

THE STRUCTURE AND PERFORMANCE OF OPTIMAL ROUTING SEQUENCES

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus Dr. D. D. Breimer,
hoogleraar in de faculteit der Wiskunde en
Natuurwetenschappen en die der Geneeskunde,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 24 juni 2003
te klokke 15.15 uur

door

Dingeman Aren van der Laan

geboren te Alphen aan den Rijn
op 8 september 1976

Samenstelling van de promotiecommissie:

promotoren: Prof. dr. A. Hordijk
Prof. dr. R. Tijdeman

referent: Dr. B. O. Gaujal (ENS Lyon, France)

overige leden: Prof. dr. ir. O. J. Boxma (TU Eindhoven)
Prof. dr. G. van Dijk
Prof. dr. L. C. M. Kallenberg
Prof. dr. G. M. Koole (VU Amsterdam)

The structure and performance
of optimal routing sequences

Dinard van der Laan

THOMAS STIELTJES INSTITUTE
FOR MATHEMATICS



Contents

1	Introduction	9
2	Optimal routing to fully loaded parallel queueing systems with deterministic interarrival and service times	17
2.1	Introduction	17
2.2	The queueing model	20
2.3	The relation between waiting time and unused capacity	22
2.4	An upper bound for the minimal long-run average waiting time	29
2.5	The structure of optimal policies	35
2.6	The optimal policy in case of rational service rates	42
2.6.1	Bounds corresponding to the mathematical programming problem	48
2.7	The optimality of regular policies	55
2.8	Algorithms to find good policies	64
3	Analysis of the performance of periodic routing sequences	73
3.1	Introduction	73
3.2	Comparing routing sequences	77
3.3	Bounding the difference in expected average waiting time between sequences	90

3.4	Routing to parallel queues	98
3.5	Billiard sequences and routing sequences	106
3.6	Appendix	116
4	Deterministic parallel queueing systems: on the average waiting time for regular routing and the corresponding lower bound	119
4.1	Introduction	119
4.2	Description of the queueing system and notation	122
4.2.1	Routing policies and words	123
4.3	Definitions and preliminary results	124
4.4	The upper bracket sequence and Farey intervals	126
4.4.1	Factorisation of the upper bracket sequence	126
4.5	The average number of customers in a single queue	128
4.6	Computation of the average waiting time	133
4.6.1	The value of $W(d)$ in best lower approximation points	133
4.6.2	Best lower approximations and the continued fraction expansion	137
4.7	Minimization over multiple queues	141
4.7.1	Properties of a minimal point	141
4.7.2	Algorithms for determining a minimal point	147
4.8	Appendix: An extension of Little's Law for routing policies	152
5	On the optimality of a stationary policy for deterministic parallel queueing systems	157
5.1	Introduction	157
5.2	The optimality of a stationary policy	158
5.2.1	The description of the Markov Decision Chain	158
5.2.2	Definitions and theory on the MDC	159

5.2.3	Sufficient conditions for the existence of an optimal stationary policy	161
5.2.4	Verification of the assumptions	162
5.3	Properties of routing sequences corresponding to optimal stationary policies	166
5.3.1	Ultimately periodic routing sequences	166
5.3.2	The existence of an optimal periodic routing sequence in case of rational service times	167
5.4	Optimal routing sequences in case of irrational service times	168
6	Multimodular functions and partial orders on routing sequences	171
6.1	Introduction	171
6.2	The multimodular order and the cone order	172
6.2.1	The cone order	173
6.2.2	The multimodular order and the cone order	176
6.2.3	Shift invariant counterparts	180
6.3	The graph order and the unbalance	182
6.4	Relations and counterexamples	188
6.4.1	The shift invariant cone order and the graph order	188
6.4.2	The shift invariant multimodular order and the shift invariant cone order	189
6.4.3	The graph order and the shift invariant multimodular order	190
6.4.4	The shift invariant orders: conclusion	190
	Index	199
	Index of notation.	201

Chapter 1

Introduction

We consider queueing systems in which arriving customers (jobs) are routed at the moment of their arrival to $N \geq 2$ parallel servers, each having its own queue with infinite buffer size. The N parallel queues are assumed to be FIFO (First In First Out) queues, i.e. in every queue the customers are served in the same order as they are routed to the queue. In this thesis we use the description parallel queueing system for such systems. Let T_n be the n -th arrival epoch, that is the moment at which the n -th customer arrives. Then T_n is also a decision epoch, since arriving customers are routed at the moment of their arrival. It is usually assumed that the system is empty at T_1 , the moment that the first customer arrives, i.e there is no load in any of the queues. In this thesis we assume in general that the interarrival times $\delta_n := T_{n+1} - T_n$, $n = 1, 2, \dots$ are independent and identically distributed (i.i.d.) random variables. We number the N parallel queues by $1, 2, \dots, N$. For $i = 1, 2, \dots, N$ and $n = 1, 2, \dots$, let σ_i^n be the service time of the n -th customer that is routed to queue i . We shall assume in general that for $i = 1, 2, \dots, N$ the σ_i^n , $n = 1, 2, \dots$, are i.i.d. random variables and that they are independent of the interarrival times. However, in general the σ_i^n , $n = 1, 2, \dots$ and σ_j^n , $j = 1, 2, \dots$ are differently distributed if $i \neq j$, i.e. the servers may be heterogeneous.

A routing policy is roughly speaking a sequence of decision rules which for every decision epoch T_n prescribes the way for choosing a server to which the arriving customer is routed. In this thesis we focus on open-loop control, which means that the routing policy (control) is independent of time dependent information on the system, such as the sizes of the queues at the decision epoch. So, in case of open-loop control there is no current state information at the decision epochs and the routing

policy depends only on the base parameters of the system. For the parallel queueing systems that we consider these base parameters are the distribution of the interarrival times δ_n and the distributions of the service times σ_i^n , $i = 1, 2, \dots, n$. This type of control using no current state information is also called static routing control (see [20]). Open-loop control is used in many telecommunication networks. In this thesis the performance criterion for the routing policy is usually the (expected) total average waiting time over all arriving customers, that is the expectation of $\limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{n=1}^t W_n$, where W_n is the waiting time of the customer arriving at arrival epoch T_n . Sometimes the performance criterion is generalized to other functions of the (expected) waiting times.

In contrast to open-loop control there is closed-loop control in which the controller may use time dependent information on the system at every decision epoch. This is also called dynamic control. See [35] and [45] for an overview of results on closed-loop control. In general the knowledge of time dependent information in closed-loop control yields a better performance than in open-loop control. However, since closed-loop control must be done online, the implementation is more difficult than for open-loop control. Therefore open-loop control is chosen in many applications.

In case of open-loop control for parallel queueing systems there is a distinction between probabilistic and deterministic routing policies. Probabilistic routing is also known as Bernoulli routing or splitting (see for example [19] and [44]). For the Bernoulli routing it is proved that equal routing probabilities are optimal in case of homogeneous servers. In case of heterogeneous servers the optimal routing probabilities for the Bernoulli routing are in general distinct.

A deterministic routing policy is by definition a policy such that for every decision epoch T_n the server $i \in \{1, 2, \dots, N\}$, to which the customer arriving at T_n is routed, is determined. Then the routing policy can be described by an infinite sequence of integers $U = (U_1, U_2, \dots)$, where U_n is the server to which the customer arriving at decision epoch T_n is routed. A deterministic routing policy and the corresponding routing sequence are called optimal for a parallel queueing system if it gives an expected total average waiting time which is minimal among all deterministic routing policies. Routing according to a deterministic routing sequence is just as Bernoulli routing some type of open loop control. However, in general optimal (or even sub-optimal) deterministic routing policies have a performance which is superior to the performance of optimal Bernoulli routing. Deterministic routing is also called semi-dynamic routing (see [69]). In this thesis we investigate this type of control and the corresponding deterministic routing sequences. In case of a parallel queueing system with homogeneous servers it is proved that round robin routing is optimal in [47].

In case of heterogeneous servers, as we consider in this thesis, routing according to a deterministic routing sequence is also known as generalized round robin routing (see [11]). In general constructing an optimal generalized round robin policy is difficult. In [2] and [36] the problem is transformed to a Markov decision process in case of exponential service times. We derive some structural results for optimal routing sequences in various parallel queueing systems. For example we prove the existence of an optimal periodic routing sequence for certain parallel queueing systems. For other systems we prove the existence of an optimal billiard sequence (see Section 2.5).

We have a particular interest in parallel queueing systems for which the interarrival times are deterministic (and thus constant) and also the service times are deterministic. The transmission of data can often be modeled by constant service times, as the data is split in packets of constant size. If generalized round robin routing is used in a parallel queueing system with deterministic interarrival and service times then the evolution of the system is completely determined by the routing sequence and the initial state of the system. So, the base parameters of the system give the controller all information and therefore there is not an essential difference between open-loop and closed-loop control. Thus in that case routing according to an optimal deterministic routing sequence gives the same performance as optimal closed-loop control.

We also consider the routing to a single server i , $i \in \{1, 2, \dots, N\}$ of a parallel queueing system with N parallel queues. Therefore, for a given routing sequence $U = (U_1, U_2, \dots)$ for N parallel queues we define for $i = 1, 2, \dots, N$ a sequence $u^i = (u_1^i, u_2^i, \dots)$ of zeros and ones by $u_n^i = 1$ if $U_n = i$ and $u_n^i = 0$ if $U_n \neq i$. An infinite sequence of zeros and ones can be considered as a routing sequence for a single server queue and in particular u^i is the routing sequence for server (queue) i . A routing sequence for a single server queue is also called a splitting sequence or admission sequence, since the ones correspond to the customers that are admitted to the queue, while the zeros correspond to customers that are not admitted to the queue. So, if you consider a single server queue i then the stream of arriving customers is split according to this sequence of zeros and ones.

For a given routing sequence $U = (U_1, U_2, \dots)$ put $N_i^t := |\{n \leq t : U_n = i\}|$, the number of times a customer is routed to server i among the first t arriving customers. If $\lim_{t \rightarrow \infty} \frac{N_i^t}{t}$ exists, then this limit is the fraction of customers routed to queue i by routing sequence U . In that case we say that splitting sequence u^i has density $d_i := \lim_{t \rightarrow \infty} \frac{N_i^t}{t}$, the density of the ones in the sequence. Considering routing sequences for N parallel queues we are interested in the existence of an

optimal routing sequence U for which every splitting sequence u^i has a density d_i . Then $\sum_{i=1}^N d_i = 1$. Note that this condition on the structure of a routing sequence is weaker than periodicity. So, this problem is particularly interesting if it is not possible to prove that there exists a periodic optimal routing sequence. For fixed densities d_1, d_2, \dots, d_N with $\sum_{i=1}^N d_i = 1$ we obtain for various parallel queueing systems bounds on the best possible performance for routing sequences with these densities (see Section 3.4 and Section 4.7).

An infinite sequence of zeros and ones $u = (u_1, u_2, \dots)$ is said to be regular with density d if every subsequence of length n contains exactly $\lfloor nd \rfloor$ or $\lceil nd \rceil$ ones. Such sequences are balanced, since the difference in the number of ones of subsequences of the same length is not greater than one. Regular sequences are also called bracket sequences, since the support of the ones in a regular sequence is given by an expression of the form $\{\lfloor \frac{n}{d} + \varphi \rfloor\}_{n=1}^{\infty}$ or $\{\lceil \frac{n}{d} + \varphi \rceil\}_{n=1}^{\infty}$, where $\varphi \in \mathbb{R}$ is called the phase of the sequence and d the density of the sequence. Given the density, the distribution of the ones (and also of the zeros) in a regular sequence is the most regular distribution that is possible. In the seminal paper [29] it is proved for an exponential queue that it is optimal to admit the customers according to a regular sequence of density d if (at least) a fraction d of the arriving customers has to be admitted to that queue. A fundamental concept in the proof is multimodularity. Multimodular functions are for functions defined on a lattice set the counterpart of convex functions. In [3] and [5] it is proved by using multimodularity that regular sequences are optimal for the routing (admission) to a single queue for generally distributed stationary sequences of interarrival and service times. Moreover, it is proved that for the routing to a parallel queueing system with $N = 2$ parallel queues there exists an optimal routing sequence $U = (U_1, U_2, \dots)$ and some $d \in [0, 1]$ such that the corresponding splitting sequences u^1 and u^2 are both regular with densities d and $1 - d$ respectively. In other words the optimal routing is such that the routing to each of the queues is regular. In some special cases this also holds if $N > 2$ (see [4]), for example if there are two sets of identical servers. However, if $N \geq 3$ then in general the optimal routing sequence is not a composition of regular sequences, since the regular sequences can not be combined to a feasible routing sequence. In fact it is a hard combinatorial problem to decide for given densities d_1, d_2, \dots, d_N with $\sum_{i=1}^N d_i = 1$ whether the set of (positive) integers can be covered by regular sets with these densities. A set of densities d_1, d_2, \dots, d_N with $\sum_{i=1}^N d_i = 1$ for which this is possible is called balanceable. In general a given set of routing densities is not balanceable and for such densities it is not possible that for every queue the corresponding routing sequence is regular. However, for various systems a lower bound on the minimal expected average waiting time is obtained by assuming that densities are always balanceable.

This lower bound is calculated by computing the expected average waiting time for each of the single server queues, given that the routing is regular.

We also want to compare the performances of routing sequences which are not regular. In particular we develop methods to compare the performance of periodic sequences of the same density. For such routing sequences we define the notion of unbalance, which is a measure for the irregularity of the sequence. Roughly speaking the unbalance of a sequence is its distance to the regular sequence. Using the notion of unbalance we obtain for any periodic sequence a bound on the difference in performance between this sequence and a regular sequence of the same density. A partial order, called the graph order, is used to generalise this result to any pair of ordered sequences. We investigate some more partial orders on routing sequences, like the cone order and multimodular order. These orders are defined such that if two sequences are ordered then the performance of the greater one is better than the performance of the smaller one. We examine the relation between these orders and the graph order.

This thesis is organised as follows. In Chapter 2 the optimal routing to parallel queueing systems with deterministic interarrival and service times is analysed. We consider systems for which the arrival rate is equal to the combined service rate of the parallel servers. In fact we assume that the interarrival is equal to 1 and that $\sum_{i=1}^n a_i = 1$, where n is the number of servers and a_i , $i = 1, 2, \dots, n$, is the service rate of server i . So, in this model we have that a_i^{-1} is the service time of a job routed to server i . Moreover, the traffic intensity $\rho := \frac{1}{\sum_{i=1}^n a_i}$ satisfies $\rho := 1$ and we say that the system is fully loaded. In case of stochastic interarrival and service times it is known that a system overflows if $\rho \geq 1$ and thus waiting times tend to ∞ . However, in case of deterministic interarrival and service times the system can just be stabilized if $\rho = 1$ and thus there exist routing sequences with finite average waiting time. We deduce for these fully loaded systems a fundamental relation between the total average waiting time and the total unused work capacity. Then we formulate a mathematical programming problem (MPP) for minimizing the total average waiting time and we show that $\frac{n-1}{2}$, where n is the number of parallel queues, is an upper bound for the minimal average waiting time. This upper bound is shown to be tight if the service rates a_i are linearly independent over \mathbb{Z} . In Section 2.5 we introduce an algorithm to construct billiard sequences with given densities. After that we show that there exists an optimal routing sequence, which is a billiard sequence with densities $d_i = a_i$ for $i = 1, 2, \dots, n$. This is a rather strong property and in Section 2.6 we show that this implies the existence of a periodic optimal routing sequence if all the service rates a_i (and thus all the densities d_i and all the service times a_i^{-1}) are rational numbers. In this rational case we show that the minimal average waiting

time and a periodic optimal billiard routing sequence achieving this can be obtained by solving some integer linear problem (ILP). By solving the linear programming (LP) relaxation of this ILP we obtain a lower bound on the minimal average waiting time. Next we show that this lower bound is attained if a routing sequence is used such that all the corresponding splitting sequences u^i are regular with density a_i . Hence the lower bound is tight if the a_i are balanceable. An explicit formula is obtained for the average waiting time in a single server queue i if the routing to that queue is regular with density d_i equal to a_i . Finally some algorithms to obtain good routing sequences are discussed.

In Chapter 3 the difference in performance between periodic routing (splitting) sequences with the same densities is analysed. We obtain bounds on the expected average waiting time for splitting sequences to one queue and for routing sequences for a parallel queueing system. These bounds are insensitive, since they are valid for any distribution of interarrival and service times. In Section 3.2 we start with a combinatorial analysis of finite and periodic sequences of zeros and ones. We introduce the combinatorial notion of unbalance of such sequences, which is a measure for the irregularity of the sequence. In fact we define a primal unbalance and a dual unbalance. Similarly we define several partial orders called the upper graph order, lower graph order and (total) graph order on the set (of conjugacy classes) of infinite periodic sequences of zeros and ones with a given density d . These partial orders are defined such that the regular sequence is smaller than all other sequences. After this combinatorial part we use a sample path comparison to obtain a bound on the difference in expected average waiting time in a single server queue of periodic splitting sequences with a given density d if these sequences are ordered in the upper or lower graph order. The obtained bound depends only on the density d , the mean interarrival time and the difference in the primal or dual unbalance, respectively. Moreover, comparing a periodic splitting sequence with a regular sequence of the same density gives both a lower bound and an upper bound on its performance. The lower bound is given by the performance of the regular sequence, while the difference between the upper bound and the lower bound is proportional to the primal unbalance of the sequence. This upper bound on the performance is shown to be tight for a fully loaded queue with deterministic arrival and service times, as we considered in Chapter 2. The results are extended to routing sequences for parallel queueing systems by defining the total (primal) unbalance as the sum of the (primal) unbalances of the splitting sequences. Subsequently we derive some properties of billiard sequences. We show that for given rational densities there exists a billiard sequence which has minimal total unbalance among all sequences with those densities.

In Chapter 4 we consider parallel queueing systems with deterministic interarrival

and service times. However, the systems are not assumed to be fully loaded as in Chapter 2. We deal with the problem of calculating a lower bound on the total average waiting time for optimal routing, where this lower bound comes from the assumption that it is always possible to use a routing sequence such that the routing to each of the queues is regular. First we consider a single queue and study the average waiting time and average number of customers in this queue for regular routing with varying densities. Using several tools from number theory such as continued fractions and Farey intervals we derive an efficient algorithm for computing the average waiting time in case of regular routing and we give some properties of the average waiting time as a function of the density. Thereafter we consider the routing to N parallel queues and analyse the problem of finding the lower bound and the densities for which this lower bound is attained. We show that if the system is not fully loaded then there exist rational densities which attain the lower bound. A corollary of this result is the existence of an optimal periodic routing sequence in case of $N = 2$ parallel queues if the system is not fully loaded. This was proved by Gaujal and Hyon in [23].

In Chapter 5 we consider parallel queueing systems with deterministic interarrival and service times as in Chapter 4. The problem of finding an optimal routing policy is transformed to a Markov Decision Chain (MDC) with average cost minimisation. Then, by showing that the corresponding MDC has some specific properties, we show that there exists an optimal (deterministic) stationary policy for controlling the MDC. Thereafter it is proved that if the N service times S_1, S_2, \dots, S_N of the N parallel queues are all rational numbers, where the constant interarrival time is set to 1 by time scaling, that a routing sequence corresponding to an optimal (deterministic) stationary policy is ultimately periodic. From this it follows that there exists an optimal periodic routing sequence in case of rational service times.

In Chapter 6 we compare the performance of (periodic) routing and admission sequences of the same density. It is known that for a given density the regular admission sequence has always the best performance. To generalize this we try for given sequences u and v of the same density to show by combinatorial properties of the sequences that u has always a better performance than v or vice versa. Therefore we introduce partial orders called the multimodular order and the cone order and we show that they are equivalent. Moreover, for periodic sequences we also define the shift invariant multimodular order and the shift invariant cone order. We show that the period cycle of the regular sequence is a minimal element for these shift invariant orders. These shift invariant orders are not only defined for (periodic) sequences of zeros and ones, but also for (periodic) sequences of nonnegative integers. The notions of graph order and unbalance are also generalized for (periodic) sequences

of nonnegative integers and we analyse the connection with the shift invariant multimodular order and the shift invariant cone order. It is shown that the unbalance (both primal and dual) is a shift invariant multimodular function. This implies that if u is smaller than v with respect to the shift invariant multimodular order or the shift invariant cone order, that then u has a smaller unbalance (both primal and dual) than v .

This thesis contains material from the following papers.

Chapter 2 is a modified version of

D. A. van der Laan (2000). Routing jobs to servers with deterministic service times. Technical Report MI no. 2000-20, Leiden University. Available on www.math.leidenuniv.nl/reports/2000-20.shtml. Submitted to Mathematics of Operations Research.

Chapter 3 has, except for some minor modifications, appeared as

A. Hordijk and D.A. van der Laan (2000). Periodic routing to parallel queues with bounds on the average waiting time. Technical Report MI 2000-44, Leiden University. Available on www.math.leidenuniv.nl/reports/2000-44.shtml. An extended abstract of this chapter has appeared as [38].

Two papers titled “The unbalance and bounds on the average waiting time for periodic routing to one queue” and “Periodic routing to parallel queues and billiard sequences” respectively, which contain the results of this chapter, have been submitted to Mathematical Methods of Operations Research.

Chapter 4 has, except for some minor modifications, appeared as

A. Hordijk and D. A. van der Laan (2002). On the average waiting time for regular routing to deterministic queues. Technical Report MI 2002-24, Leiden University. Available on www.math.leidenuniv.nl/reports/2002-24.shtml. Submitted to Mathematics of Operations Research.

Chapter 6 contains results from

B. Gaujal, A. Hordijk and D. A. van der Laan (2001). On orders and bounds for multimodular functions. Technical Report MI 2001-29, Leiden University. Available on www.math.leidenuniv.nl/reports/2001-29.shtml. A slightly modified version of this report appears in [9].

Chapter 2

Optimal routing to fully loaded parallel queueing systems with deterministic interarrival and service times

2.1 Introduction

We consider a queueing system with $n \geq 2$ parallel servers each having its own queue. Arriving jobs have to be routed to one of the servers at the moment of arrival. We assume that the arrival of jobs is deterministic with a constant rate. We also assume that the serving times are deterministic, but typically the servers have different rates. We may think of a computer system with several processors which has to perform the incoming jobs. Our goal is to minimize the long-run average waiting time. Similar queueing systems with parallel heterogeneous servers have been considered in literature, but in general Poisson arrivals are assumed and the serving times are exponentially distributed or general. Further in such stochastic models a distinction is made between dynamic and static routing policies. In the dynamic case the policy may depend on time dependent information, for example the number of jobs or the remaining workload in each queue. In the static case the policy should only depend on the base characteristics of the system, such as the arrival rate

and service time distributions. However, it is clear that for our deterministic model there is no distinction between dynamic and static policies. The stochastic model that is closest to ours is the static case with allocation according to a fixed (periodic) pattern. This is also called semi-dynamic deterministic routing. Some papers dealing with such models are [11], [69], [20] and [36]. In these papers several algorithms and heuristics are given to obtain reasonable good policies for the models considered.

