THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *09/06/2020* **par :**

SANTIAGO DURAN

Resource allocation with observable and unobservable environments

(Allocation de ressources avec environments observables et non-observables)

JURY

KONSTANTIN AVRACHENKOV	Directeur de Recherche	Rapporteur
OLIVIER BRUN	Directeur de Recherche	Examinateur
CHRISTINE FRICKER	Chargée de Recherche	Examinateur
NICOLAS GAST	Chargé de Recherche	Rapporteur

École doctorale et spécialité :
EDSYS : Informatique 4200018
Unité de Recherche :
CNRS, LAAS (UPR8001)
Directeur(s) de Thèse :
Urtzi AYESTA et *Ina Maria VERLOOP*

Acknowledgements

I would like to express my deepest gratitude to my advisors Maaike and Urtzi. Their generosity and trust during my work were the most important help I could have received. Moreover, I always felt them by my side during these three and a half years, and they provided me with the tools and challenges needed to learn and grow. I also thank them for offering me the best working environment one can expect. I would like to thank EDSYS for funding my Ph.D., INP and LAAS, for funding the conferences and the research visits, and Yezekael Hayel and Université d'Avignon, for funding the last 3 months of work.

I have special words for Philippe Robert, who hosted me a whole month during a summer in Paris, and who had to put up with me during the following visits as well. He amazed me with high-level discussions and hard exercices, and all the time we spent together was a constant and intense learning process for me.

I thank the jury members Olivier Brun, Nicolas Gast, Christine Fricker and Konstantin Avrachenkov, for reading the thesis, for giving me very valuable comments and for helping to set up a defense during such a strange period.

I would like to thank my colleagues from both labs. There are those who were there from the very beginning, such as Nicola, Oana, Dorin and Gilles. Those who joined during this journey, such as Amal, Kevin, Mohamed, Iovi and Josue. And those who had the bad luck of sharing an office with me, such as Nicolas, Guillaume, François, Justin and Romain. For all of them, thanks for being there in my everyday life, sharing meals, coffees, and the moments of stress. I also thank Maialen and Elene, my scientific sisters. To Maialen, for all the support she gave me in continuing her research projects. To Elene, for being a great partner, and for always being there, to hear and offer me a hand in anything I needed.

Je veux remercier mes ami.e.s, que j'ai rencontré et qui m'ont entouré tout au long de cette période. Un grand merci à ceux qui ont partagé le quotidien à la maison, en particulier à Thomas, Paul-Henri, Mymi, Denis, Audrey, Gonza, Elise et Kane. Un grand merci à mes collègues de flamenco, et en particulier Alice, qui a été très presente les derniers mois. Un grand merci à Charlotte, Elise et Maréva, pour leurs présences aux différentes étapes de la thèse.

Un agradecimiento enorme a mis amigos en Toulouse, que fueron, son y serán mi familia en Francia. La lista no puede ser extensiva, pero no puedo dejar de nombrarlos: gracias a Jere, Juan, Vir, Nico, Ro, Alfred, Harmonie, Fede, Lean, Pablo, Mati, Guido, Balta, el Parce, Ale, Gasti, Maxi, Fela, Pepe, Pablo, Mauro. Gracias también a la familia cuerva francesa: Caro, Agus, Pablo, Emi y Ernesto.

Quiero agradecer especialmente a María, por haber estado ahí durante toda la tesis. Las palabras no son suficientes para expresar todo lo que siento. La paciencia en los momentos de nervios y la alegría en los momentos de tranquilidad fueron claves para seguir adelante. Gracias por ser una gran persona, por dejarme acompañarte, y por maravillarme cada día. Muchas gracias.

Finalmente quiero agradecer a mi familia, que a pesar de la distancia siempre los sentí y los siento muy cerca mío. A mis viejos y a mis hermanos, con los que nos vimos a menudo y quienes hubieran venido a la defensa, en condiciones más normales. No estaría donde estoy si no fuera por ellos. Y gracias a mis amigos en Buenos Aires, en particular a los pibes, que es un grupo único e irrepetible, y que desde muy chicos me bancan en las buenas y en las malas. Infinitas gracias por su apoyo incondicional.

Abstract

This thesis studies resource allocation problems in large-scale stochastic networks. We work on problems where the availability of resources is subject to time fluctuations, a situation that one may encounter, for example, in load balancing systems or in wireless downlink scheduling systems. The time fluctuations are modelled considering two types of processes, *controllable* processes, whose evolution depends on the action of the decision maker, and *environment* processes, whose evolution is exogenous. The stochastic evolution of the *controllable* process depends on the the current state of the *environment*. Depending on whether the decision maker observes the state of the environment, we say that the environment is *observable* or *unobservable*. The mathematical formulation used is the *Markov Decision Processes* (MDPs).

The thesis follows three main research axes. In the first problem we study the optimal control of a Multi-armed restless bandit problem (MARBP) with an unobservable environment. The objective is to characterise the optimal policy for the controllable process in spite of the fact that the environment cannot be observed. We consider the large-scale asymptotic regime in which the number of bandits and the speed of the environment both tend to infinity. In our main result we establish that a set of priority policies is asymptotically optimal. We show that, in particular, this set includes the Whittle index policy of a system whose parameters are averaged over the stationary behaviour of the environment. In the second problem, we consider an MARBP with an observable environment. The objective is to leverage information on the environment to derive an optimal policy for the controllable process. Assuming that the technical condition of *indexability* holds, we develop an algorithm to compute Whittle's index. We then apply this result to the particular case of a queue with abandonments. We prove indexability, and we provide closed-form expressions of Whittle's index. In the third problem we consider a model of a large-scale storage system, where there are files distributed across a set of nodes. Each node breaks down following a law that depends on the load it handles. Whenever a node breaks down, all the files it had are reallocated to other nodes. We study the evolution of the load of a single node in the mean-field regime, when the number of nodes and files grow large. We prove the existence of the process in the mean-field regime. We further show the convergence in distribution of the load in steady state as the average number of files per node tends to infinity.

Contents

Chapter 1

Introduction

Contents

Resource allocation is the assignment of scarce resources to competing tasks or users. The problems studied in resource allocation are multidisciplinary. They have a broad range of applications spanning from health care and transportation systems to communication networks, and are studied in various domains, such as computer science, electrical engineering and control theory. In this thesis, we focus on resource allocation problems in systems that evolve over time, and where there are decisions to take that depend on the current state of the system. The mathematical formulation used will be the *Markov Decision Processes* (MDPs), a relevant modelling framework for optimisation in Markov processes.

We work on problems where the availability of resources is subject to time fluctuations. This is a situation that arises in different contexts, for example, the number of cashiers needed in a supermarket or the demand in a cloud computing system. We therefore consider MDPs whose parameters fluctuate over time, and we differentiate two types of processes: the *controllable* processes and the *environments*. The environments are exogenous Markov processes used to model time fluctuations. The state of the environments determines the value of the parameters of the controllable processes. We also refer to this as *Markov-modulated environments*.

We do not make any assumption on the correlation of the different environments. In the case of independent distributed environments, this allows us to model the effect of independent fluctuating parameters, as one may encounter, for example, in the arrival rate of new jobs in load balancing systems, or the abandonment rate of impatient customers. On the other hand, when environments are strongly correlated, one could model dependence between the controllable processes through "environmental effects", as it happens, for example, in wireless downlink scheduling.

We further differentiate whether the environments can be observed or not at the moment there is a decision to take. In the *unobservable environment* case, the decision maker does not know the actual state of the

environment, whereas in the *observable environment* case, the decision maker observes the current state of the environment in order to choose actions. Under both settings, we are interested in deriving efficient control policies.

Another topic we address is the resource allocation in *large-scale storage systems*. We consider a model of storage systems, where there are files distributed across nodes, and the nodes break down following random times. When a breakdown occurs, the system allocates the files present in the broken node to other nodes at random, according to a particular allocation policy. Since exact analysis of the system seems out of reach, we study the performance as the number of nodes grows large.

1.1 Overview of research problems

We describe now the three problems we address in the thesis, each one of them corresponds to a chapter. The results obtained in each problem are described in Section 1.3.

1.1.1 Control with unobservable environments

In Chapter 2, we study a large-scale resource allocation problem with controllable processes affected by unobservable environments. We assume that the decision maker takes actions when the controllable processes change state, and the available information is the current state of the controllable processes but not the state of the environments. Considering unobservable environments can be motivated from the fact that it might be too costly to observe the environment and/or change action each time the environment changes.

We focus on a *Multi-Armed Restless Bandit Problem* (MARBP), a specific subclass of MDPs in which there are a number of concurrent projects, known as bandits, that can be made active or passive. In Section 1.2.3 we provide more details on MARBP.

The main goal is to find policies that minimise the average cost incurred by the bandits in a long-run criterion. We do this in an asymptotic regime, where the number of bandits and the speed of the environments grow large. This regime is known as the *mean-field limit*. See Section 1.2.5 for an introduction to the mean-field theory. Among our main results, we propose a set of priority policies that is asymptotically optimal, and we describe an implementable policy that is inside this set. This policy is derived from an averaged version of *Whittle's index* as studied for the classical MARBP, see Section 1.2.4 for an introduction to Whittle's index.

We cite here related work where models on stochastic control with (partially) unobservable environments have been studied. In [21], the authors study a multi-class single-server queueing network with arrival and service rates modulated by an unobservable environment. Note that in the multi-class single-server queue without environments the standard $c\mu$ rule is optimal. Under this rule, the served class is the one maximising the product $c_k\mu_k$, where c_k is the holding cost and μ_k^{-1} is the mean service time of class-k customers [22, 31]. In the main result in [21], it is shown that an "averaged" version of the classical $c\mu$-rule is asymptotically optimal in the heavy-traffic regime. In [6], optimal load balancing is studied when the queue lengths of the servers are unobservable. A set of round-robin policies is proved to be optimal in a heavy-traffic many-server limiting regime. In [7, 50], the authors study optimal control in a multi-class queueing system where the servers' capacity varies according to a Markov-modulated random environment.

The environment can be partially observed for the activated server. As the control decision influences the observation made, the authors search for policies that achieve maximum stability.

1.1.2 Control with observable environments

In Chapter 3, we consider a control problem with environment processes that can be observed. Again, we focus on a Multi-Armed Restless Bandit problem, where each bandit is formed by two processes, a controllable process and an environment. The decision maker bases its decisions on the current state of the controllable processes *and* the current state of the environments. Whenever any of the processes changes state, the action can be changed as well. These problems arise naturally when there is an exogenous phenomenon that can be observed but cannot be controlled. For example in the context of wireless communications, the available download rate depends on weather conditions, and in cloud computing systems the arrival rates of new jobs fluctuate according to the time of the day.

For the approach of the problem, we provide an algorithm that computes Whittle's index, which gives rise to Whittle's index policy. We are able to obtain an analytical characterisation of the index in the concrete example of a multi-class abandonment queue: the queues are the controllable processes that may be served or not, and their transition rates depend on an exogenous environment. More details on our contributions are provided in Section 1.3.

Optimal control with observable environments has been studied in [1, 4, 30]. In [1, 4], the problem of scheduling tasks in a wireless setting is modelled through MARBPs with an observable environment. 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–. Both previous references consider that the environments across processes are independent. In [30] the authors consider an MDP in a mean-field regime, with independent processes evolving in a common environment. It is shown that as the number of processes tends to infinity, the optimal policy converges to the optimal policy of a deterministic discrete time system.

1.1.3 Performance analysis for large-scale storage systems

In Chapter 4, we address an allocation problem in large storage systems using a stochastic modelling approach. We consider a model where there are files distributed across a set of nodes, and each node breaks down after an exponentially distributed amount of time that depends on the number of files it handles. When a node breaks down, all the files it had are reallocated to other nodes, and the node goes back to a state with zero files. The destination node of the files is chosen at random and independently across the files.

Since the model with a fixed number of nodes involves a complex analysis, we focus on the mean-field regime, as we did in Chapter 2. We take the total number of nodes and the total number of files to infinity, with a fixed average number of files per node. We study the existence of the process that represents the load of a single node in the mean-field limit. We further consider the case where the average number of files per node tends to infinity, and we show the convergence in distribution of the load in steady state.

The problem studied here is motivated by the replication methods used in many large storage systems that rely on data redistribution, such as Hadoop Distributed File System (HDFS) [18], Google File System (GFS)

[32] and Cassandra [42]. Whenever there is a loss of content of a node due to its breakdown, the system regenerates the lost data blocks from other copies and allocates them in other functioning nodes. The fact of a node breaking down is closely related to its level of activity: the more data it stores and the more tasks it handles, the faster it presents failures.
Consistency and fault-tolerance in replicated data stores have been largely studied, see for example [67, Chapters 7 and 8] and [56, Chapter 13]. In [63] the authors study the evolution of the load of a node in an equivalent problem to the one we consider. They propose a model of nodes that breakdown after an exponentially distributed amount of time with constant rates. When a node crashes, all the files that were in the node are reallocated following different allocation policies. This model does not take into account the fact that a node breaks down according to its level of activity. The novelty of our work relies on considering nodes whose time of failure is determined by a function of the current load they have.

1.2 Methodology

We introduce here the methodology used in the thesis. In Section 1.2.1 we define the Markov Decision Processes (MDPs), a general framework used for modelling stochastic optimisation problems. In Section 1.2.2 we describe the value iteration method, a numerical technique that can be used for solving MDPs. The Multi-Armed Restless Bandit Problem (MARBP), a particular subclass of MDPs, is presented in Section 1.2.3. In Section 1.2.4 a relaxation approach for the MARBP is discussed, which allows us to derive well-performing heuristics policies. The mean-field technique used in Chapter 4 to study systems in large-scale regimes is outlined in Section 1.2.5.

1.2.1 Markov Decision Processes

In this section we introduce a powerful framework related to the study of stochastic optimisation problems, the Markov Decision Processes (MDPs). They have been largely studied in literature and are one of the main tools used in the thesis. MDPs were introduced by Bellman in 1957 ([11]) as a generalisation of Markov Processes where decisions take a role. For a full overview we refer to the textbooks [13, 38, 57, 60].
A continuous-time MDP is a tuple $(\mathcal{X}, \mathcal{A}, q, C)$, where

- $\mathcal{X}$ is the state space of the process. We will consider the discrete state space $\mathcal{X} = \mathbb{N} \cup \{0\}$.
- $\mathcal{A}$ is a finite set of actions, and $\mathcal{A}(m) \subset \mathcal{A}$ is the set of actions for state $m \in \mathcal{X}$.
- $q : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}_+$ is the transition rates function.
- $C : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is the cost function.

In an MDP, a decision maker knows the current state $m \in \mathcal{X}$ and selects an action a from the space of actions $\mathcal{A}(m)$. The transition to the next state m' is defined by the exponential transition rate $q(m'|m, a)$ that depends not only on the current state of the Markov Process but also on the action. Let φ be a policy that prescribes which action to take in each moment in time. We will denote by $X^\varphi(t)$ and $A^\varphi(t)$ the state and the action of the process at time t under policy φ, respectively.

When the process visits state m and the action taken is a, $C(m,a)$ determines the incurred cost. The objective is to find a policy that minimises some cost criterion. Typically, one studies finite-horizon, discounted infinite-horizon and average-cost criterion. We will focus on the latter, and we will consider stationary policies. Hence, our objective is to find a policy φ that minimises

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T C(X^\varphi(t), A^\varphi(t))\mathrm{d}t\right).$$

Obtaining an optimal policy of an MDPs is typically out of reach. The existence of a solution cannot always be assured, and when it exists, its complexity rapidly scales with the dimensions of the state space, the so-called "curse of dimensionality" [11].

A method for proving the existence and identifying an optimal policy of an MDP comes after the study of the average-cost optimality equations, also known as *Bellman equations*. Let $\tau(m,a) = \sum_{m'\in\mathcal{X}} q(m'|m,a)$ denote the total transition rate for state m and action a. If there is a pair $(g, V(\cdot))$ such that

$$g + \tau(m,a)V(m) = \min_{a\in\mathcal{A}(m)}\left\{C(m,a) + \sum_{m'} q(m'|m,a)V(m')\right\} \quad \forall m, \tag{1.2.1}$$

then a policy that is the argument of the minimum in (1.2.1) is average-cost optimal [60, Section V]. The value g is the incurred *optimal average cost* per unit of time, and $V(\cdot)$ is known as the *value function*, which captures the difference in cost between starting in state m and an arbitrary reference state.

1.2.2 Uniformisation and value iteration

We present now the discrete-time MDPs with time steps $n = 1, 2, \ldots$. A discrete-time MDP is a tuple $\left(\tilde{\mathcal{X}}, \tilde{\mathcal{A}}, P, \tilde{C}\right)$, where $\tilde{\mathcal{X}}$, $\tilde{A}$ and $\tilde{C}$ are the state space, the set of actions and the cost function, respectively, as in the continuous-time MDP. The function $P(m'|m,a)$ is the probability that the process makes a transition to state m' in one step given that the current state is m and the action taken is a.

We are interested in discretising a continuous-time MDP in order to get an equivalent discrete-time one. In this sense, we use a discretisation method known as *uniformisation*. Let $(\mathcal{X}, \mathcal{A}, q, C)$ be a continuous-time MDP, and let $\tau(m,a)$ be the total transition rate for state m and action a. We say the transition rates are *uniformly bounded* if there is a constant γ such that $\tau(m,a) \le \gamma$ for every m, a. In order to uniformise $(\mathcal{X}, \mathcal{A}, q, C)$, we assume it has uniformly bounded rates.

We define an associated discrete-time MDP $\left(\tilde{\mathcal{X}}, \tilde{\mathcal{A}}, P, \tilde{C}\right)$, where the time between steps represents exponentially distributed time intervals of mean $1/\gamma$ in the continuous-time MDP. We let $\tilde{\mathcal{X}} = \mathcal{X}$ and $\tilde{\mathcal{A}} = \mathcal{A}$ be defined as in the continuous-time MDP, and $\tilde{C}(m,a) = C(m,a)/\gamma$ is the cost function per time step. Then, for the probabilities, we set

$$P(m'|m,a) = \begin{cases} q(m'|m,a)/\gamma & \text{if } m' \neq m \\ 1 - \tau(m,a)/\gamma & \text{if } m' = m, \end{cases}$$

where $1-\tau(m,a)/\gamma$ is the probability of having a "dummy" transition, i.e., the chain does not change its state. The resulting discrete-time process is embedded in the original continuous-time one. As a consequence, the

optimal policy for the discrete-time model is the same as for the continuous-time MDP. For a deeper insight in uniformisation see [54, Chapter 3].

We can now introduce the discrete version of the Bellman Equation for a given discrete-time MDP. If there is a pair $(\tilde{g}, V(\cdot))$ such that

$$\tilde{g} + V(m) = \min_{a \in \mathcal{A}(m)} \left\{ \tilde{C}(m,a) + \sum_{m'} P(m'|m,a) V(m') \right\} \quad \forall m, \tag{1.2.2}$$

then the argument of the minimum in Equation (1.2.2) is an optimal average cost policy. Note that $\tilde{g} = g/\gamma$ is the average optimal cost per transition and $V(\cdot)$ is the value function, as described in the continuous-time Bellman Equation (1.2.1).

Equation (1.2.2) can be numerically solved with algorithms like *policy iteration* or *value iteration*. We focus here on the latter, which we use to characterise optimal policies in Sections 3.4.1 and 3.4.2, and to numerically determine the performance of an optimal policy in Sections 2.5 and 3.6. Value iteration works as follows. For each $m \in \mathcal{X}$ set

$$\begin{aligned} V_0(m) &= 0 \\ V_{n+1}(m) &= \min_{a \in \tilde{\mathcal{A}}(m)} \left\{ \tilde{C}(m,a) + \sum_{m'} P(m'|m,a) V_n(m') \right\} \qquad \text{for } n = 0, 1, \ldots. \end{aligned} \tag{1.2.3}$$

The function $V_n(m)$ is the minimum expected finite-horizon cost that can be achieved in n steps when starting in state m. It can also be written as $V_n(m) = \min_\varphi \mathbb{E}\left(\sum_{s=0}^{n-1} C(X^\varphi(s), A^\varphi(s)) | X_0 = m \right)$. When taking the limit in n of the sequence of functions $V_n(\cdot)$, whose existence can be assured under certain conditions, the value function $V(\cdot)$ and the average optimal cost $\tilde{g}$ from Equation (1.2.2) can be retrieved: the limits $V_n(\cdot) - n\tilde{g} \to V(\cdot)$ and $V_{n+1}(\cdot) - V_n(\cdot) \to \tilde{g}$ as $n \to \infty$ hold. The sequence of actions defined as the minimising action in each step n in (1.2.3) also converges, and the limit is an average-cost optimal policy, see [38] for details. In addition, when properties of the value function $V(\cdot)$, such as monotonicity and convexity, can be proved for the functions $V_n(\cdot)$, these carry over to $V(\cdot)$. We use this approach in the proof of Theorem 3.14 in Appendix 3.7.5 and for Proposition 3.9 in Section 3.4.1.

We present how value iteration can be numerically applied to obtain an optimal policy of a discrete-time MDP with finite state space. The algorithm consists in building the functions $V_n(\cdot)$ in a recursive fashion as defined in Equation (1.2.3). Since for any m, $V_{n+1}(m) - V_n(m)$ converges to the average optimal cost $\tilde{g}$, we will stop the recursion when the difference $\max_{m \in \mathcal{X}} (V_{n+1}(m) - V_n(m)) - \min_{m \in \mathcal{X}} (V_{n+1}(m) - V_n(m))$ is sufficiently small.

Some of the models we study have infinite state space with rates that are not uniformly bounded, namely the abandonment queue studied in Section 3.4. When we consider abandonments, each customer in the queue can abandon at a fixed rate, which means that the rates grow linearly with the number of customers. A method used to handle infinite state spaces and unbounded rates is the so-called *truncation*. We fix a state L such that the process does not visit any state $m' > L$, i.e. $q(m'|m,a) = 0$ for any m, a and $m' > L$. The rates $q(m'|m,a)$ for $m, m' \leq L$ remain unchanged.

When we use this truncation method for proving analytical properties for $V(\cdot)$ (e.g. monotonicity), boundary effects may appear. We will then make use of a smoothing rate truncation method, see [14]. We modify the rates of the original process for states $m < L$, such that they decrease to 0 as m grows large, up to state L, where there are no transitions to states $m' > L$. The obtained smoothed process, and its corresponding value function $V^L(\cdot)$, converges to the original one when $L \to \infty$, as it is stated in [14, Theorem 3.1]. The proofs in Sections 3.7.2 and 3.7.5 are based on this method.

1.2.3 Multi-Armed Restless Bandits

We present a particular subclass of MDPs, the Multi-Armed Restless Bandit Problem (MARBP) introduced in [72]. The restless bandit model has been proved to be a powerful stochastic optimisation framework to model control problems. In this thesis we study an MARBP in Chapters 2 and 3.

We consider a set of N *bandits*. The bandits are controllable stochastic processes whose transition rates depend on whether they are made active or passive, with the constraint of having at most $R < N$ active bandits at the same time. We use the same notation as in Section 1.2.1 for the continuous-time MDPs. The state space of the bandits is $\mathcal{X} = \mathbb{N} \cup \{0\}$, and the set of actions is $\mathcal{A} = \{0, 1\}$, where $a = 0$ denotes the action of making a bandit passive and $a = 1$ of making it active. Let φ be the policy that determines the action taken for each bandit, which is a function of the states of the bandits.

For a given policy φ, $X_k^\varphi(t) \in \mathcal{X}$ denotes the state of bandit k at time t, and $\vec{X}^\varphi(t) := (X_1^\varphi(t), \ldots, X_N^\varphi(t))$. $A_k^\varphi(t) \in \mathcal{A}$ denotes the action of bandit k at time t, and $\vec{A}^\varphi(t) := (A_1^\varphi(t), \ldots, A_N^\varphi(t))$. The constraint can then be expressed as

$$\sum_{k=1}^{N} A_k^\varphi(t) \leq R \quad \forall t \geq 0. \tag{1.2.4}$$

We define the set of feasible policies Φ as the set of all policies that satisfy (1.2.4).

When $X_k^\varphi(t) = m$ and $A_k^\varphi(t) = a$, bandit k makes a transition from state m to state m' at rate $q_k(m'|m, a)$. The *restless* term means that bandits may change state even if they are passive. If $C_k(m, a)$ denotes the cost per unit of time of bandit k, the objective of the optimisation problem is to find the policy φ that minimises the long-run average cost:

$$\limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\left(\int_0^T \sum_{k=1}^{N} C_k(X_k^\varphi(t), A_k^\varphi(t)) \mathrm{d}t \right). \tag{1.2.5}$$

In summary, we look for a policy φ that minimises (1.2.5) under constraint (1.2.4).

In order to solve the MARBP, there exists a methodology, explained in Section 1.2.4, that allows us to derive efficient index-based policies. An index policy assigns an *index* to each bandit depending on its current state, and makes active the R bandits having currently the largest indices. Index type policies are simple to implement. A first seminal work with an index-type solution for the multi-armed bandit problem has been proposed by Gittins in the classical non-restless model [34], i.e., where the passive bandits do not change their state.

Optimal control in restless bandit problems provides a powerful optimization framework to model dynamic scheduling of activities. In particular, regarding optimal control of computing systems, the restless bandit framework has been successfully applied in, for example, the context of wireless downlink scheduling [4,

9, 55], load balancing problems [45], systems with delayed state observation [27], broadcast systems [58], multi-channel access models [2, 47], stochastic scheduling problems [5, 36, 51] and scheduling in systems with abandonments [10, 35, 43, 52].

1.2.4 Lagrangian relaxation

We describe now the methodology pioneered by Whittle to derive well-performing heuristics for the MARBP problem [72]. A key step in the methodology is a relaxation of the constraint (1.2.4), which states that the number of active bandits must be satisfied on average, and not in every decision epoch:

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^N A_k^{\varphi}(t)\mathrm{d}t\right) \leq R. \tag{1.2.6}$$

Let Φ^{REL} denote the set of policies satisfying the relaxed constraint (1.2.6), and in particular $\Phi \subseteq \Phi^{REL}$. We consider the problem of finding a policy φ that minimises (1.2.5) under constraint (1.2.6). Using the Lagrangian multipliers approach we can rewrite the problem and obtain the following unconstrained version, which we call *relaxed optimisation problem*. Find a policy φ that minimises

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \left(\sum_{k=1}^N C_k(X_k^{\varphi}(t), A_k^{\varphi}(t)) - W\sum_{k=0}^N A_k^{\varphi}(t)\right)\mathrm{d}t\right). \tag{1.2.7}$$

The Lagrange multiplier W represents a subsidy for making each bandit passive. The Lagrange multipliers' theory states that there exists a multiplier W such that constraint (1.2.6) is satisfied. This new formulation does not have a common constraint, and hence it allows us to decompose the problem in N subproblems, one for each bandit. For a given W, we will consider the one-dimensional problem of minimising

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \left(C_k(X_k^{\varphi}(t), A_k^{\varphi}(t)) - WA_k^{\varphi}(t)\right)\mathrm{d}t\right), \tag{1.2.8}$$

for each bandit k. As a direct consequence, the optimal policy of the relaxed problem is obtained by combining the optimal policies of the N subproblems. Studying the one-dimensional problem is the key idea introduced by Whittle [72], used in order to simplify the original problem. Whittle further showed that, in case the technical condition known as indexability holds, the solution to (1.2.8) is of index type, i.e., there exists a function $W_k(m)$ such that the optimal policy is to make active a bandit if and only if $W_k(m) > W$. This in turn implies that the optimal policy to solve (1.2.7) is of index type as well, and it activates all bandits such that their respective current indices are larger than W.

The indexability condition establishes that as the subsidy for passivity W increases, the set of states in which it is optimal to be passive for bandit k (denoted by $P_k(W)$) increases as well, i.e., $W < W'$ implies $P_k(W) \subseteq P_k(W')$. In other words, indexability states that, given a state $m \in \mathcal{X}$, if it is optimal to make bandit k passive in state m for a given W, it is also optimal to make it passive for any $W' \geq W$.

If the indexability condition holds, the solution to (1.2.8) is given by

$$W_k(m) := \inf\{W : m \in P_k(W)\}. \tag{1.2.9}$$

Note that $W_k(m)$ is the smallest value for the subsidy such that it is optimal for that bandit to be passive in that state. This approach as proposed by Whittle has developed into a large research area, and as a consequence (1.2.9) is commonly known as Whittle's index.
Since the solution to the relaxed optimisation problem (1.2.7) will in general not be feasible for the original problem with constraint (1.2.4), Whittle proposed the following heuristic also based on the Whittle's index: in every decision epoch make active the R bandits having currently the largest Whittle's index. We will refer to this policy as *Whittle's index policy*. This heuristic has been proved to be well-performing in many important examples, see Niño-Mora [53], and asymptotically optimal under certain circumstances, see Weber and Weiss [71], Ji et al. [40], Ouyang et al. [55] and Verloop [69].

1.2.5 Mean-field theory

In this section we present the mean-field limit theory. We consider systems with interacting Markov processes, and we aim to describe the behaviour of the whole system as the number of processes grows large. We use this approach to study large-scale systems, in Chapters 2 and 4. For more details we refer to the lectures [66], and the works in continuous-time [16], in discrete-time [46] and in discrete-time with vanishing intensity [12].
Assume we have N continuous-time Markov processes living in the finite state space $\mathcal{X}$, that are identically distributed and indistinguishable. This means that the transition rates of a given process depend on the total number of processes present in each state $i \in \mathcal{X}$, and not on the current state of one of them. Let $X(t)$ denote the state of a randomly chosen process. We further define $N_i(t)$ for each $i \in \mathcal{X}$ as the number of processes in state i at time t. Let $\vec{N}(t) := (N_i(t) : i \in \mathcal{X})$ be the state of the whole system, which satisfies $\sum_{i\in\mathcal{X}} N_i(t) = N$ for all time t. We assume that the evolution of the system is such that there are no synchronous transitions in different processes. Under these definitions, the transition rates of $X(t)$ when the state of the system is $\vec{N}(t)$ are given by the matrix Q_{ij} defined as follows:

$$\begin{aligned} Q_{ij}(\vec{N}(t)) &= \begin{cases} \lim_{h\to 0} \frac{1}{h}\mathbb{P}\left(X(t+h) = j | X(t) = i, \vec{N}(t) = N\vec{n}\right) & \text{if } N_i(t) > 0 \\ 0 & \text{if } N_i(t) = 0 \end{cases} \\ Q_{ii}(\vec{N}(t)) &= -\sum_{j\in\mathcal{X}, j\neq i} Q_{ij}(\vec{N}(t)), \end{aligned} \tag{1.2.10}$$

where $\mathbb{P}\left(X(t+h) = j | X(t) = i, \vec{N}(t) = N\vec{n}\right)$ is the probability of the process to be in state j at time $t+h$ conditioned to the fact that at time t it is in state i and the whole system is in state $N\vec{n}$.
Let $\vec{n}(t) = \vec{N}(t)/N$ denote the normalised vector, i.e., $n_i(t)$ is the proportion of processes in state i at time t, we omit the dependence on N for the ease of the reading. The transition rate function associated to $\vec{n}(t)$ is given by

$$q_{ij}^{(N)}(\vec{n}(t)) := Q_{ij}(N \cdot \vec{n}(t)). \tag{1.2.11}$$

Although (1.2.10) and (1.2.11) describe the same transition matrix, (1.2.10) is defined at discrete points of integer valued vectors, while (1.2.11) is defined at discrete points whose coordinates are integer multiples of

$1/N$. We will focus on the latter since we are interested in studying the limit in distribution of $\vec{n}(t)$ as N goes to infinity.
An important result in mean-field limit theory ([46, 16]) states that, under certain conditions, there exists a function $q_{ij}(\vec{n}(t))$ such that $q_{ij}(\vec{n}(t)) = q_{ij}^{(N)}(\vec{n}(t))$, for all $N \geq 1$ and all $\vec{n}$ proportions out of N. Then, the normalised vector $\vec{n}(t)$ converges in distribution to a deterministic vector as N tends to infinity. Moreover, for each $i \in \mathcal{X}$ we have the following differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} n_i(t) = \sum_{j \in \mathcal{X}} n_j(t) q_{ij}(\vec{n}(t)).$$

This result can be adapted to different contexts. In Chapter 2 we refer to the limiting results in [17] for mean-field results in the presence of an environment process. In Chapter 4 we make a detailed analysis of the evolution of the load of a node in a storage system in the mean-field limit.

