)}(W)$.
Now assume that $g^{(n,0)}(W)$ and $g^{(n,m)}(W)$ are not parallel, so that Equation (23) can be used. From Properties 1) and 2) in Lemma 4, $s_2\left((n,m),(n,0)\right) = \sum_{l=0}^{m} \pi^{(n,m)}(l,2) - \pi^{(n,0)}(l,2) \geq 0$ and $s_1\left((n,m),(n,0)\right) = \sum_{l=0}^{n} \pi^{(n,m)}(l,1) - \sum_{l=0}^{n} \pi^{(n,0)}(l,1) \leq 0$. From the fact that $s_1\left((n,m),(n,0)\right)$ and $s_2\left((n,m),(n,0)\right)$ have opposite signs, and using Equation (23), $\overline{W}((n,m),(n,0))$ can not be in $(W^{(1)}, W^{(2)})$. Hence, we are in one of the following cases.
If $\overline{W}((n,m),(n,0)) \leq W^{(1)}$, then from Equation (23) we have

$$\frac{s_2((n,m),(n,0))}{s_1((n,m),(n,0)) + s_2((n,m),(n,0))} \leq 0.$$

Since $s_2((n,m),(n,0)) \geq 0$, $s_1((n,m),(n,0)) + s_2((n,m),(n,0)) < 0$, which means that $g^{(n,0)}(W)$ is steeper than $g^{(n,m)}(W)$. This implies that $g^{(n,0)}(W) \leq g^{(n,m)}(W)$ for $W \geq \overline{W}((n,m),(n,0))$, hence, in particular for $W \geq W^{(1)}$.
If $\overline{W}((n,m),(n,0)) \geq W^{(2)}$, the reasoning is analogous. From Equation (23) we have

$$\frac{s_1((n,m),(n,0))}{s_1((n,m),(n,0)) + s_2((n,m),(n,0))} \leq 0,$$

and, as it was stated before, $s_1((n,m),(n,0)) \leq 0$. Hence, $s_1((n,m),(n,0)) + s_2((n,m),(n,0)) > 0$, which means that $g^{(n,m)(W)}$ is steeper than $g^{(n,0)}(W)$. As $\overline{W}((n,m),(n,0)) \geq W^{(2)}$, for any $W \leq W^{(2)}$, $g^{(n,0)}(W) \leq g^{(n,m)}(W)$. □