For this kind of models the optimization procedure actually consist of two steps:

1. Approximate for $i = 1, 2, \dots, n$ the fraction p_i^* of jobs that should be routed to server i in the optimal pattern by fractions p_i such that $\sum_{i=1}^n p_i = 1$.
2. Construct an allocation pattern with the fractions p_i .

Usually most of the attention goes to step 1. In our model we concentrate entirely on step 2 where we assume that the arrival rate is equal to the combined service rate of the n servers, in other words, that the traffic intensity ρ satisfies $\rho = 1$. For a stochastic model the system overflows if $\rho = 1$ and waiting times will tend to ∞ . However in the deterministic model the system can just be stabilized. The fact that the fractions p_i are fixed also implies that with minimizing the long-run average waiting time we also minimize the long-run average sojourn time (which is waiting time plus service time). We think that an optimal allocation pattern for given fractions in our model will at least in the heavy traffic region perform very well for more general service time distributions and arrival processes too.

Consider the following “most regular” zero-one valued splitting sequence of asymptotic mean p :

$$\{b_k^p(\phi)\}_{k=1}^\infty = \lfloor (k+1)p + \phi \rfloor - \lfloor kp + \phi \rfloor. \quad (2.1)$$

In (2.1) $\phi \in \mathbb{R}$ is an arbitrary phase. Such a sequence is called a regular, Beatty, Sturmian, or bracket sequence with density p . Such sequences are studied in several areas of mathematics and for more about this sequence see [53], [54], [4], [66], [68], [67], [51] and [52]. Remark that the sequence is periodic if $p \in \mathbb{Q}$. For a single server let $\{b_k\}_{k=1}^\infty$ be a zero-one splitting sequence, where the k -th arriving job is routed to the server if and only if $b_k = 1$. In [29] it is shown that if a fraction p of jobs has to be routed to a single exponential server, then the long-run average queue size is minimized, if sequence (2.1) is used. So according to Little’s law the long-run average waiting time of jobs which are routed to that server is minimized by sequence (2.1). Hajek proved this for an exponential server, but it holds much more generally (see [3] and [5]) and we will see that it also holds in our model. The optimality of

the regular sequence for a single queue can be used to prove that a static routing policy is optimal if the corresponding splitting sequences for every single server i are regular sequences with given (optimized) density p_i for $i = 1, 2, \dots, n$. This is done in [4] and [5] for several models. The integer sequence corresponding to such an optimal policy is called an exactly covering sequence or balanced sequence. However, if $n > 2$ then in general an exactly covering sequence does not exist for given densities p_i . Only for $n = 2$ there exists a balanced sequence for every pair of fractions $(p, 1 - p)$, because the complement of a regular sequence with density p is a regular sequence with density $1 - p$. In [4], [68], [67] and [51] it is studied which densities are balanceable if $n > 2$. So the optimal routing policy in our model is in fact known if $n = 2$. In other cases the policy will be such that the corresponding splitting sequences are simultaneously as regular as possible in some sense.

This chapter is organised as follows. In Section 2.2 we describe the queueing model and some notation is introduced. In Section 2.3 we deduce a fundamental relation between the long-run average waiting time and the total unused work capacity of the system. In Section 2.4 we will find a mathematical programming problem (MPP) which can be used to minimize the long-run average waiting time and we deduce the upper bound $\frac{n-1}{2}$ for the minimal long-run average waiting time. In Section 2.5 we find some results on the structure of optimal policies. In particular we show that there exists an optimal policy which corresponds to a billiard sequence. Further we show that the upper bound $\frac{n-1}{2}$ for the minimal long-run average waiting time is tight if the fractions p_i are linearly independent over \mathbb{Z} . In Section 2.6 we consider the case that all the fractions p_i are rational. We show that in that case we can restrict to proportional periodic policies to find an optimal routing policy and that an optimal periodic policy can be found by solving some integer linear problem (ILP). In fact we have a linear programming problem (LP) with zero-one variables. We obtain a lower bound for the minimal long-run average waiting time by solving the LP-relaxation of this ILP. Further we give an upper bound for this case which is a fraction better than the earlier found general upper bound. In Section 2.7 we show that the lower bound on the minimal long-run average waiting time that we obtained in Section 2.6 is attained if we have a policy such that for every single queue the corresponding splitting sequence is a regular sequence with appropriate density. Thus the lower bound can be attained if the given fractions are balanceable. Further we deduce an explicit formula for the long-run average waiting time of customers routed to some server i if the corresponding splitting sequence is regular. Finally in Section 2.8 we consider some algorithms to obtain good policies.

Notations. For $t \in \mathbb{R}$ we denote by $\mathbb{R}_{>t}$ ($\mathbb{R}_{\geq t}$) the set of real numbers that are greater (or equal) than t . For the integers \mathbb{Z} and the rational numbers \mathbb{Q} we denote

such subsets in a similar way.

For $x \in \mathbb{R}$ we denote by $\lfloor x \rfloor$ and $\lceil x \rceil$ the maximal integer not larger than x and the minimum integer not smaller than x , respectively.

Moreover, \gcd denotes the greatest common divisor and lcm denotes the least common multiple.

2.2 The queueing model

We consider the following queueing system. Arriving jobs have to be routed at the moment of arrival to one of $n \geq 2$ parallel servers, each having its own queue. We assume that the arrival of jobs is deterministic with a constant rate. Starting at time $t = 0$ one job arrives every time unit. We also assume that the serving times of the n parallel servers are deterministic and that they have a total working capacity of 1. So,

$$a_1 + a_2 + \cdots + a_n = 1, \tag{2.2}$$

where a_i^{-1} is the service time per job of server i . Moreover without loss of generality we assume that $a_1 \geq a_2 \geq \cdots \geq a_n$ and we denote the system by (a_1, a_2, \dots, a_n) .

If policy ψ is applied then for $t \in \mathbb{N}$ we define $W_t = W_t(\psi)$ as the waiting time of the t -th arriving job, that is the time between arrival and beginning of the serving process of the t -th arriving job. We define the long-run average waiting time if policy ψ is applied as

$$W = W(\psi) = W(a_1, a_2, \dots, a_n, \psi) = \limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} W_t.$$

Further if policy ψ is applied then for $t \in \mathbb{N}$ we define $V_t = V_t(\psi)$ as the sojourn time of the t -th arriving job, that is the time between arrival and end of the serving process of the t -th arriving job. We define the long-run average sojourn time if policy ψ is applied as

$$V = V(\psi) = V(a_1, a_2, \dots, a_n, \psi) = \limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} V_t.$$

Our goal is to find routing policies which minimize the long-run average waiting time W . As we will show, such a policy also minimizes the long-run average sojourn time V . We introduce some more notation.

Define for all $s \in \mathbb{Z}_{\geq 0}$ the variables

- u_i^s = the total amount of time units that server i has been idle between $t = 0$ and $t = s$.
- v_i^s = the total amount of time units after $t = s$ that server i needs to finish jobs that have been routed to server i before time $t = s$ and are still in the system .
- $k_s = k_s(\psi)$ is the server to which the job arriving at time $t = s$ is routed if policy ψ is applied.

Note that we define the v_i^s in such a way that the job arriving at moment s is not considered and thus v_i^s is the remaining workload in time units for server i at moment $s- = \lim_{t \uparrow s} t$. Moreover, we define $Q_i^s := a_i \cdot v_i^s$, which is the remaining workload in amount of jobs for server i at moment $s-$.

We have the following lemma.

Lemma 2.2.1 *If policy ψ is applied we have*

$$W = \limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} v_{k_t}^t(\psi) \quad (2.3)$$

and

$$V = \limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} (v_{k_t}^t(\psi) + a_{k_t}^{-1}). \quad (2.4)$$

Proof. Note for $t \in \mathbb{N}$ that W_t, V_t are resp. the waiting and sojourn time of the job arriving at moment $t - 1$. Hence $W_t = v_{k_{t-1}}^{t-1}(\psi)$ and $V_t = W_t + a_{k_{t-1}}^{-1}$. Thus

$$W = \limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} W_t = \limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} v_{k_{t-1}}^{t-1}(\psi) = \limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} v_{k_t}^t(\psi)$$

and

$$V = \limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} V_t = \limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} (v_{k_{t-1}}^{t-1}(\psi) + a_{k_{t-1}}^{-1}) =$$

$$\limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} (v_{k_t}^t(\psi) + a_{k_t}^{-1}). \quad \square$$

Further we define

$$\widetilde{W} = \widetilde{W}(a_1, a_2, \dots, a_n) = \inf_{\psi} W(\psi)$$

as the minimal long-run average waiting time for the the given service rates and

$$\widetilde{V} = \widetilde{V}(a_1, a_2, \dots, a_n) = \inf_{\psi} V(\psi)$$

as the minimal long-run average sojourn time for the the given service rates.

2.3 The relation between waiting time and unused capacity

In this section we will find a relation between W and the total amount of unused work capacity S of the system as $t \rightarrow \infty$. For $t \in \mathbb{Z}_{\geq 0}$ put $I_t = \{0, 1, \dots, t-1\}$. Then if policy ψ is applied we define for $i \in \{1, 2, \dots, n\}$ and $t \in \mathbb{Z}_{\geq 0}$ that

$$N_i^t = N_i^t(\psi) = \sum_{\{t' \in I_t : k_{t'}(\psi) = i\}} 1.$$

Hence N_i^t is the number of jobs among the first t incoming jobs that are routed to server i . Since the remaining workload in amount of jobs for server i at moment t is equal to the number of jobs routed to server i minus the amount of jobs that have been served by server i , we have

$$Q_i^t = N_i^t - a_i(t - u_i^t) \text{ for every } t \in \mathbb{Z}_{\geq 0}. \quad (2.5)$$

Remark that for $t \in \mathbb{Z}_{\geq 0}$

$$S^t := \sum_{i=1}^n a_i \cdot u_i^t$$

represents the total unused work capacity until time t . We have the following relation between the u_i and the v_i .

Lemma 2.3.1 *For all $t \in \mathbb{Z}_{\geq 0}$ we have*

$$\sum_{i=1}^n a_i \cdot u_i^t = \sum_{i=1}^n a_i \cdot v_i^t.$$

Proof. By (2.5) we obtain

$$\sum_{i=1}^n a_i \cdot v_i^t = \sum_{i=1}^n Q_i^t = t - t \cdot \sum_{i=1}^n a_i + \sum_{i=1}^n a_i \cdot u_i^t = \sum_{i=1}^n a_i \cdot u_i^t. \quad \square$$

Since $\sum_{i=1}^n a_i \cdot v_i^t$ is the total remaining workload measured in amount of jobs at time t , we have proved that S^t is equal to the total remaining workload in jobs at time t . Further S^t is monotonically non-decreasing in t , because the u_i^t are monotonically non-decreasing for $i \in \{1, 2, \dots, n\}$. For a policy ψ we define the total unused work capacity S as follows:

$$S = S(\psi) = S(a_1, a_2, \dots, a_n, \psi) = \lim_{t \rightarrow \infty} S^t.$$

Thus we have

$$S = \sum_{i=1}^n \lim_{t \rightarrow \infty} a_i u_i^t = \lim_{t \rightarrow \infty} \sum_{i=1}^n a_i u_i^t = \lim_{t \rightarrow \infty} \sum_{i=1}^n a_i v_i^t. \quad (2.6)$$

We define the minimal total unused work capacity for the given service rates as

$$\tilde{S} = \tilde{S}(a_1, a_2, \dots, a_n) = \inf_{\psi} S(\psi).$$

The following fundamental relation exists between W and S .

Theorem 2.3.2 *For all (a_1, a_2, \dots, a_n) systems and policies ψ it holds that $W < \infty$ if and only if $S < \infty$. Further if $W < \infty$ then $\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} v_{k_t}^t$ exists and*

$$W = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} v_{k_t}^t = S - \frac{n-1}{2}.$$

Theorem 2.3.2 is the main theorem of this section. Before we prove Theorem 2.3.2 we first present some auxiliary results.

Lemma 2.3.3 *Let $f : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}$ be bounded. Suppose there exist $H \subseteq \mathbb{Z}_{\geq 0}$ and $a, b \in \mathbb{R}_{>0}$ such that*

$$\begin{aligned} f(n+1) - f(n) &= a & \text{for } n \in H, \\ f(n+1) - f(n) &= -b & \text{for } n \in \mathbb{Z}_{\geq 0} \setminus H. \end{aligned}$$

Let $H_N = \#\{n \in \mathbb{Z}_{\geq 0} : n < N, n \in H\}$,

$$A_N = \begin{cases} 0 & \text{if } H_N = N \\ \frac{1}{N-H_N} \cdot \sum_{\{n \notin H, n < N\}} f(n) & \text{if } H_N < N \end{cases},$$

$$B_N = \begin{cases} 0 & \text{if } H_N = 0 \\ \frac{1}{H_N} \cdot \sum_{\{n \in H, n < N\}} f(n) & \text{if } H_N > 0 \end{cases} \quad \text{and}$$

$$E_N = \frac{1}{N} \cdot \sum_{n=0}^{N-1} f(n)$$

for $N = 1, 2, \dots$. Then

$$\lim_{N \rightarrow \infty} (A_N - B_N) = \frac{a+b}{2} \quad \text{and} \quad \lim_{N \rightarrow \infty} (E_N - B_N) = \frac{a}{2}.$$

Proof. Because f is bounded, $\lim_{N \rightarrow \infty} H_N = \infty$ and $\lim_{N \rightarrow \infty} (N - H_N) = \infty$. So there exists an $N_0 \in \mathbb{N}$ such that $A_N = \frac{1}{N-H_N} \cdot \sum_{\{n \notin H, n < N\}} f(n)$ and $B_N = \frac{1}{H_N} \cdot \sum_{\{n \in H, n < N\}} f(n)$ for $N \geq N_0$. Let $\tilde{f} : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ be the continuous piecewise linear extension of f , $\tilde{f}(x) = f(\lfloor x \rfloor) + (x - \lfloor x \rfloor) \cdot (f(\lfloor x \rfloor + 1) - f(\lfloor x \rfloor))$ for $x \in \mathbb{R}_{\geq 0}$. Let

$$C_N = \frac{1}{N - H_N} \cdot \sum_{\{n \notin H, n < N\}} \int_n^{n+1} \tilde{f}(t) dt \quad \text{and}$$

$$D_N = \frac{1}{H_N} \cdot \sum_{\{n \in H, n < N\}} \int_n^{n+1} \tilde{f}(t) dt$$

for $N \geq N_0$. If $n \in H$ then $\int_n^{n+1} \tilde{f}(t) dt = \int_n^{n+1} (f(n) + a \cdot (t - n)) dt = f(n) + \frac{a}{2}$, hence $D_N = B_N + \frac{a}{2}$ for $N \geq N_0$. If $n \in \mathbb{Z}_{\geq 0} \setminus H$ then

$$\int_n^{n+1} \tilde{f}(t) dt = \int_n^{n+1} (f(n) - b \cdot (t - n)) dt = f(n) - \frac{b}{2},$$

hence $C_N = A_N - \frac{b}{2}$ for $N \geq N_0$. So it suffices to prove that $\lim_{N \rightarrow \infty} (C_N - D_N) = 0$.

Since f is bounded, we can choose $m, M \in \mathbb{R}$ such that $f(n) \in [m, M]$ for all $n \in \mathbb{Z}_{\geq 0}$. Put $P_N = \{f(0), f(1), \dots, f(N-1)\}$ for $N = 1, 2, \dots$. Define step functions $g_N^- : [m, M] \rightarrow \mathbb{Z}_{\geq 0}$ and $g_N^+ : [m, M] \rightarrow \mathbb{Z}_{\geq 0}$ by

$$g_N^-(x) = \begin{cases} 0 & \text{if } x \in P_N \\ \#\{t \in [0, N) : \tilde{f}(t) = x, \tilde{f}'(t) < 0\} & \text{else} \end{cases}$$

and

$$g_N^+(x) = \begin{cases} 0 & \text{if } x \in P_N \\ \#\{t \in [0, N) : \tilde{f}(t) = x, \tilde{f}'(t) > 0\} & \text{else} \end{cases}.$$

If $n \in H$ then

$$\int_{f(n)}^{f(n+1)} x dx = \int_{f(n)}^{f(n)+a} x dx = a \int_n^{n+1} \tilde{f}(t) dt,$$

whence $\int_m^M x g_N^+(x) dx = \sum_{n \in H, n < N} \int_{f(n)}^{f(n+1)} x dx = a \sum_{\{n \in H, n < N\}} \int_n^{n+1} \tilde{f}(t) dt$ for $N = 1, 2, \dots$. Similarly if $n \in H$ we have

$$\int_m^M g_N^+(x) dx = \sum_{n \in H, n < N} \int_{f(n)}^{f(n+1)} dx = a H_N.$$

So

$$D_N = \frac{\int_m^M x \cdot g_N^+(x) dx}{\int_m^M g_N^+(x) dx} \text{ if } N \geq N_0. \quad (2.7)$$

Analogously it follows that

$$C_N = \frac{\int_m^M x \cdot g_N^-(x) dx}{\int_m^M g_N^-(x) dx} \text{ if } N \geq N_0. \quad (2.8)$$

Since \tilde{f} is continuous, $|g_N^-(x) - g_N^+(x)| \leq 1$ for $N = 1, 2, \dots$ and $x \in [m, M]$. Hence

$$\begin{aligned} & \left| \int_m^M x \cdot g_N^-(x) dx - \int_m^M x \cdot g_N^+(x) dx \right| \leq \\ & \int_m^M |x| dx \leq \max(m^2, M^2) \end{aligned} \quad (2.9)$$

and

$$\left| \int_m^M g_N^-(x) dx - \int_m^M g_N^+(x) dx \right| \leq M - m \quad (2.10)$$

for $N = 1, 2, \dots$. Further

$$\lim_{N \rightarrow \infty} \int_m^M g_N^-(x) dx = \lim_{N \rightarrow \infty} b \cdot (N - H_N) = \infty \quad (2.11)$$

and

$$\lim_{N \rightarrow \infty} \int_m^M g_N^+(x) dx = \lim_{N \rightarrow \infty} a H_N = \infty. \quad (2.12)$$

From (2.7) and (2.8) it follows that

$$C_N - D_N = \frac{\int_m^M x \cdot g_N^-(x) dx - \int_m^M x \cdot g_N^+(x) dx}{\int_m^M g_N^-(x) dx} +$$

$$\frac{\frac{\int_m^M x \cdot g_N^+(x) dx}{\int_m^M g_N^+(x) dx} \cdot (\int_m^M g_N^+(x) dx - \int_m^M g_N^-(x) dx)}{\int_m^M g_N^-(x) dx}. \quad (2.13)$$

From (2.9) and (2.11) we obtain

$$\lim_{N \rightarrow \infty} \frac{\int_m^M x \cdot g_N^-(x) dx - \int_m^M x \cdot g_N^+(x) dx}{\int_m^M g_N^-(x) dx} = 0.$$

So, by (2.7),(2.10) and (2.13) we have

$$|\lim_{N \rightarrow \infty} (C_N - D_N)| \leq (M - m) \cdot \lim_{N \rightarrow \infty} \frac{D_N}{\int_m^M g_N^-(x) dx}. \quad (2.14)$$

Because $m \leq D_N \leq M$ for $N \geq N_0$ it follows from (2.11) and (2.14) that

$$\lim_{N \rightarrow \infty} (C_N - D_N) = 0.$$

□

We apply Lemma 2.3.3 to a function f which depends on the variables u_i^t and v_i^t .

Lemma 2.3.4 *If in an (a_1, a_2, \dots, a_n) system a policy ψ is applied such that for some $i \in \{1, 2, \dots, n\}$ the function $f : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}$ defined by $f(t) = v_i^t - u_i^t$ is bounded then*

$$\lim_{\tau \rightarrow \infty} \left\{ \frac{\sum_{t=0}^{\tau-1} v_i^t}{\tau} - \frac{\sum_{\{t \in I_\tau : k_t = i\}} v_i^t}{N_i^\tau} \right\} = \frac{1}{2a_i} - \frac{1}{2}.$$

Proof. Let $t \in \mathbb{Z}_{\geq 0}$ and $l \in \{1, 2, \dots, n\}$ and assume that $k_t = l$. Then

$$u_l^{t+1} = u_l^t \text{ and } v_l^{t+1} = v_l^t + a_l^{-1} - 1. \quad (2.15)$$

Moreover for all $j \neq l$ we have

$$u_j^{t+1} = u_j^t + \max(0, 1 - v_j^t) \text{ and } v_j^{t+1} = \max(0, v_j^t - 1) = v_j^t + \max(-v_j^t, -1). \quad (2.16)$$

If $k_t = i$ then we have by (2.15) that $f(t+1) - f(t) = \frac{1}{a_i} - 1$. If $k_t \neq i$ then we have by (2.16) that $f(t+1) - f(t) = -1$. So f satisfies the conditions of Lemma 2.3.3 with $H = \{t : k_t = i\}$, $a = \frac{1}{a_i} - 1$ and $b = 1$. Hence

$$\lim_{\tau \rightarrow \infty} \left\{ \frac{\sum_{t=0}^{\tau-1} (v_i^t - u_i^t)}{\tau} - \frac{\sum_{\{t \in I_\tau : k_t = i\}} (v_i^t - u_i^t)}{N_i^\tau} \right\} = \frac{1}{2a_i} - \frac{1}{2}. \quad (2.17)$$

Since $u_i^{t+1} > u_i^t$ implies that $v_i^{t+1} = 0$ it follows from the boundedness of f that $\lim_{t \rightarrow \infty} u_i^t =: L < \infty$ and thus

$$\lim_{\tau \rightarrow \infty} \frac{\sum_{t=0}^{\tau-1} u_i^t}{\tau} = \lim_{\tau \rightarrow \infty} \frac{\sum_{\{t \in I_\tau: k_t=i\}} u_i^t}{N_i^\tau} = L. \quad (2.18)$$

By (2.17) and (2.18) we have

$$\lim_{\tau \rightarrow \infty} \left\{ \frac{\sum_{t=0}^{\tau-1} v_i^t}{\tau} - \frac{\sum_{\{t \in I_\tau: k_t=i\}} v_i^t}{N_i^\tau} \right\} = \frac{1}{2a_i} - \frac{1}{2}. \quad \square$$

Corollary 2.3.5 *If in an (a_1, a_2, \dots, a_n) system a policy ψ is applied with $S(\psi) < \infty$ then we have for every server $i \in \{1, 2, \dots, n\}$ that*

$$\lim_{\tau \rightarrow \infty} \left\{ \frac{\sum_{t=0}^{\tau-1} v_i^t}{\tau} - \frac{\sum_{\{t \in I_\tau: k_t=i\}} v_i^t}{N_i^\tau} \right\} = \frac{1}{2a_i} - \frac{1}{2}.$$

Proof. Since $S < \infty$ we have for every $i \in \{1, 2, \dots, n\}$ that $\lim_{t \rightarrow \infty} a_i \cdot u_i^t := L_i \leq S < \infty$. Moreover, we have $a_i \cdot v_i^t \leq \sum_{j=1}^n a_j \cdot v_j^t = S^t \leq S$ and $a_i \cdot u_i^t \leq L_i$ for every $t \in \mathbb{Z}_{\geq 0}$. Define $f_i : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}$ by $f_i(t) = v_i^t - u_i^t$ as in Lemma 2.3.4. Then $f_i(t) \in [-\frac{L_i}{a_i}, \frac{S}{a_i}]$ for every $t \in \mathbb{Z}_{\geq 0}$ and thus f_i is bounded. Now apply Lemma 2.3.4. \square

If $\lim_{t \rightarrow \infty} \frac{N_i^t}{t}$ exists for $i \in \{1, 2, \dots, n\}$ then we define $p_i := \lim_{t \rightarrow \infty} \frac{N_i^t}{t}$ as the fraction of jobs that is routed to server i . From the following proposition it follows that these fractions exist and are equal to the capacities of the corresponding servers if the long-run average waiting time is finite.