1.3 Main contributions

In this section we present the structure of the manuscript and the main contributions obtained.
In Chapter 2, we consider a stochastic problem with controllable processes and unobservable environments, as described in Section 1.1.1. The chapter is based on our publication [**SR1**]. We assume that the decision maker takes actions based only on the current state of the controllable process. We study a continuous-time multi-armed restless bandits problem (MARBP) in an asymptotic regime, where the number of bandits grows large together with the speed of the environments. The main goal is to find policies that asymptotically minimise the long-run average cost.
Our main result is Theorem 2.5, where we provide sufficient conditions for a policy to be asymptotically optimal. We further introduce a set of priority policies that satisfy those conditions. The second main result is stated in Proposition 2.13, where an averaged version of Whittle's index policy is proved to be inside the set of asymptotic optimal policies. In Section 2.5, we evaluate numerically the performance under Whittle's index policy, outside the asymptotic regime. We study a multi-class scheduling problem in a wireless downlink with two classes. We observe that already for two bandits the performance of the averaged Whittle's index policy becomes close to optimal as the speed of the modulated environment increases.
The key step of the proofs of Chapter 2 follows from studying a system with the averaged parameters of the stochastic process, where the average is determined by the stationary measure of the environments. In particular, we consider a fluid process and a linear programming (LP) problem with the defined parameters. This is based on the results in [17] for the evolution of a particle system under a rapidly changing environment. The policies proved to be asymptotically optimal are the ones that make the fluid process converge to the equilibrium of the LP problem.
In Chapter 3, we study an MARBP with observable environments, as described in Section 1.1.2. The chapter is based on [**SR2**]. Each bandit is formed by two processes, a controllable process and an environment. The decision maker observes the current state of the environment when choosing actions for the controllable

processes. The challenge relies on finding optimal control of bandits living in a multi-dimensional state space.

We study the dynamics of one bandit, based on the relaxed version of the system as introduced in Section 1.2.4. We consider policies of threshold type, where the controllable process is made active if and only if its state is above a certain threshold. We assume the optimality of threshold policies, which is typically the case in many queuing models, and we propose Algorithm 3.4 to find the optimal threshold policy for each value of the Lagrangian multiplier. Whittle's index is then derived. In Proposition 3.8 we show that, if we consider a slow environment process, Whittle's index coincides with Whittle's index of a bandit that only sees a fixed state of the environment.

As a relevant case study, we consider a queuing system with abandonments, under linear holding cost. Our main results are Theorem 3.14, where indexability is proved, and Theorem 3.15, where closed-form expressions for Whittle's index are provided. Various techniques have been used for the proofs, such as the comparison of stochastic processes as described in [19, Chapter 9], or the smoothing truncation method, as presented in Section 1.2.1.

In Proposition 3.19 we derive simpler expressions for the queues with abandonments when the environments vary slowly, and when they vary fast. In particular, the result for the fast regime allows us to compare the observable model of Chapter 3 with the unobservable model of Chapter 2: in Chapter 3 the index remains dependent on the state of the environment, in contrast to the index that takes the averaged parameter values in Chapter 2. This is studied numerically in Section 3.6.3, where we conclude that there can be an important gain in performance when we allow the decision maker to observe the environment, enlarged in cases where the optimal policy strongly depends on the state of the environment.

We include a further comparison between the results for the unobservable and observable regimes in the conclusions of the thesis, see Chapter 5.

In Chapter 4, we study a problem of allocation in large distributed storage system, as introduced in Section 1.1.3. This chapter is based on [**SR3**]. We consider files distributed among a set of nodes. When one of the nodes breaks down all the files it has are reallocated at random to other nodes, and the node goes back to a state with zero files. The analysis is focused on the mean-field regime of the system, i.e., when the number of nodes and files grow large, assuming a fixed average number of files per node. Under this regime, the evolution of the load of a node is studied. The main difficulty of the problem relies on the fact that the rate of the breakdowns of the nodes depend on the load they currently have.

We first consider the so-called Random Weighted allocation policy, that assigns to each node a weighted probability of receiving a file that depends on its current load. Our main contribution is Theorem 4.5, where we show the existence of the process in the mean-field limit, together with the existence and uniqueness of its stationary measure. In a following result in Theorem 4.7, we make the average number of files per node grow large, and we prove the convergence in distribution of the load in steady state. Alternative proofs are presented for the Random allocation policy, where the probability of assigning a file is uniform among all servers. This is done in Theorems 4.8 and 4.10.

We work with urn models in order to describe the dynamics of the process. In an urn model, there is a set of identical balls moving across urns. Each ball waits an exponentially distributed amount of time and then it jumps to another urn. For our case, we consider a model where all the balls in an urn jump at the same time. We study the global balance equations of such a process, and using the properties of the allocation

functions, we characterise the stationary measure of the process as the solution of a fixed point equation, see Section 4.3.

Chapter 2

Optimal control with unobservable environments

Contents

In this chapter we study a large-scale resource allocation problem with controllable processes affected by unobservable environments. We focus on a multi-armed restless bandit problem, as introduced in Section 1.2.3. The bandits are the controllable processes, whose transition rates depend on the current state of the environments. We consider that the environments change state relatively fast with respect to the controllable processes. We study the mean-field limit of the processes, obtained by letting the population of bandits grow large, together with the speed of the environment.

In our first main contribution, we present sufficient conditions for a policy to be asymptotically optimal and we show that a set of priority policies satisfies these. In our second main contribution, we introduce the averaged Whittle's index policy and prove it to be inside the set of asymptotically optimal policies. We further provide a numerical evaluation of the performance of the averaged Whittle's index policy.

For our proofs we consider techniques as used in [69, 71], where MARB problems without environments are studied in a mean-field regime. In the seminal work [71], Whittle's index policy is shown to be asymptotically optimal as the number of bandits that can be made active grows proportionally to the total number of bandits. In [69], the author does not assume indexability, and, using fluid-scaling techniques, proves that a set of priority policies is asymptotically optimal.

Another technique that we use is based on convergence results for particle systems living in rapidly varying environments. We refer to [12, 17], where the authors derived that in the limit the system is described by an ODE, and the transitions rates of the particles are averaged according to the steady-state distribution of the environments. The paper [17] considers a countable state space and each particle is associated its

own modulated environment (no assumption is made on the joint evolution of the environments). On the other hand, in [12] a finite state space and one common environment is considered, resulting in less complex technical conditions.

In order to infer the state of the unobservable environment, one could apply learning techniques as the ones studied in [65]. However, we will not use such techniques, since the environments change at a much faster time scale than the controllable processes, and we assume that it will be too costly to learn from them. We follow up this discussion in Section 5.2.2, where in the context of *reinforcement learning*, we introduce a method for the case where the environments do not change relatively fast.

The chapter is organized as follows. In Section 2.1, we define the multi-class restless bandit control problem and introduce the Markov-modulated environments. Section 2.2 contains the asymptotic optimality results. In Section 2.3, we define a set of priority policies and prove it to be asymptotically optimal when the state space is finite. In Section 2.4 we define an averaged version of Whittle's index policy and prove it to be asymptotically optimal. Section 2.5 presents our numerical results.

2.1 Model description

We consider a multi-class restless bandit problem in continuous time. There are K classes of bandits and there are N_k class-k bandits present in the system. We further define $N := \sum_k N_k$ as the total number of bandits and define $\gamma_k := N_k/N$ as the fraction of class-k bandits. At any moment in time, a class-k bandit is in a certain state $j \in \{1, 2, \ldots, J_k\}$, with $J_k \leq \infty$. In particular, the state space can be countable infinite. At any moment in time, a bandit can either be kept passive or active, denoted by $a = 0$ and $a = 1$, respectively. There is the restriction that at most αN bandits can be made active at a time, $\alpha \leq 1$. The transitions of the class-k bandits depend on a process called environment, described by the Markov process $D_k(t)$ that lives in a countable state space $\mathcal{Z} = \{1, \ldots, d, \ldots\}$ and is positive recurrent. We make no further assumptions on the distribution of the joint vector $\vec{D}(t) = (D_1(t), \ldots, D_K(t))$. For example, it could be that there is one common environment for all classes of bandits, or instead, the environments per class are independently distributed. We further let $\phi(\vec{d})$ denote the stationary probability vector that the environment vector $\vec{D}(t)$ is in state $\vec{d}$. We let $\phi_k(d_k)$ denote the marginal probability of environment $D_k(t)$ to be in state d_k. We further assume that $\sum_{\vec{d}'} r(\vec{d}'|\vec{d}) < C_1$, for all $\vec{d}$, for some $C_1 < \infty$, with $r(\vec{d}'|\vec{d})$ the transition rate of $\vec{D}(t)$ from $\vec{d}$ to $\vec{d}'$.

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

$$\sum_{j=1}^{J_k} q_k^{(d)}(j|i,a) < C_2, \text{ for all } a, d, i, k, \tag{2.1.1}$$

for some $C_2 < \infty$.

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

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

Given policy φ, we then define the function $y^{\varphi,1} : \mathcal{B} \to [0,1]^{\sum_{k=1}^K J_k}$ that distinguishes the action chosen for the bandits. That is, given a policy φ, $y_{j,k}^{\varphi,1}(\vec{x})$ denotes the proportion of class-k bandits in state j that are activated when the proportion of bandits in each state is given by $\vec{x}$. Hence, $y^{\varphi,1}(\cdot)$ satisfies $y_{j,k}^{\varphi,1}(\vec{x}) \leq x_{j,k}$ and $y_{j,k}^{\varphi,1}(\vec{x}) \leq \alpha, \ \ \forall j,k$. We focus on the set of policies such that $y^{\varphi,1}(\cdot)$ is continuous. We further define $y_{j,k}^{\varphi,0}(\vec{x}) := x_{j,k} - y_{j,k}^{\varphi,1}(\vec{x})$, as the proportion of class-k bandits in state j that are kept passive.

For a given policy φ, we define $\vec{X}^{N,\varphi}(t) := (X_{j,k}^{N,\varphi}(t); k = 1, \ldots, K, j = 1, \ldots, J_k)$, with $X_{j,k}^{N,\varphi}(t)$ the number of class-k bandits that are in state j at time t.

Our performance criteria are stability and long-run average holding cost. For a given policy φ, we will call the system *stable* if the process $\vec{X}^{N,\varphi}(t)$ has a unique invariant probability distribution. We will denote by $\vec{X}^{N,\varphi}$ and $X_{j,k}^{N,\varphi}$ the random variables following the steady state distributions, assuming they exist. In case the state space is finite, the process $\vec{X}^{N,\varphi}(t)$ being unichain is a sufficient condition for stability of the policy φ. For infinite state space, whether or not the system is stable can depend strongly on the employed policy. We will only be interested in the set of stable policies.

We denote by $C_k^{(d)}(j,a) \in \mathbb{R}$, $j = 1, \ldots, J_k$, the holding cost per unit of time for having a class-k customer in state j under action a and when in environment d. We note that $C_k^{(d)}(j,a)$ can be negative, i.e., representing a reward. We define the holding cost averaged over the states of the environment as $\overline{C}_k(j,a) := \sum_{d \in \mathcal{Z}} \phi_k(d) C_k^{(d)}(j,a)$. We further introduce the following value functions for given policy φ, and initial proportion of bandits $\dfrac{\vec{X}^{N,\varphi}(0)}{N} = \vec{x} \in \mathcal{B}$:

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

and

$$V_+^{N,\varphi}(\vec{x}) := \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\left(\int_0^T \sum_{k=1}^K \sum_{j=1}^{J_k} \sum_{a=0}^{1} C_k^{(D_k(t))}(j,a) y_{j,k}^{\varphi,a}\left(\frac{\vec{X}^{N,\varphi}(t)}{N} \right) \mathrm{d}t \right). \tag{2.1.2}$$

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

$$V_+^{N,\varphi^*}(\vec{x}) \leq V_-^{N,\varphi}(\vec{x}), \quad \text{for all } \vec{x} \text{ and for all policies } \varphi, \tag{2.1.3}$$

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

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{j,k}^{\varphi,1}\left(\frac{\vec{X}^{N,\varphi}(t)}{N}\right) \leq \alpha, \text{ for all } t. \tag{2.1.4}$$

2.2 Rapidly varying modulated environments

In this section we study the system as the environment is fast changing and the number of bandits scales. In Section 2.2.1 we prove asymptotic independence between the proportion of bandits in each state and the environment. In Section 2.2.2 we establish convergence to a fluid limit and use this to lower bound the performance for any policy. This lower bound allows to prove asymptotic optimality of certain policies, which is presented in Section 2.2.3.

2.2.1 Asymptotic independence

In the lemma below we prove that the bandits are asymptotically independent of the environment, i.e., when the amount of bandits N tends to infinity. The proof can be found in the Appendix 2.6.

Lemma 2.1. *Assume φ is a stable policy for any N. Then, for any subsequence of N such that $\left(\vec{X}^{N,\varphi}/N\right)_{N\in\mathbb{N}}$ converges in distribution, we have*

$$\lim_{N\to\infty} \mathbb{E}\left(\mathrm{e}^{-s_{11}\frac{X_{11}^{N,\varphi}}{N}} \dots \mathrm{e}^{-s_{J_K K}\frac{X_{J_K K}^{N,\varphi}}{N}} \mathbf{1}_{(\vec{D}=\vec{d})}\right) = \phi(\vec{d}) \lim_{N\to\infty} \mathbb{E}\left(\mathrm{e}^{-s_{11}\frac{X_{11}^{N,\varphi}}{N}} \dots \mathrm{e}^{-s_{J_K K}\frac{X_{J_K K}^{N,\varphi}}{N}}\right). \tag{2.2.1}$$

2.2.2 Fluid control problem and lower bound

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

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

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

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

$$x^{u,a}_{j,k}(t) = y^{\varphi,a}_{j,k}(\vec{x}^{\varphi}(t)). \tag{2.2.2}$$

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

$$\begin{aligned}\frac{\mathrm{d}x^u_{j,k}(t)}{\mathrm{d}t} &= \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k} x^{u,a}_{i,k}(t)\overline{q}_k(j|i,a) - \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k} x^{u,a}_{j,k}(t)\overline{q}_k(i|j,a)\\ &= \sum_{a=0}^{1}\sum_{i=1}^{J_k} x^{u,a}_{i,k}(t)\overline{q}_k(j|i,a),\end{aligned} \tag{2.2.3}$$

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

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k} x^{u,1}_{j,k}(t) \leq \alpha, \text{ for all } t \geq 0. \tag{2.2.4}$$

In the lemma below we formally state the convergence of the stochastic process $\vec{X}^{\varphi}(t)/N$ of the proportions of bandits. This result will not be used in any of the proofs of our results, but is presented for completeness. The proof can be found in the Appendix 2.6.

For the convergence to hold, it is assumed in Lemma 2.2 that the process describing the state of a class-k bandit at time t, is tight [15] in N, that is, roughly speaking that the processes must not oscillate too wildly so that probability mass cannot disappear from compact sets. For either a finite state space ($J_k < \infty$) or when the possible transitions in each state is finite, tightness follows directly, see [17, page 12] for details.

Lemma 2.2. *Assume policy φ is such that $y^{\varphi,1}_{j,k}(\cdot)$ is uniformly Lipschitz continuous, i.e.,*

$$\sup_{j,k}|y^{\varphi^*,1}_{j,k}(\vec{x}) - y^{\varphi^*,1}_{j,k}(\vec{z})| \leq C\sup_{i,l}|x_{i,l} - z_{i,l}|, \quad \text{for all } \vec{x},\vec{z}, \tag{2.2.5}$$

with $C > 0$. For a given policy φ, if the process describing the state of a class-k bandit at time t is tight (with respect to N), then the stochastic process $\frac{\vec{X}^{N,\varphi}(Nt)}{N}$ converges to the deterministic process $x^u(t)$, with $u = \varphi$ (as defined in Equations (2.2.2) and (2.2.3)).

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

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

Setting (2.2.3) equal to zero, this gives us the following linear optimisation problem:

$$
\begin{aligned}
(LP) \qquad v^* := \min_{(x^a_{j,k})} & \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}\overline{C}_k(j,a)x^a_{j,k} \\
\text{s.t. } & 0 = \sum_{a=0}^{1}\sum_{i=1}^{J_k} x^a_{i,k}\overline{q}_k(j|i,a), \quad \forall\ j,k, && (2.2.6)\\
& \sum_{k=1}^{K}\sum_{j=1}^{J_k} x^1_{j,k} \leq \alpha, && (2.2.7)\\
& \sum_{j=1}^{J_k}\sum_{a=0}^{1} x^a_{j,k} = \gamma_k, \quad \forall\ k && (2.2.8)\\
& x^a_{j,k} \geq 0, \quad \forall\ j,k,a, && (2.2.9)
\end{aligned}
$$

Let x^* and v^* denote an optimal solution and optimal value of the (LP) problem, respectively.
In Lemma 2.4 below, we use the LP formulation to find a lower bound on the original stochastic optimisation problem. Before that, we need the following condition.

Condition 2.3. *Given a policy φ,*

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

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

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

When bandits have a finite state space ($J_k < \infty$), Condition 2.3 is true whenever $X^{N,\varphi}(t)$ is unichain.

Lemma 2.4. *Assume Condition 2.3 is satisfied. It then holds that the feasible set of the (LP) problem is non-empty and*

$$\liminf_{N\to\infty} V^{N,\varphi}_{-}(\vec{x}) \geq v^*, \tag{2.2.10}$$

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

The proof can be found in the Appendix 2.6.

2.2.3 Asymptotic optimality

In this section we present the asymptotic optimality results.

Theorem 2.5. *Let x^* be an optimal solution of the (LP) problem. Let φ^* be a policy for which the fluid process $x^{\varphi^*}(t)$ converges to x^* as $t\to\infty$, and x^* is the unique equilibrium point (global attractor property). Assume φ^* satisfies Condition 2.3 and the assumptions made in Lemma 2.2. Then, φ^* is asymptotically optimal, that is, for all $\vec{x}$ and all policies φ, it holds that*

$$\lim_{N\to\infty} V^{N,\varphi^*}(\vec{x}) \leq \liminf_{N\to\infty} V^{N,\varphi}_{-}(\vec{x}). \tag{2.2.11}$$

In Section 2.3 we will present a class of policies that satisfy (2.2.5). The global attractor property is verified numerically for the different examples presented in this chapter, see Section 2.5.

Proof of Theorem 2.5: In [17, Theorem 2.3] the mean-field limit for a particle system in a rapidly varying environment is given for the stationary regime. Since we assumed tightness of $\frac{\vec{X}^{N,\varphi^*}}{N}$ (Condition 2.3) and the fact that the fluid process $x^{\varphi^*}(t)$ has a unique global attractor x^*, we can apply their result. Also recall the discussion in the proof of Lemma 2.2. Hence, from [17, Theorem 2.3] we have, $\lim_{N\to\infty} \mathbb{P}\left(\frac{\vec{X}^{N,\varphi^*}}{N} = x^*\right) = 1$, for each state j and class k. Together with Lemma 2.1, this gives

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{\vec{X}^{N,\varphi^*}}{N} = x^*, \vec{D} = \vec{d}\right) = \phi(\vec{d}). \tag{2.2.12}$$

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

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

where the first equality follows from the ergodicity theorem, the second equality from uniform integrability of $\vec{X}/N$ (interchange of limit and summations) and (2.2.12), the third equality from the fact that $y_{j,k}^{\varphi^*,a}(\vec{x}^*) = x_{j,k}^{*,a}$ (see (2.2.2) and $\lim_{t\to\infty} x^{\varphi^*}(t) = x^*$), and the last step follows since $\overline{x}^*$ is an optimal solution of (LP). This concludes the proof of Theorem 2.5. □

2.3 Priority policies

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

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

$$y_{j,k}^{prio,1}(\vec{x}) = \min\left(\left(\alpha - \sum_{(i,l)\in S_k^{prio}(j)} x_{i,l}\right)^+, x_{j,k}\right), \tag{2.3.1}$$

where $S_k^{prio}(j)$ denotes the set of pairs $(i,l), i = 1,\ldots,J_l, l = 1,\ldots,K$ such that class-l bandits in state i have higher priority than class-k bandits in state j under policy *prio*. In the lemma below we show that this function satisfies (2.2.5) when bandits have a finite state space. The proof is in the Appendix 2.6.

Lemma 2.6. *If $J_k < \infty$, Equation* (2.2.5) *is valid for any priority policy.*

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

Definition 2.7 (Set of priority policies Π^*)**.**
We define

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

The set of priority policies Π^* is defined as

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

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

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

We can now state the asymptotic optimality result for priority policies in the class Π^*. Intuitively, this result can be explained as follows. Let φ^* be some policy from the set $\Pi(x^*) \subset \Pi^*$. It is easily verified that x^* is an equilibrium point of the fluid process $x^{\varphi^*}(t)$. If in addition, the point x^* is a global attractor, that is, for any initial point, the process $x^{\varphi^*}(t)$ converges to x^*, then using Theorem 2.5 one obtains the result.

Corollary 2.8. *Assume a finite state space, $J_k < \infty$, for all k. For a given policy $\varphi^* \in \Pi(x^*) \subset \Pi^*$, assume x^* is the global attractor of the fluid process $x^{\varphi^*}(t)$. If in addition, the process $X^{N,\varphi^*}(t)$ is unichain, then φ^* is asymptotically optimal, that is,* (2.2.11) *is satisfied.*

Proof: Lemma 2.6 gives that φ^* satisfies (2.2.5). Hence the result follows directly from Theorem 2.5. □

Remark 2.9 (Infinite state space)**.** *The assumption of finite state space in Corollary 2.8 was made in order to assure uniformly Lipschitz continuity of the function $y_{j,k}^{\varphi^*,1}(\cdot)$. In fact, when $J_k = \infty$, one can easily construct a setting in which* (2.2.5) *does not hold. For example, for $K = 1$, take $\vec{x}^{(l)}$ s.t. $x_i^{(l)} = \alpha/l$ for $i < l$, $x_l^{(l)} = 1 - \alpha$, and $x_i^{(l)} = 0$, for $i > l$. Take $\vec{z}^{(l)}$ s.t. $z_i^{(l)} = 0$ for $i < l$, $z_l^{(l)} = 1 - \alpha$, and $z_i^{(l)} = \alpha/l$, for $i > l$. Then, $y_l^{prio,1}(\vec{x}^{(l)}) = 0$ and $y_l^{prio,1}(\vec{z}^{(l)}) = 1 - \alpha$, where prio is the policy that prioritizes state 1 over state 2, and state 2 over state 3, etc. However, $\sup_j |x_j^{(l)} - z_j^{(l)}| = \alpha/l$. Since the state space is infinite, we can now take $l \to \infty$. We then directly see that* (2.2.5) *does not hold.*
We however note that in our numerical example, where we consider an infinite state space, we do observe a very close to optimal performance of the priority policies.

2.4 Averaged Whittle's index policy

In this section we introduce an averaged version of Whittle's index. The averaged Whittle's index policy is a particular case of a priority policy. It is however only defined in case the system is *indexable*, while our definition of the set of policies Π^* is well-defined for both indexable and non-indexable systems.
In Section 1.2.3 we introduced the relaxation technique and Whittle's index. We recall here the needed concepts, based on this chapter's model. We start by Whittle's relaxation. In (2.1.4) we set the constraint that at most αN bandits can be active at a time. We replace it by its time-average version:

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{j,k}^{\varphi,1}\left(\frac{\vec{X}^{N,\varphi}(t)}{N}\right)\mathrm{d}t\right) \leq \alpha, \tag{2.4.1}$$

and we consider the relaxed-constraint problem, i.e., minimise (2.1.2) under constraint (2.4.1). Using the Lagrangian approach, this can then be written as minimising

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1} C_k^{(D_k(t))}(j,a) y_{j,k}^{\varphi,a}\left(\frac{\vec{X}^{N,\varphi}(t)}{N}\right)\mathrm{d}t \right.$$
$$\left. + W\left(\int_0^T \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{j,k}^{\varphi,1}\left(\frac{\vec{X}^{N,\varphi}(t)}{N}\right)\mathrm{d}t - \alpha\right)\right),$$

where W is the Lagrange multiplier (chosen such that the time-average constraint (2.4.1) holds). The latter can be decomposed into K subproblems, one for each class of bandit, where in each subproblem one needs to minimise the cost term plus a cost W whenever the bandit is made active. Each subproblem can be written as:

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T (C_k^{(D_k(t))}(S_k^{\varphi}(t), A_l^{\varphi}(t)) + W\mathbf{1}_{(A_k^{\varphi}(t)=1)})\mathrm{d}t\right), \tag{2.4.2}$$

where $S_k^{\varphi}(t)$ denotes the state of a class-k bandit at time t and $A_k^{\varphi}(t)$ denotes the action chosen for the class-k bandit under policy φ (for the one-dimensional process).
We have then the following definitions.

Definition 2.10 (Whittle's index)**.** *Whittle's index $W_k(j)$ is defined as the least value of W for which it is optimal in* (2.4.2) *to make the class-k bandit in state j passive.*

We introduce now the indexability condition, that assures structural properties on the optimal policy of the relaxed control problem, and allows us to derive a policy from Whittle's index.

Definition 2.11 (Indexability)**.** *A bandit is indexable if the set of states in which passive is an optimal action in* (2.4.2) *is increasing in W.*

When indexability holds, the solution for the relaxed subproblem (2.4.2) is to activate all bandits that are currently in a state j such that their Whittle index $W_k(j)$ is larger than W. This solution is however not feasible for the original K-dimensional problem, since sometimes it may activate more than αN bandits at a time. For the original problem, the following heuristic was then proposed: active those αN bandits having currently the highest index value. This is what we call *Whittle's index policy*. The relaxation technique as such provides a systematic approach to get a simple index policy.
In this chapter, we study restless bandits living in a modulated environment. In the limiting regime, as the environment varies rapidly, we found that a bandit observes only the averaged (over the steady state of the environment) parameters, that is, $\overline{q}_k(\cdot)$ and $\overline{C}_k(\cdot,\cdot)$. This motivates us to define the averaged Whittle index policy.

Definition 2.12 (Averaged Whittle's index policy)**.** *The* averaged Whittle's index $\overline{W}_k(j)$ *for our restless bandit problem living in a modulated environment is defined as the Whittle index that would result from the restless bandit problem with parameters $\overline{q}_k(i|j,a)$, $\overline{C}_k(j,a)$, and no modulating environment.*
The averaged Whittle index policy *activates those αN bandits having currently the highest averaged Whittle index value $\overline{W}_k(S_{n,k}(t))$, where $S_{n,k}(t)$ is the state of the nth class-k bandit, $k = 1,\dots,K$, $n = 1,\dots,N_k$.*

Proposition 2.13. *If for the averaged version of the restless bandit problem the process describing the state of a class-k bandit is unichain, regardless of the policy employed, and in addition the averaged restless bandit problem is indexable, then there is an x^* such that the averaged Whittle's index policy is inside the set of priority policies $\Pi(x^*) \subset \Pi^*$.*
If in addition $J_k < \infty$, for all k, and x^* is the global attractor of the fluid process $x^{\varphi^*}(t)$, then the averaged Whittle index policy is asymptotically optimal.

The above proposition extends the asymptotic optimality result of Whittle's index policy as obtained in [39, 69, 71] to that of restless bandits living in rapidly varying Markov-modulated environments. The assumptions made in Proposition 2.13 are the same as those needed in [39, 69, 71].

Proof of Proposition 2.13: In [69, Proposition 5.6] it was proved that for the standard restless bandit problem, Whittle's index policy was inside some set of priority policies defined by some linear program problem $(\tilde{LP})$. Since we defined our averaged Whittle's index policy based on the standard restless bandit problem with averaged parameters, we directly have that the averaged Whittle index policy is inside the set of priority policies defined for $(\tilde{LP})$ based on the averaged parameters. However, it can be checked that the $(\tilde{LP})$ for the averaged parameters is equivalent to our (LP) problem, hence, the averaged Whittle index policy is inside $\tilde{\Pi}^*$, which is equal to Π^*.

From the unichain assumption, we obtain that Condition 2.3 is satisfied when $J_k < \infty$. The asymptotic optimality result now follows directly from Theorem 2.5. □

2.5 Numerical evaluation

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

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

This mimics opportunistic scheduling, since the more class-k users present in the system, the higher will be the maximum capacity available among the class-k users. See for further details the discussion in [45, Section 6.1].

The scheduler now needs to decide at each moment in time which of the two classes to activate. We assume

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

with $\overline{\lambda}_k = \sum_{d=1}^{2} \lambda_k^{(d)} \phi_k(d)$ and $\overline{\mu}_k = \sum_{d=1}^{2} \mu_k^{(d)} \phi_k(d)$, so any work-conserving policy makes the system stable. The objective will be to minimise the mean number of users in the system.

The performance for a given policy will be computed using the value iteration approach, as introduced in Section 1.2.2, by writing the dynamic programming equation for the process

$$(M_1^{\varphi}(t), M_2^{\varphi}(t), D_1(t), D_2(t)).$$

Note that we only consider policies that cannot observe the environments when the scheduling decision is taken.

We describe the averaged Whittle index policy in Section 2.5.1. We calculate a benchmark policy and numerically evaluate index policies for several settings: in Section 2.5.3 each class is associated its own environment in order to model Markov-modulated arrival processes. In Section 2.5.4 we will study the effect of having one common environment that influences the departure rates of the classes. In Section 2.5.5 we introduce a discussion for the observable case. Our overall conclusion is that the averaged Whittle index policy performs close to optimal when the modulated environments varies rapidly.

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

Table 2.1: Results for Example 1.