Lemma 2.3.6 *For every (a_1, a_2, \dots, a_n) system and policy ψ we have $S < \infty$ if and only if $W < \infty$. Further if $W < \infty$ then $\lim_{t \rightarrow \infty} \frac{N_i^t}{t}$ exists for all $i \in \{1, 2, \dots, n\}$ and*

$$p_i = \lim_{t \rightarrow \infty} \frac{N_i^t}{t} = a_i \quad (2.19)$$

for $i = 1, 2, \dots, n$.

Proof. Suppose $S < \infty$. Then there exists $M_0 \in \mathbb{R}$ such that $\sum_{i=1}^n a_i v_i^t < M_0$ for $t \in \mathbb{Z}_{\geq 0}$. It follows that $a_i v_i^t < M_0$ for $i = 1, 2, \dots, n$ and $t \in \mathbb{Z}_{\geq 0}$ and thus $v_i^t < M := \frac{M_0}{\min_{i=1}^n a_i}$. Hence $v_{k_t}^t < M$ for all $t \in \mathbb{Z}_{\geq 0}$ and thus $W \leq M$. Suppose $S = \infty$. Let $L(t)$ be the total number of waiting jobs at time $t \in \mathbb{R}_{\geq 0}$. It

is clear that $L(t) \geq S^{\lfloor t \rfloor} - n$ for every $t \in \mathbb{R}_{\geq 0}$ and thus $\lim_{t \rightarrow \infty} L(t) = \infty$. So the limiting time-average number of waiting jobs satisfies

$$L := \lim_{t \rightarrow \infty} \frac{1}{t} \cdot \int_0^t L(t) dt = \infty. \quad (2.20)$$

For $t \in \mathbb{R}_{\geq 0}$ let $J_k(t) = 1$, if the k -th arriving job is waiting in one of the queues at time t and else $J_k(t) = 0$. Then,

$$W_k = \int_0^\infty J_k(t) dt \text{ and } L(t) = \sum_{k=1}^\infty J_k(t). \quad (2.21)$$

Let $U(t) := \sum_{k=1}^{t+1} W_k$ be the sum of the waiting times of jobs arriving in $[0, t]$ for $t \in \mathbb{R}_{\geq 0}$. Then for all $T \in \mathbb{R}_{\geq 0}$ we have by (2.21) that

$$\int_0^T L(t) dt = \sum_{k=1}^\infty \int_0^T J_k(t) dt = \sum_{k=1}^{T+1} \int_0^T J_k(t) dt \leq \sum_{k=1}^{T+1} W_k = U(T). \quad (2.22)$$

Hence by (2.20) and (2.22) we have $\lim_{t \rightarrow \infty} \frac{1}{t} \cdot U(t) = \infty$. Thus

$$W = \limsup_{t \rightarrow \infty} \frac{1}{t} \cdot \sum_{k=1}^t W_k = \limsup_{t \rightarrow \infty} \frac{1}{\lfloor t \rfloor + 1} \cdot U(t) = \lim_{t \rightarrow \infty} \frac{1}{t} \cdot U(t) = \infty.$$

The first part of the lemma has been proved. For the second part we assume $W < \infty$ and thus $S < \infty$. Since $S = \lim_{t \rightarrow \infty} S^t < \infty$ it follows from (2.6) that $\limsup_{t \rightarrow \infty} Q_i^t = \limsup_{t \rightarrow \infty} a_i \cdot v_i^t < \infty$ and $\limsup_{t \rightarrow \infty} a_i \cdot u_i^t < \infty$. Thus $\lim_{t \rightarrow \infty} \frac{Q_i^t}{t} = \lim_{t \rightarrow \infty} \frac{a_i \cdot u_i^t}{t} = 0$. Dividing equality (2.5) by t we obtain $\frac{N_i^t}{t} = \frac{Q_i^t}{t} + a_i - \frac{a_i \cdot u_i^t}{t}$. Hence $\lim_{t \rightarrow \infty} \frac{N_i^t}{t}$ exists and $p_i = \lim_{t \rightarrow \infty} \frac{N_i^t}{t} = a_i$ for $i \in \{1, 2, \dots, n\}$. \square

Remark. The argument for proving that $W < \infty$ implies $S < \infty$ is a simplification of a proof of Little's law by Stidham in [65].

Proof of Theorem 2.3.2. The first part of the theorem follows from Lemma 2.3.6. We now assume $S < \infty$. From (2.19) and Corollary 2.3.5 it follows for every $i \in \{1, 2, \dots, n\}$ that

$$\begin{aligned} & \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \left(\sum_{t=0}^{\tau-1} a_i \cdot v_i^t - \sum_{\{t \in I_\tau : k_t = i\}} v_i^t \right) = \\ & a_i \cdot \lim_{\tau \rightarrow \infty} \left\{ \frac{\sum_{t=0}^{\tau-1} v_i^t}{\tau} - \frac{\sum_{\{t \in I_\tau : k_t = i\}} v_i^t}{N_i^\tau} \right\} = a_i \cdot \left(\frac{1}{2a_i} - \frac{1}{2} \right) = \frac{1}{2} - \frac{1}{2} a_i. \end{aligned} \quad (2.23)$$

Because S^t is monotonically non-decreasing in t and bounded it follows that

$$S = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} S^t = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} \sum_{i=1}^n a_i \cdot v_i^t = \lim_{\tau \rightarrow \infty} \sum_{i=1}^n \sum_{t=0}^{\tau-1} \frac{a_i \cdot v_i^t}{\tau}.$$

Hence, by (2.23), we see that $\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} v_{k_t}^t$ exists and that

$$S\text{-}\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} v_{k_t}^t = \lim_{\tau \rightarrow \infty} \sum_{i=1}^n \frac{1}{\tau} \cdot \left(\sum_{t=0}^{\tau-1} a_i \cdot v_i^t - \sum_{\{t \in I_\tau : k_t = i\}} v_i^t \right) = \sum_{i=1}^n \left(\frac{1}{2} - \frac{1}{2} \cdot a_i \right) = \frac{n-1}{2}.$$

Therefore

$$W = \limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} v_{k_t}^t = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} v_{k_t}^t = S - \frac{n-1}{2}. \quad \square$$

The following proposition shows that minimizing the long-run average waiting time W also minimizes the long-run average sojourn time V and vice versa.

Proposition 2.3.7 *For all (a_1, a_2, \dots, a_n) systems and policies ψ it holds that $V < \infty$ if and only if $S < \infty$, and if $S < \infty$, then $V = W + n$.*

Proof. Since $V < \infty$ if and only if $W < \infty$, the first assertion follows from Theorem 2.3.2. Let ψ be a policy such that $S < \infty$. Then, by (2.19),

$$V = \limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} (v_{k_t}^t + a_{k_t}^{-1}) = W + \lim_{\tau \rightarrow \infty} \sum_{i=1}^n a_i^{-1} \frac{N_i^\tau}{\tau} = W + \sum_{i=1}^n a_i^{-1} a_i = W + n. \quad \square$$

From now on we will only consider W . The results for V follow from Proposition 2.3.7.

2.4 An upper bound for the minimal long-run average waiting time

In this section we derive an upper bound for the minimal long-run average waiting time \widetilde{W} . Further we show that for every (a_1, a_2, \dots, a_n) system an optimal policy exists and we give a MPP such that from an optimal solution of the MPP an optimal policy can be obtained and vice versa.

Lemma 2.4.1 For all $s \in \mathbb{Z}_{\geq 0}$ we have

$$u_i^s = \max_{t \in I_{s+1}} (u_i^t - v_i^t) = \max_{t \in I_{s+1}} (t - a_i^{-1} \cdot N_i^t).$$

Proof. The proof of the first equality is done by induction on s . It holds for $s = 0$. Suppose that it holds for $s' = s$. If $k_s = i$ then by (2.15) we have $u_i^{s+1} = u_i^s$ and $v_i^{s+1} > v_i^s$. So, by the induction hypothesis, $u_i^{s+1} = u_i^s = \max_{t \in I_{s+1}} (u_i^t - v_i^t) = \max_{t \in I_{s+2}} (u_i^t - v_i^t)$. If $k_s \neq i$ and $v_i^s \leq 1$, then by (2.16) we have $u_i^{s+1} \geq u_i^s$ and $v_i^{s+1} = 0$. Thus, by the induction hypothesis, $u_i^{s+1} - v_i^{s+1} \geq u_i^s = \max_{t \in I_{s+1}} (u_i^t - v_i^t)$. Hence $u_i^{s+1} = u_i^{s+1} - v_i^{s+1} = \max_{t \in I_{s+2}} (u_i^t - v_i^t)$. Further if $k_s \neq i$ and $v_i^s > 1$ then by (2.16) we have $u_i^{s+1} = u_i^s$ and $v_i^{s+1} = v_i^s - 1 \geq 0$. So, by the induction hypothesis, $u_i^{s+1} - v_i^{s+1} \leq u_i^s = \max_{t \in I_{s+1}} (u_i^t - v_i^t)$. Hence $u_i^{s+1} = u_i^s = \max_{t \in I_{s+1}} (u_i^t - v_i^t) = \max_{t \in I_{s+2}} (u_i^t - v_i^t)$ and the induction step is complete.

For the second equality it suffices to show that

$$u_i^t - v_i^t = t - a_i^{-1} \cdot N_i^t \text{ for all } t \in \mathbb{Z}_{\geq 0}. \quad (2.24)$$

This follows directly by dividing (2.5) by a_i and then substituting v_i^t . \square

Note. The equality $u_i^s = \max_{t \in I_{s+1}} (t - a_i^{-1} \cdot N_i^t)$ is kind of a discrete version of some result about reflected Brownian motion (see [32]).

We define for $i \in \{1, 2, \dots, n\}$ the variables

$$d_i(\psi) := \lim_{t \rightarrow \infty} a_i \cdot u_i^t,$$

the total amount of unused work capacity of server i if policy ψ is applied. We use these variables for the formulation of the mathematical programming problem. It is clear that

$$S = \sum_{i=1}^n d_i(\psi). \quad (2.25)$$

The following lemma follows directly from Lemma 2.4.1.

Lemma 2.4.2 For every (a_1, a_2, \dots, a_n) system and applied policy ψ we have

$$d_i(\psi) = \sup_{t \in \mathbb{Z}_{\geq 0}} (a_i \cdot t - N_i^t) \text{ for all } i \in \{1, 2, \dots, n\}.$$

Let $d_i \in \mathbb{R}_{\geq 0}$ be given. Then it follows from Lemma 2.4.2 that $d_i(\psi) \leq d_i$ if and only if $t \cdot a_i - N_i^t(\psi) \leq d_i$ for all $t \in \mathbb{Z}_{\geq 0}$. Since $N_i^t(\psi) \in \mathbb{Z}_{\geq 0}$ it follows that $N_i^t(\psi) \geq \max(\lceil t \cdot a_i - d_i \rceil, 0)$ for all $t \in \mathbb{Z}_{\geq 0}$. So, suppose that $d_1, d_2, \dots, d_n \in \mathbb{R}_{\geq 0}$ are given. Then

$$t \geq \sum_{i=1}^n \max(\lceil t \cdot a_i - d_i \rceil, 0) \text{ for } t = 0, 1, \dots \quad (2.26)$$

is a necessary condition for the existence of a policy ψ such that $d_i(\psi) = d_i$ for $i = 1, 2, \dots, n$. Thus we have proved the following lemma.

Lemma 2.4.3 *If a policy ψ is applied to an (a_1, a_2, \dots, a_n) system then*

$$t \geq \sum_{i=1}^n \max(\lceil t \cdot a_i - d_i(\psi) \rceil, 0) \text{ for } t = 0, 1, \dots$$

On the other hand, (2.26) is not sufficient for the existence of a policy ψ with $d_i(\psi) = d_i$ for $i \in \{1, 2, \dots, n\}$. For example, if $a_i = \frac{p}{q}$ with $\gcd(p, q) = 1$ then, by Lemma 2.4.2,

$$q \cdot d_i(\psi) = q \cdot \sup_{t \in \mathbb{Z}_{\geq 0}} (a_i \cdot t - N_i^t) \in \mathbb{Z}.$$

So, in that case $q \cdot d_i \in \mathbb{Z}$ is another necessary condition. However, we will show that there exists a policy ψ with $d_i(\psi) \leq d_i$ for $i = 1, 2, \dots, n$ if d_1, d_2, \dots, d_n fulfill (2.26). To do so we use the following greedy algorithm to construct such a policy ψ . We denote this algorithm as the GG (General Greedy) algorithm.

GG Algorithm. Let $d_1, d_2, \dots, d_n \in \mathbb{R}_{\geq 0}$ be given. Then we determine ψ inductively in the following way. Suppose $k_0(\psi), k_1(\psi), \dots, k_{t-1}(\psi)$ have been determined. Then for $j = 1, 2, \dots, n$ put $q_j(t) = \min\{t' \in \mathbb{Z}_{\geq 0} : t' \cdot a_j > d_j - t \cdot a_j + N_j^{t'}(\psi)\}$. Choose $k_t(\psi)$ such that $k_t(\psi) = i \in \{1, 2, \dots, n\}$ for which $q_i(t) = \min_{j \in \{1, 2, \dots, n\}} q_j(t)$.

Theorem 2.4.4 *Let ψ be a policy obtained by applying the GG algorithm. If d_1, d_2, \dots, d_n fulfill (2.26) then $d_i(\psi) \leq d_i$ for $i = 1, 2, \dots, n$.*

Proof. Suppose there exists a $j \in \{1, 2, \dots, n\}$ with $d_j(\psi) > d_j$. Then there exist $t \in \mathbb{Z}_{\geq 0}$ for which $N_j^t(\psi) < t \cdot a_j - d_j$ and thus $t \cdot a_j > d_j$ and $q_j(t) = 0$. Let $t' = \min\{t \in \mathbb{Z}_{\geq 0} : q_j(t) = 0\}$. Then $N_j^{t'}(\psi) \leq \lceil t' \cdot a_j - d_j \rceil - 1$. Thus from $\sum_{i=1}^n N_i^{t'}(\psi) = t' \geq \sum_{i=1}^n \max(\lceil t' \cdot a_i - d_i \rceil, 0)$ it follows that there exists a $k \neq j$ with $N_k^{t'}(\psi) \geq \max(\lceil t' \cdot a_k - d_k \rceil, 0) + 1$. Hence $N_k^{t'}(\psi) \geq 1$ and $N_k^{t'}(\psi) \geq t' \cdot a_k - d_k + 1$.

Let $t_0 = \max\{0 \leq t < t' : k_t(\psi) = k\}$. Then we have $N_j^{t_0}(\psi) \leq N_j^{t'}(\psi) < t' \cdot a_j - d_j$ and thus $q_j(t_0) \leq t' - t_0$. Moreover $N_k^{t_0}(\psi) = N_k^{t'}(\psi) - 1 \geq t' \cdot a_k - d_k$ and thus $q_k(t_0) \geq t' - t_0 + 1 > q_j(t_0)$. This contradicts the fact that the GG algorithm is applied. \square

Put

$$D = D(a_1, a_2, \dots, a_n) = \{(d_1, d_2, \dots, d_n) \in \mathbb{R}^n : d_1, d_2, \dots, d_n \text{ fulfill condition (2.26)}\}.$$

According to Theorem 2.4.6 we can obtain \tilde{S} , the minimum value of the total unused work capacity, by minimizing a linear function over the set D . Moreover, by applying the following lemma we find an upper bound for this value.

Lemma 2.4.5 *Let $d_1, d_2, \dots, d_n \in \mathbb{R}$ such that*

$$\sum_{i=1}^n d_i = n - 1 \text{ and } 0 \leq d_i < 1 + a_i \text{ for } i = 1, 2, \dots, n. \quad (2.27)$$

Then $(d_1, d_2, \dots, d_n) \in D$.

Proof. Because $d_i \geq 0$ for $i = 1, 2, \dots, n$ we have $\max(\lceil -d_i \rceil, 0) = 0$ for $i = 1, 2, \dots, n$. Hence $\sum_{i=1}^n \max(\lceil -d_i \rceil, 0) = 0$. Further by $d_i < 1 + a_i$ we have for $t = 1, 2, \dots$ that

$$\max(0, \lceil t \cdot a_i - d_i \rceil) = \lceil t \cdot a_i - d_i \rceil < t \cdot a_i - d_i + 1$$

for $i = 1, 2, \dots, n$. Hence

$$\sum_{i=1}^n \max(\lceil t \cdot a_i - d_i \rceil, 0) < \sum_{i=1}^n (t \cdot a_i - d_i + 1) = t - (n - 1) + n = t + 1$$

and $\sum_{i=1}^n \max(\lceil t \cdot a_i - d_i \rceil, 0) \in \mathbb{Z}_{\geq 0}$. So, $\sum_{i=1}^n \max(\lceil t \cdot a_i - d_i \rceil, 0) \leq t$ and condition (2.26) is fulfilled. \square

Examples of d_1, d_2, \dots, d_n that fulfill (2.27) are given by $d_i = 0$ for an arbitrary $i \in \{1, 2, \dots, n\}$ and $d_j = 1$ for all $j \neq i$. Another example is $d_i = 1 - a_i$ for $i = 1, 2, \dots, n$. Hence $D \neq \emptyset$ for every (a_1, a_2, \dots, a_n) system. Further it is obvious that if $(d_1, d_2, \dots, d_n) \in D$ then $(d_1 + e_1, d_2 + e_2, \dots, d_n + e_n) \in D$ for every $e_1, e_2, \dots, e_n \in \mathbb{R}_{\geq 0}$. It follows from (2.26) with $t = 0$ that if $(d_1, d_2, \dots, d_n) \in D$ then $d_i \geq 0$ for $i = 1, 2, \dots, n$. Thus $\inf_{(d_1, d_2, \dots, d_n) \in D} \sum_{i=1}^n d_i \geq 0$.

Theorem 2.4.6 For every (a_1, a_2, \dots, a_n) system we have

$$\tilde{S} = \inf_{(d_1, d_2, \dots, d_n) \in D} \sum_{i=1}^n d_i \leq n - 1.$$

Proof. There exist $(d_1, d_2, \dots, d_n) \in D$ with $\sum_{i=1}^n d_i = n - 1$ and thus we have $\inf_{(d_1, d_2, \dots, d_n) \in D} \sum_{i=1}^n d_i \leq n - 1$. So, it remains to prove that

$\tilde{S} = \inf_{(d_1, d_2, \dots, d_n) \in D} \sum_{i=1}^n d_i$. According to Lemma 2.4.3 we have for every policy ψ that $d_1(\psi), d_2(\psi), \dots, d_n(\psi)$ fulfill condition (2.26). Hence, $S(\psi) = \sum_{i=1}^n d_i(\psi) \geq \inf_{(d_1, d_2, \dots, d_n) \in D} \sum_{i=1}^n d_i$ and thus $\tilde{S} = \inf_{\psi} S(\psi) \geq \inf_{(d_1, d_2, \dots, d_n) \in D} \sum_{i=1}^n d_i$.

On the other hand, if $(d_1, d_2, \dots, d_n) \in D$ then according to Theorem 2.4.4 there exists a policy ψ with $d_i(\psi) \leq d_i$ for $i = 1, 2, \dots, n$ and thus $S(\psi) \leq \sum_{i=1}^n d_i$. Hence,

$$\tilde{S} = \inf_{\psi} S(\psi) \leq \inf_{(d_1, d_2, \dots, d_n) \in D} \sum_{i=1}^n d_i. \quad \square$$

From Theorem 2.3.2 and Theorem 2.4.6 we deduce the following corollary which gives an upper bound for the minimal long-run average waiting time \tilde{W} for every (a_1, a_2, \dots, a_n) system.

Corollary 2.4.7 For every (a_1, a_2, \dots, a_n) system we have

$$0 \leq \tilde{W} = \tilde{S} - \frac{n-1}{2} \leq \frac{n-1}{2}. \quad (2.28)$$

We define a policy ψ to be optimal for a given (a_1, a_2, \dots, a_n) system if and only if $S(\psi) = \tilde{S}$ and thus $W(\psi) = \tilde{W}$. According to the next theorem we have for every (a_1, a_2, \dots, a_n) system that there exists a policy which achieves the optimal value \tilde{S} . So, we show that for every (a_1, a_2, \dots, a_n) system there exists some optimal policy. We prove this by using the well-known fact that a continuous function has always a minimum on a compact domain. We first derive the following lemma.

Lemma 2.4.8 $D \subseteq \mathbb{R}^n$ is closed for every (a_1, a_2, \dots, a_n) system.

Proof. For $t = 0, 1, \dots$ put

$$D^t = D^t(a_1, a_2, \dots, a_n) = \{(d_1, d_2, \dots, d_n) \in \mathbb{R}^n : t \geq \sum_{i=1}^n \max(0, \lceil t \cdot a_i - d_i \rceil)\}.$$

Then $D = \bigcap_{t=0}^{\infty} D_t$. So it suffices to show that $D_t \subseteq \mathbb{R}^n$ is closed for $t = 0, 1, \dots$. Let $t \in \mathbb{Z}_{\geq 0}$ be fixed and suppose $\underline{b} = (b_1, b_2, \dots, b_n) \in \mathbb{R}^n \setminus D^t$. Put $m_i = \lceil t \cdot a_i - b_i \rceil - 1$

for $i = 1, 2, \dots, n$. Then $m_i \in \mathbb{Z}$ and $t \cdot a_i - b_i = m_i + \delta_i$ with $\delta_i \in (0, 1]$ for $i = 1, 2, \dots, n$. Let $0 < \varepsilon < \min_{i=1}^n \delta_i$ and $\underline{c} = (c_1, c_2, \dots, c_n) \in \mathbb{R}^n$ with $\|\underline{c} - \underline{b}\|_\infty < \varepsilon$. Then

$$\lceil t \cdot a_i - c_i \rceil \geq \lceil t \cdot a_i - b_i - \varepsilon \rceil = \lceil m_i + \delta_i - \varepsilon \rceil =$$

$$m_i + 1 = \lceil t \cdot a_i - b_i \rceil \text{ for } i = 1, 2, \dots, n.$$

Hence,

$$\sum_{i=1}^n \max(\lceil t \cdot a_i - c_i \rceil, 0) \geq \sum_{i=1}^n \max(\lceil t \cdot a_i - b_i \rceil, 0) > t$$

and thus $\underline{c} \in \mathbb{R}^n \setminus D^t$. So, $\mathbb{R}^n \setminus D^t$ is open and $D^t \subseteq \mathbb{R}^n$ is closed for $t = 0, 1, \dots$. \square

Theorem 2.4.9 *For every (a_1, a_2, \dots, a_n) system we have*

$$\tilde{S} = \min_{\psi} S(\psi). \tag{2.29}$$

Proof. Let $H = \{(d_1, d_2, \dots, d_n) \in \mathbb{R}^n : \sum_{i=1}^n d_i \leq n - 1\}$ and $D' = D \cap H$. Then D' is closed, since D and H are closed. Further $D' \neq \emptyset$, because $(1 - a_1, 1 - a_2, \dots, 1 - a_n) \in D'$. Let $(d_1, d_2, \dots, d_n) \in D'$. As observed in the proof of Theorem 2.4.6 we have $d_i \geq 0$ for $i = 1, 2, \dots, n$. So, $0 \leq d_i \leq n - 1$ for $i = 1, 2, \dots, n$ and thus D' is bounded. Hence D' is compact. Because a continuous function assumes a minimum value on a compact set there exists $(d'_1, d'_2, \dots, d'_n) \in D'$ for which

$$\sum_{i=1}^n d'_i = \min_{(d_1, d_2, \dots, d_n) \in D'} \sum_{i=1}^n d_i = \min_{(d_1, d_2, \dots, d_n) \in D} \sum_{i=1}^n d_i. \tag{2.30}$$

Since $(d'_1, d'_2, \dots, d'_n) \in D$, it follows from Theorem 2.4.4 that there exists a policy ψ' with $d_i(\psi') \leq d'_i$ for $i = 1, 2, \dots, n$. So, if $d_i(\psi') \neq d'_i$ for some $i \in \{1, 2, \dots, n\}$ then $\sum_{i=1}^n d_i(\psi') < \sum_{i=1}^n d'_i$. This contradicts the minimality of $\sum_{i=1}^n d'_i$, because $(d_1(\psi'), d_2(\psi'), \dots, d_n(\psi')) \in D'$. Hence $d_i(\psi') = d'_i$ for $i = 1, 2, \dots, n$ and according to (2.30) and Theorem 2.4.6 we have

$$S(\psi') = \sum_{i=1}^n d_i(\psi') = \sum_{i=1}^n d'_i = \inf_{(d_1, d_2, \dots, d_n) \in D} \sum_{i=1}^n d_i = \tilde{S}. \quad \square$$

Consider the following MPP.

$$\begin{aligned}
& \text{minimize } S & = & \sum_{i=1}^n d_i \\
& \text{subject to:} & & \\
& \sum_{i=1}^n \max(0, \lceil t \cdot a_i - d_i \rceil) & \leq & t \quad \text{for } t = 0, 1, \dots \\
& d_i & \in & \mathbb{R}_{\geq 0} \quad \text{for } i = 1, 2, \dots, n
\end{aligned} \tag{2.31}$$

The following corollary follows from Theorem 2.4.4, Theorem 2.4.6 and Theorem 2.4.9.

Corollary 2.4.10 *For every (a_1, a_2, \dots, a_n) system an optimal solution of (2.31) exists. Let $d_1 = d_1^*, d_2 = d_2^*, \dots, d_n = d_n^*$ be an optimal solution of (2.31). Then we have that $\tilde{S}(a_1, a_2, \dots, a_n) = \sum_{i=1}^n d_i^*$ and from this optimal solution of (2.31) an optimal policy can be obtained by applying the GG algorithm and choosing $d_1 = d_1^*, d_2 = d_2^*, \dots, d_n = d_n^*$ in this algorithm. Conversely, if ψ is an optimal policy then an optimal solution of (2.31) is obtained by putting $d_1 = d_1(\psi), d_2 = d_2(\psi), \dots, d_n = d_n(\psi)$.*

Note that d_1, d_2, \dots, d_n is a feasible solution of (2.31) if and only if $(d_1, d_2, \dots, d_n) \in D$. So, by Theorem 2.4.4 it follows that for any feasible solution d_1, d_2, \dots, d_n of (2.31) a policy ψ can be obtained with $d_i(\psi) \leq d_i$ for $i = 1, 2, \dots, n$ by applying the GG algorithm. Then we have for the obtained policy ψ that

$$\tilde{S} \leq S(\psi) = \sum_{i=1}^n d_i(\psi) \leq \sum_{i=1}^n d_i. \tag{2.32}$$

Moreover, by Lemma 2.4.3 the $d_i(\psi)$ give a feasible solution of (2.31). Hence the same procedure can be applied to the $d_i(\psi)$ to obtain a policy ψ' with $d_i(\psi') \leq d_i(\psi)$ for $i = 1, 2, \dots, n$. Thus policy ψ' is at least as good as policy ψ . So, proceeding in this way we may obtain iteratively better policies. In Section 2.6 we will give an example of this.

2.5 The structure of optimal policies

In this section we study optimal solutions of (2.31). We show that for every $\varepsilon > 0$ and $i \in \{1, 2, \dots, n\}$ there exists an optimal solution d_1, d_2, \dots, d_n of (2.31) with $d_i < \varepsilon$ and $d_j < 1$ for all $j \neq i$. Moreover, we show that the upper bound of Theorem 2.4.6 is tight in case that the work capacities a_1, a_2, \dots, a_n are linearly independent over \mathbb{Z} .

Let ψ be a given policy. Then for all $s \in \mathbb{Z}_{\geq 0}$ we define the policy ψ^s by $k_t(\psi^s) = k_{t+s}(\psi)$ for $t = 0, 1, \dots$. The following lemma implies that policy ψ^s is at least as good as policy ψ .

Lemma 2.5.1 *Let ψ be a policy with $k_0(\psi) = j$. Then we have $d_j(\psi^1) \leq d_j(\psi) + 1 - a_j$ and $d_i(\psi^1) = d_i(\psi) - a_i$ for all $i \neq j$.*

Proof. By Lemma 2.4.2 we have that

$$\begin{aligned} d_j(\psi^1) &= \sup_{t \in \mathbb{Z}_{\geq 0}} (t \cdot a_j - N_j^t(\psi^1)) = \sup_{t \in \mathbb{Z}_{\geq 0}} (t \cdot a_j - (N_j^{t+1}(\psi) - 1)) = \\ &\sup_{t \in \mathbb{Z}_{\geq 1}} (t \cdot a_j - N_j^t(\psi)) + (1 - a_j) \leq \sup_{t \in \mathbb{Z}_{\geq 0}} (t \cdot a_j - N_j^t(\psi)) + (1 - a_j) = d_j(\psi) + 1 - a_j. \end{aligned}$$

Further for every $i \neq j$ we have $\sup_{t \in \mathbb{Z}_{\geq 1}} (t \cdot a_i - N_i^t(\psi)) \geq a_i$ and thus by Lemma 2.4.2 we have

$$\begin{aligned} d_i(\psi^1) &= \sup_{t \in \mathbb{Z}_{\geq 0}} (t \cdot a_i - N_i^t(\psi^1)) = \sup_{t \in \mathbb{Z}_{\geq 1}} ((t-1) \cdot a_i - N_i^t(\psi)) = \\ &\sup_{t \in \mathbb{Z}_{\geq 0}} (t \cdot a_i - N_i^t(\psi)) - a_i = d_i(\psi) - a_i. \end{aligned}$$

□

Corollary 2.5.2 *For every policy ψ we have $S(\psi^t) \leq S(\psi)$ for every $t \in \mathbb{Z}_{\geq 0}$.*

Proof. By Lemma 2.5.1 we have $S(\psi^1) = \sum_{i=1}^n d_i(\psi^1) \leq \sum_{i=1}^n (d_i(\psi) - a_i) + 1 = \sum_{i=1}^n d_i(\psi) = S(\psi)$. By induction we have $S(\psi) \geq S(\psi^1) \geq S(\psi^2) \geq \dots$ □

If policy ψ is applied in an (a_1, a_2, \dots, a_n) system then for every $t \in \mathbb{Z}_{\geq 0}$ and $i \in \{1, 2, \dots, n\}$ we put

$$h_i^t = h_i^t(\psi) = d_i(\psi) - t \cdot a_i + N_i^t(\psi).$$

Observe that by Lemma 2.4.2 we have $h_i^t(\psi) \geq 0$ for every policy ψ , $t \in \mathbb{Z}_{\geq 0}$ and $i \in \{1, 2, \dots, n\}$. Further we have the following lemma.

Lemma 2.5.3 *For every policy ψ , $t \in \mathbb{Z}_{\geq 0}$ and $i \in \{1, 2, \dots, n\}$ we have that $h_i^t \geq a_i \cdot v_i^t$. Moreover $h_i^t - a_i \cdot v_i^t$ is monotonically non-increasing in t and $\lim_{t \rightarrow \infty} (h_i^t - a_i \cdot v_i^t) = 0$.*

Proof. By (2.5) we have that

$$h_i^t - a_i \cdot v_i^t = d_i - a_i \cdot u_i^t \text{ for every } t \in \mathbb{Z}_{\geq 0}. \quad (2.33)$$

Thus the lemma follows from the facts that u_i^t is monotonically non-decreasing in t and $d_i = \lim_{t \rightarrow \infty} a_i \cdot u_i^t$. \square

Lemma 2.5.4 *If ψ is an optimal policy then for every $t \in \mathbb{Z}_{\geq 0}$ we have that ψ^t is an optimal policy and $d_i(\psi^t) = h_i^t(\psi)$ for $i = 1, 2, \dots, n$.*

Proof. By Lemma 2.5.1 and induction it follows for every $t \in \mathbb{Z}_{\geq 0}$ that $d_i(\psi^t) \leq h_i^t(\psi)$ for $i = 1, 2, \dots, n$. Suppose that there exists an $i \in \{1, 2, \dots, n\}$ for which $d_i(\psi^t) < h_i^t(\psi)$. Then

$$S(\psi^t) = \sum_{i=1}^n d_i(\psi^t) < \sum_{i=1}^n h_i^t(\psi) = S(\psi),$$

which contradicts the fact that ψ is an optimal policy. Hence $d_i(\psi^t) = h_i^t(\psi)$ for $i = 1, 2, \dots, n$ and $S(\psi^t) = S(\psi) = \tilde{S}$. \square

Corollary 2.5.5 *Let ψ be an optimal policy in an (a_1, a_2, \dots, a_n) system. Then $d_1 := h_1^t(\psi)$, $d_2 := h_2^t(\psi)$, \dots , $d_n := h_n^t(\psi)$ is an optimal solution of (2.31) for every $t \in \mathbb{Z}_{\geq 0}$.*

The following two lemmas are used in the proof of Theorem 2.5.8, which gives a rather useful property of some optimal solution of the MPP (2.31).

Lemma 2.5.6 *Let ψ be an optimal policy in an (a_1, a_2, \dots, a_n) system. Then for every $i \in \{1, 2, \dots, n\}$, $\varepsilon > 0$ and $t_0 \in \mathbb{Z}_{\geq 0}$ there exist $t \in \mathbb{Z}_{\geq t_0}$ such that $h_i^t(\psi) < \varepsilon$.*

Proof. By Lemma 2.5.4 and Lemma 2.4.2 we have that

$$\begin{aligned} \sup_{t \in \mathbb{Z}_{\geq t_0}} (d_i(\psi^{t_0}) - h_i^t(\psi)) &= \sup_{t \in \mathbb{Z}_{\geq t_0}} (h_i^{t_0}(\psi) - h_i^t(\psi)) = \\ \sup_{t \in \mathbb{Z}_{\geq t_0}} ((t - t_0)a_i + N_i^t(\psi) - N_i^{t_0}(\psi)) &= d_i(\psi^{t_0}). \end{aligned}$$

So, for every $\varepsilon > 0$ there exist $t \in \mathbb{Z}_{\geq t_0}$ such that $h_i^t(\psi) < \varepsilon$. \square

Lemma 2.5.7 *Suppose that for some policy ψ we have for a $t_0 \in \mathbb{Z}_{\geq 0}$ that there exists a $j \in \{1, 2, \dots, n\}$ such that $h_j^{t_0}(\psi) \geq 1$, $N_j^{t_0}(\psi) \geq 1$ and $k_{t_0}(\psi) = l \neq j$.*

Let $\tilde{t} = \max\{t < t_0 : k_t(\psi) = j\}$. Consider the policy φ defined by $k_t(\varphi) = k_t(\psi)$ for every $t \in \mathbb{Z}_{\geq 0}$, $t \neq \tilde{t}, t_0$, $k_{\tilde{t}}(\varphi) = l$ and $k_{t_0}(\varphi) = j$. Then $d_l(\varphi) \leq d_l(\psi)$ and $d_i(\varphi) = d_i(\psi)$ for every $i \in \{1, 2, \dots, n\} \setminus \{l\}$.

Proof. Since $N_l^t(\varphi) \geq N_l^t(\psi)$ for every $t \in \mathbb{Z}_{\geq 0}$ we have by Lemma 2.4.2 that $d_l(\varphi) \leq d_l(\psi)$. Further it is obvious for $i \in \{1, 2, \dots, n\} \setminus \{j, l\}$ that $d_i(\varphi) = d_i(\psi)$. So it remains to prove that $d_j(\varphi) = d_j(\psi)$. Because $N_j^t(\varphi) \leq N_j^t(\psi)$ for every $t \in \mathbb{Z}_{\geq 0}$ we have by Lemma 2.4.2 that $d_j(\varphi) \geq d_j(\psi)$. If $d_j(\varphi) > d_j(\psi)$ then

$$d_j(\varphi) = t_0 \cdot a_j - N_j^{t_0}(\varphi) = t_0 \cdot a_j - N_j^{t_0}(\psi) + 1 > d_j(\psi) \Rightarrow$$

$$h_j^{t_0}(\psi) = d_j(\psi) - t_0 \cdot a_j + N_j^{t_0}(\psi) < 1,$$

which is a contradiction. Thus $d_j(\varphi) = d_j(\psi)$. \square

Theorem 2.5.8 *For every $\varepsilon > 0$ and $i \in \{1, 2, \dots, n\}$ there exists an optimal solution d_1, d_2, \dots, d_n of (2.31) with $d_i < \varepsilon$ and $d_j < 1$ for all $j \neq i$.*

Proof. Without loss of generality we can assume $i = 1$. Let $d_1^*, d_2^*, \dots, d_n^*$ be an optimal solution of (2.31) and ψ an optimal policy obtained by applying the GG algorithm to this optimal solution. Then according to Corollary 2.5.5 it is sufficient to show that for every $\varepsilon > 0$ there exists a $t \in \mathbb{Z}_{\geq 0}$ such that $h_1^t(\psi) < \varepsilon$ and $h_j^t(\psi) < 1$ for $j = 2, 3, \dots, n$.

Suppose this does not hold. Then there exists an ε with $0 < \varepsilon < a_1$ such that for all $t \in \mathbb{Z}_{\geq 0}$ for which $h_1^t < \varepsilon$ there exists an $i \in \{2, 3, \dots, n\}$ with $h_i^t \geq 1$. Let $t' = \min\{t \in \mathbb{Z}_{\geq 0} : N_i^t(\psi) \geq 1\}$ for all $i \in \{2, 3, \dots, n\}$ and $J = \{t \in \mathbb{Z}_{\geq 0}, t \geq t' : h_1^t < \varepsilon\}$. Then by Lemma 2.5.6 we have that $J \neq \emptyset$ and it follows for every $t \in J$ that there exists an $i \in \{2, 3, \dots, n\}$ such that $h_i^t \geq 1$ and $N_i^t \geq 1$. By Lemma 2.5.1 we have for $t \in J$ that $k_t(\psi) = 1$, because $h_1^t < \varepsilon < a_1$ and $h_1^{t+1} \geq 0$. We construct a new policy ϕ from policy ψ by changing some of the $k_t(\psi)$ in the following way. Let $J = \{c_0, c_1, c_2, \dots\}$ with $c_0 < c_1 < c_2 < \dots$. We start with $k_t(\phi) = k_t(\psi)$ for every $t \in \mathbb{Z}_{\geq 0}$ and then we do consecutively for $s = c_0, s = c_1, \dots$ the following. Let $i \in \{2, 3, \dots, n\} : h_i^s \geq 1$ and $\tilde{t} = \max\{t < s : k_t(\psi) = i\}$. For the new policy ϕ we take $k_{\tilde{t}}(\phi) = 1 = k_s(\psi)$ and $k_s(\phi) = i = k_{\tilde{t}}(\psi)$. Then by Lemma 2.5.7 and induction we have that $d_1(\varphi) \leq d_1(\psi)$ and $d_i(\varphi) = d_i(\psi)$ for $i = 2, 3, \dots, n$. If $d_1(\varphi) < d_1(\psi)$ then $S(\varphi) = \sum_{i=1}^n d_i(\varphi) < \sum_{i=1}^n d_i(\psi) = S(\psi)$, which contradicts the fact that ψ is an optimal policy. Hence, $d_i(\varphi) = d_i(\psi)$ for $i = 1, 2, \dots, n$ and φ is also an optimal policy. Thus for every $t \in \mathbb{Z}_{\geq 0}$ we have that

$$h_1^t(\varphi) = d_1(\psi) - t \cdot a_1 + N_1^t(\varphi) \geq d_1(\psi) - t \cdot a_1 + N_1^t(\psi) = h_1^t(\psi)$$

and for every $t \in J$ we have $h_1^t(\varphi) = h_1^t(\psi) + 1 \geq 1$. So for every $t \geq t'$ we have $h_1^t(\varphi) \geq \varepsilon$. However, by Lemma 2.5.6 we have that φ is not an optimal policy, which is a contradiction. \square

From Lemma 2.5.3, Lemma 2.5.6 and Theorem 2.5.8 we obtain the following property of optimal policies.

Corollary 2.5.9 *If in an (a_1, a_2, \dots, a_n) system an optimal policy ψ is applied then for all $t \in \mathbb{Z}_{\geq 0}$, $i \in \{1, 2, \dots, n\}$ and $\varepsilon > 0$ there exist $t' \in \mathbb{Z}_{\geq 0}$, $t' > t$ such that $a_i \cdot v_i^{t'} < \varepsilon$ and $a_j \cdot v_j^{t'} < 1$ for all $j \neq i$.*

If $d_i < 1$ then $\max(0, \lceil t \cdot a_i - d_i \rceil) = \lceil t \cdot a_i - d_i \rceil$ for every $t \in \mathbb{Z}_{\geq 0}$. So, to obtain an optimal policy for an (a_1, a_2, \dots, a_n) system we can, according to Theorem 2.5.8, solve the following MPP instead of (2.31).

$$\begin{aligned}
 \text{minimize } S &= \sum_{i=1}^n d_i \\
 \text{subject to:} & \\
 \sum_{i=1}^n \lceil t \cdot a_i - d_i \rceil &\leq t && \text{for } t = 0, 1, \dots \\
 0 \leq d_i < 1 &&& \text{for } i = 1, 2, \dots, n
 \end{aligned} \tag{2.34}$$

From an optimal solution d_1, d_2, \dots, d_n of (2.34) an optimal policy ψ can be obtained by applying the GG algorithm. Further the GG algorithms can be used to search iteratively for better solutions of (2.34). We now introduce a simple greedy algorithm which in fact is a (special) GG algorithm and thus has the above properties. We denote this algorithm as the SG (Special Greedy) algorithm and in the SG algorithm we completely fix the k_t . This is contrary to the GG algorithm, where there is sometimes a freedom in the choice of k_t .

SG Algorithm. The servers are ordered such that $a_1 \geq a_2 \geq \dots \geq a_n$. Let $d_1, d_2, \dots, d_n \in \mathbb{R}_{\geq 0}$ be given. Then the SG algorithm determines ψ inductively in the following way. Suppose

$k_0(\psi), k_1(\psi), \dots, k_{t-1}(\psi)$ have been determined. Then $k_t(\psi)$ is defined as the $i \in \{1, 2, \dots, n\}$ for which $\frac{h_i^t}{a_i}$ is minimal, where $h_i^t = d_i - t \cdot a_i + N_i^t(\psi)$ for $i = 1, 2, \dots, n$. In case of a tie the server with the lowest index is chosen.

Proposition 2.5.10 *The SG algorithm is a special GG algorithm.*

Proof. For the GG algorithm we have that

$$q_i(t) = \min\{t' \in \mathbb{Z}_{\geq 0} : t' \cdot a_i > d_i - t \cdot a_i + N_i^t(\psi)\} =$$

$$\lfloor \frac{d_i - t \cdot a_i + N_i^t(\psi)}{a_i} \rfloor + 1 = \lfloor \frac{h_i^t}{a_i} \rfloor + 1 \text{ for } i = 1, 2, \dots, n.$$

So, for every $i \in \{1, 2, \dots, n\}$ for which $\frac{h_i^t}{a_i}$ is minimal also $q_i(t)$ is minimal. Thus the SG algorithm is a special GG algorithm. \square

Note that from the proof of Proposition 2.5.10 it follows that the only difference between the GG algorithm and the SG algorithm is that the GG algorithm may route the arriving job at time t to *any* server i for which $\lfloor \frac{h_i^t}{a_i} \rfloor$ is minimal, while the SG algorithm routes this arriving job to *the unique* server i (of lowest index in case of a tie) for which $\frac{h_i^t}{a_i}$ is minimal.

Let ψ be an optimal policy obtained by applying the SG algorithm to an optimal solution d_1, d_2, \dots, d_n of (2.34). Then this optimal policy ψ can be represented in the following way. For $i \in \{1, 2, \dots, n\}$ we represent server i as a point P_i on a circle that moves from $t = 0$ clockwise over the circle with an angular momentum of $a_i \cdot 2\pi$ radians per time unit. So, the angular momentum is directly proportional to the work capacity a_i of the server. At $t = 0$ the point P_i starts $d_i \cdot 2\pi$ radians before the top of the circle. Then the order $k_0(\psi), k_1(\psi), \dots$ in which the arriving jobs are routed to the servers if policy ψ is applied is the order in which the points representing the servers pass the top of the circle. If several points pass the top of the circle at exactly the same moment then the convention for the SG algorithm is used, namely that the server with the lowest index of the tied servers comes first in order (note that in the GG algorithm the choice of order is free if several points pass the top of the circle in the same time interval of unit length).

Sequences $k_0(\psi), k_1(\psi), \dots$ obtained by applying the SG algorithm to d_1, d_2, \dots, d_n where $d_i < 1$ for $i = 1, 2, \dots, n$ are known as billiard sequences (see for example [12]). Namely, if a billiard ball bounces in an n - dimensional cube of which the hyperfaces are numbered and opposite hyperfaces have the same number, then the sequence of integers denoting the order of the hyperfaces to which the ball bounces is called a billiard sequence. It is clear that the sequence obtained from the SG algorithm is a billiard sequence for some well chosen initial position and direction of the billiard ball. So, from Theorem 2.5.8 we have derived the following result on the optimality of billiard sequences.