2.5.1 Averaged Whittle's index

This model fits in the restless bandit framework with Markov-modulated environments, as presented in this chapter. In particular, there are two bandits, each bandit representing a class of users, and the state $j \in \{1, 2, \ldots\}$ of the class-k bandit represents its queue length. Hence, when $D_k = d$, the class-k bandit makes a transition from state j to state $j+1$ at rate $\lambda_k^{(d)}$, and from j to $j-1$ at rate $\mu_k^{(d)} \frac{j}{j+1} a$, where $a = 1$ when the class-k bandit is served, and $a = 0$ otherwise. At most one bandit can be activated at a time.
Using the expression of Whittle's index as derived in [45], we obtain that the *averaged* Whittle index for the class-k bandit in state n (its queue length), is given by

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

Here M_k^n denotes the stationary random variable of the process $M_k^n(t)$, where $\varphi = n$ is known as the *threshold policy* with threshold n. It is a one-dimensional birth-and-death process with birth rates $\overline{\lambda}_k$ and death rates in state j at rate $\mathbf{1}_{(j>n)} \overline{\mu}_k \frac{j}{j+1}$. $\pi_k^n(\cdot)$ denotes the stationary measure of $M_k^n(t)$. We will study threshold policies in detail in Chapter 3.
The averaged Whittle index policy serves at each moment in time the class of users having the highest index value $\overline{W}_k(M_k(t))$. From Proposition 2.13 we have that this policy is asymptotically optimal under certain conditions. One of the conditions concerns indexability, which is easily verified for this example using [45, Proposition 2]. The global attractor property is verified numerically for the different examples presented in this section. Another condition needed is that the state space is bounded, which is not the case for our example. The numerical results however do indicate a good performance.
The optimality results in this chapter are for the limiting regime where both the number of bandits (classes) as well as the speed of the modulated environments grow large. In the numerical examples we will instead keep the number of bandits equal to two, which allows to evaluate the performance of our policy outside the asymptotic regime. We then evaluate the performance of the averaged Whittle index policy, as well as other heuristics, for different speeds of the environments.

2.5.2 Benchmark policy

The modulated environments are *unobservable*. In addition, their evolution does not depend on the state of the bandits. We consider the following benchmark policy, derived from the Bellman equation:

$$V(\vec{n}) = n_1 + n_2 + \sum_{k=1}^{2}\sum_{d=1}^{2} \phi_k(d)\lambda_k^{(d)} V(\vec{n}+e_k) + \min_{a\in\{1,2\}} \left[\sum_{k=1}^{2}\sum_{d=1}^{2} \mathbf{1}_{(a=k)}\phi_k(d)\mu_k^{(d)} V((\vec{n}-e_k)^+) \right], \quad (2.5.1)$$

that is, in every state $\vec{n} = (n_1, n_2)$, the only information available to the decision maker is the steady-state distribution of the environment. In the next sections we compare the performance of our heuristics to that of the performance under the actions as defined in (2.5.1).

2.5.3 Markov-modulated arrival processes

Our first numerical example studies Markov-modulated arrival processes. That is, environments $D_1(t)$ and $D_2(t)$ are two independent Markov processes. Each environment can be in two states $\{1, 2\}$ and environment $D_k(t)$ makes a transition from state d to state d' at rate $r_k(d'|d)$. When class k sees environment d, the arrival rate is $\lambda_k^{(d)}$ (while the departure rates remain unchanged).

Example 1: We set the parameters as follows: $\lambda_1^{(1)} = 5$, $\lambda_1^{(2)} = 0.1$, $\lambda_2^{(1)} = 0.5$, $\lambda_2^{(2)} = 5$, and $\mu_1^{(d)} = 7.5$, $\mu_2^{(d)} = 9$, for $d = 1, 2$. We take $r_k(2|1) = 0.001 \cdot \gamma$, $r_k(1|2) = 0.009 \cdot \gamma$, $k = 1, 2$, and let γ vary from 1 up to 5000, in order to study the effect of the speed of the modulated environments. We then have $\phi_1(1) = \phi_2(1) = 0.9$, hence $\rho \approx 0.67$.

In Table 2.1 we show the average performance under the averaged Whittle index policy and that of the benchmark policy. We also show the performance under three other index policies: we consider the index policy $\sum_{d=1}^{2} \phi_k(d) W_k^{(d)}(n)$, which is the Whittle index averaged over the different environments, and we consider the index policy $W_k^{(d)}(n)$ for $d = 1, 2$, which is the Whittle index in case the environment would be always in state d. We denote by g^B the performance of the policy as defined in Section 2.5.2 and let g^φ denote the average cost under policy φ. We define by $Rel(\varphi) := \frac{g^\varphi - g^B}{g^B} * 100\%$ the loss gap (in %). We observe that the averaged Whittle index policy $\overline{W}$ is 2.5% away from the benchmark policy for $\gamma = 25$, i.e., when the transition rates of the environment are $r_k(2|1) = 0.025$ and $r_k(1|2) = 0.225$. Hence, already for a normal scaled environment, the performance of the averaged Whittle index policy is very high. For slow speed, $\gamma = 1$, the averaged Whittle index policy is only 16% away from the benchmark policy. The index policy $\sum_{d=1}^{2} \phi_k(d) W_k^{(d)}(n)$ gives slightly worse performance than that of $\overline{W}$.

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

2.5.4 One common environment affecting the departure rates

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

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

Table 2.2: Results for Example 2.

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

Table 2.3: Results for Example 3.

Example 2: In this example, we chose the parameters such that, when in environment 1, class 1 has a high departure rate and class 2 a low departure rate, while in environment 2, we have the opposite. The values for the parameters are: $\lambda_1^{(d)} = 0.6$, $\lambda_2^{(d)} = 1.2$, for $d = 1, 2$, and $\mu_1^{(1)} = 4$, $\mu_1^{(2)} = 0.5$, $\mu_2^{(1)} = 0.1$, $\mu_2^{(2)} = 6$. We take $r(2|1) = 0.004 \cdot \gamma$, $r(1|2) = 0.006 \cdot \gamma$. We then have $\phi_1(1) = \phi_2(1) = 0.6$, hence $\rho \approx 0.72$.
Note that for these parameters, $\lambda_1^{(2)} > \mu_1^{(2)}$ and $\lambda_2^{(1)} > \mu_2^{(1)}$. That is, if $D(t) = 1$, then class 2 is in overload, while if $D(t) = 2$, then class 1 is in overload. In particular, this implies that the indices $W_k^{(d)}(n)$ are not well-defined. We therefore only simulate the performance under the averaged Whittle index policy $\overline{W}$.
The results can be found in Table 2.2. We observe that the averaged Whittle index policy is 6.4% away from the benchmark policy when $\gamma = 1000$, i.e., $r(2|1) = 4$ and $r(1|2) = 6$. Hence, we observe that already for a normal scaled environment, the performance of the averaged Whittle index policy is very high. When $\gamma = 2500$, i.e., $r(2|1) = 10$ and $r(1|2) = 15$, the gap reduces to 0.8%. For $\gamma = 500$, i.e., $r(2|1) = 2$ and $r(1|2) = 3$, and smaller γ, the loss gap becomes significantly large.
Example 3: In this example, we chose the parameters such that the departure rate of class 1 is always lower than that of class 2, in each environment. In addition, the departure rate for class 1 is considerably higher in environment 2 compared to its rate in environment 1.
The values for the parameters are: $\lambda_1^{(d)} = 1$, $\lambda_2^{(d)} = 3.5$, for $d = 1, 2$, and $\mu_1^{(1)} = 1.5$, $\mu_1^{(2)} = 10$, $\mu_2^{(1)} = 12$, $\mu_2^{(2)} = 11$. We take $r(2|1) = 0.002 \cdot \gamma$, $r(1|2) = 0.008 \cdot \gamma$. We then have $\phi_1(1) = \phi_2(1) = 0.8$, hence $\rho \approx 0.61$.
The results can be found in Table 2.3. We observe that the averaged Whittle index policy $\overline{W}$ is 7% away from the benchmark policy for $\gamma = 5000$, i.e., the transition rates of the environment are $r_k(2|1) = 10$ and $r_k(1|2) = 40$. For slow speed, $\gamma = 1$, the averaged Whittle index policy is 29% away from the benchmark policy. The index policy $\sum_{d=1}^{2} \phi_k(d) W_k^{(d)}(n)$ gives worse performance than that of $\overline{W}$.
For any speed of the environment, we observe that the index policy $W_k^{(1)}(n)$ outperforms the averaged Whittle index policy, while the index policy $W_k^{(2)}(n)$ gives very bad performance ranging between a loss

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

2.5.5 Observable environments

Until now, in the numerical examples, we observed that already for a rather normal speed of the environment, the averaged Whittle index policy works well. Hence, in case the decision maker aims for a policy that is robust without observing the environment, our heuristic seems to be a good choice.
We introduce now an analysis considering that the decision maker can observe the environments, as it will be studied in Chapter 3. For Example 3 we have calculated the performance under the optimal policy when the environment can be observed, and where decision epochs are moments when one of the bandits changes state, or the environment changes state. Note that for the observable setting, the stability condition strongly depends on the policy employed. Numerically, we derived that being able to observe the environment gives an improvement of 16% when $\gamma = 1$, and of 60% when $\gamma = 5000$. The large improvement, especially when $\gamma = 5000$, comes from the fact that the policy will schedule in an opportunistic manner. Recall that the departure rate of class 1 is much larger in environment 2 (compared to environment 1). Hence, in environment 2, an optimal policy will give more preference to class 1. However, in environment 1, class 2 has a much higher departure rate compared to class 1, hence, more priority will be given to class 2. Observing the environment, and being able to change the action when the environment changes, makes this large improvement in the performance possible.

2.6 Appendix

Proof of Lemma 2.1: For ease of notation, we remove the superscripts φ and N in the proof.
For the random vector $\frac{\vec{X}}{N}(t)$, we define the following Probability Generating Function (2.6.1) and the Moment Generating Function (2.6.2), conditioned on the environment vector $\vec{d}$:

$$\begin{aligned} g^{(\vec{d})}(\vec{z}) &:= \mathbb{E}\left(z_{11}^{\frac{X_{11}}{N}} z_{21}^{\frac{X_{21}}{N}} \dots z_{jk}^{\frac{X_{jk}}{N}} \dots z_{J_K K}^{\frac{X_{J_K K}}{N}} \mathbf{1}_{(\vec{D}=\vec{d})} \right) \\ &= \sum_{\vec{x}} z_{11}^{\frac{x_{11}}{N}} z_{21}^{\frac{x_{21}}{N}} \dots z_{jk}^{\frac{x_{jk}}{N}} \dots z_{J_K K}^{\frac{x_{J_K K}}{N}} \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x}, \vec{D} = \vec{d} \right), \end{aligned} \tag{2.6.1}$$

and

$$\tilde{g}^{(\vec{d})}(\vec{s}) = g^{(\vec{d})}(\mathrm{e}^{-\vec{s}}) = \mathbb{E}\left(\mathrm{e}^{-\vec{s}\frac{\vec{X}}{N}} \mathbf{1}_{(\vec{D}=\vec{d})} \right). \tag{2.6.2}$$

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

$$\begin{aligned} &\mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x}, \vec{D} = \vec{d} \right) \left[\sum_{\vec{d'} \in \mathcal{Z}^K} r(\vec{d'}|\vec{d}) + \sum_{k,a,i,j} \frac{1}{N} q_k^{(d_k)}(j|i,a) y_{i,k}^a(\vec{x}) \cdot N \right] \\ &= \sum_{\vec{d'} \in \mathcal{Z}^K} r(\vec{d}|\vec{d'}) \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x}, \vec{D} = \vec{d'} \right) + \sum_{\substack{k,a,j \\ i/x_{i,k}>0 \\ j/x_{j,k}<1}} \frac{1}{N} q_k^{(d_k)}(i|j,a) y_{j,k}^a(\vec{x} + \frac{\vec{e}_{j,k}}{N} - \frac{\vec{e}_{i,k}}{N}) \\ &\qquad \cdot N \cdot \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x} + \frac{\vec{e}_{j,k}}{N} - \frac{\vec{e}_{i,k}}{N}, \vec{D} = \vec{d} \right). \end{aligned} \tag{2.6.3}$$

Note that we have to restrict the sum on the rhs in order to consider the boundary cases. Multiplying (2.6.3) by $z_{11}^{x_{11}} \dots z_{J_K K}^{x_{J_K K}}$, summing over $\vec{x}$ on both sides of the equality, we then obtain the following expression:

$$\begin{aligned} &g^{(\vec{d})}(\vec{z}) \sum_{\vec{d'} \in \mathcal{Z}^K} P(\vec{d'}|\vec{d}) \; + \sum_{\vec{x}} z_{11}^{x_{11}} \dots z_{J_K K}^{x_{J_K K}} \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x}, \vec{D} = \vec{d} \right) \sum_{k,a,i,j} q_k^{(d_k)}(j|i,a) y_{i,k}^a(\vec{x}) \\ &= \sum_{\vec{d'} \in \mathcal{Z}^K} P(\vec{d}|\vec{d'}) \; g^{(\vec{d'})}(\vec{z}) \quad + \sum_{\vec{x}} z_{11}^{x_{11}} \dots z_{J_K K}^{x_{J_K K}} \sum_{\substack{k,a,j \\ i/x_{i,k}>0 \\ j/x_{j,k}<1}} q_k^{(d_k)}(i|j,a) y_{j,k}^a(\vec{x} + \frac{\vec{e}_{j,k}}{N} - \frac{\vec{e}_{i,k}}{N}) \\ &\qquad \cdot \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x} + \frac{\vec{e}_{j,k}}{N} - \frac{\vec{e}_{i,k}}{N}, \vec{D} = \vec{d} \right). \end{aligned} \tag{2.6.4}$$

Below we will show that the second terms in both the LHS and the RHS of (2.6.4) are equal when taking limits. We begin rewriting the second term in the LHS. By changing variables $\vec{z} \to \mathrm{e}^{-\vec{s}}$ (since in the limit

we have that $\frac{\vec{X}}{N}$ is a continuous variable), we have that it is equal to

$$\sum_{\vec{x}} \mathrm{e}^{-s_{11}x_{11}} \dots \mathrm{e}^{-s_{J_K K}x_{J_K K}} \sum_{k,a,i,j} q_k^{(d_k)}(j|i,a) y_{i,k}^a(\vec{x}) \cdot \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{x}, \vec{D} = \vec{d}\right). \tag{2.6.5}$$

We make the same change of variables in the second term in the RHS. Furthermore, the sums in k, a, i, j are bounded, because of the hypothesis (2.1.1), the fact that $y_{j,k}^a(.)$ is a proportion and
mathbbP is a probability. Thus, as a consequence of Fubini's theorem, we can interchange the order of summations:

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

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

$$\sum_{k,a,i,j} \sum_{\substack{\vec{y}/y_{i,k}<1 \\ y_{j,k}>0}} \mathrm{e}^{-s_{11}y_{11}} \dots \mathrm{e}^{-s_{J_K K}y_{J_K K}} \frac{\mathrm{e}^{-\frac{\sum_k s_{i,k}}{N}}}{\mathrm{e}^{-\frac{\sum_k s_{j,k}}{N}}} \cdot q_k^{(d_k)}(i|j,a) y_{j,k}^a(\vec{y}) \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{y}, \vec{D} = \vec{d}\right)$$
$$= \sum_{\vec{y}} \mathrm{e}^{-s_{11}y_{11}} \dots \mathrm{e}^{-s_{J_K K}y_{J_K K}} \frac{\mathrm{e}^{-\frac{s_i}{N}}}{\mathrm{e}^{-\frac{s_j}{N}}} \cdot \sum_{\substack{k,a,j \\ i/y_{i,k}<1 \\ j/y_{j,k}>0}} q_k^{(d_k)}(i|j,a) y_{j,k}^a(\vec{y}) \mathbb{P}\left(\frac{\vec{X}}{N} = \vec{y}, \vec{D} = \vec{d}\right). \tag{2.6.6}$$

Comparing (2.6.5) and (2.6.6) there are two aspects that remain different. The first one is the restriction of summing only when $y_{j,k} > 0$ and $y_{i,k} < 1$ in (2.6.6), but we can add the boundary cases which only sum 0: note that when $y_{j,k} = 0$, there would be a $y_{j,k}^a(\vec{y}) = 0$ multiplying, and when $y_{i,k} = 1$, as it's a proportion in class k bandits, this means that $y_{j,k} = 0$ and again there would be a $y_{j,k}^a(\vec{y}) = 0$ multiplying. The second aspect is the factor $\frac{\mathrm{e}^{-\frac{\sum_k s_{i,k}}{N}}}{\mathrm{e}^{-\frac{\sum_k s_{j,k}}{N}}}$, which converges to 1 as $N \to \infty$. Since $y_{j,k}^a(\vec{x}) \leq 1$, in the limit both terms are the same.
So we can conclude from (2.6.4) that

$$\lim_{N\to\infty} \tilde{g}^{(\vec{d})}(\vec{s}) \sum_{\vec{d'} \in \mathcal{Z}^K} r(\vec{d'}|\vec{d}) = \sum_{\vec{d'} \in \mathcal{Z}^K} r(\vec{d}|\vec{d'}) \lim_{N\to\infty} \tilde{g}^{(\vec{d'})}(\vec{s}), \tag{2.6.7}$$

that is, $\lim_{N\to\infty} \tilde{g}^{(\vec{d})}(\vec{s})$ satisfies the balance equations for $\vec{D}$. Thus, $\lim_{N\to\infty} \tilde{g}^{(\vec{d})}(\vec{s}) = c(\vec{s})\phi(\vec{d})$, where $c(\vec{s})$ does not depend on $\vec{d}$. When summing both sides over $\vec{d}$, we obtain $c(\vec{s}) = \lim_{N\to\infty} \mathbb{E}\left(\mathrm{e}^{-\frac{\vec{s}\vec{X}}{N}}\right)$, that is, Equation (2.2.1) holds true. □

Proof of Lemma 2.2 In [17], the mean-field limit is obtained for a particle system living in a rapidly

varying environment. In particular, in [17, Theorem 2.2] the convergence result in the transient regime is obtained, which would prove our lemma. Hence, to use the results obtained in [17] we need to verify several the assumptions **A0-A8** as made in that paper. We will do so below.

Each bandit represents a particle, and a transition of a class-k bandit/particle from state j to state i when in environment d occurs at rate

$$\frac{1}{N}\sum_{a=0}^{1} q_k^{(d)}(i|j,a)y_{j,k}^{\varphi,a}\left(\frac{\vec{X}(t)}{N}\right), \tag{2.6.8}$$

where $y_{j,k}^{\varphi,a}\left(\frac{\vec{X}(t)}{N}\right)$ needs to be interpreted as the probability for a class-k bandit in state j to see action a. Expression (2.6.8) is the equivalence of [17, Equation (3)].

In [17] a discrete-time setting is considered. In the model we consider, the transition rates are uniformly bounded[1], hence we can uniformise our system to obtain a discrete-time model [54, Section 2.6]. Furthermore, we consider a multi-class particle system, while the results in [17] are for an exchangeable system. As noted in [17, page 38], this is easily generalized to a multi-class model, when adding a class description to the state of the bandit and having the class of the vector of bandits at time 0 be determined by an exchangeable random vector.

We are left with verifying Assumptions **A0-A8** as stated in [17]. Most of them are true by definition, except for **A0** and **A2**, which we discuss in the remainder of the proof.

[A0.] states a weak correlation between the bandits' transitions. That is, let $A_{n,k}(t)$ denotes the event of the n-th class-k bandit to have a transition at time t. Then in A0 of [17] it is assumed that $\mathbb{P}\left(A_{n_1,k_1}(t)A_{n_2,k_2}(t)\right) \leq \rho(N)/N$, with $\rho(N) \to 0$. For our model this is satisfied, since bandits behave independently from each other, hence this probability is equal to zero.

[A2.] states that the transition rates of the bandits are uniformly Lipschitz, using the total variation between two empirical measures. That is, there exists a constant C such that for every $\vec{x}, \vec{z}$,

$$\begin{aligned}&\sup_{j,k,d}\left\{\sum_{i=1}^{J_k}|q_k^{(d)}(i|j,0)y_{j,k}^{\varphi,0}(\vec{x}) + q_k^{(d)}(i|j,1)y_{j,k}^{\varphi,1}(\vec{x}) - q_k^{(d)}(i|j,0)y_{j,k}^{\varphi,0}(\vec{z}) - q_k^{(d)}(i|j,1)y_{j,k}^{\varphi,1}(\vec{z})|\right\}\\ &\leq C\sup_{j,k}|x_{j,k} - z_{j,k}|.\end{aligned} \tag{2.6.9}$$

[1] The transition rate out of state $(\vec{X}, \vec{D})$ are $\sum_{i,k} X_{j,k}q_k^{(d)}(i|j,a)/N + \sum_{\vec{d}'} r(\vec{d}'|\vec{D}) \leq C_1 + \max_{j,k}\sum_i q_k^{(d)}(i|j,a) < C_1 + C_2$, where we use the assumption made on the environment in Section 2.1 and (2.1.1).

The LHS is equal to

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

where we used (2.1.1) and (2.2.5). That is, (2.6.9) is satisfied.

□

Proof of Lemma 2.4: Let φ be a policy that satisfies Condition 2.3. We first prove that the feasible set of the (LP) problem is non-empty.

Note that we have relative compactness for the sequence of random variables $\left(\vec{X}^{N,\varphi}/N\right)_{N\in\mathbb{N}}$. In case $J_k<\infty$ for every class k, this is valid because the random vectors live on a compact space. In case $J_k=\infty$ for some k, relative compactness comes from Condition 2.3 and [15, Theorem 6.1]. As a consequence, we can consider a subsequence of N (for ease of notation still denoted as N) such that $\vec{X}^{N,\varphi}/N$ converges in distribution when $N\to\infty$. Together with Condition 2.3 we can then define the following limiting point $\vec{y}:=(y_{j,k}^{\varphi,a})$, where

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

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

$$\lim_{t\to\infty}\frac{X_{j,k}^{N,\varphi}(t)}{t}=0,\ \text{for all } j,k. \tag{2.6.10}$$

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

$$\begin{aligned}
X_{j,k}^{N,\varphi}(t)=&X_{j,k}^{N,\varphi}(0)+\sum_{a=0}^{1}\sum_{i\neq j}\sum_{d\in\mathcal{Z}}N^{\frac{q_k^{(d)}(j|i,a)}{N}}\left(\int_0^t\mathbf{1}_{(D_k(s)=d)}y_{i,k}^{\varphi,a}\left(\frac{\vec{X}^{N,\varphi}(s)}{N}\right)\cdot N\mathrm{d}s\right)\\
&-\sum_{a=0}^{1}\sum_{i\neq j}\sum_{d\in\mathcal{Z}}N^{\frac{q_k^{(d)}(i|j,a)}{N}}\left(\int_0^t\mathbf{1}_{(D_k(s)=d)}y_{j,k}^{\varphi,a}\left(\frac{\vec{X}^{N,\varphi}(s)}{N}\right)\cdot N\mathrm{d}s\right),
\end{aligned}\tag{2.6.11}$$

where $N^{q_k^{(d)}(j|i,a)/N}(t)$ are independent Poisson processes having as rates $q_k^{(d)}(j|i,a)/N$, $i,j=1,\dots,J_k$, $k=1,\dots,K$, $a=0,1$. By the ergodic theorem [23] and because of $\vec{X}^N$ having an invariant distribution (see Condition 2.3, item a), we obtain that $\frac{1}{t}\int_0^t \mathbf{1}_{(D_k(s)=d)} y_{j,k}^{\varphi,a}(\frac{\vec{X}^{N,\varphi}(s)}{N})\cdot N\mathrm{d}s$ converges to $\mathbb{E}\left(\mathbf{1}_{(D_k=d)} y_{j,k}^{\varphi,a}(\frac{\vec{X}^{N,\varphi}}{N})\cdot N\right) < \infty$ as $t\to\infty$, for all j,k,a,d,N. Because of Lemma 2.1 we further have

$$\lim_{N\to\infty}\mathbb{E}\left(\mathbf{1}_{(D_k=d)} y_{j,k}^{\varphi,a}\left(\frac{\vec{X}^{N,\varphi}}{N}\right)\right) = \phi_k(d) y_{j,k}^{\varphi,a}. \tag{2.6.12}$$

Now, when dividing both sides in (2.6.11) by t, taking $t\to\infty$, and using that $N^\theta(at)/t \to a\theta$ and (2.6.10) hold, we obtain

$$\begin{aligned}
0 &= \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k}\sum_{d\in\mathcal{Z}} \frac{q_k^{(d)}(j|i,a)}{N}\mathbb{E}\left(\mathbf{1}_{(D_k=d)} y_{i,k}^{\varphi,a}\left(\frac{\vec{X}^{N,\varphi}}{N}\right)\cdot N\right) \\
&\quad - \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k}\sum_{d\in\mathcal{Z}} \frac{q_k^{(d)}(i|j,a)}{N}\mathbb{E}\left(\mathbf{1}_{(D_k=d)} y_{j,k}^{\varphi,a}\left(\frac{\vec{X}^{N,\varphi}}{N}\right)\cdot N\right) \\
&= \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k}\sum_{d\in\mathcal{Z}} q_k^{(d)}(j|i,a)\mathbb{E}\left(\mathbf{1}_{(D_k=d)} y_{i,k}^{\varphi,a}\left(\frac{\vec{X}^{N,\varphi}}{N}\right)\right) \\
&\quad - \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k}\sum_{d\in\mathcal{Z}} q_k^{(d)}(i|j,a)\mathbb{E}\left(\mathbf{1}_{(D_k=d)} y_{j,k}^{\varphi,a}\left(\frac{\vec{X}^{N,\varphi}}{N}\right)\right).
\end{aligned} \tag{2.6.13}$$

Letting $N\to\infty$, together with (2.6.12), we then obtain

$$\begin{aligned}
0 &= \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k}\sum_{d\in\mathcal{Z}} q_k^{d}(j|i,a)\phi_k(d) y_{i,k}^{\varphi,a} - \sum_{a=0}^{1}\sum_{i=0,i\neq j}^{J_k}\sum_{d\in\mathcal{Z}} q_k^{d}(i|j,a)\phi_k(d) y_{j,k}^{\varphi,a}, \\
&= \sum_{a=0}^{1}\sum_{i=1,i\neq j}^{J_k} \overline{q}_k(j|i,a) y_{i,k}^{\varphi,a} - \sum_{a=0}^{1}\sum_{i=0,i\neq j}^{J_k} \overline{q}_k(i|j,a) y_{j,k}^{\varphi,a},
\end{aligned} \tag{2.6.14}$$

a.s., that is, $\vec{y}^{\varphi}$ satisfies Equation (2.2.6). By definition, $\vec{y}^{\varphi}$ satisfies $\sum_{k,j} y_{j,k}^{\varphi,1} \le \alpha$ and $y_{j,k}^{\varphi,a}\ge 0$. Hence, $\vec{y}^{\varphi}$ is a feasible solution of (LP). Note that the interchange of limit and summations to go from (2.6.13) to (2.6.14) is possible because of the uniform integrability given by Condition 2.3.
It is left to prove Inequality (2.2.10). We have,

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

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

$$\liminf_{N\to\infty}\liminf_{T\to\infty}\frac{1}{T}\int_0^T\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{a=0}^{1}C_k^{(D_k(t))}(j,a)y_{j,k}^{\varphi,a}(\frac{\vec{X}^{N,\varphi}(t)}{N})\mathrm{d}t\geq v^*, \tag{2.6.15}$$

almost surely.
Consider a fixed realization ω of the process. We first assume that the LHS in (2.6.15) is finite. We then obtain that

$$\begin{aligned}
&\liminf_{N\to\infty}\liminf_{T\to\infty}\frac{1}{T}\int_0^T y_{j,k}^{\varphi,a}\left(\frac{\vec{X}^{N,\varphi}(t)}{N}\right)\mathbf{1}_{(D_k(t)=d)}\mathrm{d}t\\
&=\liminf_{N\to\infty}\mathbb{E}\left(\mathbf{1}_{(D_k=d)}y_{j,k}^{\varphi,a}\left(\frac{\vec{X}^{N,\varphi}}{N}\right)\right) &(2.6.16)\\
&=\lim_{N_i\to\infty}\mathbb{E}\left(\mathbf{1}_{(D_k=d)}y_{j,k}^{\varphi,a}\left(\frac{\vec{X}^{N_i,\varphi}}{N_i}\right)\right)\\
&=\mathbb{E}\left(\lim_{N_i\to\infty}\mathbf{1}_{(D_k=d)}y_{j,k}^{\varphi,a}\left(\frac{\vec{X}^{N_i,\varphi}}{N_i}\right)\right)\\
&=\phi_k(d)y_{j,k}^{\varphi,a}, &(2.6.17)
\end{aligned}$$

with N_i the subsequence corresponding to the liminf sequence and such that $\frac{\vec{X}^{N,\varphi}}{N}$ converges in distribution. In the third step we used that $\vec{X}^{N,\varphi}/N$ is uniformly integrable and in the fourth step we used (2.6.12) (this equation holds for any weakly converging subsequence of $\frac{\vec{X}^{N,\varphi}}{N}$). As a consequence we obtain that

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

where the last inequality holds because $\vec{y}$ is a feasible solution of the (LP) problem.
In particular, the above shows that $v^*<\infty$, since we assumed in Section 2.1 that there exists a policy for which the LHS in (2.6.15) is finite.
Assume now that the LHS of (2.6.15) is infinite. Then inequality (2.6.15) follows directly, since $v^*<\infty$. This concludes the proof. □

Proof of Lemma 2.6: We consider four possible cases:
(1) If $\left(\alpha-\sum_{(i,l)\in S_k^{\varphi}(j)}x_{i,l}\right)<0$, then by definition (2.3.1) we have $y_{j,k}^{prio,1}(\vec{x})=0$.

If as well $\left(\alpha - \sum_{(i,l)\in S_k^{\varphi}(j)} z_{i,l}\right) < 0$, then the LHS of (2.2.5) equals zero, hence (2.2.5) holds. If instead $\left(\alpha - \sum_{(i,l)\in S_k^{\varphi}(j)} z_{i,l}\right) \geq 0$, then by definition (2.3.1),

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

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

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

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

$$|y_{j,k}^{prio,1}(\vec{x}) - y_{j,k}^{prio,1}(\vec{z})| = \left| x_{j,k} - \alpha + \sum_{(i,l)\in S_k^{prio}(j)} z_{i,l} \right|. \tag{2.6.18}$$

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

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

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

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

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

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

Setting $C = \sum_{k=1}^{K} J_k$, we proved (2.2.5). □

Chapter 3

Optimal control with observable environments

Contents

In this chapter we study an optimisation problem with controllable processes affected by observable environments. We consider a multi-armed restless bandit problem, where each bandit is formed by two processes, a controllable process and an environment. Thus, the state descriptor of the bandits is two-dimensional. The decision maker has full information of the state the controllable process and the environment are in. We further consider the particular case of queues with abandonments under linear holding costs, where impatient customers can leave the system, regardless of whether they are being served or not. There, 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.

In our first 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. For the problem of a multi-class abandonment queue, we show that threshold policies are optimal. We prove indexability and obtain closed-form expressions for Whittle's index. These are further simplified in case the environments vary slowly, and when they vary fast. We further propose a heuristic for a multi-class queue without abandonments, derived from the results for the problem with abandonments. 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 is small. Numerically, we also assess the performance of the averaged Whittle index policy.

The latter was proved to be asymptotically optimal for *unobservable* environments, see Chapter 2, but we show that it can be far from optimal in the observable setting.
Most of the references in MARBP consider unidimensional bandits, which is a critical assumption in order to establish indexability, and in turn to calculate Whittle's index. On the other hand, literature on multidimensional MARBPs, as considered here, is scarce. The main difficulty lies in establishing indexability, i.e., ordering the states, in a multidimensional space. Important exceptions are [1, 4], as we described in Section 1.1.2.
The multi-class abandonment queue has been largely studied in recent literature, see for example the Special Issue in Queuing Systems on queuing systems with abandonments [37] and the survey [25] in the many-server settings. In particular, we mention the work of [44], where the authors formulate the abandonment queue as an MARBP and derive a closed-form expression for Whittle's index. In this chapter, we extend this work by considering the abandonment queue with rates that depend on environments.
The chapter is organised as follows. In Section 3.1 we describe the model. In Section 3.2 we present the relaxation of the original problem and Whittle's index. In Section 3.3 we introduce the threshold policies and, assuming they are optimal, we provide the algorithm to determine Whittle's index. We establish an analytical solution of Whittle's index when the dynamics of the environment are much slower than the controllable process. In Section 3.4 we study the multi-class queue with abandonments. Optimality of threshold policies is proved in Section 3.4.1. In Section 3.4.2 we prove indexability and obtain expressions for Whittle's index. The proofs of the theorems are in Section 3.4.3. In Section 3.5 we derive Whittle's index for a multi-class queue without abandonments living in a Markov-modulated environment. Finally, in Section 3.6 we numerically evaluate the performance of Whittle's index policy. For ease of the reading, many proofs are presented in the appendix.