Theorem 2.5.11 *For every (a_1, a_2, \dots, a_n) system there exists an optimal routing*

policy ψ such that the corresponding routing sequence $k_0(\psi), k_1(\psi), \dots$ is a billiard sequence.

Observe that in the circle representation the constraints $\sum_{i=1}^n [t \cdot a_i - d_i] \leq t$ of (2.34) imply that the starting points on the circle corresponding with the d_i must be such that the total number of passages through the top of the circle in the time-interval $[0, t)$ is at most t for $t = 0, 1, \dots$

According to the following theorem the upper bound of Corollary 2.4.6 is tight if the work capacities a_1, a_2, \dots, a_n are linearly independent over \mathbb{Z} .

Theorem 2.5.12 *If the work capacities a_1, a_2, \dots, a_n are linearly independent over \mathbb{Z} then $S(\psi) = \sum_{i=1}^n d_i(\psi) \geq n - 1$ for every policy ψ .*

Proof. Suppose that $\sum_{i=1}^n d_i(\psi) \leq n - 1$ for some policy ψ . Then there exist $z_i \in \mathbb{R}_{\geq 0}$ such that $\sum_{i=1}^n z_i = n - 1$ and $d_i = d_i(\psi) \leq z_i$ for $i = 1, 2, \dots, n$. Let $\varepsilon > 0$ be an arbitrary small positive number. We have that $1, a_2, a_3, \dots, a_n$ are linearly independent over \mathbb{Z} , because $\sum_{i=1}^n a_i = 1$. So, according to Kronecker's theorem on inhomogeneous simultaneous Diophantine approximation, there exist $m_i \in \mathbb{Z}$, $t \in \mathbb{N}$ such that $|t \cdot a_i - m_i - (z_i + \frac{\varepsilon}{2})| < \frac{\varepsilon}{2}$ for $i = 2, 3, \dots, n$ where ε is some positive number less than $\frac{1}{n}$. Then

$$t \cdot a_i = m_i + z_i + \theta_i \quad (2.35)$$

with $\theta_i \in (0, \varepsilon)$ for $i = 2, 3, \dots, n$. Further

$$t \cdot a_1 = t - \sum_{i=2}^n t \cdot a_i = t - \sum_{i=2}^n m_i - \sum_{i=2}^n z_i - \sum_{i=2}^n \theta_i = m_1 + z_1 + \theta_1 \quad (2.36)$$

where $m_1 := t - M_1 - n + 1 \in \mathbb{Z}$ and $\theta_1 := -\sum_{i=2}^n \theta_i \in (-\delta, 0)$ with $\delta = (n - 1) \cdot \varepsilon$. By Lemma 2.4.2 we have that $t \cdot a_1 - N_1^t \leq d_1 \leq z_1$. Hence by (2.36) we have $m_1 + z_1 + \theta_1 - N_1^t \leq z_1$ and thus

$$N_1^t \geq m_1. \quad (2.37)$$

For $i = 2, 3, \dots, n$ we have $t \cdot a_i - N_i^t \leq d_i \leq z_i$ by Lemma 2.4.2. Thus by (2.35) $m_i + z_i + \theta_i - N_i^t \leq z_i$. So, $N_i^t \geq m_i + \theta_i$ whence

$$N_i^t \geq m_i + 1 \text{ for } i = 2, 3, \dots, n. \quad (2.38)$$

Since $\sum_{i=1}^n m_i = \sum_{i=1}^n (t \cdot a_i - z_i - \theta_i) = t - n + 1$, we have by (2.37) and (2.38) that

$$t = \sum_{i=1}^n N_i^t \geq m_1 + \sum_{i=2}^n (m_i + 1) = \sum_{i=1}^n m_i + n - 1 = t.$$

Hence $N_1^t = m_1$ and $N_i^t = m_i + 1$ for $i = 2, 3, \dots, n$. From $N_1^t = m_1$ and (2.36) it follows that $t \cdot a_1 - N_1^t = z_1 + \theta_1$ with $\theta_1 \in (-\delta, 0)$ where δ can be made arbitrarily small. So, by Lemma 2.4.2 $d_1 = z_1$. Analogously it follows that $d_i = z_i$ for $i = 2, 3, \dots, n$ and thus $\sum_{i=1}^n d_i = n - 1$. \square

Corollary 2.5.13 *If a_1, a_2, \dots, a_n are linearly independent over \mathbb{Z} then $\tilde{S} = n - 1$ and $\tilde{W} = \frac{n-1}{2}$. In that case all d_1, d_2, \dots, d_n that fulfill (2.27) are optimal solutions of (2.31). If furthermore $d_i < 1$ for $i = 1, 2, \dots, n$ then those d_i are also an optimal solution for (2.34).*

Conjecture 2.5.14 *Let $q := \dim_{\mathbb{Q}}(a_1\mathbb{Q} + a_2\mathbb{Q} + \dots + a_n\mathbb{Q})$ be the irrationality degree of a_1, a_2, \dots, a_n . Then $\tilde{S} \geq \frac{n+q}{2} - 1$.*

By Theorem 2.3.2, $\tilde{W} \geq 0$ and Corollary 2.5.13 the conjecture is true in the extremal cases $q = 1$ and $q = n$.

2.6 The optimal policy in case of rational service rates

In this section we consider the case that all the work capacities a_i are rational, i.e. $a_i \in \mathbb{Q}$ for $i = 1, 2, \dots, n$. Then there exist $T \in \mathbb{N}$, $x_i \in \mathbb{N}$ such that $a_i = \frac{x_i}{T}$ for $i = 1, 2, \dots, n$ and $\gcd(x_1, x_2, \dots, x_n) = 1$. By (2.2) we have that $T = \sum_{i=1}^n x_i$. We denote T as the period of the system, because T possesses the following property.

Theorem 2.6.1 $T = \min\{t \in \mathbb{N} : \text{there exist periodic policies with period } t \text{ which are optimal}\}$.

Proof. In the previous section we have shown that a policy ψ obtained from applying the SG algorithm to an optimal solution of (2.34) is optimal and we have given a circle representation for such a policy ψ . From the circle representation it follows that if $a_i = \frac{x_i}{T}$ then in T time-units the point P_i representing server i makes exactly $a_i \cdot T = x_i$ turns around the circle. So, at $t = T$ the point P_i has passed the top of the circle exactly x_i times and is at the same place on the circle as where it started

at $t = 0$. Hence, if $a_i = \frac{x_i}{T}$ for $i = 1, 2, \dots, n$ then the optimal policy ψ is periodic with period T and from T consecutively arriving jobs exactly x_i are routed to server i for $i = 1, 2, \dots, n$. In other words, for policy ψ we have for all $t \in \mathbb{Z}_{\geq 0}$ that

$$k_{t+T}(\psi) = k_t(\psi) \quad (2.39)$$

and

$$N_i^{t+T}(\psi) = N_i^t(\psi) + x_i \text{ for } i = 1, 2, \dots, n. \quad (2.40)$$

So, indeed there exist optimal policies with period T .

It remains to show that there do not exist optimal policies with period $t' < T$ exist. Suppose ψ' is an optimal policy with period $t' < T$. Then there exist $m_i \in \mathbb{N}$ such that $N_i^{t+t'}(\psi') = N_i^t(\psi') + m_i$ for $i = 1, 2, \dots, n$ and all $t \in \mathbb{Z}_{\geq 0}$. Hence

$$p_i = \frac{m_i}{t'} \text{ for } i = 1, 2, \dots, n. \quad (2.41)$$

Because ψ' is optimal it follows from Corollary 2.4.6 that $S(\psi') \leq n - 1$. So, $S(\psi') < \infty$ and by Lemma 2.3.6 we have $p_i = a_i = \frac{x_i}{T}$ for $i = 1, 2, \dots, n$. So, by (2.41), it follows that $\frac{m_i}{t'} = \frac{x_i}{T}$ and thus $x_i = \frac{m_i \cdot T}{t'}$ for $i = 1, 2, \dots, n$. For a prime number p and $t \in \mathbb{N}$ let $\text{ord}_p(t) = \max\{k \in \mathbb{Z}_{\geq 0} : p^k \text{ divides } t\}$. Because $T > t'$ there exists a prime number p such that $\text{ord}_p(T) > \text{ord}_p(t')$. Hence we have that $\text{ord}_p(x_i) = \text{ord}_p(m_i) + \text{ord}_p(T) - \text{ord}_p(t') \geq \text{ord}_p(T) - \text{ord}_p(t') \geq 1$ and thus p divides x_i for $i = 1, 2, \dots, n$. Hence $\text{gcd}(x_1, x_2, \dots, x_n) \geq p > 1$, which is a contradiction. \square

So, if all the work capacities a_i are rational, $a_i = \frac{x_i}{T}$ with $x_i, T \in \mathbb{N}$ and $\text{gcd}(x_1, x_2, \dots, x_n) = 1$, then we can restrict to policies ψ for which (2.39) and (2.40) hold to obtain optimal policies. We will denote policies for which (2.39) and (2.40) hold as proportional periodic (p.p.) policies. Observe that a p.p. policy ψ is characterized by $k_0(\psi), k_1(\psi), \dots, k_{T-1}(\psi)$ and therefore we denote ψ as $(k_0(\psi), k_1(\psi), \dots, k_{T-1}(\psi))^\infty$.

Further we will render such a system with rational work capacities as (x_1, x_2, \dots, x_n) instead of (a_1, a_2, \dots, a_n) . The following lemma shows that the performance of p.p. policies is invariant under shifts.

Lemma 2.6.2 *For an p.p. policy ψ in an (x_1, x_2, \dots, x_n) system we have that*

$$S(\psi^t) = S(\psi) \text{ for every } t \in \mathbb{Z}_{\geq 0}.$$

Proof. A p.p. policy ψ is the same policy as ψ^{mT} for $m \in \mathbb{Z}_{\geq 0}$. Hence, $S(\psi) = S(\psi^T) = S(\psi^{2T}) = \dots$ and the lemma follows from Corollary 2.5.2. \square

It is convenient to multiply some of the earlier introduced variables with the period T to obtain integers. If policy ψ is applied to an (x_1, x_2, \dots, x_n) system then we define for $i = 1, 2, \dots, n$ the variables

$$s_i = s_i(\psi) := T \cdot d_i(\psi) = \lim_{t \rightarrow \infty} x_i \cdot u_i^t.$$

From this definition and Lemma 2.4.2 we derive

$$s_i = \max_{t \in \mathbb{Z}_{\geq 0}} (t \cdot x_i - T \cdot N_i^t). \quad (2.42)$$

Moreover, if $s_i < \infty$ then $s_i \in \mathbb{Z}_{\geq 0}$ and thus

$$S \cdot T = T \cdot \sum_{i=1}^n d_i = \sum_{i=1}^n s_i \in \mathbb{Z}_{\geq 0}. \quad (2.43)$$

By Corollary 2.4.7 and Theorem 2.3.2 it follows for any (x_1, x_2, \dots, x_n) system that $2T \cdot W \in \mathbb{Z}_{\geq 0}$ if $W < \infty$, $\tilde{S} \cdot T \in \mathbb{Z}_{\geq 0}$ and $2T \cdot \tilde{W} \in \mathbb{Z}_{\geq 0}$. Note that the factors 2 in the preceding expressions may be deleted if n is odd.

We have for $t \in \mathbb{Z}_{\geq 0}$ that $S^t \cdot T = \sum_{i=1}^n (a_i \cdot T) \cdot v_i^t = \sum_{i=1}^n x_i \cdot v_i^t$. Therefore it is convenient to define for $t \in \mathbb{Z}_{\geq 0}$ and $i \in \{1, 2, \dots, n\}$ the variables

$$w_i^t = x_i \cdot v_i^t \text{ which is } T \text{ times the remaining workload in server } i \text{ at time } t.$$

We note the following properties of w_i^t :

w1 $S^t \cdot T = \sum_{i=1}^n w_i^t$.

w2 $w_1^0 = w_2^0 = \dots = w_n^0 = 0$.

w3 If $k_{t-1} = l$, then $w_l^t = w_l^{t-1} + T - x_l$ and $w_k^t = \max(w_k^{t-1} - x_k, 0) = w_k^{t-1} + \max(-x_k, -w_k^{t-1})$ for $k \neq l$.

w4 $w_i^t \in \mathbb{Z}_{\geq 0}$ for $i \in \{1, 2, \dots, n\}$ and $t \in \mathbb{Z}_{\geq 0}$.

w5 $S^t \cdot T - S^{t-1} \cdot T = \sum_{k \neq l} \max(x_k - w_k^{t-1}, 0)$.

Proof. The properties w1, w2 and w3 are obvious and property w4 follows from w3. So it only remains to prove property w5. We have

$$S^t \cdot T - S^{t-1} \cdot T \stackrel{\text{w1}}{=} \sum_{i=1}^n (w_i^t - w_i^{t-1}) \stackrel{\text{w3}}{=} T - x_l + \sum_{k \neq l} \max(-x_k, -w_k^{t-1}) =$$

$$\sum_{k \neq l} \{x_k + \max(-x_k, -w_k^{t-1})\} = \sum_{k \neq l} \max(x_k - w_k^{t-1}, 0). \quad \square$$

From properties w1 and w4 it follows that $S^t \cdot T \in \mathbb{Z}_{\geq 0}$ for all $t \in \mathbb{Z}_{\geq 0}$. If a p.p. policy ψ is applied then the w_i^t have some interesting properties. To illustrate this we calculate the w_i^t for the situation that $n = 3$, $(x_1, x_2, x_3) = (4, 3, 2)$, so $T = 9$, and we apply the policy $\psi = (1, 2, 1, 3, 2, 1, 2, 1, 3)^\infty$.

Table 1

t	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
k_t	1	2	1	3	2	1	2	1	3	1	2	1	3	2	1	2	1	3
w_1^t	0	5	1	6	2	0	5	1	6	2	7	3	8	4	0	5	1	6
w_2^t	0	0	6	3	0	6	3	9	6	3	0	6	3	0	6	3	9	6
w_3^t	0	0	0	0	7	5	3	1	0	7	5	3	1	8	6	4	2	0
$S^t \cdot T$	0	5	7	9	9	11	11	11	12	12	12	12	12	12	12	12	12	12

Table 1 contains all the calculated w_i^t until the end of the second period. Note that the w_i^t for $t = 8$ and $t = 17$ are the same. Since the k_t have also period 9, the w_i^t are periodical with period $T = 9$ from moment $t = 8$ on. Further it follows that $S^t \cdot T = \sum_{i=1}^3 w_i^t$ is constant from $t = 8$. In this example $S^t \cdot T = 12$ and thus $S^t = \frac{4}{3}$ for $t \geq 8$. So $S = \frac{4}{3}$. More generally, we have the following theorem.

Theorem 2.6.3 *If a p.p. policy ψ is applied to an (x_1, x_2, \dots, x_n) system then for all $t \in \mathbb{Z}_{\geq T-1}$ we have $S = S^t$, $d_i(\psi) = a_i \cdot u_i^t$, $s_i(\psi) = x_i \cdot u_i^t$,*

$$v_i^t = v_i^{t+T} \text{ and } w_i^t = w_i^{t+T} \quad (2.44)$$

for $i = 1, 2, \dots, n$. Further for $t \in \{T, 2T, 3T, \dots\}$ we have

$$d_i(\psi) = a_i \cdot v_i^t \text{ and } s_i(\psi) = w_i^t. \quad (2.45)$$

Proof. We consider an arbitrary server i . Because a p.p. policy is applied we know that in the first period exactly x_i jobs are routed to server i . Assume in the first period an arriving job is routed to server i at moments $t_i^1, t_i^2, \dots, t_i^{x_i}$ with $0 \leq t_i^1 < t_i^2 < \dots < t_i^{x_i} < T$. After the first period this is periodically repeated, so the arriving jobs are routed to server i at moments $t_i^1 + m \cdot T$, $t_i^2 + m \cdot T$, \dots , $t_i^{x_i} + m \cdot T$ with $m \in \mathbb{Z}_{\geq 0}$. We first show that from time $t = t_i^{x_i}$ server i is constantly in use for

processing jobs. Let $t > t_i^{x_i}$ be an arbitrary moment and $t_0 = \max(0, t - T)$. Then exactly x_i jobs have been routed to server i in the interval $[t_0, t)$, because a p.p. policy is applied and $t > t_i^{x_i}$. Since server i could not start earlier than at time t_0 with the processing of those jobs, server i is busy with those jobs at least until time $t_0 + x_i \cdot a_i^{-1} = t_0 + x_i \cdot \frac{T}{x_i} = t_0 + T \geq t$. So server i is busy at any arbitrary moment $t > t_i^{x_i}$, whence $u_i^{t_i^{x_i}} = u_i^{t_i^{x_i}+1} = \dots$. Because $t_i^{x_i} \leq T - 1$ for $i = 1, 2, \dots, n$, we have for $i = 1, 2, \dots, n$ that $u_i^{T-1} = u_i^T = \dots$. Hence $S^{T-1} = S^T = \dots = S$ and it follows for all $t \in \mathbb{Z}_{\geq T-1}$ that $S = S^t$ and $d_i(\psi) = a_i \cdot u_i^t$, $s_i(\psi) = x_i \cdot u_i^t$ for $i = 1, 2, \dots, n$. Further if $t \geq t_i^{x_i}$, $t \in \mathbb{N}$ and $k_t \neq i$, then $v_i^t \geq 1$ and thus $w_i^t \geq x_i$. Hence we have by property w3 for $t \in \mathbb{N}$, $t \geq t_i^{x_i}$ that

$$\begin{aligned} w_i^{t+1} &= w_i^t + T - x_i & \text{if } k_t = i, \\ w_i^{t+1} &= w_i^t - x_i & \text{if } k_t \neq i \end{aligned}$$

and $w_i^{t+T} = w_i^t + x_i(T - x_i) - (T - x_i)x_i = w_i^t$. This implies that $w_i^{T+t} = w_i^t$, $v_i^{T+t} = v_i^t$ for $i \in \{1, 2, \dots, n\}$ and $t \geq T - 1$.

Thus $v_i^T = v_i^{2T} = v_i^{3T} = \dots$ and we only have to show that $a_i \cdot v_i^T = d_i(\psi)$ for $i = 1, 2, \dots, n$ to complete the proof. From the proof of Lemma 2.4.1 we have for $t \in \mathbb{Z}_{\geq 0}$ and $i \in \{1, 2, \dots, n\}$ that $a_i \cdot u_i^t - a_i \cdot v_i^t = a_i \cdot t - N_i^t(\psi)$. Hence $a_i \cdot u_i^T - a_i \cdot v_i^T = a_i \cdot T - N_i^T(\psi) = 0$, because ψ is a p.p. policy. So $a_i \cdot v_i^T = a_i \cdot u_i^T = d_i(\psi)$ and $v_i^T = T \cdot a_i \cdot v_i^T = T \cdot d_i(\psi) = s_i(\psi)$ for $i = 1, 2, \dots, n$. \square

By Theorem 2.3.2 we have the following corollary.

Corollary 2.6.4 *If a p.p. policy is applied in an (x_1, x_2, \dots, x_n) system then for all $t \in \mathbb{Z}_{\geq T-1}$ we have*

$$W = S^t - \frac{n-1}{2}.$$

By Theorem 2.6.3 and Corollary 2.6.4 we have in the $(4, 3, 2)$ example that $s_1 = 2$, $s_2 = 3$, $s_3 = 7$ and that the long-run average waiting time W equals $S^8 - 1 = \frac{4}{3} - 1 = \frac{1}{3}$. We can compute W also in a more direct way. The total waiting time for jobs that arrive consecutively at the moments $t_0, t_0+1, \dots, t_0+t_1$ with $t_0, t_1 \in \mathbb{Z}_{\geq 0}$ equals $\sum_{t=t_0}^{t_0+t_1} v_{k_t}^t = \sum_{t=t_0}^{t_0+t_1} \frac{w_{k_t}^t}{x_{k_t}}$. If a p.p. policy is applied, then it follows from Theorem 2.6.3 that $\sum_{t=t_0}^{t_0+T-1} \frac{w_{k_t}^t}{x_{k_t}}$ is independent of t_0 if $t_0 \geq T - 1$. Thus from $t = T - 1$ on the total waiting time of T consecutively arriving jobs is constant if a p.p. policy is applied. Hence W can be computed by the formula

$$T \cdot W = \sum_{t=t_0}^{t_0+T-1} \frac{w_{k_t}^t}{x_{k_t}} \text{ for } t_0 \geq T - 1. \quad (2.46)$$

In (2.46) it is convenient to take $t_0 = T$. In the (4, 3, 2) example we find that $T \cdot W = \sum_{t=9}^{17} \frac{w_{k_t}^t}{x_{k_t}} = 3$ (see Table 1). So $W = \frac{3}{T} = \frac{1}{3}$, as expected.

If a p.p. policy ψ is applied in an (x_1, x_2, \dots, x_n) system then for $i \in \{1, 2, \dots, n\}$ we put

$$t_i = \min\{t \in \mathbb{Z}_{\geq 0} : u_i^t = \lim_{t \rightarrow \infty} u_i^t\}. \quad (2.47)$$

From the proof of Theorem 2.6.3 it follows that $t_i \in \{0, 1, \dots, T-1\}$ for $i = 1, 2, \dots, n$.

Proposition 2.6.5 *If a p.p. policy is applied to an (x_1, x_2, \dots, x_n) system, then for each*

$i \in \{1, 2, \dots, n\}$ there exists at least one moment t' in each period for which $w_i^{t'} = 0$.

Proof. Let $i \in \{1, 2, \dots, n\}$ be arbitrary. It is clear that $w_i^{t_i} = 0$. Further from the proof of Theorem 2.6.3 it follows that $w_i^{t+T} = w_i^t$ for all $t \geq t_i$. So, for all $m \in \mathbb{Z}_{\geq 0}$ we have $w_i^{t_i+m \cdot T} = w_i^{t_i} = 0$. \square

For an optimal p.p. policy we have the following stronger result.