3.1 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, $R \in \mathbb{N}$. 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 $M_k^\varphi(t) = m \in \mathcal{X}$ denote the state of the controllable process of bandit k at time t, $k = 1, \ldots, N$. Note that $(M_k^\varphi(t), D_k(t))$ describes the two-dimensional state of bandit k. The full description

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

is then a Markov process.
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 can be expressed as

$$\sum_{k=1}^{N} A_k^\varphi(t) \leq R, \quad \forall t \geq 0. \tag{3.1.1}$$

We define the set of feasible policies Φ, as the set of all Markovian policies that satisfy (3.1.1).
When action a is applied to bandit k, it makes a transition from state (m, d) to state (m', d) 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 (3.1.1), i.e., solve:

$$\min_{\varphi \in \Phi} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left(\int_0^T \sum_{k=1}^{N} C_k(M_k^\varphi(t), D_k(t), A_k^\varphi(t)) \mathrm{d}t \right). \tag{3.1.2}$$

Finding a solution for the constrained problem (3.1.2) is PSPACE-hard, hence infeasible, see [33, 49, 73]. However, we refer to the Lagrangian relaxation method, as we introduced in Section 1.2.4, that allows to obtain efficient index policies for the original problem. In Section 3.2 this is applied to the MARBP with observable environments.

3.2 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 [72], is to solve an unconstrained problem obtained via a Lagrangian relaxation approach, instead of solving problem (3.1.2) under constraint (3.1.1). This leads to a significant simplification of the problem, where the multi-dimensional setting can be replaced by a a one-dimensional setting. The solution for the relaxed problem can be then described by the Whittle's index.

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:

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^{N} A_k^{\varphi}(t)\mathrm{d}t\right) \leq R. \tag{3.2.1}$$

We denote the set of policies satisfying constraint (3.2.1) by Φ^{REL}, and we note that $\Phi \subseteq \Phi^{REL}$. We now consider the problem of finding a policy φ that minimises (3.1.2) under constraint (3.2.1). We use the Lagrangian multipliers approach to rewrite the following unconstrained version of the relaxed problem:

$$\min_{\varphi} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \left(\sum_{k=1}^{N} C_k(M_k^{\varphi}(t), D_k(t), A_k^{\varphi}(t)) - W\sum_{k=0}^{N} A_k^{\varphi}(t)\right)\mathrm{d}t\right). \tag{3.2.2}$$

The Lagrange multiplier W can be viewed as a subsidy for making each bandit passive. The key observation made by Whittle is that problem (3.2.2) can be decomposed in N subproblems, one for each bandit, due to the fact that there is no longer a common constraint. Thus, the solution to (3.2.2) is obtained by combining the solution to N separate optimisation problems, that is,

$$\min_{\varphi} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \left(C_k(M_k^{\varphi}(t), D_k(t), A_k^{\varphi}(t)) - W A_k^{\varphi}(t)\right)\mathrm{d}t\right), \tag{3.2.3}$$

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

$$g_k^{\varphi}(W) := \mathbb{E}\left(C_k(M^{\varphi}, D_k, A_k^{\varphi})\right) - W\mathbb{E}\left(A_k^{\varphi}\right), \tag{3.2.4}$$

where $M_k^{\varphi}, D_k, A_k^{\varphi}$ are the respective steady-state variables for a given bandit under policy φ. Then, problem (3.2.3) 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 3.1. *Bandit k is indexable if $W < W'$ implies $P_k(W) \subseteq P_k(W')$.*

In other words, if for $W = W_0$ it is optimal to be passive in state (m,d), it is also optimal to be passive for any value for the subsidy $W \geq W_0$. The Whittle's index in state (m,d) is now defined as follows: it is the smallest value for the subsidy such that it is optimal to be passive in that state.

Definition 3.2. *When bandit k is indexable, the 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 (3.2.2) 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 (3.2.1) is satisfied. Since the solution to the relaxed optimisation problem will in general not be feasible for the original problem (3.1.2) with constraint (3.1.1), Whittle proposed a heuristic also based on the Whittle's index. We will refer to this policy as *Whittle's index policy*.

Definition 3.3 (Whittle's index policy)**.** *At time t, the Whittle's index policy prescribes to make active the R bandits having currently the largest value for their Whittle's index $W_k(M_k(t), D_k(t))$.*

3.3 Calculation of Whittle's index

In this section we present our main results for the general model. In Section 3.3.1, we provide an algorithm to calculate the Whittle's index, and in Section 3.3.2, 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 (3.2.3), we omit the dependence on k for ease of notation throughout this section.

3.3.1 Threshold policies

We assume throughout the chapter that threshold policies (defined below) are an optimal solution of (3.2.3). This is typically the case in many queueing models. For the case study of an abandonment queue, we prove it to hold in Section 3.4.1. In this section, we further assume the bandit is indexable, as defined in Definition 3.1.
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 [45, Section 3.2] for a case where both types of threshold policies are considered.
In order to obtain Whittle's index, one needs to find the optimal threshold policies for any value of W. First recall that the average cost (see (3.2.4)) under a threshold policy $\vec{n}$ is given by

$$g^{\vec{n}}(W) := \sum_{d\in\mathcal{Z}} \sum_{m=0}^{\infty} C(m,d,a)\, \pi^{\vec{n}}(m,d) - W \sum_{d\in\mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d). \tag{3.3.1}$$

Since we assumed that an optimal solution is of threshold type, the optimal average cost can be written as

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

In Figure 3.1, 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 is determined by $-\sum_{d\in\mathcal{Z}}\sum_{m=0}^{n_d}\pi^{\vec{n}}(m,d)$, 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. Now one can simply look to the first value of the subsidy, $\hat{W}_0$, where a linear function $g^{\vec{n}}(W)$ crosses $g^{(-1,-1,\ldots)}(W)$, and is optimal for $W\in[\hat{W}_0,\hat{W}_1]$. Let $\vec{n}^0$ be the corresponding threshold policy of this linear function. 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 $n_0^d\geq 0$, since $\hat{W}_0$ is the smallest value such that it is optimal to be passive in those states. This happens because threshold policy $(-1,-1,\ldots)$ is optimal for $W\leq\hat{W}_0$ and $\vec{n}^0$ is optimal for $W\in[\hat{W}_0,\hat{W}_1]$, with $\hat{W}_1$ still unknown. Similarly, one can now determine $\hat{W}_1$, and find the optimal threshold policy in the point $\hat{W}_1$, and as such, the Whittle index for states (m,d), with $n_d^0<m\leq n_d^1$.

To formalize the above procedure, we introduce 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 (3.3.1) is given by:

$$\overline{W}(\vec{n},\vec{n}') = \frac{\sum_{d\in\mathcal{Z}}\sum_{m=0}^{\infty}C(m,d,a)\,\pi^{\vec{n}}(m,d)-\sum_{d\in\mathcal{Z}}\sum_{m=0}^{\infty}C(m,d,a)\,\pi^{\vec{n}'}(m,d)}{\sum_{d\in\mathcal{Z}}\sum_{m=0}^{n_d}\pi^{\vec{n}}(m,d)-\sum_{d\in\mathcal{Z}}\sum_{m=0}^{n'_d}\pi^{\vec{n}'}(m,d)}. \tag{3.3.2}$$

Note that in case there are two threshold policies that, besides $(-1,-1,\ldots)$, are optimal in $\hat{W}_0$, $g^{\vec{n}}(\hat{W}_0)=g^{\vec{n}'}(\hat{W}_0)$, then the one having the steepest slope will be the one that minimises when $W=\hat{W}_0+\Delta$, with Δ small enough, hence, we choose that one in the case of a tie.

In Algorithm 3.4 we summarise the above. Under the assumption of threshold optimality and indexability, the output of Algorithm 3.4 are the threshold policies $\vec{n}$ that optimizes (3.2.3) for each W, and Whittle's index.

Algorithm 3.4. *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*

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

and compute

$$\hat{W}_j = \inf_{\vec{n}\in E_j}\overline{W}(\vec{n},\vec{n}^{j-1}) \tag{3.3.3}$$

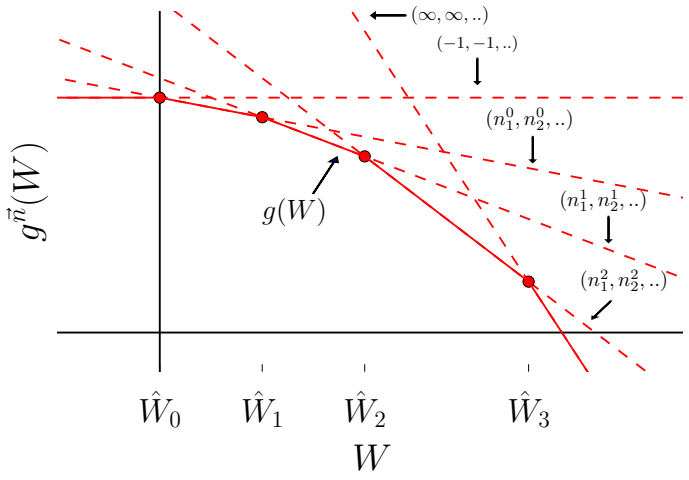

Figure 3.1: Lower envelop $g := \min_{\vec{n}} g^{\vec{n}}$. Algorithm 3.4 finds $\hat{W}_j$ for $j \geq 0$.

Denote by $\vec{n}^j$ the minimiser of (3.3.3). In case of a tie, choose the minimiser that has the steepest slope $-\sum_{d\in\mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d)$. 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.

Remark 3.5. *The numerical computation of this algorithm requires the set where the minimisation is done to be finite. Hence, in case $|\mathcal{X}| = \infty$ or $|\mathcal{Z}| = \infty$, the results obtained are an approximation.*

3.3.2 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 $(p^{n_d,(d)}(m))_{d\in\mathcal{Z}}$ denote the corresponding steady-state probability under threshold policy n_d. We further let $W^{(d)}(m)$ be the 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 two states, i.e., $|\mathcal{Z}| = 2$. Whether the results extend to $|\mathcal{Z}| > 2$ is left as further research.

When the environment changes state at a much slower time scale than the controllable process, the conditional (on the environment) steady-state behaviour of the bandit is that of a bandit whose environment never changes. This is stated formally below. The proof can be found in Appendix 3.7.1.

Lemma 3.6. *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*

$$\lim_{\beta\to 0} \pi^{\vec{n}}(m,d) = \phi(d) p^{n_d,(d)}(m), \qquad \forall m \in \mathbb{N}_0. \tag{3.3.4}$$

The following lemma states that in the slow regime, the index value given by Algorithm 3.4 can be found by changing the threshold value in only one of the two environments. For a given β, we denote by $\vec{n}^j(\beta) = (n_1^j(\beta), n_2^j(\beta))$ the values of $\vec{n}^j$ as defined in Algorithm 3.4. The proof can be found in Appendix 3.7.1.

Lemma 3.7. *Let $|\mathcal{Z}| = 2$ and the transitions of the environment be scaled by β, $\beta r^{(dd')}$. Assume $(\hat{n}_1^j, \hat{n}_2^j) := \lim_{\beta\to 0}(n_1^j(\beta), n_2^j(\beta))$ exists, for all j, and that the family $\{M^{\vec{n}}, \beta\}$ is uniform integrable, for any threshold policy $\vec{n}$. Then,*

$$\lim_{\beta\to 0} \inf_{\vec{n}\in E_j} \frac{\sum_{d=1}^2 \sum_{m=0}^\infty C(m,d,a)\, \pi^{\vec{n}}(m,d) - \sum_{d=1}^2 \sum_{m=0}^\infty C(m,d,a)\, \pi^{\vec{n}^{j-1}(\beta)}(m,d)}{\sum_{d=1}^2 \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) - \sum_{d=1}^2 \sum_{m=0}^{n_d^{j-1}(\beta)} \pi^{\vec{n}^{j-1}(\beta)}(m,d)} \tag{3.3.5}$$

$$= \min_{d=1,2} \inf_{n>\hat{n}_d^{j-1}} \frac{\sum_{m=0}^\infty C(m,d,a)\, p^{n,(d)}(m) - \sum_{m=0}^\infty C(m,d,a)\, p^{\hat{n}_d^{j-1},(d)}(m)}{\sum_{m=0}^{n} p^{n,(d)}(m) - \sum_{m=0}^{\hat{n}_d^{j-1}} p^{\hat{n}_d^{j-1},(d)}(m)} \tag{3.3.6}$$

$$= \frac{\sum_{m=0}^\infty C(m,d_0,a)\, p^{\hat{n}_{d_0}^j,(d_0)}(m) - \sum_{m=0}^\infty C(m,d_0,a)\, p^{\hat{n}_{d_0}^{j-1},(d_0)}(m)}{\sum_{m=0}^{\hat{n}_{d_0}^j} p^{\hat{n}_{d_0}^j,(d_0)}(m) - \sum_{m=0}^{\hat{n}_{d_0}^{j-1}} p^{\hat{n}_{d_0}^{j-1},(d_0)}(m)}, \tag{3.3.7}$$

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 3.8. *Let $|\mathcal{Z}| = 2$ and the transitions of the environment be scaled by β, $\beta r^{(dd')}$. Under the assumptions of Lemma 3.7, 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 3.4, we give an alternative proof for the above, which does not require to verify uniform integrability, see Proposition 3.19.

Proof of Proposition 3.8: The index $W(m_0,d_0)$ is as obtained from the Algorithm 3.4, 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

$$\lim_{\beta\to 0} W(m_0,d_0)$$
$$= \lim_{\beta\to 0} \inf_{\vec{n}\in E_j} \frac{\sum_{d=1}^2 \sum_{m=0}^\infty C(m,d,a)\, \pi^{\vec{n}}(m,d) - \sum_{d=1}^2 \sum_{m=0}^\infty C(m,d,a)\, \pi^{\vec{n}^{j-1}(\beta)}(m,d)}{\sum_{d=1}^2 \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) - \sum_{d=1}^2 \sum_{m=0}^{n_d^{j-1}(\beta)} \pi^{\vec{n}^{j-1}(\beta)}(m,d)}.$$

For d_0 it holds that $\hat{n}^j_{d_0} \neq \hat{n}^{j-1}_{d_0}$. Hence, by Lemma 3.7, the latter is equal to

$$\lim_{\beta\to 0} W(m_0, d_0) = \inf_{n>\hat{n}^{j-1}_{d_0}} \frac{\sum_{m=0}^{\infty} C(m, d_0, a)\, p^{n,(d_0)}(m) - \sum_{m=0}^{\infty} C(m, d_0, a)\, p^{\hat{n}^{j-1}_{d_0},(d_0)}(m)}{\sum_{m=0}^{n} p^{n,(d_0)}(m) - \sum_{m=0}^{\hat{n}^{j-1}_{d_0}} p^{\hat{n}^{j-1}_{d_0},(d_0)}(m)}. \tag{3.3.8}$$

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 3.4 to the bandit that always sees environment d_0. This concludes the proof. □

3.3.3 Asymptotic optimality of Whittle's index policy

In [71], 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 results 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, Whittle's index policy is then also asymptotically optimal. The result does not directly carry over to correlated environments. This is left as future research.

3.4 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. For ease of notation, we define $r^{(d)}_k := r^{(d,3-d)}_k$ 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 $\lambda^{(d)}_k$. They require an exponentially distributed amount of service with parameter $\mu^{(d)}_k$. A class-k job abandons the system after an exponentially distributed amount of time with parameter $\theta^{(d)}_k$. 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

$$C_k(m, d, 0) - C_k((m-1)^+, d, 0) \leq C_k(m+1, d, 1) - C_k(m, d, 1) \leq C_k(m+1, d, 0) - C_k(m, d, 0), \tag{3.4.1}$$

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 minimize the time-average holding cost, see (3.1.2).
Similar to [43], 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:

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

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 3.4, we first show in Section 3.4.1 that an optimal solution of the relaxed problem is of threshold type. Then, in Section 3.4.2, we obtain a closed-form expression for Whittle's index in the case of linear holding cost.

3.4.1 Threshold policies

In this section, we prove that an optimal solution of the relaxed problem (3.2.3) for the abandonment model is of threshold type. We henceforth focus on one bandit. For ease of exposition, we remove the subscript k.

Proposition 3.9. *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* (3.2.3).

Proof: The value function $V(m, d)$ satisfies the Bellman's optimality equation for average costs [57], which in this case is

$$\begin{aligned}
&(\mu^{(d)} + m\theta^{(d)} + \lambda^{(d)} + r^{(d)})V(m, d) + g = \\
&\lambda^{(d)}V(m + 1, d) + m\theta^{(d)}V((m - 1)^+, d) + r^{(d)}V(m, 3 - d) \\
&+ \min\left\{C(m, d, 0) - W + \mu^{(d)}V(m, d), C(m, d, 1) + \mu^{(d)}V((m - 1)^+, d)\right\},
\end{aligned} \tag{3.4.2}$$

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 (3.4.2), 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 (3.4.1) 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 3.7.2 we show that $V^L(m, d)$ is a convex function, and in Appendix 3.7.2 that sufficient conditions hold in order to apply [14, 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 $\vec{n}$.
Given threshold policy $\vec{n}$ with $n_i \geq 0$ for $i = 1, 2$, we have the following rate conservation property for a given environment.

Lemma 3.10. *Under threshold policy $\vec{n}$ it holds that*

$$\lambda^{(d)}\phi(d) \quad + \quad r^{(3-d)}\mathbb{E}\left(M^{\vec{n}}\mathbf{1}_{(D=3-d)}\right) = \left(\theta^{(d)} + r^{(d)}\right)\mathbb{E}\left(M^{\vec{n}}\mathbf{1}_{(D=d)}\right) + \mu^{(d)}\sum_{m=n_d+1}^{\infty}\pi^{\vec{n}}(m, d), \tag{3.4.3}$$

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

Proof: The rate conservation in a stable system is due to the balance between the arrivals and the departures in the long-term process. We refer to [20, Section 2.4.7], where the Principle of Set Balance states that almost surely, the number of exits equals the number of entrances as t tends to infinity. We consider that each environment defines a stable subsystem. Then, the mentioned balance is represented in both sides of Equation (3.4.3) for a given environment d. This can be explained as follows. On average, the rate for arrivals under environment d is $\lambda^{(d)}\phi(d)$, the rate for departures due to service is $\mu^{(d)}\sum_{m=n_d+1}^{\infty}\pi^{\vec{n}}(m,d)$, and due to abandonments is $\theta^{(d)}\mathbb{E}\left(M^{\vec{n}}\mathbf{1}_{(D=d)}\right)$. When the environment of the bandit is in state $D(t) = d$, the subsystem in environment d has all the customers, while the subsystem in environment $3-d$ is empty. When the environment changes state, a batch departure of every customer is done from one subsystem to the other. In this sense, for a fixed environment d, the term $r^{(3-d)}\mathbb{E}\left(M^{\vec{n}}\mathbf{1}_{(D=3-d)}\right)$ represents arrivals to the subsystem, and $r^{(d)}\mathbb{E}\left(M^{\vec{n}}\mathbf{1}_{(D=d)}\right)$ the departures. □

Lemma 3.11 proves monotonicity properties on $\sum_{m=0}^{n_d}\pi^{\vec{n}}(m,d)$. This quantity represents the probability that a class is not being served and observes environment d.

Lemma 3.11.

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 one is keeping this class passive in more states, while the drift towards these passive states remains the same. Hence, the probability of being passive increases. If instead n_{3-d} increases, the states where the bandit is passive in environment d are the same. However, the downward drift in environment $3-d$ decreases in some states, which explains that the probability of being passive in environment d decreases, i.e., the second property. The proofs can be found in Appendix 3.7.3.

3.4.2 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 3.4.3.

We introduce the following notation:

$$W^{(d)} := c\mu^{(d)}\frac{\theta^{(3-d)} + r^{(1)} + r^{(2)}}{\theta^{(1)}\theta^{(2)} + r^{(1)}\theta^{(2)} + r^{(2)}\theta^{(1)}}. \tag{3.4.4}$$

The following two conditions are used in order to prove indexability, Theorem 3.14. Numerical simulations however suggest that indexability holds for any set of parameters.

Condition 3.12.

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

Condition 3.13.

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

Theorem 3.14. *Assume Conditions 3.12 and 3.13 hold. Then all bandits are indexable.*

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

Theorem 3.15. *Assume the bandit is indexable, and w.l.o.g., that* $\frac{\mu^{(1)}}{\theta^{(1)}+r^{(1)}+r^{(2)}} \leq \frac{\mu^{(2)}}{\theta^{(2)}+r^{(1)}+r^{(2)}}$. *We define the subsequence* $(n_j)_{j\geq -1}$ *as follows:* $n_{-1}=-1, n_0=0$, *and for* $j\geq 1$ *let*

$$n_j := \min_{n>n_{j-1}} \overline{W}((n_{j-1},0),(n,0)). \tag{3.4.5}$$

In case there is more than one minimiser, 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*

$$W(m,d)=\begin{cases} 0 & \text{for } m=0 \\ \overline{W}((n_{j-1},0),(n_j,0)) & \text{for } n_{j-1}<m\leq n_j, \text{ and } d=1 \\ W^{(2)} & \text{for } m\geq 1 \text{ and } d=2. \end{cases} \tag{3.4.6}$$

Moreover, $W(m,1)\leq W^{(1)}\leq W^{(2)}$ *for all* m.
If $\frac{\mu^{(1)}}{\theta^{(1)}+r^{(1)}+r^{(2)}} = \frac{\mu^{(2)}}{\theta^{(2)}+r^{(1)}+r^{(2)}}$, *then* $W(m,1)=W^{(1)}=W^{(2)}$ *for all* m.

Remark 3.16. *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 3.23. As a consequence, in that case, for each step* j *in Theorem 3.15, 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 $j\geq -1$, or equivalently, the sequence $\overline{W}((n-1,0),(n,0))$ is strictly increasing in n.

Corollary 3.17. *If* $\frac{\mu^{(1)}}{\theta^{(1)}+r^{(1)}+r^{(2)}} \leq \frac{\mu^{(2)}}{\theta^{(2)}+r^{(1)}+r^{(2)}}$ *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* (3.4.7).

$$W(m,d)=\begin{cases} 0 & \text{for } m=0 \\ \overline{W}((m-1,0),(m,0)) & \text{for } d=1 \\ W^{(2)} & \text{for } d=2 \text{ and } m\geq 1. \end{cases} \tag{3.4.7}$$

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

Remark 3.18. *The queuing model with abandonments without an environment has been studied in [44]. 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 3.15 equals* $c\mu/\theta$, *which is in agreement with [44, Section 6.1].*

We now scale the transition rates of the environments, to further characterize Whittle's index in the two extreme cases of very slow changing or very fast changing environments. As in Section 3.3.2, β is the scaling parameter and $\beta r^{(d)}$ is the transition rate of the environment. This proof can be found in Appendix 3.7.4.

Proposition 3.19. *Assume the transition rates of the environment are scaled as $\beta r^{(d)}$. It holds that*

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

When $\dfrac{\mu^{(1)}}{\theta^{(1)}+r^{(1)}+r^{(2)}} \leq \dfrac{\mu^{(2)}}{\theta^{(2)}+r^{(1)}+r^{(2)}}$, *it holds that*

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

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's index when the bandit always sees environment d, given by $\frac{c\mu^{(d)}}{\theta^{(d)}}$, see [43].
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*, in Chapter 2.

Hence, we see that when the environment changes very fast, it takes the form of departure rate over the averaged abandonment rate. In case the environment is unobservable, the averaged value is taken in the index.

3.4.3 Proof of Theorems 3.14 and 3.15

In this section, we set out the proofs of Theorems 3.14 and 3.15. In order to do so, we state several lemmas and propositions, whose proofs can be found in Appendix 3.7.5.
Throughout this section, we assume that $W^{(1)} \leq W^{(2)}$, or equivalently $\frac{\mu^{(1)}}{\theta^{(1)}+r^{(1)}+r^{(2)}} \leq \frac{\mu^{(2)}}{\theta^{(2)}+r^{(1)}+r^{(2)}}$.
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 (3.3.2), we obtain

$$\overline{W}(\vec{n},\vec{n}') = c\cdot \frac{\sum_{d=1}^{2}\sum_{m=0}^{\infty} m\pi^{\vec{n}}(m,d) - \sum_{d=1}^{2}\sum_{m=0}^{\infty} m\pi^{\vec{n}'}(m,d)}{\sum_{d=1}^{2}\sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) - \sum_{d=1}^{2}\sum_{m=0}^{n'_d} \pi^{\vec{n}'}(m,d)}, \tag{3.4.8}$$

given that the denominator in (3.4.8) is not equal to 0.
The rate conservation property of Lemma 3.10 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 3.20. *It holds that*

$$\overline{W}(\vec{n},\vec{n}') \quad = \quad tW^{(1)} + (1-t)W^{(2)}, \tag{3.4.9}$$

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 thus

$$\overline{W}((\infty,n_2),(\infty,n_2')) = W^{(2)}. \tag{3.4.10}$$

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

Lemma 3.21. *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* $n_2 = \infty$.

We can now characterise optimal threshold policies that solve the relaxed optimisation problem for (3.3.1) as a function of the subsidy W.

Proposition 3.22.

1. *The threshold policy* $(-1,-1)$ *is an optimal solution of* (3.3.1) *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 3.12 and 3.13 hold. For* $W \in \left[0, W^{(2)}\right)$ *the optimal threshold solutions of* (3.3.1) *are of the form* $(n,0)$ *with* $n \geq -1$.

3. *The threshold policy* (∞,∞) *is an optimal solution of* (3.3.1) *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 3.23. *Assume Condition 3.12 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 3.14.

Proof of Theorem 3.14: By Proposition 3.9, an optimal solution of (3.3.1) 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

$$n_1(W) \leq n_1(W+\Delta) \;\text{ and }\; n_2(W) \leq n_2(W+\Delta), \tag{3.4.11}$$

with $\Delta > 0$, then, the bandit is indexable. Equation (3.4.11) will be proved below.
First assume $W < 0$. Then, by Proposition 3.22, 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 $W \leq 0$, (3.4.11) holds.
Now assume $0 \leq W < W^{(2)}$. If $W + \Delta > W^{(2)}$, then the inequality (3.4.11) is trivially true. Now assume $W + \Delta < W^{(2)}$. By Proposition 3.22, 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 3.1. Furthermore, in Proposition 3.23 we proved that the slope of the linear functions $g^{(n,0)}(W)$ is non-increasing in n. Therefore, $n \leq n'$, and (3.4.11) holds.
Finally, for $W \geq W^{(2)}$, the optimal threshold policy is (∞,∞), hence $n_d(W) = \infty$ for both $d = 1,2$, and (3.4.11) is trivially true. □

In order to prove Theorem 3.15, we first show that threshold policy $(\infty,0)$ is an optimal solution of (3.3.1) when $W \in \left[W^{(1)}, W^{(2)}\right]$.

Proposition 3.24. *The threshold policy $(\infty,0)$ is an optimal solution of (3.3.1) 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.*

Using Propositions 3.22 and 3.24 we can prove Theorem 3.15.
Proof of Theorem 3.15: Recall that the Whittle's 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 3.22, 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 3.24, 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 3.4 characterises Whittle's index by iteratively defining $\hat{W}_j$. From Proposition 3.22, 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 $\overline{W}((n_j,0),(n_{j-1},0)) \leq W^{(1)}$ for every j. To do so, recall from (3.4.9) that

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

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 3.11 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 minimises for $W \leq \overline{W}$. In this case, since $g^{(n_{j-1},0)}(W)$ minimises for $W \leq \overline{W}((n_j,0),(n_{j-1},0))$ and $g^{(n_j,0)}(W)$ minimises for $W \geq \overline{W}((n_j,0),(n_{j-1},0))$, $g^{(n_j,0)}$ 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, $\overline{W}((n_j,0),(n_{j-1},0)) \leq W^{(1)}$ for every j. □

3.5 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 3.4 where now the abandonment rates are set equal to zero. First, we discuss the stability conditions of the model. Secondly, we derive Whittle's index, based on the results obtained in the previous section for an abandonment queue (by letting $\theta \to 0$). The latter is used to propose an index policy.

3.5.1 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 [62].

Proposition 3.25 (Proposition 1, Section 4, in [62])**.** *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

$$\gamma_{k\vec{d}} \geq 0 \text{ and } \sum_{k=1}^{N} \gamma_{k\vec{d}} \leq \phi(\vec{d}), \text{ for all } k, \vec{d}, \tag{3.5.1}$$

and

$$\sum_{d\in\mathcal{Z}} \lambda_k^{(d)} \phi_k(d) \leq \sum_{\vec{d}\in\mathcal{Z}^N} \mu_k^{(d_k)} \gamma_{k\vec{d}}, \quad \forall 1 \leq k \leq N. \tag{3.5.2}$$

If there exists a vector $\vec{\gamma}$ such that (3.5.1) *and*

$$\sum_{d\in\mathcal{Z}} \lambda_k^{(d)} \phi_k(d) < \sum_{\vec{d}\in\mathcal{Z}^N} \mu_k^{(d_k)} \gamma_{k\vec{d}}, \quad \forall 1 \leq k \leq N, \tag{3.5.3}$$

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 $\gamma^{\varphi}_{kd\vec{}}$ be the average fraction of time that the environment is $\vec{d}$ and bandit k is active. The obtained vector $\vec{\gamma}^{\varphi}$ satisfies (3.5.1) by definition. It also satisfies Equation (3.5.2), 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 (3.5.1) and (3.5.3) 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 (3.5.3) is larger than its arrival rate. This implies stability. □

3.5.2 Whittle's index policy

For the multi-class queue without abandonments, one cannot directly apply Algorithm 3.4 in order to get Whittle's index. The reason for this is as follows. Let us focus on one bandit, and assume $\mu^{(1)} < \mu^{(2)}$. Then, we believe that as the subsidy grows from $-\infty$ to ∞, threshold policies of the form $(n, 0)$, $n = 1, 2, \ldots$, will be optimal, and then, for larger 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 [44, 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's index in the presence of abandonments (as derived in Section 3.4.2), with $\theta_1 = \theta_2 = \theta$, with $\theta > 0$. It is direct from Theorem 3.15 that if $\mu^{(1)} < \mu^{(2)}$, then

$$\lim_{\theta\to 0} \theta W^{\theta}(m, 2) = c\mu^{(2)}, \text{ for all } m \geq 1. \tag{3.5.4}$$

That is, in environment 2, one needs to consider the scaled index. In environment 1, we observed numerically that no scaling was needed in order to obtain a non-trivial limit. In fact, we believe the following to be true:

Conjecture 3.26.

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

In order to prove this conjecture, one needs to study $\overline{W}((n_{j-1}, 0), (n_j, 0))$, which by (3.4.9) depends on the steady-state distributions. A perturbation approach as presented in [3, Theorem 2], would allow to write the steady state as an expansion in θ. However, the results of [3] 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 $\lim_{\theta\to 0}\overline{W}_k^{\theta}(m, 3-d_k)$ as defined in Conjecture 3.26.