Theorem 2.6.6 *If an optimal p.p. policy is applied in an (x_1, x_2, \dots, x_n) system, then for each $i \in \{1, 2, \dots, n\}$ there exists at least one moment t' in each period such that $w_i^{t'} = 0$ and $w_j^{t'} \leq T-1$ for all $j \neq i$.*

Proof. By property w2 the statement holds for the first period. So, by (2.44) we only have to prove that for each $i \in \{1, 2, \dots, n\}$ there exist $t' \in \{T-1, T, T+1, \dots\}$ such that $w_i^{t'} = 0$ and $w_j^{t'} \leq T-1$ for all $j \neq i$. Let $0 < \varepsilon < \frac{1}{T}$. Then by Corollary 2.5.9 there exist $t' \in \{T-1, T, T+1, \dots\}$ such that $a_i \cdot v_i^{t'} < \varepsilon$ and $a_j \cdot v_j^{t'} < 1$ for all $j \neq i$. Hence $w_i^{t'} < T \cdot \varepsilon < 1$ and $w_j^{t'} < T$ for all $j \neq i$ and from property w4 it follows that $w_i^{t'} = 0$ and $w_j^{t'} \leq T-1$ for all $j \neq i$. \square

At the end of Section 2.4 we said that a GG algorithm may be used to obtain iteratively better policies. This is practical in case of rational a_i , since we can calculate the $d_i(\psi)$ quickly for p.p. policies. We have by Theorem 2.6.3 that $d_i(\psi) = a_i \cdot v_i^T$ for $i = 1, 2, \dots, n$ and v_i^T can be determined quickly. We use the SG algorithm to get p.p. policies. We start with a feasible solution d_1, d_2, \dots, d_n of (2.31) for which $d_i < 1$ for $i = 1, 2, \dots, n$. According to (2.39) and (2.40) the SG algorithm yields a p.p. policy ψ in that case. Then we can apply the SG algorithm to the feasible solution $d_1(\psi), d_2(\psi), \dots, d_n(\psi)$ of (2.31). If at some point in this procedure the SG algorithm is applied to a feasible solution d_1, d_2, \dots, d_n of (2.31) for which $d_i \geq 1$

for some $i = 1, 2, \dots, n$ then it is possible that the obtained policy ψ is not a p.p policy. However, it is easily seen that the obtained policy ψ is always ultimately proportional periodic, i.e there exists a $t \in \mathbb{Z}_{\geq 0}$ such that ψ^t is a p.p policy. In this case we think it is best to take $d_1(\psi^t), d_2(\psi^t), \dots, d_n(\psi^t)$ as new feasible solution of (2.31) instead of $d_1(\psi), d_2(\psi), \dots, d_n(\psi)$. Namely, ψ^t is a p.p policy and according to Corollary 2.5.2 we have that policy ψ^t is at least as good (and probably better) than policy ψ . We given an example to illustrate this procedure of obtaining iteratively better policies by using the SG algorithm.

Example. We apply the SG algorithm to a $(7, 6, 3, 2, 1)$ system and we start with the feasible solution $d_1 = d_2 = \dots = d_5 = \frac{18}{19}$ of (2.31). We denote $(d_1(\psi), d_2(\psi), \dots, d_5(\psi))$ by $d(\psi)$. Applying the SG algorithm we find consecutively $\psi_1 = (1, 2, 1, 3, 2, 1, 4, 2, 1, 3, 2, 1, 2, 1, 5, 4, 3, 2, 1)^\infty$ with $d(\psi_1) = (\frac{12}{19}, \frac{7}{19}, \frac{10}{19}, \frac{12}{19}, \frac{14}{19})$. $\psi_2 = (2, 1, 3, 2, 1, 4, 1, 2, 3, 1, 2, 1, 2, 5, 1, 4, 3, 2, 1)^\infty$ with $d(\psi_2) = (\frac{12}{19}, \frac{7}{19}, \frac{10}{19}, \frac{11}{19}, \frac{13}{19})$. $\psi_3 = (2, 1, 3, 2, 1, 4, 1, 2, 3, 1, 2, 1, 5, 2, 4, 1, 3, 2, 1)^\infty$ with $d(\psi_3) = (\frac{12}{19}, \frac{7}{19}, \frac{10}{19}, \frac{10}{19}, \frac{12}{19})$. $\psi_4 = (2, 1, 3, 2, 1, 4, 1, 2, 3, 1, 2, 5, 1, 2, 4, 1, 3, 2, 1)^\infty$ with $d(\psi_4) = (\frac{12}{19}, \frac{7}{19}, \frac{10}{19}, \frac{10}{19}, \frac{11}{19})$. Finally applying the SG algorithm to this last feasible solution of (2.31) yields again policy ψ_4 and the iterative process stops. We say that policy ψ_4 is stable under applying the SG algorithm. However, policy ψ_4 is not an optimal policy for this system. In fact $s_1 = 0, s_2 = 8, s_3 = 18, s_4 = 6, s_5 = 15$ is an optimal solution of (2.51). Hence $\tilde{S} = \frac{47}{19}$ for this system, but $S(\psi_4) = \frac{50}{19}$.

Remark. In general if you have a good start solution then this iterative process takes fewer iterations and finds a better stable policy. In Section 2.8 we will give several algorithms to find a reasonable good policy and thus a good start solution. It is also possible to apply the iterative process to several feasible solutions of (2.31) (which can be obtained by different algorithms) and then see which one yields the best stable policy.

2.6.1 Bounds corresponding to the mathematical programming problem

In this subsection we transform the mathematical programming problem (2.31) to an integer linear problem (ILP) in case of rational work capacities a_i . Moreover, we improve some of the lower bounds and upper bounds for the minimal average waiting time for this case.

Note that for given $s_i \in \mathbb{Z}_{\geq 0}$ with $i \in \{1, 2, \dots, n\}$ we have

$$\max(0, \lceil t \cdot a_i - d_i \rceil) = \max(0, \lceil t \cdot a_i - \frac{s_i}{T} \rceil) =$$

$$\max(0, \lceil \frac{t \cdot x_i - s_i}{T} \rceil) \text{ for all } t \in \mathbb{Z}_{\geq 0}. \quad (2.48)$$

So, by Lemma 2.42 we can solve the following MPP instead of (2.31) to obtain an optimal policy for an (x_1, x_2, \dots, x_n) system.

$$\begin{aligned} \text{minimize } S \cdot T &= \sum_{i=1}^n s_i \\ \text{subject to:} & \\ \sum_{i=1}^n \max(0, \lceil \frac{t \cdot x_i - s_i}{T} \rceil) &\leq t && \text{for } t = 0, 1, \dots \\ s_i &\in \mathbb{Z}_{\geq 0} && \text{for } i = 1, 2, \dots, n \end{aligned} \quad (2.49)$$

If an optimal p.p policy ψ is applied then $s_1 := s_1(\psi) = w_1^T$, $s_2 := s_2(\psi) = w_2^T$, \dots , $s_n := s_n(\psi) = w_n^T$ is an optimal solution of (2.49). In fact, by (2.33), Corollary 2.5.5 and Theorem 2.6.3 the following lemma holds.

Lemma 2.6.7 *If an optimal p.p policy is applied then for all $t \in \{T-1, T, T+1, \dots\}$ we have that $s_1 := w_1^t$, $s_2 := w_2^t$, \dots , $s_n := w_n^t$ is an optimal solution of (2.49).*

By Theorem 2.6.6 and Lemma 2.6.7 we have the following analogue of Theorem 2.5.8.

Theorem 2.6.8 *For every $i \in \{1, 2, \dots, n\}$ there exists an optimal solution s_1, s_2, \dots, s_n of (2.49) with $s_i = 0$ and $s_j \leq T - 1$ for all $j \neq i$.*

Corollary 2.6.9 *For every (x_1, x_2, \dots, x_n) system we have*

$$\tilde{S} \leq n - 1 - \frac{n - 1}{T}.$$

The upper bound of Corollary 2.6.9 is a fraction better than the general upper bound of Theorem 2.4.6. In Theorem 2.6.11 we will improve the upper bound a little more in case there exists an $i \in \{1, 2, \dots, n\}$ with $\gcd(x_i, T) > 1$.

If $s_i \leq T - 1$ then for all $t, m \in \mathbb{Z}_{\geq 0}$ we have

$$\max(0, \lceil \frac{(t + m \cdot T) \cdot x_i - s_i}{T} \rceil) = m \cdot x_i + \lceil \frac{t \cdot x_i - s_i}{T} \rceil, \quad (2.50)$$

hence $\sum_{i=1}^n \max(0, \lceil \frac{t \cdot x_i - s_i}{T} \rceil) \leq t$ if and only if $\sum_{i=1}^n \lceil \frac{(t + m \cdot T) \cdot x_i - s_i}{T} \rceil \leq t + m \cdot T$. So, by Theorem 2.6.8 and (2.50) we can obtain an optimal policy for an (x_1, x_2, \dots, x_n) system by solving the following MPP instead of (2.49).

$$\begin{aligned}
\text{minimize } ST &= \sum_{i=1}^n s_i \\
\text{subject to:} & \\
\sum_{i=1}^n \lceil \frac{t \cdot x_i - s_i}{T} \rceil &\leq t && \text{for } t = 0, 1, \dots, T-1 \\
s_i &\in \mathbb{Z}_{\geq 0} && \text{for } i = 1, 2, \dots, n \\
s_i &\leq T-1 && \text{for } i = 1, 2, \dots, n
\end{aligned} \tag{2.51}$$

From an optimal solution s_1, s_2, \dots, s_n of (2.51) we obtain an optimal p.p policy ψ such that $w_i^T = s_i(\psi) = s_i$ for $i = 1, 2, \dots, n$ by applying the SG algorithm to $d_i = \frac{s_i}{T}$ for $i = 1, 2, \dots, n$.

Lemma 2.6.10 *Let s_1, s_2, \dots, s_n be an optimal solution of (2.51). Then*

$$\gcd(x_i, T) \mid s_i \text{ for } i = 1, 2, \dots, n. \tag{2.52}$$

Proof. There exists an optimal p.p policy ψ such that $s_i(\psi) = s_i$ for $i = 1, 2, \dots, n$. For every policy it follows from properties w2 and w3 that $\gcd(x_i, T) \mid w_i^t$ for all $t \in \mathbb{Z}_{\geq 0}$ and $i \in \{1, 2, \dots, n\}$. Hence by (2.45) we have that $\gcd(x_i, T) \mid w_i^T = s_i$ for $i = 1, 2, \dots, n$. \square

Theorem 2.6.11 *For every (x_1, x_2, \dots, x_n) system we have*

$$\tilde{S} \leq n - 1 - \frac{D(x_1, x_2, \dots, x_n)}{T}$$

$$\text{with } D(x_1, x_2, \dots, x_n) = \sum_{i=1}^n \gcd(x_i, T) - \min_{i \in \{1, 2, \dots, n\}} \gcd(x_i, T).$$

Proof. Let $j \in \{1, 2, \dots, n\}$ such that $\gcd(x_j, T) = \min_{i \in \{1, 2, \dots, n\}} \gcd(x_i, T)$. Then by Theorem 2.6.8 there exists an optimal solution s_1, s_2, \dots, s_n of (2.51) for which $s_j = 0$. Further for all $i \neq j$ we have by Lemma 2.6.10 that $\gcd(x_i, T) \mid s_i$ and thus $s_i \leq T - \gcd(x_i, T)$. Hence

$$\tilde{S} \cdot T = \sum_{i=1}^n s_i \leq \sum_{i \neq j} (T - \gcd(x_i, T)) = T \cdot (n - 1) - D(x_1, x_2, \dots, x_n).$$

$$\text{So, } \tilde{S} \leq n - 1 - \frac{D(x_1, x_2, \dots, x_n)}{T}. \quad \square$$

Next we show that (2.51) is equivalent to a linear problem with zero-one variables and thus an ILP. We denote in the sequel by $z \pmod T$ the nonnegative integer

$a \leq T - 1$ for which T is a divisor of $z - a$. For $i = 1, 2, \dots, n$, $t = 0, 1, \dots, T - 1$ and $j = 0, 1, \dots, T - 1$ we put $e_{it}(j) = (j - t \cdot x_i) \pmod{T}$. Moreover, we define for $i = 1, 2, \dots, n$ and $j = 0, 1, \dots, T - 1$ zero-one variables λ_{ij} by

$$\lambda_{ij} = \begin{cases} 1 & \text{if } s_i = j \\ 0 & \text{if } s_i \neq j \end{cases}.$$

The λ_{ij} correspond to the variables s_i of (2.51). If $s_i \in \{0, 1, \dots, T - 1\}$ for $i = 1, 2, \dots, n$ then we have $\sum_{i=1}^n s_i = \sum_{i=1}^n \sum_{j=0}^{T-1} j \cdot \lambda_{ij}$ and

$$\sum_{i=1}^n \{(s_i - t \cdot x_i) \pmod{T}\} = \sum_{i=1}^n e_{it}(s_i) = \sum_{i=1}^n \sum_{j=0}^{T-1} e_{it}(j) \cdot \lambda_{ij} \text{ for } t = 0, 1, \dots, T - 1.$$

Hence (2.51) is equivalent to the following ILP.

$$\begin{aligned} \text{minimize } ST &= \sum_{i=1}^n \sum_{j=0}^{T-1} j \cdot \lambda_{ij} \\ \text{subject to:} & \\ \sum_{i=1}^n \sum_{j=0}^{T-1} (e_{it}(j) - j) \cdot \lambda_{ij} &\leq 0 && \text{for } t = 0, 1, \dots, T - 1 \\ \sum_{j=0}^{T-1} \lambda_{ij} &= 1 && \text{for } i = 1, 2, \dots, n \\ \lambda_{ij} &\in \{0, 1\} && \text{for } i = 1, 2, \dots, n \\ &&& \text{and } j = 0, 1, \dots, T - 1 \end{aligned} \tag{2.53}$$

Thus solving this ILP yields an optimal policy in case the a_i are rational. In some special cases (for example if $n = 2$ or T is small) it is easy to obtain an optimal solution of (2.53) and thus an optimal policy. However, in general if n and T are not fixed then we think that it is NP - hard to obtain an optimal policy. A reason for this is that the somewhat similar Periodic Maintenance Scheduling problem is shown to be NP - complete in [14].

By solving the LP - relaxation of (2.53) we get a lower bound for \widetilde{S} and thus for \widetilde{W} (see Corollary 2.6.14). The LP - relaxation of (2.53) is the following LP -problem.

$$\begin{aligned}
& \text{minimize} && \sum_{i=1}^n \sum_{j=0}^{T-1} j \cdot \lambda_{ij} \\
& \text{subject to:} && \\
& \sum_{i=1}^n \sum_{j=0}^{T-1} (e_{it}(j) - j) \cdot \lambda_{ij} && \leq 0 \quad \text{for } t = 0, 1, \dots, T-1 \\
& \sum_{j=0}^{T-1} \lambda_{ij} && = 1 \quad \text{for } i = 1, 2, \dots, n \\
& \lambda_{ij} && \geq 0 \quad \text{for } i = 1, 2, \dots, n \\
& && \text{and } j = 0, 1, \dots, T-1
\end{aligned} \tag{2.54}$$

Put $F_i = \{0, \gcd(x_i, T), 2 \gcd(x_i, T), \dots, (T - \gcd(x_i, T))\}$ for $i = 1, 2, \dots, n$. Then, by considering the dual program of (2.54) and applying von Neumann's duality theorem of linear programming we obtain the following theorem.

Theorem 2.6.12 *For every (x_1, x_2, \dots, x_n) system an optimal solution of (2.54) is given by*

$$\lambda_{ij}^* = \begin{cases} 0 & \text{if } j \notin F_i \\ \frac{\gcd(x_i, T)}{T} & \text{if } j \in F_i \end{cases} \tag{2.55}$$

and the value of (2.54) is

$$\frac{nT}{2} - \frac{C(x_1, x_2, \dots, x_n)}{2}, \text{ where } C(x_1, x_2, \dots, x_n) = \sum_{i=1}^n \gcd(x_i, T).$$

Proof. Let $t \in 0, 1, \dots, T-1$ be fixed. Then for every $i \in \{1, 2, \dots, n\}$ we have that

$$\{e_{it}(j) : j \in F_i\} = \{0, \gcd(x_i, T), 2 \gcd(x_i, T), \dots, (T - \gcd(x_i, T))\} = F_i$$

and thus $\sum_{j \in F_i} e_{it}(j) = \sum_{j \in F_i} j$. Hence

$$\sum_{i=1}^n \sum_{j=0}^{T-1} (e_{it}(j) - j) \cdot \lambda_{ij}^* = \sum_{i=1}^n \sum_{j \in F_i} (e_{it}(j) - j) \cdot \frac{\gcd(x_i, T)}{T} = \sum_{i=1}^n 0 = 0$$

for $t = 0, 1, \dots$. Further for every $i \in \{1, 2, \dots, n\}$ we have that

$$\sum_{j=0}^{T-1} \lambda_{ij}^* = \sum_{j \in F_i} \frac{\gcd(x_i, T)}{T} = \frac{T}{\gcd(x_i, T)} \cdot \frac{\gcd(x_i, T)}{T} = 1.$$

Thus (2.55) is a feasible solution of (2.54) and the value of this solution is

$$\begin{aligned} \sum_{i=1}^n \sum_{j=0}^{T-1} j \cdot \lambda_{ij}^* &= \sum_{i=1}^n \sum_{j \in F_i} j \cdot \frac{\gcd(x_i, T)}{T} = \\ &= \sum_{i=1}^n \left\{ \frac{\gcd(x_i, T)}{T} \cdot \left(\sum_{k=0}^{\frac{T}{\gcd(x_i, T)} - 1} k \cdot \gcd(x_i, T) \right) \right\} = \\ &= \sum_{i=1}^n \frac{\gcd(x_i, T)}{2} \cdot \left(\frac{T}{\gcd(x_i, T)} - 1 \right) = \frac{nT}{2} - \frac{1}{2} \cdot C(x_1, x_2, \dots, x_n). \end{aligned}$$

We consider the following dual problem of (2.54).

$$\begin{aligned} &\text{maximize} && \sum_{i=1}^n z_i \\ &\text{subject to:} && \\ &\sum_{t=0}^{T-1} \{(j - e_{it}(j)) \cdot y_t\} + z_i &\leq j & \text{ for } i = 1, 2, \dots, n \\ & && \text{and } j = 0, 1, \dots, T-1 \\ &y_t &\geq 0 & \text{ for } t = 0, 1, \dots, T-1 \end{aligned} \quad (2.56)$$

We claim that an optimal solution of (2.56) is given by

$$\begin{aligned} y_t^* &= \frac{1}{T} && \text{for } t = 0, 1, \dots, T-1 \text{ and} \\ z_i^* &= \frac{T}{2} - \frac{\gcd(x_i, T)}{2} && \text{for } i = 1, 2, \dots, n. \end{aligned} \quad (2.57)$$

Namely, let $i \in \{1, 2, \dots, n\}$ be fixed. Then for every $j \in \{0, 1, \dots, T-1\}$ we have that

$$\begin{aligned} \sum_{t=0}^{T-1} (e_{it}(j) \cdot y_t^*) &\geq \frac{1}{T} \cdot \sum_{t=0}^{T-1} \{-t \cdot x_i \pmod{T}\} = \frac{\gcd(x_i, T)}{T} \cdot \sum_{k=0}^{\frac{T}{\gcd(x_i, T)} - 1} k \cdot \gcd(x_i, T) = \\ &= \frac{\gcd(x_i, T)}{2} \cdot \left(\frac{T}{\gcd(x_i, T)} - 1 \right) = \frac{T}{2} - \frac{\gcd(x_i, T)}{2}. \end{aligned}$$

Hence

$$\sum_{t=0}^{T-1} (j - e_{it}(j)) \cdot y_t^* + z_i^* = j - \sum_{t=0}^{T-1} (e_{it}(j) \cdot y_t^*) + \frac{T}{2} - \frac{\gcd(x_i, T)}{2} \leq j$$

for $i = 1, 2, \dots, n$ and $j = 0, 1, \dots, T-1$. So, (2.57) is a feasible solution of (2.56) and the value of this solution is

$$\sum_{i=1}^n z_i^* = \sum_{i=1}^n \left(\frac{T}{2} - \frac{\gcd(x_i, T)}{2} \right) = \frac{nT}{2} - \frac{1}{2} \cdot C(x_1, x_2, \dots, x_n).$$

This lower bound for the value of (2.56) is equal to the earlier obtained upper bound for the value of (2.54). So, by the duality theorem of linear programming, due to von Neumann, we have that (2.55) is an optimal solution of (2.54), (2.57) is an optimal solution of (2.56) and the optimal value of both (2.54) and (2.56) is $\frac{nT}{2} - \frac{1}{2} \cdot C(x_1, x_2, \dots, x_n)$. \square

By Theorem 2.6.12 and Theorem 2.3.2 we have the following lower bound on \widetilde{S} and \widetilde{W} in case of rational a_i .

Corollary 2.6.13 *For every (x_1, x_2, \dots, x_n) system we have*

$$\widetilde{S} \geq \frac{n}{2} - \frac{C(x_1, x_2, \dots, x_n)}{2T} \quad (2.58)$$

and

$$\widetilde{W} \geq \frac{1}{2} - \frac{C(x_1, x_2, \dots, x_n)}{2T}, \text{ where } C(x_1, x_2, \dots, x_n) = \sum_{i=1}^n \gcd(x_i, T). \quad (2.59)$$

Corollary 2.6.14 shows that equalities can always be attained if $n = 2$ and Theorem 2.7.12 provides a criterion to have equality when $n > 2$.

Corollary 2.6.14 *For every (x_1, x_2) system $\widetilde{S} = 1 - \frac{1}{T}$ and $\widetilde{W} = \frac{1}{2} - \frac{1}{T}$.*

Proof. By definition we have that $\gcd(x_1, x_2) = 1$ and $T = x_1 + x_2$. Hence $\gcd(x_1, T) = \gcd(x_2, T) = 1$. Corollary 2.6.13 yields that $\widetilde{S} \geq 1 - \frac{1}{T}$. By Corollary 2.6.9 we have $\widetilde{S} \leq 1 - \frac{1}{T}$. So $\widetilde{S} = 1 - \frac{1}{T}$, $\widetilde{W} = 1 - \frac{1}{T}$ and every optimal policy attains the lower bounds of Corollary 2.6.13. \square

Remark. By Theorem 2.6.8 and Corollary 2.6.14 we have for every (x_1, x_2) system that both $s_1 = 0, s_2 = T - 1$ and $s_1 = T - 1, s_2 = 0$ are optimal solutions of (2.51). In fact it is easily seen that all the optimal solutions of (2.51) are the pairs of $s_1, s_2 \in \mathbb{Z}_{\geq 0}$ such that $s_1 + s_2 = T - 1$. Thus there are exactly T optimal solutions of (2.51) and it is easily seen that every optimal solution corresponds to exactly one optimal p.p policy. Indeed, there is never freedom of choice when the GG algorithm is applied to such a solution. So, if ψ is such an optimal p.p policy for an (x_1, x_2) system then all the optimal policies for that system are given by $\psi^0, \psi^1, \dots, \psi^{T-1}$. These T shifts of policy ψ are distinct and by Lemma 2.6.2 they are all optimal. Thus in an (x_1, x_2) system all the optimal policies are cyclic shifts of each other.

2.7 The optimality of regular policies

In this section we define regular sequences and corresponding policies. We will show that in order to minimize W the applied policy ψ should be as close to regular (balanced) as possible. In Theorem 2.7.11 we obtain the value of the long-run average waiting time in a single queue if the routing to that queue is regular. Theorem 2.7.12 shows that a regular policy is optimal by showing that it attains the lower bound of Corollary 2.6.13.