3.6 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., $C_k(m, d, k) = m$.

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 3.6.1, we generate a large number of parameters and using boxplots we show the suboptimality gaps under different policies. In Section 3.6.2 we focus on one particular set of parameters. In Section 3.6.3, we discuss how much one can gain by observing the state of the environments.

3.6.1 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 at random which bandit to serve 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 3.2 (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

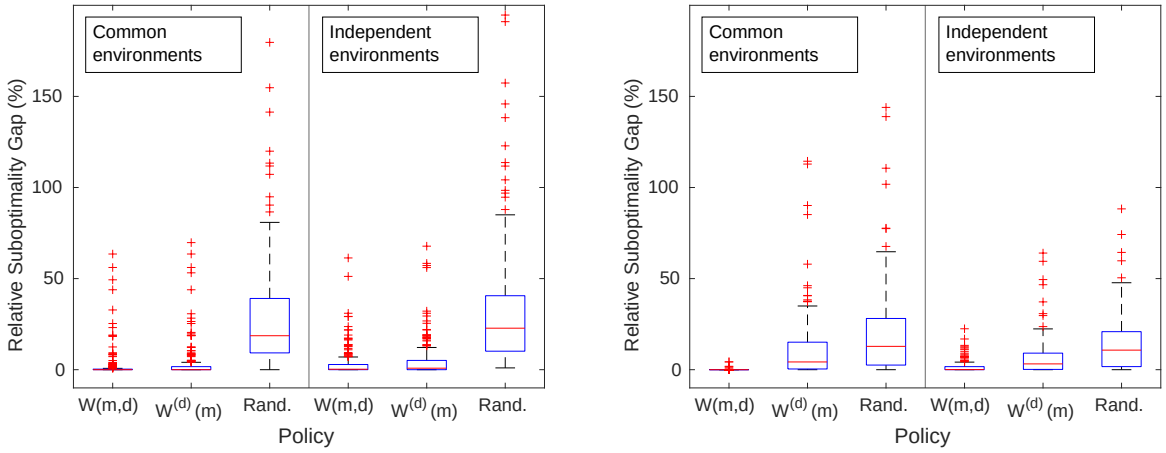

Figure 3.2: Suboptimality gap with (left) random parameters and (right) constrained parameters.

$d = 1$, 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 3.2 (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 (vs. 0% both values for Whittle's index policy), and the median is 3% and the 75th percentile is 9% for the independent environments setting (vs. 0% and 2% for Whittle's index policy).

3.6.2 Particular example

From the above plots, 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

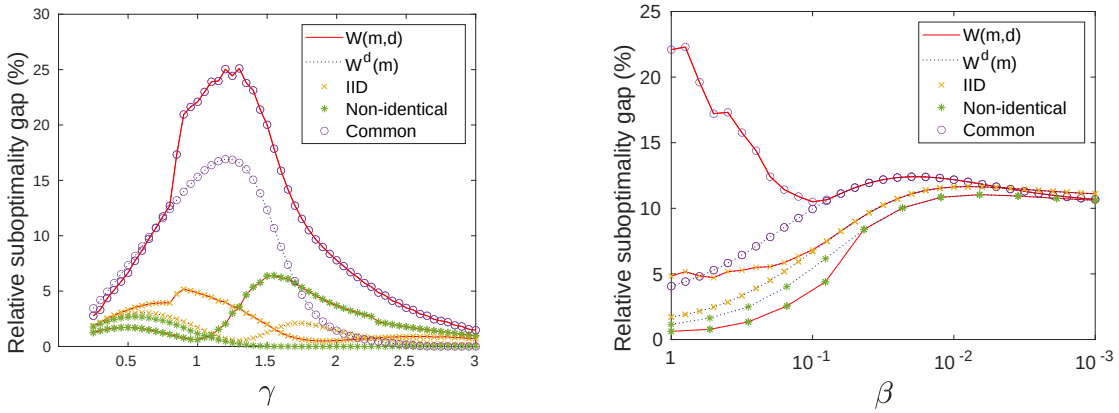

Figure 3.3: Suboptimality gap of Whittle's index policy and the index $W^{(d)}(m)$ as a function of (left) the scaling parameter of arrival rates (γ) and (right) the scaling parameter of the transition rates for the environment (β).

his 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".

- 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 3.12 and Condition 3.13, 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.

Scaling arrivals

We first study the suboptimality gap as the load in the system changes. We set $\beta = 1$. In Figure 3.3 (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%.

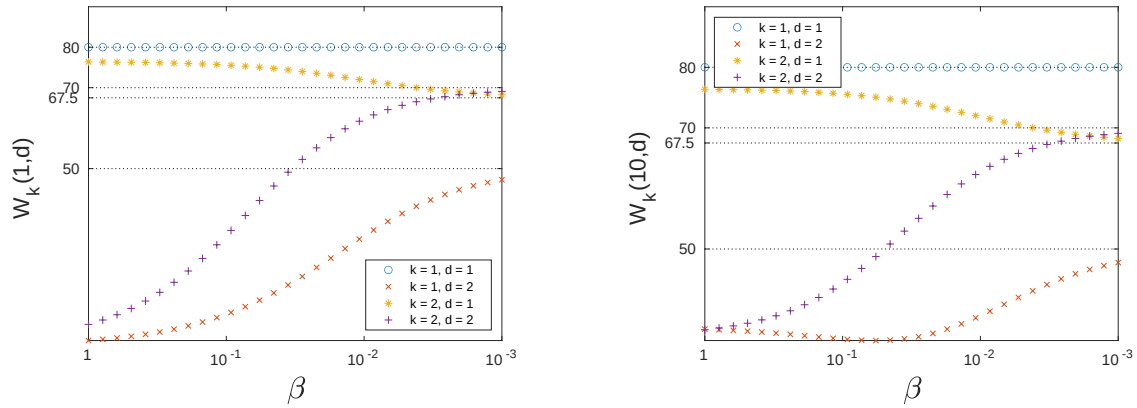

Figure 3.4: Whittle's index as a function of the scaling parameter of the transition rates for the environment. (left) state $m = 1$, (right) state $m = 10$.

Scaling speed of the environments

We now study the suboptimality gap as the speed of the environments changes.
In Figure 3.3, (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 3.8 that Whittle's index $W(m, d)$ converges as $\beta \to 0$ to $W^{(d)}(m)$. This explains why in Figure 3.3, (right) the performance of Whittle's index policy converges towards the fixed Whittle's index policy. 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 3.4, we plot $W_k(1, d)$, $W_k(10, d)$ and the values $c\mu_k^{(d)}/\theta_k^{(d)}$ and see the convergence.

3.6.3 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 Chapter 2 it was shown that the averaged Whittle index policy is 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 Chapter 2 together with [44], 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 $\overline{\mu}_k := \sum_{d \in \mathcal{Z}} \phi_k(d)\mu_k^{(d)}$.

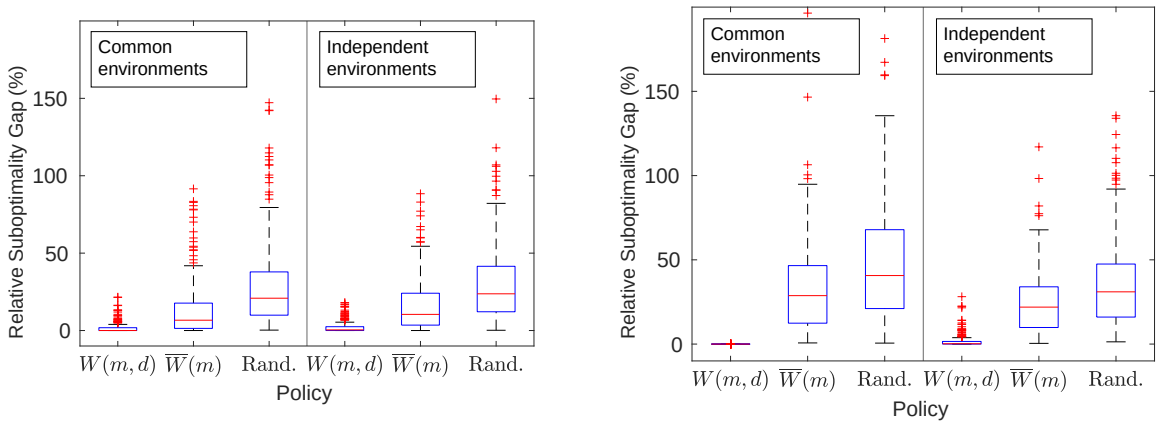

Figure 3.5: Suboptimality gap for unobservable policies with (left) random parameters and (right) constrained parameters.

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 3.5 (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 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 3.5 (right), we narrow our set of randomly generated parameters. In particular, we choose parameters such that the optimal action depends on the state the environment is in. 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 3.5 (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 3.5, 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 we proved in Chapter 2 that the averaged Whittle's index policy $\overline{W}(m)$ is optimal in a rapidly varying unobservable environment. In the boxplots of Figure 3.6 the parameters are chosen as in Figure 3.5, except for the rates of the environment, where we take

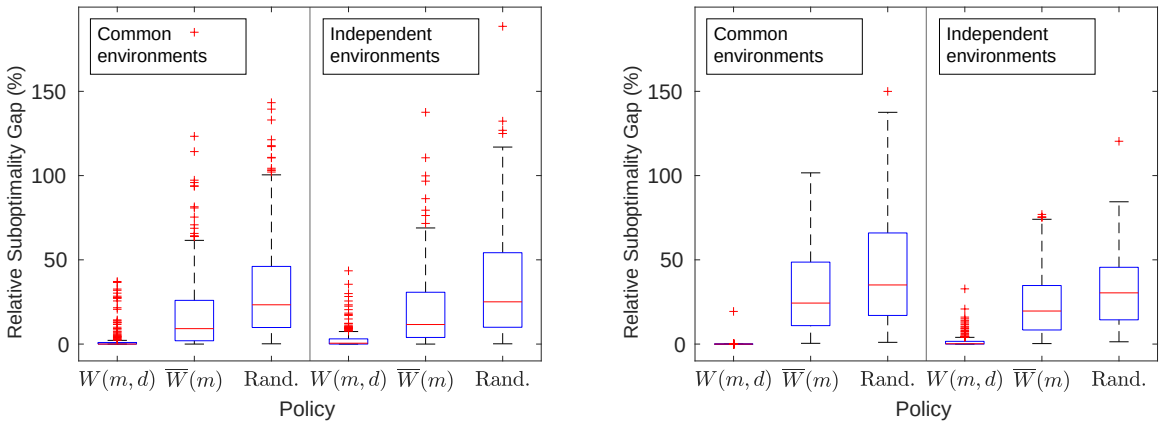

Figure 3.6: Suboptimality gap for unobservable policies with a fast environment, with (left) random parameters and (right) constrained parameters.

$r_k^{(d)} = r^{(d)} \in [100,\ 2000]$. 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. We obtained medians above 7% when the parameters are chosen at random, and above 20% when the optimal policy depends strongly on the state of the environment.

3.7 Appendix

3.7.1 Proof of Lemmas of Section 3.3.2.

Proof of Lemma 3.6:

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:

$$\begin{aligned} &\left(\lambda^{(d)} + m\theta^{(d)} + \mathbf{1}_{(m>n_d)}\mu^{(d)}\right) p^{n_d,(d)}(m) \\ &= \mathbf{1}_{(m>0)}\lambda^{(d)} p^{n_d,(d)}(m-1) + \left[(m+1)\theta^{(d)} + \mathbf{1}_{(m+1>n_d)}\mu^{(d)}\right] p^{n_d,(d)}(m+1), \end{aligned} \tag{3.7.1}$$

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

$$\begin{aligned} &(\lambda^{(d)} + m\theta^{(d)} + \mathbf{1}_{(m>n_d)}\mu^{(d)} + \beta r^{(d)})\pi^{\vec{n}}(m,d) \\ &= \mathbf{1}_{(m>0)}\lambda^{(d)}\pi^{\vec{n}}(m-1,1) + \left[(m+1)\theta^{(d)} + \mathbf{1}_{(m+1>n_d)}\mu^{(d)}\right]\pi^{\vec{n}}(m+1,d) \\ &\quad + \sum_{d' \neq d} \beta r^{(dd')}\pi^{\vec{n}}(m,d'), \end{aligned} \tag{3.7.2}$$

for all $m \geq 0$, d. The stationary probability measure that satisfies the balance equations is unique. As $\beta \to 0$, (3.7.2) is equal to (3.7.1) 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 $m \geq 0$. □

Proof of Lemma 3.7:

Recall that for a given β, $\vec{n}^j(\beta)$, $j = 0, \ldots,$ is the minimisation vector obtained in (3.3.3), 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

$$f^j(\vec{n}) := \lim_{\beta\to 0} \frac{\sum_{d=1}^{2}\sum_{m=0}^{\infty} C(m,d,a)\,\pi^{\vec{n}}(m,d) - \sum_{d=1}^{2}\sum_{m=0}^{\infty} C(m,d,a)\,\pi^{\vec{n}^j(\beta)}(m,d)}{\sum_{d=1}^{2}\sum_{m=0}^{n_d}\pi^{\vec{n}}(m,d) - \sum_{d=1}^{2}\sum_{m=0}^{n_d^j(\beta)}\pi^{\vec{n}^j(\beta)}(m,d)}, \tag{3.7.3}$$

for $\vec{n} \in E_j$. By Lemma 3.6 and since $\vec{M}^{\vec{n}(\beta)}$ is uniform integrable, we have that $f^j(\vec{n})$ is equal to

$$\frac{\sum_{d=1}^{2}\sum_{m=0}^{\infty} C\left(m,d,a\right)\phi(d)p^{n_d,(d)}(m) - \sum_{d=1}^{2}\sum_{m=0}^{\infty} C\left(m,d,a\right)\phi(d)p^{\hat{n}_d^j,(d)}}{\sum_{d=1}^{2}\sum_{m=0}^{n_d}\phi(d)p^{n_d,(d)} - \sum_{d=1}^{2}\sum_{m=0}^{\hat{n}_d^j}\phi(d)p^{\hat{n}_d^j,(d)}}.$$

Hence, equivalently, we can write $f^j(\vec{n}) = \dfrac{\mathbf{1}_{(n_1>\hat{n}_1^j)}\phi(1)\hat{A}_1^j(n_1) + \mathbf{1}_{(n_2>\hat{n}_2^j)}\phi(2)\hat{A}_2^j(n_2)}{\mathbf{1}_{(n_1>\hat{n}_1^j)}\phi(1)\hat{B}_1^j(n_1) + \mathbf{1}_{(n_2>\hat{n}_2^j)}\phi(2)\hat{B}_2^j(n_2)}$.
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 minimized in $\vec{\hat{n}}^{j+1}$, that is,

$$f^j(\vec{\hat{n}}^{j+1}) = \inf_{\vec{n}\in E_j} f^j(\vec{n}). \tag{3.7.4}$$

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

$$\frac{\phi(1)\hat{A}_1^j(\hat{n}_1^{j+1}) + \phi(2)\hat{A}_2^j(\hat{n}_2^{j+1})}{\phi(1)\hat{B}_1^j(\hat{n}_1^{j+1}) + \phi(2)\hat{B}_2^j(\hat{n}_2^{j+1})} \leq \min\left(\frac{\hat{A}_1^j(\hat{n}_1^{j+1})}{\hat{B}_1^j(\hat{n}_1^{j+1})}, \frac{\hat{A}_2^j(\hat{n}_2^{j+1})}{\hat{B}_2^j(\hat{n}_2^{j+1})}\right). \tag{3.7.5}$$

Now, assume there is a strict inequality in (3.7.5), and assume $\dfrac{\hat{A}_1^j(\hat{n}_1^{j+1})}{\hat{B}_1^j(\hat{n}_1^{j+1})} \leq \dfrac{\hat{A}_2^j(\hat{n}_2^{j+1})}{\hat{B}_2^j(\hat{n}_2^{j+1})}$. The strict inequality in (3.7.5) implies

$$\hat{B}_1^j(\hat{n}_1^{j+1})\hat{A}_2^j(\hat{n}_2^{j+1}) < \hat{A}_1^j(\hat{n}_1^{j+1})\hat{B}_2^j(\hat{n}_2^{j+1}), \quad \text{that is,} \quad \frac{\hat{A}_2^j(\hat{n}_2^{j+1})}{\hat{B}_2^j(\hat{n}_2^{j+1})} < \frac{\hat{A}_1^j(\hat{n}_1^{j+1})}{\hat{B}_1^j(\hat{n}_1^{j+1})}. \tag{3.7.6}$$

This gives contradiction with the assumption that $\dfrac{\hat{A}_1^j(\hat{n}_1^{j+1})}{\hat{B}_1^j(\hat{n}_1^{j+1})} \leq \dfrac{\hat{A}_2^j(\hat{n}_2^{j+1})}{\hat{B}_2^j(\hat{n}_2^{j+1})}$. Hence, by contradiction we proved that the inequality in (3.7.5) is an equality, that is,

$$f^j(\vec{\hat{n}}^{j+1}) = \min(f^j(\hat{n}_1^{j+1},\hat{n}_2^j), f^j(\hat{n}_1^j,\hat{n}_2^{j+1})). \tag{3.7.7}$$

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

$$f^j(\hat{n}_1^{j+1},\hat{n}_2^j) = \inf_{n>\hat{n}_1^j} \frac{\sum_{m=0}^{\infty} C\left(m,1,a\right) p^{n,(1)}(m) - \sum_{m=0}^{\infty} C\left(m,1,a\right) p^{\hat{n}_1^j,(1)}(m)}{\sum_{m=0}^{n} p^{n,(1)}(m) - \sum_{m=0}^{\hat{n}_1^j} p^{\hat{n}_1^j,(1)}(m)}. \tag{3.7.8}$$

Let n_1^* be the n such that the infimum is taken on the RHS. Hence, the RHS can equivalently be written as $f^j(n_1^*,\hat{n}_2^j)$. We prove (3.7.8) by contradiction. That is, assume $f^j(\hat{n}_1^{j+1},\hat{n}_2^j) > f^j(n_1^*,\hat{n}_2^j)$. But since $f^j(\vec{\hat{n}}^{j+1}) = f^j(\hat{n}_1^{j+1},\hat{n}_2^j) > f^j(n_1^*,\hat{n}_2^j)$, we have contradiction with the fact that f^j is minimized in $\vec{\hat{n}}^{j+1}$.
Combining (3.7.4), (3.7.7), and (3.7.8), we conclude the proof. □

3.7.2 Proof of Proposition 3.9:

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

$$q^L(m+1|m,d,a) := \lambda^{(d)} \left(1 - \frac{m}{L}\right)^+, \tag{3.7.9}$$

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

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 uniformisation 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

$$2V_t^L(m,d) \leq V_t^L((m-1)^+, d) + V_t^L(m+1, d), \text{ for } 0 \leq m \leq L-1, \tag{3.7.10}$$

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

$$V_{t+1}^L(m+1,d) - V_{t+1}^L(m,d) \geq 0 \text{ for } 0 \leq m \leq L-1. \tag{3.7.11}$$

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 (3.7.11), i.e., V_t^L is non-decreasing.
We consider now equation (3.7.10). 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 (3.7.10) for $t+1$. Note that

$$\begin{aligned} 2V_{t+1}^L(m,d) =& 2\left(1-\frac{m}{L}\right)\lambda^{(d)}V_t^L(m+1,d) + 2\frac{m}{L}\lambda^{(d)}V_t^L(m,d) \\ &+ 2r^{(d)}V_t^L(m,3-d) \\ &+ 2m\theta^{(d)}V_t^L(m-1,d) + 2(L-m)\theta^{(d)}V_t^L(m,d) \\ &+ 2\min\{-W + C(m,d,0) + \mu^{(2)}V_t^L(m,d), C(m,d,1) + \mu^{(2)}V_t^L(m-1,d)\} \\ &+ 2\left(\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}\right)V_t^L(m,d) \end{aligned} \tag{3.7.12}$$

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

$$\begin{aligned} &\lambda^{(d)}\left(1-\frac{m-1}{L}\right)V_t^L(m,d) + \lambda^{(d)}\left(1-\frac{m+1}{L}\right)V_t^L(\min\{m+2,L\},d) \\ &+ \lambda^{(d)}\frac{m-1}{L}V_t^L(m-1,d) + \lambda^{(d)}\frac{m+1}{L}V_t^L(m+1,d) \\ &+ r^{(d)}V_t^L(m-1,3-d) + r^{(3-d)}V_t^L(m-1,d) + r^{(d)}V_t^L(m+1,3-d) + r^{(3-d)}V_t^L(m+1,d) \\ &+ (m-1)\theta^{(d)}V_t^L((m-2)^+,d) + (m+1)\theta^{(d)}V_t^L(m,d) \\ &+ (L-m+1)\theta^{(d)}V_t^L(m-1,d) + (L-m-1)\theta^{(d)}V_t^L(m+1,d) \\ &+ \min\{-W + C(m-1,d,0) + \mu^{(2)}V_t^L(m-1,d), C(m-1,d,1) + \mu^{(2)}V_t^L((m-2)^+,d)\} \\ &+ \min\{-W + C(m+1,d,0) + \mu^{(2)}V_t^L(m+1,d), C(m+1,d,1) + \mu^{(2)}V_t^L(m,d)\} \\ &+ \left(\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}\right)V_t^L(m-1,d) \\ &+ \left(\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}\right)V_t^L(m+1,d). \end{aligned} \tag{3.7.13}$$

We first compare the two terms multiplied by $\lambda^{(d)}$ in (3.7.12) to check they are smaller than or equal to the ones multiplied by $\lambda^{(d)}$ in (3.7.13). First assume $1 \leq m \leq L-2$. The terms multiplied by $\lambda^{(d)}$ in (3.7.12) are

$$
\begin{aligned}
&2\left(1-\frac{m}{L}\right)V_t^L(m+1,d)+2\frac{m}{L}V_t^L(m,d)\\
&=2\left(1-\frac{m+1}{L}\right)V_t^L(m+1,d)+2\frac{m}{L}V_t^L(m,d)+\frac{2}{L}V_t^L(m,d)\\
&\le\left(1-\frac{m+1}{L}\right)V_t^L(m,d)+\left(1-\frac{m+1}{L}\right)V_t^L(m+2,d)+2\frac{m}{L}V_t^L(m,d)+\frac{2}{L}V_t^L(m+1,d)\\
&=\left(1-\frac{m-1}{L}\right)V_t^L(m,d)-\frac{2}{L}V_t^L(m,d)+\left(1-\frac{m+1}{L}\right)V_t^L(m+2,d)\\
&\quad+2\frac{m}{L}V_t^L(m,d)+\frac{2}{L}V_t^L(m+1,d)\\
&=\left(1-\frac{m-1}{L}\right)V_t^L(m,d)+\left(1-\frac{m+1}{L}\right)V_t^L(m+2,d)\\
&\quad+2\frac{m-1}{L}V_t^L(m,d)+\frac{2}{L}V_t^L(m+1,d)
\end{aligned}
\tag{3.7.14}
$$

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 (3.7.14) can be bounded, and we obtain:

$$
\begin{aligned}
&2\left(1-\frac{m}{L}\right)V_t^L(m+1,d)+2\frac{m}{L}V_t^L(m,d)\\
&\le\left(1-\frac{m-1}{L}\right)V_t^L(m,d)+\left(1-\frac{m+1}{L}\right)V_t^L(m+2,d)\\
&\quad+\frac{m-1}{L}\left(V_t^L(m-1,d)+V_t^L(m+1,d)\right)+\frac{2}{L}V_t^L(m+1,d)\\
&\le\left(1-\frac{m-1}{L}\right)V_t^L(m,d)+\left(1-\frac{m+1}{L}\right)V_t^L(m+2,d)\\
&\quad+\frac{m-1}{L}V_t^L(m-1,d)+\frac{m+1}{L}V_t^L(m+1,d),
\end{aligned}
\tag{3.7.15}
$$

which is the same as the terms multiplied by $\lambda^{(d)}$ in (3.7.13). Now assume $m=L-1$, then inequality (3.7.10) 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 (3.4.1), 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 (3.7.10) for $t+1$, hence function V_t^L is convex. By [57, 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. □

Hypothesis needed for [14, 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 [14, 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 [14, Theorem 3.1], we need the following definition.

Definition 3.27. *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 [14, 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

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

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

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

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

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

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

$$\lambda^{(d)}\left(1-\frac{m}{L}\right)^+ (e^{\epsilon}-1) + (m\theta^{(d)} + a\mu^{(d)})(e^{-\epsilon}-1) \le -\alpha, \tag{3.7.16}$$

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 (3.7.16) 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. □

3.7.3 Proof of Lemma 3.11

Proof of Property 1) in Lemma 3.11:

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 [19, 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

$$G^{(n_1,n_2)}\leq G^{(n_1+1,n_2)}. \tag{3.7.17}$$

Since the result 9.3.2 in [19, Chapter 9] requires a uniformisable 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 uniformisation 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 uniformisation approach [68, 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 [19, Chapter 9], we will prove that

$$G^{(n_1,n_2),L} \leq G^{(n_1+1,n_2),L}. \tag{3.7.18}$$

In order to apply result 9.3.2 in [19, Chapter 9], we need to prove that for all states (m,d) and $t \geq 0$,

$$\begin{aligned} & C^{(n_1+1,n_2)}(m,d,a) - C^{(n_1,n_2)}(m,d,a) \\ & + \sum_{(m',d')} \left[q^{(n_1+1,n_2),L}((m',d')|(m,d),a) - q^{(n_1,n_2),L}((m',d')|(m,d),a)\right] \\ & \qquad \cdot \left[V_t^{(n_1,n_2),L}(m',d') - V_t^{(n_1,n_2),L}(m,d)\right] \\ & \geq 0. \end{aligned} \tag{3.7.19}$$

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 (3.7.19) 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,

$$V_t^{(n_1,n_2),L}(n_1,1) - V_t^{(n_1,n_2),L}(n_1+1,1) \leq \frac{1}{\mu^{(1)}}. \tag{3.7.20}$$

By induction, we can prove the following more general result. For ease of reading, its proof appears in Appendix 3.7.3.

Lemma 3.28.

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

With this Lemma, the proof for the truncated processes is done and (3.7.18) holds. To generalize this for the original processes with unbounded rates we use the following result from [26, 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 (3.7.18), 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 (3.7.17) and the proof is done. □

Proof of Lemma 3.28.

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. □

Proof of Property 2) in Lemma 3.11:

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. □

3.7.4 Proof of Proposition 3.19:

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 3.15, 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 Lemma 3.20 that $\overline{W}((n_{j-1},0),(n_j,0))$ can be 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 3.6 we can deduce that for any $d \in \mathcal{Z}$ and any pair of policies $\vec{n}, \vec{n}'$ such that $n_d = n'_d$,

$$\lim_{\beta\to 0} \pi^{\vec{n}}(m,d) = \lim_{\beta\to 0} \pi^{\vec{n}'}(m,d) \quad \forall m \geq 0. \tag{3.7.21}$$

In particular,

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

Then, following from (3.4.9), $\lim_{\beta\to 0} \overline{W}((n_{j-1},0),(n_j,0)) = W^{(1)} = c\frac{\mu^{(1)}}{\theta^{(1)}}$, 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 $d = 2$. □

3.7.5 Proof of results in Section 3.4.3.

Proof of Lemma 3.20: 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 3.10 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

$$\begin{aligned}\sum_{m=0}^{\infty}m\pi^{\vec{n}}(m,d) \quad = \quad & \left(\left(\lambda^{(d)}-\mu^{(d)}\right)\phi(d)\left(\theta^{(3-d)}+r^{(3-d)}\right)+\left(\lambda_{3-d}-\mu^{(3-d)}\right)\phi(3-d)r^{(3-d)}\right.\\ & \left.+\mu^{(3-d)}r^{(3-d)}\sum_{m=0}^{n_{3-d}}\pi^{\vec{n}}(m,3-d)+\mu^{(d)}\left(\theta^{(3-d)}+r^{(3-d)}\right)\sum_{m=0}^{n_d}\pi^{\vec{n}}(m,d)\right)\\ & \cdot\frac{1}{\theta^{(1)}\theta^{(2)}+\theta^{(2)}r^{(1)}+\theta^{(1)}r^{(2)}},\end{aligned} \tag{3.7.22}$$

for $d=1,2$. In the numerator in (3.4.8), the terms can be regrouped per environment, i.e.

$$\sum_{d=1}^{2}c\left(\sum_{m=0}^{\infty}m\pi^{\vec{n}}(m,d)-\sum_{m=0}^{\infty}m\pi^{\vec{n}'}(m,d)\right). \tag{3.7.23}$$

From (3.7.22), 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

$$\sum_{m=0}^{\infty}m\pi^{\vec{n}}(m,d)-\sum_{m=0}^{\infty}m\pi^{\vec{n}'}(m,d)=\frac{\mu^{(d)}\left(\theta^{(3-d)}+r^{(3-d)}\right)s_d(\vec{n},\vec{n}')+\mu^{(3-d)}r^{(3-d)}s_{3-d}(\vec{n},\vec{n}')}{\theta^{(1)}\theta^{(2)}+\theta^{(2)}r^{(1)}+\theta^{(1)}r^{(2)}}. \tag{3.7.24}$$

Summing we obtain the following expression for (3.7.23):

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

Since the denominator in (3.4.8) is $s_1(\vec{n},\vec{n}')+s_2(\vec{n},\vec{n}')$, Lemma 3.20 is proved. □

Proof of Lemma 3.21: comparison to policy (∞,∞).

Since $\sum_{m=0}^{n_i}\pi^{\vec{n}}(m,i)\leq\phi(i)$ for $i=1,2$, it follows that $s_i((\infty,\infty),\vec{n})=\sum_{m=0}^{\infty}\pi^{(\infty,\infty)}(m,i)-\sum_{m=0}^{n_i}\pi^{\vec{n}}(m,i)=\phi(i)-\sum_{m=0}^{n_i}\pi^{\vec{n}}(m,i)\geq 0$. This implies that $0\leq t\leq 1$ as defined in Lemma 3.20, hence together with (3.4.9), it follows that $\overline{W}((\infty,\infty),\vec{n})\in[W^{(1)},W^{(2)}]$.
Note that $\sum_{m=0}^{n_1}\pi^{\vec{n}}(m,1)+\sum_{m=n_1+1}^{\infty}\pi^{\vec{n}}(m,1)=\phi(1)$ and $\sum_{m=n_1+1}^{\infty}\pi^{\vec{n}}(m,1)>0$ for $n_1<\infty$. As a consequence, $\sum_{m=0}^{n_1}\pi^{\vec{n}}(m,1)<\phi(1)$, so $s_1((\infty,\infty),\vec{n})=\phi(1)-\sum_{m=0}^{n_1}\pi^{\vec{n}}(m,1)>0$ and $t>0$. Equation (3.4.9) 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. □

Proof of Proposition 3.22

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 (3.3.1) 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 3.7.5.

Lemma 3.29. *Assume Conditions 3.12 and 3.13 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 3.29, 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 3.21 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 (3.3.1) 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 3.21). Hence, when $W=W^{(2)}$, optimal threshold policies are of the form (∞,n). □

Proof of Lemma 3.29.

We want to show the optimal policies in environment $d = 2$ for $0 \leq W < W^{(2)}$. Recall from (3.4.2) that Bellman's optimality equation is given by:

$$\begin{aligned}&(\mu^{(d)} + m\theta^{(d)} + \lambda^{(d)} + r^{(d)})V(m,d) + g = \\ &cm + \lambda^{(d)}V(m+1,d) + m\theta^{(d)}V((m-1)^+,d) + r^{(d)}V(m,3-d) \\ &+ \min\left\{-W + \mu^{(d)}V(m,d), \mu^{(d)}V((m-1)^+,d)\right\}.\end{aligned} \tag{3.7.25}$$

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,

$$V(m,2) - V(m-1,2) \geq W/\mu^{(2)}, \quad \forall m \geq 1. \tag{3.7.26}$$

Since we will use the value iteration method, we formulate an analogous property for the truncated value function V^L, see Lemma 3.30. We truncate the state space at L and smooth the transition rates from state m to state $m+1$ as in (3.7.9). 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 uniformisation constant is taken as

$$\gamma := \lambda^{(1)} + \lambda^{(2)} + \mu^{(1)} + \mu^{(2)} + r^{(1)} + r^{(2)} + L\theta^{(1)} + L\theta^{(2)}, \tag{3.7.27}$$

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 3.30. *Assume Conditions 3.12 and 3.13 hold. For* $0 \leq W < W^{(2)}$ *there exists an* L_0 *large enough such that*

$$V^L(m,2) - V^L(m-1,2) \geq W/\mu^{(2)} \quad \forall L \geq L_0 \text{ and } 1 \leq m \leq L. \tag{3.7.28}$$

We can conclude now the proof of Lemma 3.29. From Lemma 3.30 there exists an L_0 such that inequality (3.7.28) holds for all $L \geq L_0$. Since $V^L \to V$, as stated in Section 3.7.2, inequality (3.7.26) is proved, which concludes the proof. □

Proof of Lemma 3.30:

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 $t \geq t_0$,

$$C_t^{L,(2)}(m) \geq W/\mu^{(2)} \quad \text{for } 1 \leq m \leq L, \tag{3.7.29}$$

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

$$\begin{aligned} C_{t+1}^{L,(d)}(m) &= V_{t+1}^L(m,d) - V_{t+1}^L(m-1,d) \qquad (3.7.30)\\ &= \frac{c}{\gamma} + \left(1 - \frac{m}{L}\right)^+ \frac{\lambda^{(d)}}{\gamma} C_t^{L,(d)}(m+1) + \frac{r^{(d)}}{\gamma} C_t^{L,(3-d)}(m) + \frac{(m-1)\theta^{(d)}}{\gamma} C_t^{L,(d)}(m-1)\\ &\quad + \frac{1}{\gamma}\min\left\{-W + \mu^{(d)} V_t(m,d), \mu^{(d)} V_t(m-1,d)\right\}\\ &\quad - \frac{1}{\gamma}\min\left\{-W + \mu^{(d)} V_t(m-1,d), \mu^{(d)} V_t((m-2)^+,d)\right\}\\ &\quad + \frac{m-1}{L}\frac{\lambda^{(d)}}{\gamma} C_t^{L,(d)}(m) + \frac{(L-m)\theta^{(d)}}{\gamma} C_t^{L,(d)}(m)\\ &\quad + \frac{\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}}{\gamma} C_t^{L,(d)}(m). \end{aligned}$$

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

Lemma 3.31. *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 3.12, 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 3.32. *Assume Condition 3.12 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 (3.7.29) holds in $t = t_0$ for all $L \geq L_1$. Note that for $t > t_0$, Condition 3.13 will be sufficient to prove that (3.7.29) holds for $t \geq t_0$ and for all m.

Lemma 3.33. *Assume Condition 3.13 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 3.33 proves that (3.7.29) holds for $0 \leq W < W^{(2)}$ and $L \geq L_0$. Equivalently, (3.7.28) holds and the proof of Lemma 3.30 is finished. □

We provide now the proofs of Lemmas 3.31, 3.32 and 3.33.

Proof of Lemma 3.31: 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 (3.7.30) for $m \geq 1$ and for $W \geq 0$ passive is also optimal for $m = 0$. Hence,

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

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

$$\begin{aligned} C_{t+1}^{L,(d)}(m) &= \frac{c}{\gamma} + \left(1 - \frac{m}{L}\right)\frac{\lambda^{(d)}}{\gamma}C_t^{L,(d)} + \frac{r^{(d)}}{\gamma}C_t^{L,(3-d)} + \frac{(m-1)\theta^{(d)}}{\gamma}C_t^{L,(d)} \\ &\quad + \frac{\mu^{(d)}}{\gamma}C_t^{L,(d)} + \frac{m-1}{L}\frac{\lambda^{(d)}}{\gamma}C_t^{L,(d)} + \frac{(L-m)\theta^{(d)}}{\gamma}C_t^{L,(d)} \\ &\quad + \frac{\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}}{\gamma}C_t^{L,(d)} \\ &= \frac{c}{\gamma} + \left(1 - \frac{1}{L}\right)\frac{\lambda^{(d)}}{\gamma}C_t^{L,(d)} + \frac{r^{(d)}}{\gamma}C_t^{L,(3-d)} + \frac{\mu^{(d)}}{\gamma}C_t^{L,(d)} \\ &\quad + \frac{(L-1)\theta^{(d)}}{\gamma}C_t^{L,(d)} + \frac{\lambda^{(3-d)} + \mu^{(3-d)} + r^{(3-d)} + L\theta^{(3-d)}}{\gamma}C_t^{L,(d)} \\ &= \frac{c}{\gamma} + \frac{\gamma - \theta^{(d)} - r^{(d)} - \lambda^{(d)}/L}{\gamma}C_t^{L,(d)} + \frac{r^{(d)}}{\gamma}C_t^{L,(3-d)}, \end{aligned} \tag{3.7.31}$$

where the last inequality follows from the definition of γ, (3.7.27). 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 (3.7.31), $C^{L,(1)}$ and $C^{L,(2)}$ are the solutions of the following system of equations.

$$\begin{cases} C^{L,(1)} = \dfrac{c}{\gamma} + \dfrac{\gamma - \theta^{(1)} - r^{(1)} - \frac{\lambda^{(1)}}{L}}{\gamma} C^{L,(1)} + \dfrac{r^{(1)}}{\gamma} C^{L,(2)} \\ C^{L,(2)} = \dfrac{c}{\gamma} + \dfrac{\gamma - \theta^{(2)} - r^{(2)} - \frac{\lambda^{(2)}}{L}}{\gamma} C^{L,(2)} + \dfrac{r^{(2)}}{\gamma} C^{L,(1)} \end{cases} \tag{3.7.32}$$

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 3.32: Let $\Delta_t = \mu^{(2)} C_t^{L,(2)} - \mu^{(1)} C_t^{L,(1)}$. Using (3.7.31) 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 $\dfrac{c(\mu^{(2)} - \mu^{(1)})}{\gamma} \geq 0$, because $\mu^{(2)} - \mu^{(1)} \geq 0$ by Condition 3.12. 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 3.33: 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 $m \geq 1$:

$$\begin{cases} C_t^{L,(2)}(m) \geq \dfrac{W}{\mu^{(2)}} \\ C_t^{L,(1)}(m) \geq \dfrac{(\theta^{(2)} + r^{(2)} + \lambda^{(2)}/L)W - c\mu^{(2)}}{r^{(2)}\mu^{(2)}}. \end{cases} \tag{3.7.33}$$

Since from Condition 3.13, $W < W^{(2)} \leq \dfrac{c\mu^{(2)}}{\theta^{(2)} + r^{(2)}}$, 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 (3.7.30) 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 3.31 and 3.32 we proved that $C_{t_0}^{L,(2)}(m) \geq \dfrac{W}{\mu^{(2)}}$, holds for $L \geq L_1$ and $t = t_0$. Then we take $L_0 = \max(L_1, L_2)$, we assume inequalities in (3.7.33) 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 (3.7.30):

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

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 (3.7.33) for $t \geq t_0$, and then Lemma 3.33 is proved as well. □

Proof of Proposition 3.23.

The proof here is similar to the one of Property 1) in Lemma 3.11 in Section 3.7.3, in the sense that we will compare the truncated processes with limited capacity $L > n$, using the comparison result 9.3.2 in [19, Chapter 9]. Here we compare the truncated process under threshold policy $(n, 0)$ with the one under threshold policy $(n+1, 0)$. For the uniformisation, 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 (3.7.19) 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

$$V_t^{(n,0),L}(n,1) - V_t^{(n,0),L}(n+1,1) \leq \frac{1}{\mu^{(1)}}, \tag{3.7.34}$$

for $t \geq 0$. As before, we prove a more general result, as stated in Lemma 3.34.

Lemma 3.34. *Under Condition 3.12,*

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

Now, result 9.3.2 in [19, 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 3.11, we obtain that this property holds as well for the original (untruncated) process.

□

Proof of Lemma 3.34.
The proof is similar to the one in Section 3.7.3. 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 $(0,2)$.

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

By applying the inductive hypothesis, we obtain

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

The inequality $\dfrac{H + \mu^{(1)} - \theta^{(2)} - \mu^{(2)}}{H\mu^{(1)}} \leq \dfrac{1}{\mu^{(1)}}$ holds if and only if $\mu^{(1)} - \mu^{(2)} < \theta^{(2)}$. The latter holds from Condition 3.12 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. □

Proof of Proposition 3.24.

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 $g^{(\infty,0)}(W) \leq g^{(n,m)}(W)$ for $W \in [W^{(1)}, W^{(2)}]$, $g^{(\infty,0)}(W) < g^{(n,m)}(W)$ for $W \in (W^{(1)}, W^{(2)})$, and 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, we first consider a threshold policy $(n,m) = (\infty, m)$. From Equation (3.4.10), 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 3.11, 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 $W < W^{(2)}$, $g^{(\infty,0)}(W) < g^{(\infty,m)}(W)$. For the if and only if, 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 3.21.

We now consider a threshold policy (n,m) with $n \neq \infty$. 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 3.7. 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 (3.4.9)

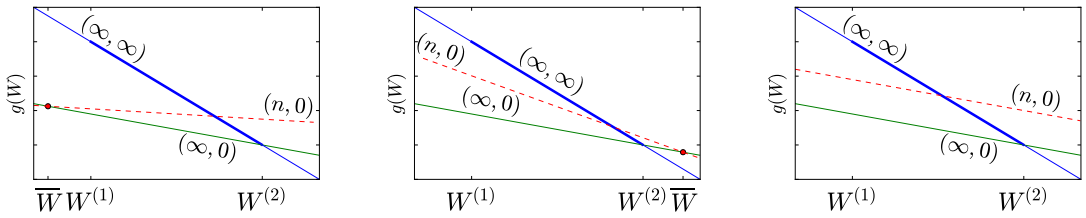

Figure 3.7: Proof of Proposition 3.24. Comparing slopes of functions $g^{(n,0)}$ and $g^{(\infty,0)}$.

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 3.11), 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 3.21. 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 $W \in \left[W^{(1)}, W^{(2)}\right]$, $g^{(n,0)}(W) \leq g^{(n,m)}(W)$.
First note that, from Properties 1) and 2) in Lemma 3.11, $s_2\left((n,m),(n,0)\right) = \sum_{l=0}^{m}\pi^{(n,m)}(l,2) - \pi^{(n,0)}(l,2) \geq 0$ 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 (or they are equal to 0), and following from Equation (3.4.9), $\overline{W}((n,m),(n,0))$ 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 (3.4.9) 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 (3.4.9) 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 $W \leq W^{(2)}$, $g^{(n,0)}(W) \leq g^{(n,m)}(W)$.

□

Chapter 4

Allocation in Large Scale Systems

Contents

In this chapter, we present a model of replication in large distributed storage systems using a stochastic modelling approach. A *storage system* consists in a set of nodes used to ensure data availability in large scale networks. Each node contains a large number of files as well as copies of those files. The nodes are interconnected in such a way that data blocks can be replicated and reallocated.

A relevant problem among these systems is the fact that the nodes break down: it is common to see nodes crashing due to issues like a high level of activity. Addressing this problem in order to ensure that the system continues operating properly is known as fault-tolerance, and in particular we are interested in the reliability feature. Reliability is the property of not losing information whenever there is a failure. A very extended method used for guaranteeing reliability is to store copies of the same file in different nodes [70]. Moreover, in order to minimise the probability of losing data, *an automatic mechanism of replication* works whenever a node breaks down. The mechanism consists in regenerating a new copy of every file that was in that node just before the breakdown, using copies present in other nodes, and reallocating the new copies to different functioning nodes as the system continues working.

Many important systems such as Hadoop Distributed File System (HDFS) [18], Google File System (GFS) [32] and Cassandra [42] use replication mechanisms as the one previously described to ensure data availability. For some references in replication and fault-tolerance in large data stores we also refer to [67, Chapters 7 and 8] and [56, Chapter 13].

We study a version of this mechanism, and we focus our attention on the load distribution of each node, that is a key feature among these systems. The load distribution can be studied by representing the process as an urn model, which consists in a set of urns and a set of identical balls jumping across the urns along time. Each ball waits a randomly distributed amount of time and then jumps to another urn, following a

particular allocation policy, which may be deterministic or randomly distributed. Urn models are frequently used in theoretical computer science to study the efficiency of algorithms, for example, for the allocation of copies of files in large distributed systems [42, 63], or the allocation of tasks to processors in large scale computing networks [41, 48].

We refer to the urn model studied in [64], where after an exponentially distributed amount of time *all the balls* in one urn are reallocated to different urns. Each ball is reallocated to an urn chosen by a random variable, and the random variables used to choose the urns are independent. Under this setting, in [64] the authors study the local mean-field interactions between urns when the number of urns grows large, and they prove the convergence of the distribution of the load of each urn under certain conditions.

Another relevant work we cite is [63], where a model of replication for large storage systems is studied. The authors consider nodes that break down after an exponential distributed amount of time with constant rate, and whenever a breakdown occurs, all the files in that node are reallocated to other nodes. Each file goes to a node chosen at random according to an allocation policy, and the random variables that select the nodes for each file are independent. Different allocation policies are analysed, in particular one that consists in choosing uniformly at random two nodes, and reallocating the file to the least loaded node among them. This policy is known as the Power of Choice, and it has been largely studied, see [59] for a survey. In [63] two limits are considered, in first term the number of nodes grows large, and in second term the average load per node is taken to infinite. Their striking result is that under the Power of Choice policy the distribution of the load of a node has a finite support.

We remark that modelling breakdowns with constant rates, as it is done in [63], does not consider that the level of activity of a node directly affects its lifespan. The novelty of our work relies in considering times of failures determined by the amount of data a node handles. The more tasks it has to process, the faster it breaks down. To the best of our knowledge a few works are done under this setting. We refer here to Schechner's work [61], inspired in Coleman's load-sharing model [24]. The author studies nodes with breakdown rates that are linear in their current charge, but it is not applied in replication mechanisms.

In this chapter we consider the following system. There are *files* distributed among *nodes*, where each file remains in its node until the node breaks down. Each node breaks down at a rate that depends on the number of files it has. When this happens, all the files present in the broken node are reallocated, and the node goes back to the state with zero files. For the ease of the tractability of the model we assume that the reallocation step is done at once for all the files in the node and it does not take time.

When a breakdown occurs, each file is allocated following a random variable. The random variables used for the reallocated files are independent and identically distributed, and we will state the model for any arbitrary distribution. For our main results, we will consider the *Random Weighted policy*, that assigns to each node a weighted probability of receiving a file according to the load it currently has. We will further study the *Random policy* as a particular case of the Random Weighted policy, where all the nodes have the same weight. In this case we provide a deeper insight.

Under these dynamics, we analyse the evolution of the storage load of a single node in the long term. We consider the case where the total number of nodes and the total number of files grow large. We make the scaling assumption that the average number of files per node converges to a real value, denoted by β. This is known as the *mean-field limit*. We will not study the convergence to the mean-field limit, but we focus on it, and in particular, on the evolution of the process in the mean-field regime.

We introduce our main results. In our first main contribution, we prove the existence of the process describing the mean-field limit for a fixed β. In this same theorem we prove the existence and uniqueness of the stationary measure of the process. In our second contribution, we let β tend to infinity, and we prove the convergence in distribution of the load in steady state. Finally, for the Random policy we obtain an expression of the generating function of the load in steady state. Our main results are done for the case where the rate of the breakdowns is linear in the number of files each node has.
The chapter is organised as follows. In Section 4.1 the stochastic model and the allocation policies are described for a fixed number of nodes. The mean-field model used to study the large scale system is presented in Section 4.2. Our main results are introduced in Section 4.3. In Section 4.3.1 we consider the Random Weighted policy, and in Section 4.3.2 we state the results for the Random policy.

4.1 Model description

We describe the model for a fixed number of nodes. Let $\mathbb{N}_0 = \{0, 1, \ldots\}$ denote the set of non-negative integers. We consider a system with N nodes indexed by $i = 1, \ldots, N$, and F_N files distributed among the nodes, with $F_N \gg N$.
The state of the system is described by a vector $\vec{l} = (l_1, \ldots, l_N) \in \mathcal{S}_N$, with

$$\mathcal{S}_N := \left\{ \vec{l} \in \mathbb{N}_0^N : l_1 + \ldots + l_N = F_N \right\}.$$

For $i = 1, \ldots, N$, l_i is the amount of files in node i.
Let $f : \mathbb{N}_0 \to \mathbb{R}_+$ be an increasing function, which determines the rates of the breakdowns of the nodes. If $\vec{l}$ is the current state of the system, node i waits an exponentially distributed amount time of rate $f(l_i)$ until it fails. For our main results in Section 4.3 we will assume that the rate of the breakdowns is linear, i.e., $f(n) = cn$, for some $c > 0$.
When node i breaks down, all the l_i files present in the node are reallocated to other nodes at random, and node i goes back to state 0. The reallocation step does not take time and it is done at the same time for all the l_i files.
For each reallocated file we consider a random variable that selects the destination node with a probability that depends on the current state of the system $\vec{l}$. We assume that the random variables used to select the destination nodes are independent and identically distributed. If the state of the system is $\vec{l}$, let $p_{ij}^N(\vec{l})$ denote the probability that a given file in node i is reallocated to node j, and $P_i^N(\vec{l}) := \left(p_{i1}^N(\vec{l}), \ldots, p_{iN}^N(\vec{l}) \right)$ the probabilities of reallocation from node i to the rest of the nodes. Note that when node i breaks down and the state of the system is $\vec{l}$, the reallocation of the files present in node i are l_i independent events. Then, if $Z_{ij}^N(\vec{l})$ denotes the number of files of node i allocated to node j, $Z_{ij}^N(\vec{l})$ follows a binomial distribution with parameters l_i and $p_{ij}^N(\vec{l})$.
We introduce the notation for the process. The state of the system at time t is denoted by a càdlàg process $\vec{L}^N(t)$, with

$$\vec{L}^N(t) := \left(L_1^N(t), \ldots, L_N^N(t) \right) \in \mathcal{S}_N,$$

$L_i^N(t)$ is the number of copies on server i at time t. We let $\Gamma^N(t)$ be the *empirical distribution* of the process, defined as

$$\Gamma^N(t)(A) := \frac{1}{N}\sum_{i=1}^{N} \delta_{L_i^N(t)}(A),$$

for any $A \subset \mathbb{N}_0$. $\Gamma^N(t)(A)$ is the fraction of nodes having a number of files in set A at time t.
The evolution of the process $\vec{L}^N(t)$ depends on the times of the breakdowns of the nodes, and the rates of the breakdowns depend themselves on the state of $\vec{L}^N(t)$. As a consequence, a full description of the evolution of $\vec{L}^N(t)$ involves the use of complex non-stationary Poisson processes (see [68, Chapter 1.3] for a definition), which we do not address in this chapter. Nevertheless, we describe here the possible transitions of the process $L_i^N(t)$ for a given i, based on the exponential rates it follows.
Fix $i \in \{1, \dots, N\}$, and consider the process $L_i^N(t)$ in $\mathbb{N}_0$. The process has transitions given by two classes of events: *a)* when node i breaks down, and *b)* when node i receives files from the breakdowns of other nodes.
a) If at a given time t, the process $L_i^N(t) = l_i$, node i breaks down at rate $f(l_i)$ and it spreads all its files. In other words, the transition $l_i \to 0$ occurs at rate $f(l_i)$.
b) If at a given time t, the whole system is in state $\vec{L}^N(t) = \vec{l}$, any of the other nodes $m = 1, \dots, N$, $m \neq i$, may break down an allocate one or more of its l_m files to node i. The breakdown of node m occurs at rate $f(l_m)$, and when this happens, $Z_{mi}^N(\vec{l})$ files are allocated to node i, where $Z_{mi}^N(\vec{l})$ is a binomial random variable with parameters l_m and $p_{mi}^N(\vec{l})$. Then, the process $L_i^N(t)$ makes the transition $l_i \to l_i + Z_{mi}^N(\vec{l})$ at rate $f(l_m)$, for all $m = 1, \dots, N$, $m \neq i$.

4.1.1 Allocation policies

For our main results, we will consider two allocation policies, the *Random Weighted* allocation policy and the *Random* policy.
For the Random Weighted allocation policy we set a non-increasing weight function $W : \mathbb{N}_0 \to (0, K]$ for some constant $K > 0$. Then, for a given node i and state $\vec{l}$, we define the probabilities $p_{ij}^N(\vec{l})$ as:

$$p_{ij}^N(\vec{l}) = W(l_j) \Big/ \sum_{k=1, k\neq i}^{N} W(l_k), \tag{4.1.1}$$

if $j \neq i$, and $p_{ii}^N(\vec{l}) = 0$. Note that $p_{ij}^N(\vec{l})$ is well defined because $\sum_{k=1,k\neq i}^{N} W(l_k) > 0$. Since $\sum_{j=1}^{N} p_{ij}^N(\vec{l}) = 1$, the vector $P_i^N(\vec{l})$ defines a probability on the set $\{1, \dots, N\}$.
The *Random* policy is a particular case of the Random Weighted policy where W is constant. In other words, for a given node i and state $\vec{l}$, the probabilities $p_{ij}^N(\vec{l})$ are

$$p_{ij}^N(\vec{l}) = \frac{1}{N-1}, \tag{4.1.2}$$

if $j \neq i$, and $p_{ii}^N(\vec{l}) = 0$.

4.2 Mean-field limit

The focus of our work is on the mean-field limit of the process, i.e., the regime where N tends to infinity. We make the scaling assumption that the average number of files per node converges to a value $\beta > 0$, i.e.:

$$\beta := \lim_{N\to\infty} \frac{F_N}{N}. \tag{4.2.1}$$

We will not study the convergence to the mean-field limit, but the limiting process of the load of a single node in such regime. In Section 4.2.1 we provide the conditions assumed in order to consider the mean-field regime. In Section 4.2.2 we describe the limiting process.

4.2.1 Conditions for convergence

The following condition is regarding the initial state $\vec{L}^N(0)$.

Condition 4.1.

- *For all N, the distribution of the vector $\vec{L}^N(0)$ satisfies the following exchangeability relation:*

$$\left(L_1^N(0), \ldots, L_N^N(0)\right) \overset{\text{dist.}}{=} \left(L_2^N(0), \ldots, L_N^N(0), L_1^N(0)\right).$$

- *The second moment of $L_1^N(0)$ is finite for all N, i.e.,*

$$\sup_{N\geq 1} \mathbb{E}\left(L_1^N(0)^2\right) < \infty.$$

- *The sequence of initial distributions $(\vec{L}^N(0))_{N\in\mathbb{N}}$ converges to a probability distribution on $\mathbb{N}_0$.*

We describe now the asymptotic behaviour of the probabilities of reallocation $P_1^N(\cdot)$ as N grows large. We define an *allocation function* $\Psi : M(\mathbb{N}_0)\times\mathbb{N}_0 \to \mathbb{R}$, where $M(\mathbb{N}_0)$ denotes the set of probability distributions on $\mathbb{N}_0$. For a given $\sigma \in M(\mathbb{N}_0)$ and $l \in \mathbb{N}_0$, if the load of the nodes is identically distributed by σ, the probability of allocating a file to a node currently having l files is of the order of $\Psi(\sigma, l)/N$, as it is stated in the following condition.

Condition 4.2. *The sequence $\left(P_1^N(\cdot)\right)_{N\in\mathbb{N}} = \left(p_{1i}^N(\cdot) : 1 \leq i \leq N\right)_{N\in\mathbb{N}}$ satisfies the relation*

$$\lim_{N\to\infty} \sup_{1\leq i\leq N, \vec{l}\in\mathcal{S}_N} \left| N p_{1i}^N(\vec{l}) - \Psi\left(\frac{1}{N}\sum_{j=1}^{N} \delta_{l_j}, l_i\right)\right| = 0, \tag{4.2.2}$$

where Ψ is a non-negative bounded function on $M(\mathbb{N}_0)\times\mathbb{N}_0$, such that, for any $\sigma \in M(\mathbb{N}_0)$,

$$\sum_{k\geq 0} \Psi(\sigma, k)\sigma(k) = 1. \tag{4.2.3}$$

Relation (4.2.3) is a mass conservation condition: when a node breaks down, every file is reallocated with probability 1.

Proposition 4.3. *For the Random Weighted policy, the allocation function Ψ_{RW} is*

$$\Psi_{RW}(\sigma, l) = W(l) \Big/ \sum_{k\geq 0} W(k)\sigma(k). \tag{4.2.4}$$

For the Random policy, the allocation function Ψ_R is

$$\Psi_R(\sigma, l) = 1 \Big/ \sum_{k\geq 0} \sigma(k). \tag{4.2.5}$$

Proof: Following from the definition of the allocation probabilities in Equations (4.1.1) and (4.1.2), we show that both functions Ψ_{RW}, Ψ_R satisfy Condition 4.2.
In the Random Weighted case, for the expression in Equation (4.2.2) we have

$$\begin{aligned}
&\sup_{1\leq i\leq N, \vec{l}\in\mathcal{S}_N} \left| Np_{1i}^N(\vec{l}) - \Psi_{RW}\left(\frac{1}{N}\sum_{j=1}^N \delta_{l_j}, l_i\right)\right| \\
&\quad = \sup_{1\leq i\leq N, \vec{l}\in\mathcal{S}_N} \left| \frac{NW(l_i)}{\sum_{k=1,k\neq i}^N W(l_k)} - \frac{W(l_i)}{\sum_{k\geq 0} W(k)\frac{1}{N}\sum_{j=1}^N \delta_{l_j}(k)}\right| \\
&\quad = \sup_{1\leq i\leq N, \vec{l}\in\mathcal{S}_N} \left| \frac{NW(l_i)}{\sum_{k=1,k\neq i}^N W(l_k)} - \frac{NW(l_i)}{\sum_{k=1}^N W(l_k)}\right|,
\end{aligned} \tag{4.2.6}$$

where the last inequality follows from the fact that, for a given $\vec{l}$, $\delta_{l_j}(k) = 1$ if and only if $k = l_j$. Since the total number of nodes grows large together with the number of files, the limit in N in Equation (4.2.6) equals 0. It is easy to see that Equation (4.2.3) holds for Ψ_{RW}.
In the Random case, in Equation (4.2.2) we have

$$\begin{aligned}
\sup_{1\leq i\leq N, \vec{l}\in\mathcal{S}_N} \left| Np_{1i}^N(\vec{l}) - \Psi_R\left(\frac{1}{N}\sum_{j=1}^N \delta_{l_j}, l_i\right)\right| &= \sup_{1\leq i\leq N, \vec{l}\in\mathcal{S}_N} \left| \frac{N}{N-1} - \frac{1}{\sum_{k\geq 0}\frac{1}{N}\sum_{j=1}^N \delta_{l_j}(k)}\right| \\
&= \sup_{1\leq i\leq N, \vec{l}\in\mathcal{S}_N} \left| \frac{N}{N-1} - 1\right|,
\end{aligned} \tag{4.2.7}$$

where for the last inequality we use that $\frac{1}{N}\sum_{j=1}^N \delta_{l_j}$ is a probability distribution, and it is evaluated over $\mathbb{N}_0$. The limit in N in Equation (4.2.7) equals 0. It is straightforward to see that Equation (4.2.3) holds for Ψ_R. □

4.2.2 Limiting process

We introduce now our main problem. We study the existence of the process as the possible limit of $(L_1^N(t))_{N\in\mathbb{N}}$ as N grows large. This process will be denoted by $(X_\beta(t))$. We first provide an informal

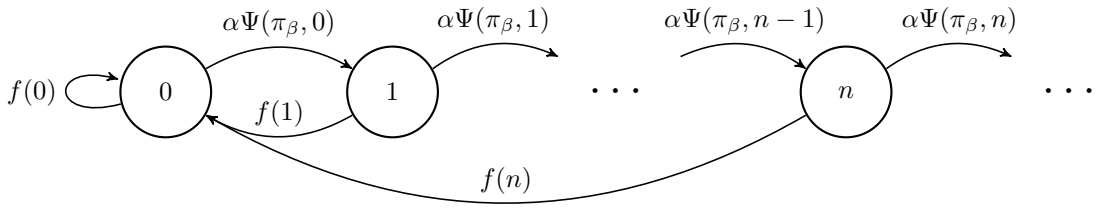

Figure 4.1: diagram of process $(X_\beta(t))$.

description of the evolution of $(X_\beta(t))$, and, in second term, we formally formulate the problem. From this section, we assume that Condition 4.1 and Condition 4.2 are satisfied.

As it was stated for $(L_1^N(t))$, the process $(X_\beta(t))$ lives in $\mathbb{N}_0$ and it has two types of transitions. *a)* when the node breaks down, and *b)* when the node receives files from the breakdowns of other nodes.

The transitions of type *a)* are the same as for $(L_1^N(t))$, thus $(X_\beta(t))$ makes transitions $n \to 0$ at rate $f(n)$ for all state $n \geq 0$.

For the transitions of type *b)*, intuitively, as the total number of nodes N grows large, the reallocation step of a single breakdown does not allocate more than one file to the same node. In other words, for N large enough, and given nodes i, j and a state $\vec{l}$,

$$\mathbb{P}\left(Z_{ij}^N(\vec{l}) \geq 2\right) = 0,$$

where $Z_{ij}^N(\vec{l})$ is a binomial random variable with parameters l_i and $p_{ij}^N(\vec{l})$. As a consequence, the transitions of type *b)* in $(X_\beta(t))$ are birth transitions $n \to n+1$ for all $n \geq 0$. We want to determine the rate of these transitions, i.e., the rate at which node 1 receives a file from any of the other nodes. In order to do so, we consider the process in equilibrium.

Let us assume that the process $(X_\beta(t))$ is ergodic. We denote by π_β its stationary measure and $\overline{X}_\beta$ the random variable distributed by π_β. Note that, from the scaling assumption (4.2.1), $\mathbb{E}(\overline{X}_\beta) = \beta$. At equilibrium, the load of an arbitrary node is given by $\overline{X}_\beta$, and the rate at which it breaks down is $f(\overline{X}_\beta)$. Therefore, the output flow of files from any node is given by the following parameter α:

$$\alpha := \mathbb{E}\left(\overline{X}_\beta f(\overline{X}_\beta)\right).$$

The probability that a file is assigned to a node depends on its current state n, and this is determined by the function Ψ as defined in Condition 4.2. Then, the rate at which a files arrives to node 1 is given by $\alpha\Psi(\pi_\beta, n)$.

In summary, provided that all the conditions needed for its existence are satisfied, $(X_\beta(t))$ is a jump process whose Q-matrix is given by $q(n, 0) = f(n)$ and $q(n, n+1) = \alpha\Psi(\pi_\beta, n)$, for all $n \geq 0$. The evolution of $(X_\beta(t))$ is represented in Figure 4.1. We can now formulate our problem.