Let ψ be the applied policy. Then for $i \in \{1, 2, \dots, n\}$ we define the splitting sequence $\underline{\delta}_i$ of the routing to server i by

$$\underline{\delta}(i) = (\delta_t(i))_{t=0}^{\infty} \text{ where } \delta_t(i) = \begin{cases} 1 & \text{if } k_t(\psi) = i \\ 0 & \text{if } k_t(\psi) \neq i \end{cases}.$$

Further we define the support of this sequence by the set

$$A(i) := \{t \in \mathbb{Z}_{\geq 0} : \delta_t(i) = 1\}.$$

Let $t^* \in \mathbb{Z}_{\geq 0}$ be such that for some $i \in \{1, 2, \dots, n\}$ and some $t \in \mathbb{Z}_{\geq 0}$ we have that $t^* \leq t$, $v_i^{t^*} = 0$ and $u_i^{t^*} = u_i^t$. Then, by dividing equation (2.5) by a_i and subtraction we obtain

$$v_i^t = v_i^t - v_i^{t^*} = \frac{1}{a_i} \cdot (N_i^t - N_i^{t^*}) - (t - t^*) = \frac{1}{a_i} \cdot \sum_{t'=t^*}^{t-1} \delta_{t'}(i) - (t - t^*). \quad (2.60)$$

For any $N \in \mathbb{N}$ we define an N -block as a set of N consecutive non-negative integers. We say that the splitting sequence $\underline{\delta}(i)$ is a regular (the term Beatty is also used in the literature) sequence with density p if for every N -block ω we have that $\sum_{t \in \omega} \delta_t(i)$, the number of ones in the N -block, equals $\lfloor Np \rfloor$ or $\lceil Np \rceil$. In that case we also say that the corresponding set $A(i)$ is a regular set. We have the following lemma (see e.g [68]).

Lemma 2.7.1 *If $\underline{\delta}(i)$ is a regular sequence with density d then $A(i)$ is of the form $\{\lfloor k \cdot \frac{1}{d} + \beta \rfloor\}_{k=1}^{\infty}$ for some $-1/d \leq \beta < 0$ or of the form $\{\lceil k \cdot \frac{1}{d} + \beta \rceil\}_{k=1}^{\infty}$ for some $-1 - 1/d < \beta \leq -1$. Conversely, if the support is of this form then the corresponding zero-one sequence is a regular sequence.*

Corollary 2.7.2 *$\underline{\delta}(i)$ is a regular sequence with density d if and only if there exists a $\theta \in \mathbb{R}$ such that $\delta_t(i) = \lfloor (t+1)d + \theta \rfloor - \lfloor td + \theta \rfloor$ or $\delta_t(i) = \lceil (t+1)d + \theta \rceil - \lceil td + \theta \rceil$ for $t = 0, 1, \dots$*

We say the splitting sequence is balanced if the number of ones in any two N -blocks differ by at most one. The corresponding set $A(i)$ is then a balanced set. It is obvious that every regular sequence is balanced. The converse is not always true. For example the sequence $1(10)^\infty$ is balanced, but not regular. However, the converse is true for periodic sequences and for sequences with irrational density (see [68]). Moreover, it is shown that every balanced sequence is eventually regular, i.e. if the sequence is balanced then there exists a $t' \in \mathbb{Z}_{\geq 0}$ such that the sequence coincides with a regular sequence for every integer $t \geq t_0$.

In general the density of the splitting sequence $\underline{\delta}(i)$ is defined as $\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} \delta_t(i)$ if this limit exists. It is obvious that this is just p_i , the fraction of jobs that are routed to server i . Hence, Lemma 2.3.6 has the following corollary.

Corollary 2.7.3 *Let ψ be a policy applied to an (a_1, a_2, \dots, a_n) system with $W(\psi) < \infty$. Suppose that the splitting sequence $\underline{\delta}(i)$ is a regular sequence for some $i \in \{1, 2, \dots, n\}$. Then $\underline{\delta}(i)$ is a regular sequence with density $d = a_i$.*

In the sequel we call the splitting sequence $\underline{\delta}(i)$ regular or balanced for the applied policy ψ in an (a_1, a_2, \dots, a_n) system, if the sequence is regular or balanced, respectively, with the appropriate density $d = a_i$.

If for $i \in \{1, 2, \dots, n\}$ the splitting sequence $\underline{\delta}(i)$ is regular then the following theorem yields a good bound for the maximal waiting time of jobs routed to server i .

Theorem 2.7.4 *Let $i \in \{1, 2, \dots, n\}$. If the splitting sequence $\underline{\delta}(i)$ is regular then $v_i^t < 1$ for every $t \in A(i)$.*

Proof. Suppose that $v_i^t > 0$ for some $t \in A(i)$ and let $t^* = \max\{t' < t : v_i^{t'} = 0\}$. Then $v_i^{t^*} = 0$, $u_i^{t^*} = u_i^{t^*+1} = \dots = u_i^t$ and thus $v_i^t = \frac{1}{a_i} \cdot \sum_{t'=t^*}^{t-1} \delta_{t'}(i) - (t - t^*)$ by (2.60). Because $\underline{\delta}(i)$ is regular we have that $\sum_{t'=t^*}^t \delta_{t'}(i) \leq \lceil a_i \cdot (t - t^* + 1) \rceil$ and thus $\sum_{t'=t^*}^{t-1} \delta_{t'}(i) \leq \lceil a_i \cdot (t - t^* + 1) \rceil - 1 < a_i \cdot (t - t^* + 1)$. So, $v_i^t < t - t^* + 1 - (t - t^*) = 1$. \square

Corollary 2.7.5 *Let $i \in \{1, 2, \dots, n\}$. If the splitting sequence $\underline{\delta}(i)$ is regular then $v_i^t < \frac{1}{a_i}$ for every $t \in \mathbb{Z}_{\geq 0}$.*

The following lemma implies a bound on the total time a server is idle when the splitting sequence is regular.

Lemma 2.7.6 *Let $i \in \{1, 2, \dots, n\}$. If a policy ψ is applied such that the splitting sequence $\underline{\delta}(i)$ is regular then $d_i(\psi) \leq 1$.*

Proof. Since $\underline{\delta}(i)$ is regular we have for every $t \in \mathbb{Z}_{\geq 0}$ that $N_i^t \geq [t \cdot a_i]$. Hence by Lemma 2.4.2 we have that $d_i \leq 1$. \square

If for an (a_1, a_2, \dots, a_n) system there exists a policy ψ such that for every $i \in \{1, 2, \dots, n\}$ the splitting sequence $\underline{\delta}(i)$ is balanced then we say that the (a_1, a_2, \dots, a_n) system is balanceable. The policy ψ is then called a balanced policy. Because every balanced sequence is eventually a regular sequence we have for a balanceable system also that there exists a policy ϕ such that for every $i \in \{1, 2, \dots, n\}$ the splitting sequence $\underline{\delta}(i)$ is a regular sequence. Namely, one can take $\phi = \psi^s$ for some $s \in \mathbb{Z}_{\geq 0}$, s large enough. We define such a policy ϕ to be a regular policy. Observe that for a regular policy the set of regular sets $\{A(i)\}_{i=1}^n$ is an exact cover of $\mathbb{Z}_{\geq 0}$. Therefore it is also said that there exists an exact cover set for balanceable densities (a_1, a_2, \dots, a_n) .

If Q is a finite set, a so-called alphabet, and $I \subseteq \mathbb{Z}$ is a subinterval of \mathbb{Z} then a mapping $W : I \rightarrow Q$ is called a I -word or just word. Thus a word is a sequence of letters from the alphabet Q . A subword of W is the restriction of W to some interval $J \subseteq I$. For a finite (sub)word W we denote the length of W by $|W|$. Further for a letter $a \in Q$ the number of a 's in W is denoted by $|W|_a$ and the support set of letter a is denoted by $S_a := \{t \in I : W(t) = a\}$. For an I -word W we say that letter a has density d if for any non-decreasing sequence of finite subintervals I_i with union I we have that $\frac{|I_i \cap S_a|}{|I_i|}$ tends to d . A policy ψ for an (a_1, a_2, \dots, a_n) system can be identified with a $\mathbb{Z}_{\geq 0}$ -word or \mathbb{N} -word on the alphabet $\{1, 2, \dots, n\}$ by taking $W(t) = k_t(\psi)$ for $t = 0, 1, \dots$ or $W(t) = k_{t-1}(\psi)$ for $t = 1, 2, \dots$ respectively. In the literature the terms balanced word, regular word and Beatty word are also used. Each of these terms has the obvious meaning. For example a balanced word is a word such that for every letter $a \in Q$ it holds for any two finite subwords W and W' with $|W| = |W'|$ that $||W|_a - |W'|_a| \leq 1$. Thus a balanced policy corresponds to a balanced word (with the appropriate densities of the letters) and in the same way a regular policy corresponds to a regular (Beatty) word. In [60] the notion m -balanced word is used, which generalizes the notion of balanced word. We note that all these kinds of words can be constructed by the SG algorithm.

Recall that $I_t = \{0, 1, \dots, t-1\}$ and $A(i) := \{t \in \mathbb{Z}_{\geq 0} : \delta_t(i) = 1\}$. For $i = 1, 2, \dots, n$ and $t \in \mathbb{Z}_{\geq 0}$ we put $A^t(i) := A(i) \cap I_t$. Note that $|A^t(i)| = N_i^t$. If $\lim_{t \rightarrow \infty} \frac{\sum_{\tau \in A^t(i)} v_i^\tau}{N_i^t}$ exists for $i \in \{1, 2, \dots, n\}$ then we define the long-run average waiting time of jobs routed to server i as

$$z_i := \lim_{t \rightarrow \infty} \frac{\sum_{\tau \in A^t(i)} v_i^\tau}{N_i^t}.$$

Lemma 2.7.7 *If for an (a_1, a_2, \dots, a_n) system the applied policy ψ is such that $\lim_{t \rightarrow \infty} \frac{\sum_{\tau \in A^t(i)} v_i^\tau}{N_i^t}$ exists and is finite for $i = 1, 2, \dots, n$ then $W(\psi) = \sum_{i=1}^n a_i \cdot z_i$.*

If $a_i = \frac{p}{q}$ with $p, q \in \mathbb{N}$ then we say that the routing to server i is proportional periodic with period q if from every q consecutively arriving jobs exactly p jobs are routed to server i . If the routing to server i is proportional periodic with period q then we have by the argument used in the proof of Theorem 2.6.3 that server i has no idle intervals from some moment $t \in \mathbb{Z}_{\geq 0}$, $t < q$ on so that we can define t_i as in (2.47). We obtain the following lemma.

Lemma 2.7.8 *Let $i \in \{1, 2, \dots, n\}$. If the routing to server i is proportional periodic with period q then we have that $t_i \in \mathbb{Z}_{\geq 0}$, $t_i < q$, $v_i^{t_i} = 0$ and that for every $t \geq t_i$*

$$v_i^t = \frac{q}{p} \cdot \sum_{t'=t_i}^{t_i+t-1} \delta_{t'}(i) - (t - t_i) \text{ and } v_i^{t+q} = v_i^t. \quad (2.61)$$

Corollary 2.7.9 *If $a_i = \frac{p}{q}$ and the routing to server i is proportional periodic with period q then*

$$\lim_{t \rightarrow \infty} \frac{\sum_{\tau \in A^t(i)} v_i^\tau}{N_i^t} \text{ exists and } z_i = \lim_{t \rightarrow \infty} \frac{\sum_{\tau \in A^t(i)} v_i^\tau}{N_i^t} = \frac{1}{p} \cdot \sum_{t \in B^{t'}} v_i^t \quad (2.62)$$

for any $t' \geq t_i$, where $B^{t'} = \{t', t' + 1, \dots, t' + q - 1\} \cap A(i)$.

By combining Lemma 2.7.7 and Corollary 2.7.9 we obtain the following theorem.

Theorem 2.7.10 *If a $p.p$ policy ψ is applied to an (x_1, x_2, \dots, x_n) system then*

$$\lim_{t \rightarrow \infty} \frac{\sum_{\tau \in A^t(i)} v_i^\tau}{N_i^t}$$

exists and is finite for $i = 1, 2, \dots, n$ and $W(\psi) = \sum_{i=1}^n a_i \cdot z_i$.

We have the following theorem which gives an explicit formula for the long-run average waiting time of customers routed to some server i if the routing to that server i is regular.

Theorem 2.7.11 *If policy ψ is applied to an (a_1, a_2, \dots, a_n) system such that for some $i \in \{1, 2, \dots, n\}$ the splitting sequence $\underline{\delta}(i)$ is a regular sequence then*

$$\lim_{t \rightarrow \infty} \frac{\sum_{\tau \in A^t(i)} v_i^\tau}{N_i^t}$$

exists. Moreover, if $a_i = \frac{p}{q}$ with $\gcd(p, q) = 1$ then $z_i = \frac{1}{2} - \frac{1}{2p}$, and if a_i is irrational then $z_i = \frac{1}{2}$.

Proof. We first treat the case that a_i is rational, $a_i = \frac{p}{q}$ with $\gcd(p, q) = 1$. Since $[q \cdot a_i] = [q \cdot a_i] = p$ we have by definition that every q -block of the splitting sequence $\underline{\delta}(i)$ contains exactly p ones. Thus the routing to server i is proportional periodic with period q . So, by Corollary 2.7.9 we have that $\lim_{t \rightarrow \infty} \frac{\sum_{\tau \in A^t(i)} v_i^\tau}{N_i^t}$ exists and (2.62) holds. Let $f : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ be the function given by $f(t) = v_i^t$ for $t \in \mathbb{Z}_{\geq 0}$. We have for $t' \geq t_i$ that

$$z_i = \frac{1}{p} \cdot \sum_{t \in B^{t'}} v_i^t = \frac{1}{p} \cdot \sum_{t \in B^{t'}} f(t). \quad (2.63)$$

Let $C = \{0, \frac{1}{p}, \dots, \frac{p-1}{p}\}$. We claim that f maps $B^{t'}$ bijectively to C for $t' \geq t_i$. By this claim and (2.63) we have that

$$z_i = \sum_{k=0}^{p-1} \frac{1}{p} \cdot \frac{k}{p} = \frac{1}{p^2} \cdot \frac{1}{2} p \cdot (p-1) = \frac{1}{2} - \frac{1}{2p}.$$

Hence, for the rational case it suffices to prove the claim. We first prove that $f(t) \in C$ for every $t \in A(i)$. By (2.15), (2.16) and induction we have that $p \cdot v_i^t \in \mathbb{Z}_{\geq 0}$ for every $t \in \mathbb{Z}_{\geq 0}$. Further by Theorem 2.7.4 we have that $v_i^t < 1$ for every $t \in A(i)$. Hence $f(t) = v_i^t \in C$ for every $t \in A(i)$.

Since $|B^{t'}| = |C| = p$ it suffices to show that f is injective on $B^{t'}$. Suppose $t_i \leq t_a < t_b$ and $v_i^{t_a} = v_i^{t_b}$. Let $P = N_i^{t_b} - N_i^{t_a}$. By (2.61) we have that $\frac{q}{p} \cdot P = t_b - t_a$. Since $\gcd(p, q) = 1$ it follows that there exists a positive integer k such that $P = k \cdot p$ and $t_b - t_a = k \cdot q \geq q$. Because $|t_1 - t_0| < q$ for $t_0, t_1 \in B^{t'}$ it follows that f is injective on $B^{t'}$ and the claim is proved.

Now we consider the case that a_i is irrational. If the routing to server i is according to some arbitrary splitting sequence $b = (b_0, b_1, b_2, \dots)$ of zeros and ones then it is clear that v_i^t depends on b_0, b_1, \dots, b_{t-1} , but not on b_t, b_{t+1}, \dots . We define for $t = 1, 2, \dots$ the functions $f_t : \{0, 1\}^t \rightarrow \mathbb{R}_{\geq 0}$ by $f_t = f_t(b_0, b_1, \dots, b_{t-1}) := v_i^t$ if the routing to server i is according to the splitting sequence (b_0, b_1, \dots) . Further we put $f_0 = 0$. It is clear that these functions f_t have the following properties:

$$f_t(b_0, b_1, \dots, b_{t-1}) \geq f_{t-1}(b_1, b_2, \dots, b_{t-1}) \quad (2.64)$$

for every $t \in \mathbb{N}$ and

$$f_{t+k}(0, 0, \dots, 0, b_0, b_1, \dots, b_{t-1}) = f_t(b_0, b_1, \dots, b_{t-1}) \quad (2.65)$$

for every $k, t \in \mathbb{N}$.

Since $\underline{\delta}(i)$ is a regular sequence we have by Corollary 2.7.2 that there exists a $\theta \in \mathbb{R}$ such that $\delta_t(i) = \lfloor (t+1)a_i + \theta \rfloor - \lfloor t \cdot a_i + \theta \rfloor$ or $\delta_t(i) = \lceil (t+1)a_i + \theta \rceil - \lceil t \cdot a_i + \theta \rceil$ for $t = 0, 1, \dots$. We assume that for some $\theta \in \mathbb{R}$ we have that

$$\delta_t(i) = \lfloor (t+1)a_i + \theta \rfloor - \lfloor t \cdot a_i + \theta \rfloor \text{ for } t = 0, 1, \dots \quad (2.66)$$

Let $b_t(\theta) = \lfloor (t+1)a_i + \theta \rfloor - \lfloor t \cdot a_i + \theta \rfloor$ for $t \in \mathbb{Z}$. Then

$$b_t(\theta) = b_t(\theta + 1) \text{ for every } t \in \mathbb{Z}, \theta \in \mathbb{R} \quad (2.67)$$

and

$$b_{t-m}(\theta + m \cdot a_i) = b_t(\theta) \text{ for every } t, m \in \mathbb{Z}, \theta \in \mathbb{R}. \quad (2.68)$$

If the routing to server i is according to (2.66) then $v_i^t = f_t(b_0(\theta), b_1(\theta), \dots, b_{t-1}(\theta))$ for every $t \in \mathbb{Z}_{\geq 0}$. For $t \in \mathbb{Z}_{\geq 0}, \theta \in \mathbb{R}$ we define $g(\theta, t) := f_t(b_0(\theta), b_1(\theta), \dots, b_{t-1}(\theta))$ and

$g'(\theta, t) := f_t(b_{-t}(\theta), b_{-t+1}(\theta), \dots, b_{-1}(\theta))$. Then $g(\theta, t) = g'(\theta + t \cdot a_i, t)$ by (2.68). Further $g(\theta, t)$ and $g'(\theta, t)$ are both periodic in θ with period 1 in view of (2.67). Moreover, by (2.64), $g'(\theta, t)$ is non-decreasing in t . Hence $g'_\infty(\theta) := \lim_{t \rightarrow \infty} g'(\theta, t)$ exists and, by Corollary 2.7.5, $0 \leq g'_\infty(\theta) \leq \frac{1}{a_i}$ for every $\theta \in \mathbb{R}$. Further $g'_\infty(\theta)$ is also periodic in θ with period 1. If the routing to server i is according to (2.66) for some fixed $\theta = \theta_0$ then

$$\begin{aligned} \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} v_i^t &= \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} g(\theta_0, t) = \\ \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} g'(\theta_0 + t \cdot a_i, t) &= \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} g'_\infty(\theta_0 + t \cdot a_i). \end{aligned}$$

Since $g'_\infty(\theta)$ is periodic in θ with period 1 and a_i is irrational, we have by the ergodic theorem of Weyl and Von Neumann (see for example [64]) that

$$\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} g'_\infty(\theta_0 + t \cdot a_i) = \int_0^1 g'_\infty(\theta) d\theta.$$

So, if the routing to server i is according to (2.66) for some $\theta \in \mathbb{R}$, then $\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} v_i^t = \int_0^1 g'_\infty(\theta) d\theta$ and in particular $\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} v_i^t$ exists and is independent of θ . By Lemma 2.7.6 $\lim_{t \rightarrow \infty} u_i^t = \frac{d_i}{a_i} \leq \frac{1}{a_i}$. Further by Corollary 2.7.5, $v_i^t < \frac{1}{a_i}$ for every $t \in \mathbb{Z}_{\geq 0}$. Thus $v_i^t - u_i^t \in [-\frac{1}{a_i}, \frac{1}{a_i}]$ for every $t \in \mathbb{Z}_{\geq 0}$ and thus the condition

of Lemma 2.3.4 is satisfied for the considered server i . Since $\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} v_i^t$ exists, Lemma 2.3.4 implies that $z_i = \lim_{t \rightarrow \infty} \frac{\sum_{\tau \in A^t(i)} v_i^\tau}{N_i^t}$ exists and

$$z_i = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} v_i^t - \frac{1}{2a_i} + \frac{1}{2} = \int_0^1 g'_\infty(\theta) d\theta - \frac{1}{2a_i} + \frac{1}{2}. \quad (2.69)$$

Now we consider the case that, for some $\theta \in \mathbb{R}$,

$$\delta_t(i) = \lceil (t+1)a_i + \theta \rceil - \lceil t \cdot a_i + \theta \rceil \text{ for } t = 0, 1, \dots \quad (2.70)$$

Then it follows, analogously to the case that the routing to server i is according to (2.66), that $\lim_{t \rightarrow \infty} \frac{\sum_{\tau \in A^t(i)} v_i^\tau}{N_i^t}$ exists and that z_i is independent of θ .

Choose a $\theta \in \mathbb{R}$ such that $1, a_i, \theta$ are linearly independent over \mathbb{Z} . Then $\lceil t \cdot a_i + \theta \rceil = \lfloor t \cdot a_i + \theta \rfloor + 1$ for every $t \in \mathbb{Z}_{\geq 0}$ and thus

$$\lceil (t+1)a_i + \theta \rceil - \lceil t \cdot a_i + \theta \rceil = \lfloor (t+1)a_i + \theta \rfloor - \lfloor t \cdot a_i + \theta \rfloor \text{ for } t = 0, 1, \dots$$

Hence the routing to server i according to (2.66) is exactly the same as the routing to server i according to (2.70) for such θ . So, splitting sequences according to (2.70) have the same z_i as splitting sequences according to (2.66) and thus (2.69) also holds for splitting sequences according to (2.70). Thus $\lim_{t \rightarrow \infty} \frac{\sum_{\tau \in A^t(i)} v_i^\tau}{N_i^t}$ exists if the splitting sequence $\underline{\delta}(i)$ is regular and z_i is the same for every regular splitting sequence.

To complete the proof it suffices to show that $z_i = \frac{1}{2}$ for a particular regular splitting sequence with density a_i . Let the routing to server i be such that $N_i^t = \lceil t \cdot a_i \rceil$ for every $t \in \mathbb{Z}_{\geq 0}$. Then we have for the corresponding splitting sequence $\underline{\delta}(i)$ that

$$\delta_t(i) = N_i^{t+1} - N_i^t = \lceil (t+1) \cdot a_i \rceil - \lceil t \cdot a_i \rceil \text{ for } t = 0, 1, \dots$$

Hence $\underline{\delta}(i)$ is according to (2.70) with $\theta = 0$ and thus $\underline{\delta}(i)$ is indeed a regular sequence. By Lemma 2.4.2 we have for the applied policy ψ that $d_i(\psi) = 0$ and thus $u_i^t = 0$ for every $t \in \mathbb{Z}_{\geq 0}$. So, by (2.24),

$$v_i^t = \frac{N_i^t}{a_i} - t = \frac{\lceil t \cdot a_i \rceil - t \cdot a_i}{a_i} \text{ for every } t \in \mathbb{Z}_{\geq 0}. \quad (2.71)$$

Since a_i is irrational we can apply the ergodic theorem of Weyl and von Neumann again and from (2.71) we deduce that

$$\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} v_i^t = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} \frac{\lceil t \cdot a_i \rceil - t \cdot a_i}{a_i} = \int_0^1 \frac{x}{a_i} dx = \frac{1}{2a_i}. \quad (2.72)$$

Hence, by (2.69), $z_i = \frac{1}{2}$. □

The following Theorem gives a criterion for the attainability of the lower bounds of Corollary 2.6.13.