Formulation of the problem

We study the existence of the process $(X_\beta(t))$. This process exists if there exist a probability measure π_β and an $\alpha \in \mathbb{R}_+$ that satisfy the following equations.

$$\begin{aligned} \pi_\beta(n)(\alpha\Psi(\pi_\beta, n) + f(n)) &= \pi_\beta(n-1)\alpha\Psi(\pi_\beta, n-1), \quad \text{for } n \geq 1, & (4.2.8)\\ \alpha &= \sum_{n\geq 0} nf(n)\pi_\beta(n), & (4.2.9) \end{aligned}$$

where (4.2.8) are the balance equations of $(X_\beta(t))$, and (4.2.9) is the output flow from an arbitrary node. In Proposition 4.4, we further characterise π_β as the solution of a fixed point equation. This characterisation will be used in the proofs of our main results.

Proposition 4.4. *Assume the process $(X_\beta(t))$ exists, i.e., there exist π_β and α that satisfy Equations* (4.2.8) *and* (4.2.9)*. Let $\overline{X}_\beta$ be the random variable distributed by π_β. Then π_β satisfies the fixed point equation $\nu(\pi_\beta) = \pi_\beta$ in $M(\mathbb{N}_0)$, where ν is defined for $\sigma \in M(\mathbb{N}_0)$ as*

$$\nu(\sigma) = \nu(\sigma)(n) \quad := \quad \frac{\mathbb{E}\left(f(\overline{X}_\beta)\right)}{\alpha(\sigma)\Psi(\sigma, n) + f(n)} \prod_{k=0}^{n-1} \frac{\alpha(\sigma)\Psi(\sigma, k)}{\alpha(\sigma)\Psi(\sigma, k) + f(k)}, \quad \textit{for } n \geq 0, \tag{4.2.10}$$

where $\alpha(\sigma) = \sum_{n\geq 0} nf(n)\sigma(n)$ and we use the convention $\prod_{k=i}^{j} a_k = 1$ if $j < i$.

Proof: Following from Equation (4.2.8), for any $n \geq 1$, $\pi_\beta(n)$ satisfies the following equation:

$$\pi_\beta(n) = \pi_\beta(n-1)\frac{\alpha(\pi_\beta)\Psi(\pi_\beta, n-1)}{\alpha(\pi_\beta)\Psi(\pi_\beta, n) + f(n)}.$$

We deduce inductively from last expression that, for $n \geq 1$,

$$\pi_\beta(n) = \pi_\beta(0)\frac{\alpha(\pi_\beta)\Psi(\pi_\beta, 0)}{\alpha(\pi_\beta)\Psi(\pi_\beta, n) + f(n)} \prod_{k=1}^{n-1} \frac{\alpha(\pi_\beta)\Psi(\pi_\beta, k)}{\alpha(\pi_\beta)\Psi(\pi_\beta, k) + f(k)}. \tag{4.2.11}$$

For $\pi_\beta(0)$, from Relation (4.2.3) on Ψ we have that $\sum_{n\geq 0} \pi_\beta(n)\Psi(\pi_\beta, n) = 1$, which can be rewritten as

$$\sum_{n\geq 1} \pi_\beta(n)\Psi(\pi_\beta, n) = 1 - \pi_\beta(0)\Psi(\pi_\beta, 0). \tag{4.2.12}$$

Then, we sum both sides of equation (4.2.8) for $n \geq 1$, and we get

$$\alpha(\pi_\beta)\left[1 - \pi_\beta(0)\Psi(\pi_\beta, 0)\right] + \mathbb{E}\left(f(\overline{X}_\beta)\right) - \pi_\beta(0)f(0) = \alpha(\pi_\beta),$$

which is equivalent to

$$\pi_\beta(0) = \frac{\mathbb{E}\left(f(\overline{X}_\beta)\right)}{\alpha(\pi_\beta)\Psi(\pi_\beta, 0) + f(0)}.$$

This proves the expression for $n = 0$. Furthermore, following from (4.2.11), we obtain

$$\pi_\beta(n) = \frac{\mathbb{E}\left(f(\overline{X}_\beta)\right)}{\alpha(\pi_\beta)\Psi(\pi_\beta, n) + f(n)} \prod_{k=0}^{n-1} \frac{\alpha(\pi_\beta)\Psi(\pi_\beta, k)}{\alpha(\pi_\beta)\Psi(\pi_\beta, k) + f(k)}, \quad \text{for } n \geq 1,$$

and the proposition is proved. □

4.3 Main results

In this section we present our results. We assume a linear rate for the breakdowns of the nodes, i.e., $f(n) = c_1 n$, $c_1 \in \mathbb{R}_{>0}$.

4.3.1 The Random Weighted allocation

Let us consider the Random Weighted allocation policy, where the allocation function Ψ_{RW} is given by Equation (4.2.4). For a fixed β, we denote the studied process as defined in Section 4.2.2 by $(X_\beta^{RW}(t))$. As a first step, we prove the process $(X_\beta^{RW}(t))$ exists. Then, we study the convergence in distribution of the process in equilibrium as β tends to infinity.

Theorem 4.5. *For a given $\beta > 0$, under the Random Weighted allocation policy and linear rate of breakdowns, the process $(X_\beta^{RW}(t))$ exists. In other words, there exist π_β and α that satisfy Equations* (4.2.8) *and* (4.2.9).

Proof: Following from Proposition 4.4, we consider $\nu(\sigma)$ as defined in (4.2.10). We study the existence and uniqueness of the solution of the fixed point equation $\pi_\beta = \nu(\pi_\beta)$, and of $\alpha(\pi_\beta)$, which we will denote α as an abuse of notation.
Let $\overline{X}_\beta^{RW}$ denote the random variable distributed by π_β, and recall that $\mathbb{E}(\overline{X}_\beta^{RW}) = \beta$. Since we consider Ψ_{RW} and $f(x) = c_1 x$, for $\pi_\beta(n) = \nu(\pi_\beta)(n)$ we obtain

$$\pi_\beta(n) = \frac{c_1\beta\gamma}{\alpha W(n) + c_1 n\gamma} \prod_{k=0}^{n-1} \frac{\alpha W(k)}{\alpha W(k)+c_1 k\gamma}, \quad \text{for } n \geq 0, \tag{4.3.1}$$

where $\alpha := c_1\, \mathbb{E}\left(\left(\overline{X}_\beta^{RW}\right)^2\right)$ and $\gamma := \sum_{k\geq 0} W(k)\pi_\beta(k)$. Therefore, the proof reduces to show that, under these definitions, there exist unique values α and γ.
We consider the following two equations.

$$\frac{1}{\beta} = \frac{c_1\gamma}{\alpha W(0)} + \sum_{n\geq 1} \frac{1}{n(n+1)} \left(1 - \prod_{k=0}^{n} \frac{\alpha W(k)}{\alpha W(k)+c_1 k\gamma}\right), \tag{4.3.2}$$

$$1 = \sum_{n\geq 0} \frac{W(n)c_1\beta}{\alpha W(n)+c_1 n\gamma} \prod_{k=0}^{n-1} \frac{\alpha W(k)}{\alpha W(k)+c_1 k\gamma}. \tag{4.3.3}$$

Equation (4.3.2) follows from the fact that $\sum_{n\geq 0} \pi_\beta(n) = 1$, a detailed computation can be found in Appendix 4.4. Equation (4.3.3) follows from the definition of γ and Equation (4.3.1).
Let $x_0 := \gamma/\alpha$, then from Equation (4.3.2) we obtain

$$\frac{1}{\beta} = \frac{c_1 x_0}{W(0)} + \sum_{n\geq 1} \frac{1}{n(n+1)} \left(1 - \prod_{k=0}^{n} \frac{W(k)}{W(k)+c_1 k x_0}\right). \tag{4.3.4}$$

Since $\prod_{k=0}^{n} \frac{W(k)}{W(k)+c_1 k x_0}$ is strictly decreasing as a function of $x_0 > 0$, the right hand side of Equation (4.3.4) is strictly increasing in x_0. Furthermore, the limit of the right hand side of Equation (4.3.4) as $x_0 \to 0^+$ is 0, and its limit as $x_0 \to +\infty$ is $+\infty$. Therefore, for a given $\beta > 0$, Equation (4.3.4) has a unique solution $x_0 > 0$.

Then, note that an equivalent formulation of (4.3.3) is

$$\alpha = c_1 \beta \sum_{n \geq 0} \prod_{k=0}^{n} \frac{\alpha W(k)}{\alpha W(k)+c_1 k \gamma} = c_1 \beta \sum_{n \geq 0} \prod_{k=0}^{n} \frac{W(k)}{W(k)+c_1 k x_0}. \tag{4.3.5}$$

Since for each $\beta > 0$ there exists a unique $x_0 > 0$, from Equation (4.3.5) we deduce that for each $\beta > 0$ there exists a unique $\alpha > 0$. Finally, for each $\beta > 0$, there exists a unique $\gamma = x_0 \alpha > 0$. We conclude that $(X_\beta^{RW}(t))$ exists, and its stationary measure is given by (4.3.1). □

Corollary 4.6. *With the notation of Theorem 4.5, we have*

$$\lim_{\beta \to +\infty} \gamma/\alpha = 0.$$

Proof: This result follows from the proof of Theorem 4.5. Recall that $\gamma/\alpha = x_0$, and for each $\beta > 0$ there exists a unique $x_0 > 0$ that satisfies Equation (4.3.4). Furthermore, as β tends to infinity, the lhs of Equation (4.3.4) tends to 0, and the x_0 that solves the equation tends to 0 as well. □

Convergence as $\beta \to \infty$ for the case $W(n)=c_2/n^\delta$.

For the remainder of the section, we assume a weight function given by

$$W(n)=\begin{cases} c_2 & \text{for } n = 0 \\ c_2/n^\delta & \text{for } n \geq 1, \end{cases} \tag{4.3.6}$$

with $\delta, c_2 > 0$. In Theorem 4.7 we prove the convergence in distribution of $\overline{X}_\beta^{RW}$ as β grows large. We recall that γ and α are parameters that depend on β, although we omit it in the notation for ease of the reading.

Theorem 4.7. *Let $W(n)$ be defined as in* (4.3.6)*. For the convergence in distribution of the variable $\overline{X}_\beta^{RW}$ we have*

$$\lim_{\beta \to \infty} \left(\frac{c_1 \gamma}{c_2 \alpha} \right)^{1/(\delta+2)} \overline{X}_\beta^{RW} = \overline{X}_\infty^{RW},$$

where $\overline{X}_\infty^{RW}$ is a random variable with density

$$\frac{x^\delta}{C} \exp\left(-\frac{x^{\delta+2}}{\delta+2} \right), \; x \geq 0, \quad \textit{with } C := (\delta+2)^{-1/(\delta+2)} \, \Gamma\left(\frac{\delta+1}{\delta+2} \right), \tag{4.3.7}$$

on $\mathbb{R}_+$, and Γ is the classical Gamma function.

Proof: Following from Equation (4.3.1), we have that the distribution of $\overline{X}_\beta^{RW}$ for $n \geq 1$ is

$$\begin{aligned}
\pi_\beta(n) &= \frac{c_1\beta\gamma n^\delta}{c_2\alpha + c_1\gamma n^{\delta+1}} \prod_{k=0}^{n-1} \frac{c_2\alpha}{c_2\alpha + c_1\gamma k^{\delta+1}} \\
&= \frac{1}{c_2\alpha} \frac{c_1\beta\gamma n^\delta}{1 + \frac{c_1\gamma}{c_2\alpha} n^{\delta+1}} \prod_{k=0}^{n-1} \frac{1}{1 + \frac{c_1\gamma}{c_2\alpha} k^{\delta+1}} \\
&= \frac{c_1\beta\gamma}{c_2\alpha} n^\delta \prod_{k=0}^{n} \frac{1}{1 + \frac{c_1\gamma}{c_2\alpha} k^{\delta+1}},
\end{aligned}$$

and $\pi_\beta(0) = \dfrac{c_1\beta\gamma}{c_2\alpha}$. We simplify the notation in order to continue the proof. Let $a := c_1\gamma/c_2\alpha$, $Z_a := (c_1\beta\gamma/c_2\alpha)^{-1}$ is the normalisation constant, and

$$U_a(n) := \begin{cases} 1 & \text{for } n = 0 \\ n^\delta \prod_{k=0}^{n} \frac{1}{1+ak^{\delta+1}} & \text{for } n \geq 1. \end{cases}$$

Then, the expression of the distribution reduces to $\pi_\beta(n) = U_a(n)/Z_a$ for all $n \geq 0$. Note that from Corollary 4.6, $\lim_{\beta\to\infty} a = 0$.

Using this notation, we prove now the convergence in distribution. If ϕ is a bounded continuous function on $\mathbb{R}_+$, define

$$A_a(\phi) := \sum_{n\geq 0} \phi\left(a^{1/(\delta+2)} n\right) \frac{U_a(n)}{N_a^{\delta+1}}, \text{ with } N_a := \left\lfloor \frac{1}{a^{1/(\delta+2)}} \right\rfloor.$$

Since the normalisation constant can also be written as $Z_a = \sum_{n\geq 0} U_a(n)$, we have that $A_a(1) = Z_a/N_a^{\delta+1}$, and

$$\mathbb{E}\left(\phi\left(a^{1/(\delta+2)}\overline{X}_\beta^{RW}\right)\right) = \frac{A_a(\phi)}{A_a(1)}.$$

We have to prove that the previous expression converges to $\mathbb{E}\left(\phi\left(\overline{X}_\infty^{RW}\right)\right)$ as a goes to 0. We prove this in two steps. Firstly we show that, if $K>0$ is sufficiently large, the sum of terms with the indices greater than KN_a in the definition of $A_a(1)$ can be made uniformly small with respect to a. Secondly, we prove the convergence of $(A_a(\phi))$ when ϕ has a compact support.

For $K>0$, let

$$\begin{aligned}
R_K(a) &:= \sum_{n\geq KN_a} \frac{U_a(n)}{N_a^{\delta+1}} \\
&= \frac{1}{N_a} \sum_{n\geq KN_a} \left(\frac{n}{N_a}\right)^\delta \exp\left(\ln\left(\prod_{k=0}^{n} \left(1+ak^{\delta+1}\right)^{-1}\right)\right) \\
&= \frac{1}{N_a} \sum_{n\geq KN_a} \left(\frac{n}{N_a}\right)^\delta \exp\left(-\sum_{k=0}^{n} \ln\left(1+\frac{1}{N_a} aN_a^{\delta+2}\left(\frac{k}{N_a}\right)^{\delta+1}\right)\right).
\end{aligned}$$

By definition of N_a, the expression $aN_a^{\delta+2}$ is converging to 1 as a goes to 0. We assume that a is sufficiently small so that $aN_a^{\delta+2}$ is lower bounded by $1/2$. We then have

$$\begin{aligned} R_K(a) &\leq \frac{1}{N_a} \sum_{n \geq KN_a} \left(\frac{n}{N_a}\right)^{\delta} \exp\left(-\sum_{k=0}^{n} \ln\left(1+\frac{1}{2N_a}\left(\frac{k}{N_a}\right)^{\delta+1}\right)\right) \\ &\leq \int_K^{+\infty} (u+1)^{\delta} \exp\left(-\sum_{k=0}^{\lfloor uN_a \rfloor} \ln\left(1+\frac{1}{2N_a}\left(\frac{k}{N_a}\right)^{\delta+1}\right)\right) \mathrm{d}u \\ &\leq \int_K^{+\infty} (u+1)^{\delta} \exp\left(-\int_0^{\lfloor uN_a \rfloor/N_a} N_a \ln\left(1+\frac{1}{2N_a}x^{\delta+1}\right) \mathrm{d}x\right) \mathrm{d}u. \end{aligned} \tag{4.3.8}$$

The change of variables in second inequality is $n \to \lfloor uN_a \rfloor$, and in third inequality is $k \to xN_a$.
If $v{>}0$, we consider the function $w{\mapsto} w\ln(1{+}v/w)$. This function is increasing on $\mathbb{R}_+$, following from its derivative, which, with the change of variables $p = \frac{w}{w+v}$, is

$$\ln(1{+}v/w) - \frac{v}{w{+}v} = -\ln(p){-}1{+}p \geq 0 \quad \text{for all } p{\in}(0,1].$$

We fix some $b{>}0$ sufficiently small. For $a{\in}(0,b)$, we have $N_a \geq N_b$, and following from (4.3.8), we get

$$R_K(a) \leq \int_K^{+\infty} (u+1)^{\delta} \exp\left(-\int_0^{\lfloor uN_a \rfloor/N_a} N_b \ln\left(1+\frac{1}{2N_b}x^{\delta+1}\right) \mathrm{d}x\right) \mathrm{d}u,$$

and therefore, by Lebesgue's Theorem, we get the relation

$$\limsup_{a \searrow 0} R_K(a) \leq \int_K^{+\infty} (u+1)^{\delta} \exp\left(-\int_0^{u} N_b \ln\left(1+\frac{1}{2N_b}x^{\delta+1}\right) \mathrm{d}x\right) \mathrm{d}u.$$

One can thus choose K sufficiently large so that the right hand side of the last expression is arbitrarily small, and therefore, all the $R_K(a)$ with a in a neighbourhood of 0 are sufficiently small as well. The first step is established.
Assume now that ϕ has compact support in $[0,K]$ with $K{\in}\mathbb{N}$, $A_a(\phi)$ can be written as

$$A_a(\phi) = \frac{1}{N_a} \sum_{n=0}^{KN_a} \phi\left(a^{1/(\delta+2)}n\right) \left(\frac{n}{N_a}\right)^{\delta} \exp\left(-\sum_{k=0}^{n} \ln\left(1+\frac{1}{N_a}aN_a^{\delta+2}\left(\frac{k}{N_a}\right)^{\delta+1}\right)\right).$$

As before, the convergence of $(aN_a^{\delta+2})$ to 1 as a goes to 0 shows that, for a sufficiently small, $A_a(\phi)$ is in an interval of the type $[B_a^{\eta_1}(\phi), B_a^{\eta_2}(\phi)]$ where $\eta_2{<}1{<}\eta_1$ and η_1 and η_2 are arbitrarily close to 1, with the definition

$$B_a^{\eta}(\phi) := \frac{1}{N_a} \sum_{n=0}^{\lceil KN_a \rceil} \phi(n/N_a) \left(\frac{n}{N_a}\right)^{\delta} \exp\left(-\sum_{k=0}^{n} \ln\left(1+\frac{1}{N_a}\eta\left(\frac{k}{N_a}\right)^{\delta+1}\right)\right).$$

We study the convergence of $B_a^{\eta}(\phi)$ as $a \to 0$. We first write the uniform estimation, for $0 \leq n \leq \lceil KN_a \rceil$,

$$-\sum_{k=0}^{n}\ln\left(1+\frac{1}{N_a}\eta\left(\frac{k}{N_a}\right)^{\delta+1}\right)=-\int_0^n\ln\left(1+\frac{1}{N_a}\eta\left(\frac{u}{N_a}\right)^{\delta+1}\right)\mathrm{d}u+O\left(\frac{1}{N_a}\right).$$

Then $B_a^\eta(\phi)$ is equivalent to

$$\begin{aligned}&\frac{1}{N_a}\sum_{n=0}^{\lceil KN_a\rceil}\phi(n/N_a)\left(\frac{n}{N_a}\right)^\delta\exp\left(-\int_0^n\ln\left(1+\frac{1}{N_a}\eta\left(\frac{u}{N_a}\right)^{\delta+1}\right)\mathrm{d}u\right)\\ =\quad&\frac{1}{N_a}\sum_{n=0}^{\lceil KN_a\rceil}\phi(n/N_a)\left(\frac{n}{N_a}\right)^\delta\exp\left(-\int_0^{n/N_a}N_a\ln\left(1+\frac{1}{N_a}\eta u^{\delta+1}\right)\mathrm{d}u\right).\end{aligned}\tag{4.3.9}$$

The elementary inequality $|\ln(1+x)-x|\le\frac{x^2}{2}$, for $x\in[0,1]$, gives us the uniform estimation

$$N_a\ln\left(1+\frac{1}{N_a}\eta u^{\delta+1}\right)=\eta u^{\delta+1}+O\left(\frac{1}{N_a}\right),$$

for u in some fixed finite interval. Following from (4.3.9), we get

$$\begin{aligned}B_a^\eta(\phi)\quad&\sim\quad\frac{1}{N_a}\sum_{n=0}^{\lceil KN_a\rceil}\phi(n/N_a)\left(\frac{n}{N_a}\right)^\delta\exp\left(-\int_0^{n/N_a}\eta u^{\delta+1}\,\mathrm{d}u\right)\\&=\quad\frac{1}{N_a}\sum_{n=0}^{\lceil KN_a\rceil}\phi(n/N_a)\left(\frac{n}{N_a}\right)^\delta\exp\left(-\eta\frac{(n/N_a)^{\delta+2}}{\delta+2}\right)\end{aligned}$$

Finally, with the uniform estimation of the integral and Lebesgue's Theorem, we get the relation

$$\lim_{a\searrow 0}B_a^\eta(\phi)=\int_0^K\phi(x)x^\delta\exp\left(-\eta\frac{x^{\delta+2}}{\delta+2}\right)\mathrm{d}x,$$

and, therefore, the desired convergence for $(A_a(\phi))$.
The constant C in (4.3.7) is the normalisation constant needed for the function to be a density, and it can be computed from the expression of $\Gamma\left(\frac{\delta+1}{\delta+2}\right)$. The theorem is proved. □

4.3.2 The Random allocation

In this section we consider the Random allocation policy, where the allocation function Ψ_R is given by Equation (4.2.5). It is a particular case of the Random Weighted allocation policy, taking a constant weight function $W(n)=K$ with $K>0$, for all $n\ge 0$. We denote the studied process as $(X_\beta^R(t))$, and $\overline{X}_\beta^R$ the variable distributed by the stationary measure. The novelty of the section relies on the explicit expression of the generating function of $\overline{X}_\beta^R$.
We assume a breakdown rate given by $f(x)=x$.

Theorem 4.8. *For a given $\beta > 0$, under the Random allocation policy and with a rate of breakdowns given by $f(x) = x$, the process $(X^R_\beta(t))$ exists. In other words, there exist π_β and α that satisfy Equations* (4.2.8) *and* (4.2.9).
Furthermore, the generating function of the variable $\overline{X}^R_\beta$ has the following expression.

$$\mathbb{E}\left(z^{\overline{X}^R_\beta}\right) = \left(1+\alpha z \int_0^1 (1-u)^\alpha e^{\alpha z u}\, \mathrm{d}u\right) \Big/ \left(1+\alpha \int_0^1 (1-u)^\alpha e^{\alpha u}\, \mathrm{d}u\right). \tag{4.3.10}$$

Proof: Following from the definition of Ψ_R, and inductively from the balance equations (4.2.8), for $n \geq 1$ we have

$$\pi_\beta(n) = \pi_\beta(0) \prod_{k=1}^{n} \frac{\alpha}{\alpha+k}, \tag{4.3.11}$$

where $\alpha := \mathbb{E}\left(\left(\overline{X}^R_\beta\right)^2\right)$. Since $\pi_\beta(0)$ is the normalisation constant, it can be expressed as

$$\pi_\beta(0) = \left(1+\sum_{n\geq 1}\prod_{k=1}^{n}\frac{\alpha}{\alpha+k}\right)^{-1}. \tag{4.3.12}$$

In this case, the existence and uniqueness of π_β reduces to show the existence and uniqueness of α such that Equations (4.3.11) and (4.3.12) are well defined. In order to obtain an expression for $\mathbb{E}(\overline{X}^R_\beta)$, we rewrite

$$\begin{aligned}
\sum_{n\geq 1} n \prod_{k=1}^{n}\frac{\alpha}{\alpha+k} &= \sum_{n\geq 1}(n+\alpha-\alpha)\prod_{k=1}^{n}\frac{\alpha}{\alpha+k} \\
&= \sum_{n\geq 1}\alpha\prod_{k=1}^{n-1}\frac{\alpha}{\alpha+k} - \alpha\sum_{n\geq 1}\prod_{k=1}^{n}\frac{\alpha}{\alpha+k} = \sum_{n\geq 1}\alpha\left(S_{n-1}-S_n\right) \\
&= \alpha\left(S_0 - \lim_n S_n\right) = \alpha.
\end{aligned}$$

Consequently,

$$\mathbb{E}(\overline{X}^R_\beta) = \sum_{n\geq 1} n\pi_\beta(n) = \alpha \Big/ \left(1+\sum_{n\geq 1}\prod_{k=1}^{n}\frac{\alpha}{\alpha+k}\right). \tag{4.3.13}$$

Denote, for $x \in \mathbb{R}_+$,

$$f_\alpha(x) := \sum_{n\geq 1}\left(\prod_{k=1}^{n}\frac{1}{\alpha+k}\right)x^{n+\alpha}, \tag{4.3.14}$$

it is easy to see that

$$f'_\alpha(x) = x^\alpha + f_\alpha(x) \text{ and } f_\alpha(0) = 0, \quad \text{hence} \quad f_\alpha(x) = \int_0^x u^\alpha e^{x-u}\, \mathrm{d}u.$$

From Equations (4.3.13) and (4.3.14), and the fact that $\beta = \mathbb{E}(\overline{X}^R_\beta)$, we obtain that

$$\beta = \frac{\alpha}{1+\alpha^{-\alpha}f_\alpha(\alpha)},$$

or, equivalently, after the substitution $u \to \alpha(1-u)$ in the integral in $f_\alpha(\alpha)$, we get

$$\frac{1}{\beta} = \frac{1}{\alpha} + \int_0^1 (1-u)^\alpha e^{\alpha u}\, \mathrm{d}u. \tag{4.3.15}$$

Note that the integrating factor $h(\alpha, u) = e^{\alpha(\ln(1-u)+u)}$ is decreasing in α, since $\ln(1-u)+u \leq 0$ for $u \in (0,1)$. Then the right hand side of the expression is strictly decreasing in $\alpha > 0$, and it varies between $+\infty$ and 0. Hence, for any $\beta > 0$ there exists a unique $\alpha > 0$ satisfying Equation (4.3.15), and the proof of the existence is ended.
For the generating function we have

$$\mathbb{E}\left(z^{\overline{X}_\beta^R}\right) = \pi_\beta(0) + \sum_{n\geq 1} \pi_\beta(n) z^n = \pi_\beta(0) + \pi_\beta(0) \sum_{n\geq 1} \left(\prod_{k=1}^n \frac{\alpha}{\alpha+k}\right) z^n,$$

and, in consequence,

$$\frac{\mathbb{E}\left(z^{\overline{X}_\beta^R}\right)}{\pi_\beta(0)} = 1 + \sum_{n\geq 1} \left(\prod_{k=1}^n \frac{\alpha}{\alpha+k}\right) z^n. \tag{4.3.16}$$

For $\pi_\beta(0)$ we have that

$$\begin{aligned} \frac{1}{\pi_\beta(0)} &= 1 + \sum_{n\geq 1} \prod_{k=1}^n \frac{\alpha}{\alpha+k} \\ &= \frac{\alpha}{\mathbb{E}\left(\overline{X}_\beta^R\right)} \\ &= 1 + \alpha \int_0^1 (1-u)^\alpha e^{\alpha u}\, \mathrm{d}u, \end{aligned}$$

where the second equality follows from Equation (4.3.13) and the third one from the fact that $\mathbb{E}(\overline{X}_\beta^R) = \beta$ and from Equation (4.3.15). Then, recall the definition of $f_\alpha(x)$ in (4.3.14), in (4.3.16) we obtain

$$\begin{aligned} \left(1 + \alpha \int_0^1 (1-u)^\alpha e^{\alpha u}\, \mathrm{d}u\right) \mathbb{E}\left(z^{\overline{X}_\beta^R}\right) &= 1 + \sum_{n\geq 1} \left(\prod_{k=1}^n \frac{\alpha}{\alpha+k}\right) z^n \\ &= 1 + \frac{f_\alpha(\alpha z)}{(\alpha z)^\alpha} = 1 + \int_0^{\alpha z} \left(\frac{u}{z\alpha}\right)^\alpha e^{\alpha z - u}\, \mathrm{d}u, \end{aligned}$$

or, equivalently,

$$\mathbb{E}\left(z^{\overline{X}_\beta^R}\right) = \left(1 + \alpha z \int_0^1 (1-u)^\alpha e^{\alpha z u}\, \mathrm{d}u\right) \Big/ \left(1 + \alpha \int_0^1 (1-u)^\alpha e^{\alpha u}\, \mathrm{d}u\right),$$

and the proof is ended. □

Convergence as $\beta \to \infty$. We present now the convergence in distribution of $\overline{X}_\beta^R$ as $\beta \to \infty$. In order to do so, we introduce Proposition 4.9, which is a technical result regarding the growth rate of $\alpha(\beta)$.

Proposition 4.9. *Under the Random allocation policy,*

$$\lim_{\beta\to+\infty}\frac{\alpha(\beta)}{\beta^2}=\frac{\pi}{2}.$$

Proof: We rewrite the integral in (4.3.15) as:

$$\begin{aligned}\int_0^1(1-u)^\alpha e^{\alpha u}\,\mathrm{d}u&=\int_0^1\exp(\alpha(\ln(1-u)+u))\,\mathrm{d}u\\&=\frac{1}{\sqrt{\alpha}}\int_0^{\sqrt{\alpha}}\exp\left(\alpha\left(\ln\left(1-\frac{u}{\sqrt{\alpha}}\right)+\frac{u}{\sqrt{\alpha}}\right)\right)\mathrm{d}u.\end{aligned}$$

Using the relation $\ln(1-x)\leq -x-x^2/2$ for $x\in(0,1)$, obtained from its Taylor polynomial, and Lebesgue's convergence theorem, we get

$$\lim_{\alpha\to+\infty}\sqrt{\alpha}\int_0^1(1-u)^\alpha e^{\alpha u}\,\mathrm{d}u=\int_0^{+\infty}e^{-u^2/2}\,\mathrm{d}u=\sqrt{\frac{\pi}{2}}.\tag{4.3.17}$$

Then, following from (4.3.15),

$$\frac{\alpha}{\beta^2}=\alpha\left(\frac{1}{\beta}\right)^2=\frac{1}{\alpha}+2\int_0^1(1-u)^\alpha e^{\alpha u}\,\mathrm{d}u+\left(\sqrt{\alpha}\int_0^1(1-u)^\alpha e^{\alpha u}\,\mathrm{d}u\right)^2.$$

Taking limit as $\beta\to\infty$, and from (4.3.17), we obtain

$$\lim_{\beta\to+\infty}\frac{\alpha}{\beta^2}=\frac{\pi}{2}.$$

□

Theorem 4.10. *For the convergence in distribution of the variable $\overline{X}_\beta^R$ we have*

$$\lim_{\beta\to\infty}\frac{\overline{X}_\beta^R}{\beta}=\overline{X}_\infty^R,$$

where $\overline{X}_\infty^R$ is a random variable with density

$$f(u)=\frac{e^{-\frac{u^2}{\pi}}}{\pi/2}.$$

Proof: We make the change of variables in the generating function, as given by (4.3.10), to get the Laplace transform of $\frac{\overline{X}_\beta^R}{\beta}$ in the limit:

$$\begin{aligned}\phi_\beta(s)&=\mathbb{E}\left(e^{\frac{s}{\beta}\overline{X}_\beta^R}\right)=\left(1+\alpha e^{-\frac{s}{\beta}}\int_0^1(1-u)^\alpha e^{\alpha e^{-\frac{s}{\beta}}u}\,\mathrm{d}u\right)\Big/\left(1+\alpha\int_0^1(1-u)^\alpha e^{\alpha u}\,\mathrm{d}u\right)\\&=\left(\frac{1}{\sqrt{\alpha}}+\sqrt{\alpha}e^{-\frac{s}{\beta}}\int_0^1(1-u)^\alpha e^{\alpha e^{-\frac{s}{\beta}}u}\,\mathrm{d}u\right)\Big/\left(\frac{1}{\sqrt{\alpha}}+\sqrt{\alpha}\int_0^1(1-u)^\alpha e^{\alpha u}\,\mathrm{d}u\right).\end{aligned}\tag{4.3.18}$$

In a separate analysis, we study the following term:

$$\sqrt{\alpha}\int_0^1 (1-u)^\alpha e^{\alpha e^{-\frac{s}{\beta}}u}\,\mathrm{d}u = \int_0^{\sqrt{\alpha}} \exp\left(\alpha\left(\ln\left(1-\frac{u}{\sqrt{\alpha}}\right)+\frac{u}{\sqrt{\alpha}}e^{-\frac{s}{\beta}}\right)\right)\mathrm{d}u.$$

Using again the inequality $\ln(1-x)\leq -x-x^2/2$ for $x\in(0,1)$, we get

$$\ln(1-x)+xe^{-\frac{s}{\beta}}\leq -\left(1-e^{-\frac{s}{\beta}}\right)x-x^2/2.$$

Then,

$$\begin{aligned}
\lim_{\beta\to\infty}\sqrt{\alpha}\int_0^1 (1-u)^\alpha e^{\alpha e^{-\frac{s}{\beta}}u}\,\mathrm{d}u &\leq \lim_{\beta\to\infty}\int_0^{\sqrt{\alpha}}\exp\left(\alpha\left(-\left(1-e^{-\frac{s}{\beta}}\right)\frac{u}{\sqrt{\alpha}}-\frac{u^2}{2\alpha}\right)\right)\mathrm{d}u\\
&= \lim_{\beta\to\infty}\int_0^{\sqrt{\alpha}}\exp\left(-\left(1-e^{-\frac{s}{\beta}}\right)\sqrt{\alpha}u-\frac{u^2}{2}\right)\mathrm{d}u\\
&= \lim_{\beta\to\infty}\int_0^{\sqrt{\alpha}}\exp\left(-\left(\frac{s}{\beta}+o(\frac{s}{\beta})\right)\sqrt{\alpha}u-\frac{u^2}{2}\right)\mathrm{d}u\\
&= \lim_{\beta\to\infty}\int_0^{\sqrt{\alpha}} e^{-\frac{\sqrt{\alpha}}{\beta}su-\frac{u^2}{2}+o(\frac{s}{\beta})\sqrt{\alpha}u}\,\mathrm{d}u\\
&= \int_0^{\infty} e^{-\sqrt{\frac{\pi}{2}}su-\frac{u^2}{2}}\,\mathrm{d}u,
\end{aligned}$$

where the second equality is due to the Taylor expansion of $1-e^{-\frac{s}{\beta}}$, and the last one to Proposition 4.9. The inequality is in fact an equality, because of Lebesgue's convergence theorem. Then, following from (4.3.18), we get the convergence of the Laplace transform,

$$\begin{aligned}
\lim_{\beta\to\infty}\phi_\beta(s) &= \int_0^\infty e^{-\sqrt{\frac{\pi}{2}}su-\frac{u^2}{2}}\,\mathrm{d}u\Big/\sqrt{\frac{\pi}{2}}\\
&= \int_0^\infty e^{-su}\frac{e^{-\frac{u^2}{\pi}}}{\pi/2}\,\mathrm{d}u,
\end{aligned}$$

and the density of $\overline{X}^R_\infty$ is deduced. □

Remark 4.11. *The result in Theorem 4.10 can be deduced from Theorem 4.7, taking the limit $\delta\to 0$, together with Proposition 4.9. In order to obtain the same expression for the density functions, a change of variables is needed.*

4.4 Appendix

Calculation of Equation (4.3.2): From Equation (4.3.1) we have that

$$\pi_\beta(n) = \frac{c_1\beta\gamma}{\alpha W(n) + c_1 n\gamma} \prod_{k=0}^{n-1} \frac{\alpha W(k)}{\alpha W(k) + c_1 k\gamma}.$$

The normalisation relation for π_β states that $\sum_{n\geq 0} \pi_\beta(n) = 1$, then

$$\begin{aligned}
\frac{1}{\beta} &= \sum_{n\geq 0} \frac{c_1\gamma}{\alpha W(n) + c_1 n\gamma} \prod_{k=0}^{n-1} \frac{\alpha W(k)}{\alpha W(k) + c_1 k\gamma} \\
&= \frac{c_1\gamma}{\alpha W(0)} + \sum_{n\geq 1} \frac{n}{n} \frac{c_1\gamma}{\alpha W(n) + c_1 n\gamma} \prod_{k=0}^{n-1} \frac{\alpha W(k)}{\alpha W(k) + c_1 k\gamma} \\
&= \frac{c_1\gamma}{\alpha W(0)} + \sum_{n\geq 1} \frac{1}{n} \left(\frac{c_1 n\gamma + \alpha W(n) - \alpha W(n)}{\alpha W(n) + c_1 n\gamma} \right) \prod_{k=0}^{n-1} \frac{\alpha W(k)}{\alpha W(k) + c_1 k\gamma} \\
&= \frac{c_1\gamma}{\alpha W(0)} + \sum_{n\geq 1} \frac{1}{n} \left(\prod_{k=0}^{n-1} \frac{\alpha W(k)}{\alpha W(k) + c_1 k\gamma} - \prod_{k=0}^{n} \frac{\alpha W(k)}{\alpha W(k) + c_1 k\gamma} \right) \\
&= \frac{c_1\gamma}{\alpha W(0)} + \sum_{n\geq 1} \frac{1}{n} \left(\left(1 - \prod_{k=0}^{n} \frac{\alpha W(k)}{\alpha W(k) + c_1 k\gamma} \right) - \left(1 - \prod_{k=0}^{n-1} \frac{\alpha W(k)}{\alpha W(k) + c_1 k\gamma} \right) \right) \\
&= \frac{c_1\gamma}{\alpha W(0)} + \sum_{n\geq 1} \left(\frac{1}{n} - \frac{1}{n+1} \right) \left(1 - \prod_{k=0}^{n} \frac{\alpha W(k)}{\alpha W(k) + c_1 k\gamma} \right) \\
&= \frac{c_1\gamma}{\alpha W(0)} + \sum_{n\geq 1} \frac{1}{n(n+1)} \left(1 - \prod_{k=0}^{n} \frac{\alpha W(k)}{\alpha W(k) + c_1 k\gamma} \right).
\end{aligned}$$

□

Chapter 5

Conclusions

Contents

Throughout this thesis we studied resource allocation problems in large-scale stochastic systems. The mathematical framework used was the Markov Decision Processes (MDPs), and, in particular, we focused on a subclass of MDPs known as Multi-armed restless bandits problems (MARBP). We worked on problems with conditions that fluctuate over time. This was modelled considering two types of processes, controllable processes and environments. The current state of the environment determines the value of the parameters of the controllable processes. We differentiated whether the environment can be observed or not, and under both settings we looked for efficient control policies in a long-term average criterion. We further studied a model of large-scale storage systems with files distributed across nodes. After a node breaks down, all the files it had are reallocated to other nodes. We analysed the evolution of the load of a single node.

Our main contributions encompasses a range of results on optimal control and on the performance of allocation policies. For our first problem we considered an MARBP with unobservable environments. We studied the asymptotic regime, as the number of bandits grows large together with speed of the environment. In Theorem 2.5 we provided sufficient conditions for a policy to be asymptotically optimal, and we defined a set of priority policies that satisfies these. When the technical condition known as indexability holds, an averaged version of Whittle's index is inside the set of priority policies, see Proposition 2.13. The second problem we considered is an MARBP with observable environments. We focused on the relaxed version of the problem, and we proposed Algorithm 3.4 to determine Whittle's index. For the problem of a queue with abandonments we prove that indexability holds in Theorem 3.14, and we derive the Whittle's index in closed-form in Theorem 3.15. In both problems we delivered a numerical evaluation to show the performance of the obtained policies. For the problem of large-scale storage systems, we focused on the mean-field regime as the number of nodes and files grows large. By studying the global balance equations, we proved the existence of the process in the mean-field regime, in Theorem 4.5. We further showed the

convergence in distribution of the load in steady state as the average number of files per node tends to infinity, in Theorem 4.7.

5.1 Impact of observability

One of the motivations of this thesis was to study the implications of observability or unobservability on Whittle's index. Numerically we analysed this impact in Sections 2.5.5 and 3.6.3. Table 5.1 summarises our main theoretical characterisations in each of the cases.

In the unobservable case, we proved in Proposition 2.13 that the averaged Whittle's index policy is asymptotically optimal in a fast environment as the number of bandits grows large. In the case of queues with abandonments, the averaged Whittle's index reduces to the index $\overline{\mu}_k/\overline{\theta}_k$, where $\overline{\mu}_k$ is the averaged service rate and $\overline{\theta}_k$ is the averaged abandonment rate of bandit k. We leave for future work the problem of optimal control with environments that vary relatively slow or with normal speed. In Section 5.2.2, we present some ideas in this direction.

In the observable case, we considered policies that can depend on the current state of the environment $d \in \mathcal{Z}$, and this is reflected in our results. Algorithm 3.4 provides expressions for the Whittle's index in the general setting. In particular, we proved in Proposition 3.8 that when the environment is slow the Whittle's index coincides with the Whittle's index of a bandit that only sees environment d. We denoted this index by $W_k^{(d)}(m)$ for bandit k in state m. For the model of queues with abandonments, Theorem 3.15 provides the Whittle's index in closed-form. In Proposition 3.19, we obtain simpler expressions for the slow and fast regime, as shown in the table. For the slow regime, the index is $\mu_k^{(d)}/\theta_k^{(d)}$, where $\mu_k^{(d)}$ is the service rate and $\theta_k^{(d)}$ the abandonment rate of bandit k when the state of the environment is d. For the fast regime, the index is $\mu_k^{(d)}/\overline{\theta}_k$. We make the remark that this index remains dependent on state d of the environment, in contrast to the index in the unobservable model, which takes both averaged parameter values.

As a general conclusion, the results for the observable case are more complete than for the unobservable case. We comment further on this in the next section.

		Speed of the environment		
		Slow	Normal	Fast
Unobservable	General problem	*Future work*	*Future work*	Averaged Whittle's index
	Queues with abandonments			$\overline{\mu}_k/\overline{\theta}_k$
Observable	General problem	$W_k^{(d)}(m)$	Algorithm 3.4	Algorithm 3.4
	Queues with abandonments	$\mu_k^{(d)}/\theta_k^{(d)}$	Theorem 3.15	$\mu_k^{(d)}/\overline{\theta}_k$

Table 5.1: Index policies obtained for different models and regimes of the speed of the environment.

5.2 Future work

We outline various lines of research that follow from the works presented in this thesis.

5.2.1 Activation constraint depending on the environment

In Chapters 2 and 3, we considered a unique sample-path constraint on the number of active bandits that does not depend on the state of the environments. Following the notation of Chapter 3, this was stated in Equation (3.1.1). An alternative approach can be considered, by proposing a sample-path constraint per state of the environments. For given values $\left\{R_{\vec{d}} : \vec{d} \in \mathcal{Z}^N\right\}$, we consider the following set of constraints:

$$\sum_{k=1}^{N} A_k^{\varphi}(t)\mathbf{1}_{\left(\vec{D}(t)=\vec{d}\right)} \leq R_{\vec{d}}\mathbf{1}_{\left(\vec{D}(t)=\vec{d}\right)}, \quad \forall t \geq 0 \text{ and } \forall \vec{d} \in \mathcal{Z}^N, \tag{5.2.1}$$

where $\vec{D}(t) := (D_1(t), \ldots, D_N(t))$ denotes the state of the environment processes of all the bandits. The relaxed version of constraints (5.2.1) states that the number of active bandits must be satisfied on average, and not in every decision epoch.

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\sum_{k=1}^{N}\int_0^T A_k^{\varphi}(t)\mathbf{1}_{\left(\vec{D}(t)=\vec{d}\right)}\mathrm{d}t\right) \leq R_{\vec{d}}\phi(\vec{d}) \quad \forall \vec{d} \in \mathcal{Z}^N,$$

where $\phi(\vec{d})$ is the stationary probability of $\vec{D}(t)$ to be in state $\vec{d}$. The Lagrangian multipliers approach applied for several constraints leads us to consider several multipliers, $W_{\vec{d}}$, one for each state $\vec{d}$ the environments can be in:

$$\min_{\varphi}\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \left(\sum_{k=1}^{N} C_k(M_k^{\varphi}(t), D_k(t), A_k^{\varphi}(t)) - W_{\vec{d}}\mathbf{1}_{\left(\vec{D}(t)=\vec{d}\right)}\sum_{k=0}^{N} A_k^{\varphi}(t)\right)\mathrm{d}t\right) \quad \forall \vec{d} \in \mathcal{Z}^N.$$

We leave for a future work the study of Whittle's index obtained after this relaxation. In particular, we believe that for the observable setting, Whittle's index will not only depend on the state of the bandit (m, d), as we have in Chapter 3, but also on the environment of the other bandits.
In [28], a relaxation of an MARBP problem with more than one sample-path constraint is proposed, resulting in multiple Lagrangian multipliers. The authors formulate a related Whittle's index.

5.2.2 Unobservable environments

As it is stated in Table 5.1, it remains as future work to study optimal control for bandits affected by unobservable environments, in cases where the environment does not change relatively fast. In addition to the environment changing fast, another assumption in our analysis was that the number of bandits tends to infinity. Thus, the mathematical analysis of the unobservable case remains largely unexplored.
The problem with unobservable environments can be considered as a MARBP with *partially observable states*. A reference in Markov Processes with incomplete state information is [8]. One possibility to get more insights would be to resort to learning algorithms, in particular *reinforcement learning* (RL). This is

an area of research that recently received a lot of attention, a full overview can be found in the textbook [65].

Q-learning

We refer here to *Q-learning*, a technique used to find optimal control for MDPs without specific knowledge of its transition rules. In Q-learning, the expected one-step reward is estimated for each state and action, by repeating many episodes of the process and updating the followed policy according to observations.
Such technique can be explored for an MARBP with unobservable environments in order to learn the optimal policy, at least numerically. As the experiment is run several times, under ergodicity assumptions, all the states of the environments are visited. Learning from these transitions allows to obtain an estimation of the evolution of the process.
For a theoretical approach, a reference in this subject is [29]. The authors adapt the Q-learning technique to obtain Whittle's index in a MARBP with unknown transition matrices. They show that the derived policy performs close to the fully-informed Whittle's index policy.

5.2.3 Observable environments: Whittle's index in slow regime

In Proposition 3.8, we proved that, when the environment is slow, Whittle's index coincides with the index of a bandit that only sees a fixed environment $d \in \mathcal{Z}$. In the proposition, and in previous results, we used the technical assumption that the environment visits only two states, i.e., $|\mathcal{Z}| = 2$. Although we were not able to extended the result for $|\mathcal{Z}| > 2$, we believe it holds, and it remains as further research.

5.2.4 Observable environments: optimality for correlated environments

In the classical MARBP theory, the bandits are assumed to be independent. In the model in Chapter 3, the bandits are formed by the controllable processes and the environments. When the environments are independent, we can conclude that the bandits are independent, and their state is a two-dimensional state of a classical bandit. This means that the classical MARBP theory applies, and the associated results hold, such as the asymptotic optimality of Whittle's index policy proved in [71].
However, in case the environments are not independent, the previous statement is not true, and some research questions arise.

- A first question to address is how efficient is Whittle's index policy in the MARBP. In Section 3.6.1 we presented a numerical evaluation with random samples to study the performance of Whittle's index policy in such a setting. Our results suggest that it has a good performance. We would like to further study whether Whittle's index policy is asymptotically optimal, as the number of bandits that can be made active grows proportionally to the total number of bandits.

- An alternative line of research follows if we consider a non-relaxed approach to solve the problem. A course of action on this direction would be to study the equilibrium solutions of the linear-programming problem derived from the fluid dynamics of the system, as it is done for the MARBP in [69].

5.2.5 Large-scale storage systems

We propose some extensions for our main result in Chapter 4, Theorem 4.5. As a first extension, we can consider a quadratic rate for the breakdowns of the nodes, instead of a linear rate. A possible approach to solve this problem is to express the solution of the fixed point equation (4.2.10) as the solution of an ODE of second order. A simpler version of this method is already used in the proof of Theorem 4.8, with an ODE of first order.

Another extension is to consider allocation policies that involve a more complex analysis than the Random Weighted policy. In particular, we are interested in the so-called *Power of d-Choices* policy, where, for allocating a file after a breakdown, the system chooses d nodes at random, and allocates the file to the least loaded node among this subset. We would like to generalise the results in [63], where the Power of d-Choices policy is considered for a system with breakdowns at a constant rate. The authors show that in the mean-field regime the process behaves as a non-homogeneous Markov process, and the stationary measure is then derived.

Self References

[SR1] S. Duran, I. M. Verloop. “Asymptotic optimal control of Markov-modulated restless bandits". In: *Proceedings of ACM SIGMETRICS* (2018), vol. 2, pp. 1-25.

[SR2] S. Duran, U. Ayesta, I. M. Verloop. “On the Whittle’s Index of Markov-modulated restless bandits". *Submitted to Queueing Systems* (2020).

[SR3] S. Duran, P. Robert. “Analysis of a stochastic model of replication with linear breakdown rates". *Preprint* (2020).

Bibliography

[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*. 2015, pp. 57–69 (cit. on pp. 3, 38).

[2] S.H.A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari. "Optimality of myopic sensing in multi-channel opportunistic access". In: *IEEE Transactions on Information Theory* 55 (2009), pp. 4040–4050 (cit. on p. 8).

[3] E. Altman, K.E. Avrachenkov, and R. Núnez-Queija. "Perturbation analysis for denumerable Markov chains with application to queueing models". In: *Advances in Applied Probability* 36.3 (2004), pp. 839–853 (cit. on p. 53).

[4] 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 (cit. on pp. 3, 7, 38).

[5] 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". In: *Mathematical Methods of Operations Research* 57 (2003), pp. 21–39 (cit. on p. 8).

[6] J. Anselmi. "Asymptotically optimal open-loop load balancing". In: *Queueing Systems* 87 (2017), pp. 245–267 (cit. on p. 2).

[7] A. Asanjarani and Y. Nazarathy. "The role of information in system stability with partially observable servers". In: *Arxiv report* 1610.02781v1 (2016) (cit. on p. 2).

[8] K.J. Åström. "Optimal control of Markov processes with incomplete state information". In: *Journal of Mathematical Analysis and Applications* 10.1 (1965), pp. 174–205 (cit. on p. 105).

[9] U. Ayesta, M. Erausquin, and P. Jacko. "A Modeling Framework for Optimizing the Flow-Level Scheduling with Time-Varying Channels". In: *Performance Evaluation* 67 (2010), pp. 1014–1029 (cit. on p. 8).

[10] U. Ayesta, P. Jacko, and V. Novak. "A Nearly-Optimal Index Rule for Scheduling of Users with Abandonment". In: *Proceedings of IEEE INFOCOM*. Hong Kong, 2011 (cit. on p. 8).

[11] R. Bellman. "A Markovian decision process". In: *Journal of mathematics and mechanics* (1957), pp. 679–684 (cit. on pp. 4, 5).

[12] M. Benaïm and J-Y Le Boudec. "A class of mean field interaction models for computer and communication systems". In: *Performance Evaluation* 65 (2008), pp. 823–838 (cit. on pp. 9, 13, 14).

[13] D.P. Bertsekas. *Dynamic programming and optimal control*. Vol. 1. 2. Athena scientific Belmont, MA, 1995 (cit. on p. 4).

[14] 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". In: *Queueing Systems* 76.4 (2014), pp. 425–446 (cit. on pp. 7, 46, 67).

[15] P. Billingsley. *Convergence of probability measures*. Wiley, New York NY, 1968 (cit. on pp. 17, 31).

[16] A. Bobbio, M. Gribaudo, and M. Telek. "Analysis of large scale interacting systems by mean field method". In: *2008 Fifth International Conference on Quantitative Evaluation of Systems*. IEEE. 2008, pp. 215–224 (cit. on pp. 9, 10).

[17] C. Bordenave, D. McDonald, and A. Proutiére. "A particle system in interaction with a rapidly varying environment: Mean field limits and applications". In: *Networks and heterogeneous media* 5.1 (2010), pp. 31–62 (cit. on pp. 10, 13, 17, 19, 29, 30).

[18] D. Borthakur et al. "HDFS architecture guide". In: *Hadoop Apache Project* 53.1-13 (2008), p. 2 (cit. on pp. 3, 85).

[19] R.J. Boucherie and N.M. Van Dijk. *Queueing networks: a fundamental approach*. Vol. 154. Springer Science & Business Media, 2010 (cit. on pp. 11, 68, 69, 79, 80).

[20] P.H. Brill. *Level crossing methods in stochastic models*. Vol. 13. Springer, 2008 (cit. on p. 47).

[21] A. Budhiraja, A. Ghosh, and X. Liu. "Scheduling control for Markov-modulated single-server multiclass queueing systems in heavy traffic". In: *Queueing Systems* 78.1 (2014), pp. 57–97 (cit. on p. 2).

[22] C. Buyukkoc, P. Varaiya, and J. Walrand. "The cμ rule revisited". In: *Advances in applied probability* 17.1 (1985), pp. 237–238 (cit. on p. 2).

[23] E. Çinlar. *Introduction to Stochastic Processes*. New Jersey: Prentice-Hall, 1975 (cit. on p. 32).

[24] B.D. Coleman. "A stochastic process model for mechanical breakdown". In: *Transactions of the Society of Rheology* 1.1 (1957), pp. 153–168 (cit. on p. 86).

[25] J.G. Dai and S. He. "Many-server queues with customer abandonment: A survey of diffusion and fluid approximations". In: *Journal of Systems Science and Systems Engineering* 21.1 (2012), pp. 1–36 (cit. on p. 38).

[26] N.M. van Dijk. "Approximate uniformization for continuous-time Markov chains with an application to performability analysis". In: *Stochastic processes and their applications* 40.2 (1992), pp. 339–357 (cit. on p. 69).

[27] N. Ehsan and M. Liu. "On the optimality of an index policy for bandwidth allocation with delayed state observation and differentiated services". In: *Proceedings of IEEE INFOCOM*. Hong Kong, 2004 (cit. on p. 8).

[28] J. Fu, B. Moran, and P.G. Taylor. "Restless bandits in action: Resource allocation, competition and reservation". In: *arXiv preprint arXiv:1804.02100* (2018) (cit. on p. 105).

[29] J. Fu, Y. Nazarathy, S. Moka, and P.G. Taylor. "Towards Q-learning the Whittle Index for Restless Bandits". In: *2019 Australian & New Zealand Control Conference (ANZCC)*. IEEE. 2019, pp. 249–254 (cit. on p. 106).

[30] N. Gast and B. Gaujal. "A mean field approach for optimization in discrete time". In: *Discrete Event Dynamic Systems* 21.1 (2011), pp. 63–101 (cit. on p. 3).

[31] J.M. George and J.M. Harrison. "Dynamic control of a queue with adjustable service rate". In: *Operations research* 49.5 (2001), pp. 720–731 (cit. on p. 2).

[32] S. Ghemawat, H. Gobioff, and S-T Leung. "The google file system". In: *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles* (2003) (cit. on pp. 4, 85).

[33] J. Gittins, K. Glazebrook, and R. Weber. *Multi-Armed Bandit Allocation Indices*. Chichester: John Wiley & Sons, 1989 (cit. on p. 39).

[34] J.C. Gittins. "Bandit processes and dynamic allocation indices". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2 (1979), pp. 148–164 (cit. on p. 7).

[35] K.D. Glazebrook, C. Kirkbride, and J. Ouenniche. "Index policies for the admission control and routing of impatient customers to heterogeneous service stations". In: *Operations Research* 57 (2009), pp. 975–989 (cit. on p. 8).

[36] K.D. Glazebrook and H.M. Mitchell. "An index policy for a stochastic scheduling model with improving/ deteriorating jobs". In: *Naval Research Logistics* 49 (2002), pp. 706–721 (cit. on p. 8).

[37] J. Hasenbein and D. Perry (Eds.) *Special Issue on Queueing Systems with Abandonments*. 2013 (cit. on p. 38).

[38] O. Hernández-Lerma and J.B. Lasserre. *Discrete-time Markov control processes: basic optimality criteria*. Vol. 30. Springer Science & Business Media, 2012 (cit. on pp. 4, 6).

[39] D.J. Hodge and K. D. Glazebrook. "On the asymptotic optimality of greedy index heuristics for multi-action restless bandits". In: *Advances in Applied Probability* 47 (2015), pp. 652–667 (cit. on p. 22).

[40] 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". In: *IEEE/ACM Transactions on Networking* 23.3 (2014), pp. 880–893 (cit. on p. 9).

[41] A. Karthik, A. Mukhopadhyay, and R.R. Mazumdar. "Choosing among heterogeneous server clouds". In: *Queueing Systems* 85.1-2 (2017), pp. 1–29 (cit. on p. 86).

[42] A. Lakshman and P. Malik. "Cassandra: a decentralized structured storage system". In: *ACM SIGOPS Operating Systems Review* 44.2 (2010), pp. 35–40 (cit. on pp. 4, 85, 86).

[43] 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 (cit. on pp. 8, 45, 49).

[44] M. Larrañaga, U. Ayesta, and I.M. Verloop. "Asymptotically optimal index policies for an abandonment queue with convex holding cost". In: *Queueing Systems* 81.2-3 (2015), pp. 99–169 (cit. on pp. 38, 48, 53, 57).

[45] M. Larrañaga, U. Ayesta, and I.M. Verloop. "Dynamic control of birth-and-death restless bandits: application to resource-allocation problems". In: *IEEE/ACM Transactions on Networking* 24.6 (2016), pp. 3812–3825 (cit. on pp. 8, 23, 24, 41).

[46] J-Y Le Boudec, D. McDonald, and J. Mundinger. "A generic mean field convergence result for systems of interacting objects". In: *Fourth international conference on the quantitative evaluation of systems (QEST 2007)*. IEEE. 2007, pp. 3–18 (cit. on pp. 9, 10).

[47] K. Liu and Q. Zhao. "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access". In: *IEEE Transactions on Information Theory* 56 (2010), pp. 5547–5567 (cit. on p. 8).

[48] S.T. Maguluri, R. Srikant, and L. Ying. "Stochastic models of load balancing and scheduling in cloud computing clusters". In: *2012 Proceedings IEEE Infocom*. IEEE. 2012, pp. 702–710 (cit. on p. 86).

[49] 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.* Springer-Verlag, 2007, pp. 121–308 (cit. on p. 39).

[50] Y. Nazarathy, T. Taimre, A. Asanjarani, J. Kuhn, B. Patch, and A. Vuorinen. "The challenge of stabilizing control for queueing systems with unobservable server states". In: *IEEE Proceedings of the 5th Australian Control Conference*. 2015 (cit. on p. 2).

[51] J. Niño-Mora. "Dynamic priority allocation via restless bandit marginal productivity indices". In: *TOP* 15 (2007), pp. 161–198 (cit. on p. 8).

[52] J. Niño-Mora. "Marginal productivity index policies for admission control and routing to parallel multi-server loss queues with reneging". In: *Lecture Notes in Computer Science* 4465 (2007), pp. 138–149 (cit. on p. 8).

[53] 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)*. IEEE. 2011, pp. 1–8 (cit. on p. 9).

[54] J. R. Norris. *Markov chains*. Vol. 2. Cambridge Series in Statistical and Probabilistic Mathematics. Reprint of 1997 original. Cambridge University Press, Cambridge, 1998, pp. xvi+237 (cit. on pp. 6, 30).

[55] W. Ouyang, A. Eryilmaz, and N.B. Shroff. "Asymptotically optimal downlink scheduling over Markovian fading channels". In: *2012 Proceedings IEEE INFOCOM*. IEEE. 2012, pp. 1224–1232 (cit. on pp. 8, 9).

[56] M.T. Özsu and P. Valduriez. *Principles of distributed database systems*. Springer Science & Business Media, 2011 (cit. on pp. 4, 85).

[57] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: Wiley, 1994 (cit. on pp. 4, 46, 67).

[58] V. Raghunathan, V. Borkar, M. Cao, and P.R. Kumar. "Index policies for real-time multicast scheduling for wireless broadcast systems". In: *Proceedings of IEEE INFOCOM*. 2008 (cit. on p. 8).

[59] A.W. Richa, M. Mitzenmacher, and R. Sitaraman. "The power of two random choices: A survey of techniques and results". In: *Combinatorial Optimization* 9 (2001), pp. 255–304 (cit. on p. 86).

[60] S.M. Ross. *Introduction to stochastic dynamic programming*. Academic press, 2014 (cit. on pp. 4, 5).

[61] Z. Schechner. "A load-sharing model: The linear breakdown rule". In: *Naval research logistics quarterly* 31.1 (1984), pp. 137–144 (cit. on p. 86).

[62] A.L. Stolyar. "Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic". In: *The Annals of Applied Probability* 14.1 (2004), pp. 1–53 (cit. on p. 52).

[63] W. Sun, V. Simon, S. Monnet, P. Robert, and P. Sens. "Analysis of a stochastic model of replication in large distributed storage systems: A mean-field approach". In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1.1 (2017), p. 24 (cit. on pp. 4, 86, 107).

[64] W. Sun, P. Robert, et al. "Analysis of large urn models with local mean-field interactions". In: *Electronic Journal of Probability* 24 (2019) (cit. on p. 86).

[65] R.S. Sutton and A.G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018 (cit. on pp. 14, 106).

[66] A-S Sznitman. "Topics in propagation of chaos". In: *Ecole d'été de probabilités de Saint-Flour XIX—1989*. Springer, 1991, pp. 165–251 (cit. on p. 9).

[67] A.S. Tanenbaum and M. Van Steen. *Distributed systems: principles and paradigms*. Prentice-Hall, 2007 (cit. on pp. 4, 85).

[68] H.C. Tijms. *Stochastic modelling and analysis: a computational approach*. John Wiley & Sons, Inc., 1986 (cit. on pp. 68, 88).

[69] I.M. Verloop. "Asymptotically optimal priority policies for indexable and nonindexable restless bandits". In: *The Annals of Applied Probability* 26.4 (2016), pp. 1947–1995 (cit. on pp. 9, 13, 22, 106).

[70] J. Von Neumann. "Probabilistic logics and the synthesis of reliable organisms from unreliable components". In: *Automata studies* 34 (1956), pp. 43–98 (cit. on p. 85).

[71] R.R. Weber and G. Weiss. "On an index policy for restless bandits". In: *Journal of Applied Probability* 27.03 (1990), pp. 637–648 (cit. on pp. 9, 13, 22, 45, 106).

[72] P. Whittle. "Restless bandits: Activity allocation in a changing world". In: *Journal of applied probability* 25.A (1988), pp. 287–298 (cit. on pp. 7, 8, 40).

[73] P. Whittle. *Optimal Control, Basics and Beyond*. John Wiley & Sons, 1996 (cit. on p. 39).