Theorem 2.7.12 *If for an (x_1, x_2, \dots, x_n) system the fractions $(\frac{x_1}{T}, \frac{x_2}{T}, \dots, \frac{x_n}{T})$ are balanceable then the lower bounds of Corollary 2.6.13 can be attained, i.e.*

$$\tilde{S} = \frac{n}{2} - \frac{C(x_1, x_2, \dots, x_n)}{2T} \text{ and } \tilde{W} = \frac{1}{2} - \frac{C(x_1, x_2, \dots, x_n)}{2T}, \quad (2.73)$$

where $C(x_1, x_2, \dots, x_n) = \sum_{i=1}^n \gcd(x_i, T)$. Moreover, every regular policy attains the lower bounds of Corollary 2.6.13 and is optimal.

Proof. Let ψ be a regular policy for the system. If for some $i \in \{1, 2, \dots, n\}$ we write $a_i = \frac{x_i}{T}$ as $\frac{p}{q}$ with $\gcd(p, q) = 1$ then $p = \frac{x_i}{\gcd(x_i, T)}$. Hence, by Theorem 2.7.11, $z_i = \frac{1}{2} - \frac{\gcd(x_i, T)}{2x_i}$ for $i = 1, 2, \dots, n$. So, by Lemma 2.7.7,

$$W(\psi) = \sum_{i=1}^n a_i \cdot z_i = \frac{1}{2} \cdot \sum_{i=1}^n a_i - \frac{\sum_{i=1}^n \gcd(x_i, T)}{2T} = \frac{1}{2} - \frac{C(x_1, x_2, \dots, x_n)}{2T}.$$

Thus ψ attains the lower bounds of Corollary 2.6.13 and ψ is an optimal policy. □

In [3] and [5] the following is shown for the more general case of stochastically independent stationary interarrival and service times. Let p be fixed. Then, if you want to route at least a fraction p of the arriving jobs (customers) to a certain server, then the best way to route to that server is according to a regular splitting sequence of density p . Substantial in the proof of this result is to show that the expected remaining workload at an arrival epoch is a multimodular function of the splitting sequence. For this coupling arguments are used. The fact that multimodular functions are minimized by regular sequences originates from [29], in which the optimality of the regular splitting sequence is proved in case of an exponential server queue.

In case of stochastic interarrival and service times it is generally assumed that the traffic density is less than one, since the queue is otherwise not stable. For that case it is shown in [4] that not only the regular splitting sequences are optimal, but also the slightly more general balanced sequences. Indeed, with probability one the queue becomes empty after some time. Hence only the “right tail” of the splitting sequence matters and thus a balanced sequence is just as good as a regular sequence since every balanced sequence is eventually regular.

If the traffic density equals 1 as in our case the situation is different. For example in case of $a_i = \frac{1}{2}$ and thus service time 2, compare for server i the regular splitting sequence $(10)^\infty$ with the balanced (but not regular) splitting sequence $1(10)^\infty$. For the regular sequence $v_i^t = 0$ if $t \in \mathbb{Z}_{\geq 0}$ is even and $v_i^t = 1$ if $t \in \mathbb{Z}_{\geq 0}$ is odd. Hence $\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} v_i^t = \frac{1}{2}$ and $z_i = 0$. For the irregular (but balanced) sequence $v_i^0 = 0$, $v_i^t = 2$ for $t \in \mathbb{N}$ even and $v_i^t = 1$ for $t \in \mathbb{N}$ odd. Hence $\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=0}^{\tau-1} v_i^t = \frac{3}{2}$ and $z_i = 1$. So, we see that for the irregular sequence the queue never becomes empty after $t = 0$ and that the irregular sequence is worse than the regular sequence of the same density.

If the fractions (a_1, a_2, \dots, a_n) are balanceable then a policy corresponding to a regular word with these fractions is optimal. A natural question is which sets of fractions (a_1, a_2, \dots, a_n) with $\sum_{i=1}^n a_i = 1$ are balanceable. This problem in general is studied in [4] and [51] and for several special cases it is studied in [26], [63], [68] and [67] and various other papers. See [67] for other references. We mention here some of the results.

In case $n = 2$ every pair of fractions $(p, 1 - p)$ is balanceable, since if a sequence of zeros and ones is balanced with density p then the complementary sequence is balanced with density $1 - p$. In the rational case Corollary 2.6.14 implies that the regular policies attain the lower bound. In the irrational case we have $\widetilde{W} = \frac{1}{2}$ by Corollary 2.5.13 and it follows from Theorem 2.7.11 that the regular policies are optimal indeed. So, for $n = 2$ the problem is completely solved and we know the optimal policies for our queueing system.

For $n > 2$ it is rather rare that a set of fractions is balanceable, especially if most of the fractions are distinct. If there are at most two distinct fractions then they are balanceable. If they are all equal to $\frac{1}{n}$ then the corresponding regular optimal policy is just the round robin policy which has average waiting time $W = 0$ for the corresponding $(1, 1, \dots, 1)$ system. Further if the fractions are balanceable and one of them is irrational then all have to be irrational and all the regular words in this case are classified in [26]. Conjecture 2.5.14, Theorem 2.3.2, Theorem 2.7.11 and Lemma 2.7.7 imply that if irrational fractions (a_1, a_2, \dots, a_n) are balanceable then $\dim_{\mathbb{Q}}(a_1\mathbb{Q} + a_2\mathbb{Q} + \dots + a_n\mathbb{Q}) = 2$. This is confirmed by the classification in [26]. Further it is known that if for $n > 2$ the fractions are distinct and balanceable then they all must be rational. In fact for every $n > 2$ there is exactly one set of distinct balanceable fractions known, viz. the fractions $\{\frac{2^{n-1}}{2^n-1}, \frac{2^{n-2}}{2^n-1}, \dots, \frac{1}{2^n-1}\}$ corresponding to a $(2^{n-1}, 2^{n-2}, \dots, 1)$ system. For example for $n = 4$ the fractions $(\frac{8}{15}, \frac{4}{15}, \frac{2}{15}, \frac{1}{15})$ are balanceable and $(abacabadabacaba)^\infty$ is a regular word on the alphabet $\{a, b, c, d\}$ with those fractions. Hence the correspond-

ing policy $(1, 2, 1, 3, 1, 2, 1, 4, 1, 2, 1, 3, 1, 2, 1)^\infty$ is optimal for an $(8, 4, 2, 1,)$ system. The following conjecture is known as Fraenkel's conjecture.

Conjecture 2.7.13 *A set of distinct fractions $\{a_1, a_2, \dots, a_n\}$ with $n \geq 3$ and $\sum_{i=1}^n a_i = 1$ is balanceable if and only if*

$$\{a_1, a_2, \dots, a_n\} = \left\{ \frac{2^{n-1}}{2^n - 1}, \frac{2^{n-2}}{2^n - 1}, \dots, \frac{1}{2^n - 1} \right\}.$$

For $n = 3$ the conjecture is proved in [51] and [66] and for $n = 4$ in [4]. In [68] and [67] the conjecture is proved for $n = 5$ and $n = 6$.

Finally we remark that a set of fractions (a_1, a_2, \dots, a_n) is balanceable if and only if there exists a solution d_1, d_2, \dots, d_n of (2.34) such that all the inequalities $\sum_{i=1}^n [t \cdot a_i - d_i] \leq t$ for $t = 0, 1, \dots$ hold with equality. If such a solution exists, a regular word with the given fractions can be obtained by applying the SG algorithm to the solution.

2.8 Algorithms to find good policies

For large systems it is hard to find an optimal policy by solving the mathematical programming problem. In this section we will consider some greedy algorithms that can be used to obtain rather good (but not necessarily optimal) policies. We apply these algorithms in the rational case, but they can also be applied in the general case. However, in the rational case the obtained policy ψ will be (ultimately) periodic and $W(\psi)$ can be determined, while in case of irrational a_i the performance $W(\psi)$ can only be estimated. Further in the rational case we can try to improve the obtained policy by iteratively using the SG algorithm, as we showed in Section 2.6.

The first greedy algorithm that we consider is based on the fact that to minimize W we have to minimize S . It selects at time $t_0 \in \mathbb{Z}_{\geq 0}$ a server k_{t_0} such that S^{t_0+1} is minimized. We call an algorithm that follows this rule an OSSM (One Step S Minimalization) algorithm. We have (see property w5 of the w_i^t for the rational case and the irrational case is similar) that

$$S^{t_0+1} - S^{t_0} = \sum_{k \neq k_{t_0}} \max(a_k - a_k \cdot v_k^{t_0}, 0). \quad (2.74)$$

So, defining

$$r_i^{t_0} = \max(a_i - a_i \cdot v_i^{t_0}, 0), \quad (2.75)$$

for $i \in \{1, 2, \dots, n\}$ it follows from (2.74) that at time t_0 an OSSM algorithm routes the job to a server l for which

$$r_l^{t_0} = \max_{i \in \{1, 2, \dots, n\}} r_i^{t_0}. \quad (2.76)$$

Applying this rule we still have to choose to which server the job is routed if $r_i^{t_0}$ is maximal for several servers i . First we consider the case that $r_1^{t_0} = r_2^{t_0} = \dots = r_n^{t_0} = 0$. Then $S^{t_0+1} = S^{t_0}$ no matter to which server the job is routed. In this case we let the algorithm minimize S^{t_0+2} instead of S^{t_0+1} . This can be done by redefining the variables $r_i^{t_0}$ as $r_i^{t_0} = \max(2a_i - a_i \cdot v_i^{t_0}, 0)$ and then choosing the server for which the new $r_i^{t_0}$ is maximal. If for the new $r_i^{t_0}$ it again holds that $r_1^{t_0} = r_2^{t_0} = \dots = r_n^{t_0} = 0$, then the $r_i^{t_0}$ can be redefined as $r_i^{t_0} = \max(3a_i - a_i \cdot v_i^{t_0})$, etc. It remains to consider the case that $r_i^{t_0}$ is maximal for the servers i_1, i_2, \dots, i_m with $m > 1$ and $r_{i_1}^{t_0} = r_{i_2}^{t_0} = \dots = r_{i_m}^{t_0} > 0$. For this case we introduce two variants of the OSSM algorithm, denoted by OSSM1 and OSSM2. Assume without loss of generality that $a_{i_1} \geq a_{i_2} \geq \dots \geq a_{i_m}$. Then OSSM 1 routes the job to server i_1 , the fastest server among them, while OSSM2 routes the job to server i_m , the slowest server among them.

Example. The OSSM2 algorithm yields the policy $(1, 2, 1, 3, 2, 1, 2, 1, 3)^\infty$ for the $(4, 3, 2)$ system that we used in Section 2.6 as example (see Table 1), while the OSSM1 algorithm yields the policy $(1, 2, 1, 3, 2, 1, 1, 2, 3)^\infty$ for this system.

We have the following result for OSSM algorithms.

Theorem 2.8.1 *In case $n = 2$ a policy ψ obtained by applying an OSSM algorithm is optimal.*

Proof. For both the rational and the irrational case it suffices to show that $S^t(\psi) < 1$ for every $t \in \mathbb{Z}_{\geq 0}$. In the rational case we have that $S(\psi) \cdot T \leq T - 1$. Hence ψ is optimal by Corollary 2.6.14. In the irrational case it implies that $S(\psi) \leq 1$ and thus ψ is optimal by Corollary 2.5.13. So if ψ is not an optimal strategy then there exists a $t_0 \in \mathbb{N}$ such that $S^{t_0} = a_1 \cdot v_1^{t_0} + a_2 \cdot v_2^{t_0} \geq 1$, but $S^{t_0-1} = a_1 \cdot v_1^{t_0-1} + a_2 \cdot v_2^{t_0-1} < 1$. We may assume that $k_{t_0-1} = 1$ without loss of generality. Then $S^{t_0} = S^{t_0-1} + r_2^{t_0-1}$ and $r_1^{t_0-1} \geq r_2^{t_0-1} > 0$. Thus $v_1^{t_0-1} < 1$ and $v_2^{t_0-1} < 1$. Hence $1 \leq a_1 \cdot v_1^{t_0} + a_2 \cdot v_2^{t_0} = a_1 \cdot v_1^{t_0-1} + 1 - a_1 < 1$, which is a contradiction. \square

The next greedy algorithm we consider follows the rule that the incoming job is routed to the server that is the first that can start with the processing of that job. So at time t the algorithm routes the job to a server i for which v_i^t (or $\frac{w_i^t}{x_i}$ in the

rational case) is minimal. If v_i^t is minimal for several servers i_1, i_2, \dots, i_m then the job will be routed to a server which is the fastest of those m servers. So, if those m servers are ordered such that $a_{i_1} \geq a_{i_2} \geq \dots \geq a_{i_m}$ then the job is routed to server i_1 . We will denote this as the SWT (Shortest Waiting Time) algorithm. The SWT algorithm gives an optimal policy in case $n = 2$ too.

Theorem 2.8.2 *In case $n = 2$ a policy ψ obtained by applying the SWT algorithm is optimal.*

Proof. It suffices again to show that $S^t(\psi) < 1$ for every $t \in \mathbb{Z}_{\geq 0}$. We prove this by induction on t . For $t = 0$ it is obvious. So suppose it holds for $t = t'$. We may assume that $k_{t'} = 1$ without loss of generality. Suppose that the induction hypothesis does not hold for $t = t' + 1$ and thus $S^{t'+1} \geq 1$. Then analogously to the proof of Theorem 2.8.1 it follows that $\max(a_2 - a_2 \cdot v_2^{t'}, 0) = r_2^{t'} > 0$ and thus $v_2^{t'} < 1$. Because the SWT algorithm is applied we have that $v_1^{t'} \leq v_2^{t'} < 1$. Hence $1 \leq a_1 \cdot v_1^{t'+1} + a_2 \cdot v_2^{t'+1} = a_1 \cdot v_1^{t'} + 1 - a_1 < 1$, which is a contradiction. \square

Remark. The SWT and the SG algorithm are closely related. If $h_i^t = a_i \cdot v_i^t$ for $i = 1, 2, \dots, n$ then the SWT algorithm and the SG algorithm route the job arriving at time t to the same server. By Lemma 2.5.3, $h_i^t - a_i \cdot v_i^t$ is monotonically non-increasing in t and $\lim_{t \rightarrow \infty} (h_i^t - a_i \cdot v_i^t) = 0$ for $i = 1, 2, \dots, n$. So, if ψ is a policy obtained by using the SWT algorithm and φ is the policy obtained by applying the SG algorithm to $d_1(\psi), d_2(\psi), \dots, d_n(\psi)$, then ψ and φ practically coincide after some time. In the rational case we can say more. Then we have for $i = 1, 2, \dots, n$ that $h_i^t = a_i \cdot v_i^t$ for every $t \geq t_i$, where $t_i \leq T - 1$. Hence $k_t(\varphi) = k_t(\psi)$ for every $t \in \mathbb{Z}_{\geq t'}$ for some $0 \leq t' \leq T - 1$. However, if φ' is the policy obtained by applying the SG algorithm to $d_1(\varphi), d_2(\varphi), \dots, d_n(\varphi)$, then there might not exist a $t \in \mathbb{Z}_{\geq 0}$ such that $\varphi' = \psi^t$.

The last algorithm we present is based on the fact that for some server i the splitting sequence $\underline{\delta}_i$ should be as regular as possible. We denote this algorithm as the GR (Greedy Regular) algorithm. We assume without loss of generality that the servers are ordered such that $a_1 \geq a_2 \geq \dots \geq a_n$. The GR algorithm is such that $\underline{\delta}(1)$, the splitting sequence of the routing to server 1, is a regular sequence. Namely, the GR algorithm constructs a policy ψ such that $N_1^t(\psi) = \lceil t \cdot a_1 \rceil$ for every $t \in \mathbb{Z}_{\geq 0}$. Then $\underline{\delta}(1)$ is a regular sequence. We construct $\underline{\delta}(2), \underline{\delta}(3), \dots$ inductively by imposing the following property. Let W be the word on the alphabet $\{1, 2, \dots, n\}$ corresponding to policy ψ and W_i be the word on the alphabet $\{i, i + 1, \dots, n\}$ obtained from W

by omitting the letters $1, 2, \dots, i - 1$. Then for $i = 2, 3, \dots, n$ the support of letter i in the word W_i is a regular set.

GR Algorithm. Policy ψ is inductively determined in the following way. Suppose $k_0(\psi), k_1(\psi), \dots, k_{t-1}(\psi)$ have been determined. Then

$$k_t(\psi) := \min \left\{ i \in \{1, 2, \dots, n\} : \frac{a_i}{\sum_{j=i}^n a_j} \cdot \left(1 + \sum_{j=i}^n N_j^t(\psi) \right) > N_i^t(\psi) \right\}.$$

Note that for $i = n$ it always holds that $\frac{a_i}{\sum_{j=i}^n a_j} \cdot (1 + \sum_{j=i}^n N_j^t(\psi)) > N_i^t(\psi)$ and thus $k_t(\psi)$ is well defined.

Remark. The GR algorithm is very similar to Algorithm 2 in [11] which is denoted as the CGRR (Conditional Generalized Round Robin) policy and which is such that

$$k_t(\psi) := \min \left\{ i \in \{1, 2, \dots, n\} : \frac{a_i}{\sum_{j=i}^n a_j} \cdot \sum_{j=i}^n N_j^t(\psi) \geq N_i^t(\psi) \right\}.$$

Both algorithms share the property that the routing to server 1 is regular and that for every other server i the routing is regular with respect to the servers with index number $\geq i$. These properties are proved in [11]. Further if a_1, a_2, \dots, a_n are rational then both algorithms construct a p.p policy. It is likely that both algorithms have the same performance on average. The main difference is that for the GR algorithm we have that $N_1^t = \lceil t \cdot a_1 \rceil$ for every $t \in \mathbb{Z}_{\geq 0}$, while for the CGRR policy $N_1^t = \lfloor t \cdot a_1 \rfloor$ or $N_1^t = \lceil t \cdot a_1 \rceil$.

In case $n = 2$ it is obvious that a policy obtained by the GR algorithm is regular. So, from the previous section we have the following theorem.

Theorem 2.8.3 *In case $n = 2$ a policy ψ obtained by applying the GR algorithm is optimal.*

Example. We apply the OSSM1, OSSM2, SWT and GR algorithm to an $(7, 5, 3, 2)$ system. For these algorithms we give the obtained policy ψ , $W(\psi)$, the policy ψ' we get after applying the SG algorithm iteratively to $d_1(\psi), d_2(\psi), d_3(\psi), d_4(\psi)$ and $W(\psi')$.

For OSSM1 we get $\psi = 1, 2, 1, 3, 2, 1, 4, 1, (2, 3, 1, 2, 1, 1, 2, 3, 4, 1, 2, 1, 3, 2, 1, 1, 4)^\infty$ and

$W(\psi) = \frac{1}{2}$. Further $\psi' = (1, 2, 1, 3, 2, 1, 4, 1, 2, 3, 1, 2, 1, 2, 1, 3, 4)^\infty$ and $W(\psi') = \frac{1}{2}$.
 For OSSM2 we get $\psi = 1, 2, 1, 3, 2, 1, 4, 1, 2, 3, 1, 2, 1, (4, 3, 1, 2, 1, 2, 1, 3, 1, 2, 4, 1, 2, 3, 1, 1, 2)^\infty$ and $W(\psi) = \frac{23}{34}$.
 Further $\psi' = (1, 2, 1, 3, 1, 2, 4, 1, 3, 2, 1, 1, 2, 4, 3, 1, 2)^\infty$ and $W(\psi') = \frac{21}{34}$. For the
 SWT algorithm we get $\psi = 1, 2, 3, 1, 4, (2, 1, 3, 2, 1, 1, 2, 4, 3, 1, 2, 1, 2, 1, 3, 4, 1)^\infty$ and
 $W(\psi) = \frac{19}{34}$. Further $\psi' = (2, 1, 3, 4, 1, 2, 1, 3, 2, 1, 1, 2, 4, 3, 1, 2, 1)^\infty = \psi^T$ and
 $W(\psi') = \frac{1}{2}$. Finally for the GR algorithm we have that
 $\psi = (1, 2, 1, 3, 1, 2, 3, 1, 2, 1, 4, 2, 1, 3, 1, 2, 4)^\infty$ with $W(\psi) = \frac{23}{34}$.
 Further $\psi' = (1, 2, 1, 3, 1, 2, 1, 2, 3, 1, 4, 2, 1, 3, 1, 2, 4)^\infty$ with $W(\psi') = \frac{23}{34}$.

From this example we see that even in the rational case a policy ψ obtained by applying one of the OSSM algorithms or the SWT algorithm is in general not a p.p policy. However, ψ is eventually a p.p policy since the v_i^t get in a loop. In fact ψ^T is always a p.p policy and therefore it suffices to determine $k_t(\psi)$ for $t = 0, 1, \dots, 2T - 1$ to obtain ψ and $W(\psi)$. It is also possible in the rational case to modify the algorithms such that always a p.p policy is obtained. This can be done by imposing that at moment $t \leq T - 1$ the incoming job should not be routed to server i if $N_i^t = x_i$. In this way we get for the obtained policy ψ that $N_i^T(\psi) = x_i$ for every $i \in \{1, 2, \dots, n\}$. Next we determine ψ by $\psi := (k_0(\psi), k_1(\psi), \dots, k_{T-1}(\psi))^\infty$. Then ψ is a p.p policy and the algorithm can be stopped after $k_0(\psi), k_1(\psi), \dots, k_{T-1}(\psi)$ have been determined. Another advantage is that the $d_i(\psi)$ and $s_i(\psi)$ can be determined by Theorem 2.6.3 if ψ is a p.p policy. Further the performance of the obtained policy does not change by this modification since the $d_i(\psi)$ will not change.

For several systems with rational a_i we have calculated the long-run average waiting time W for policies obtained by applying the several algorithms. We denote with $W(OS1)$, $W(OS2)$, $W(SWT)$, $W(GR)$ the long-run average waiting time for the policy obtained by applying the *OSSM1*, *OSSM2*, *SWT*, *GR* algorithm respectively. In fact we give $T \cdot W(OS1)$ where T is the period of the system. Next to $T \cdot W(OS1)$ we give between parentheses T times the average waiting time of the policy we get after applying the SG algorithm iteratively on the solution of (2.51) obtained by applying the OSSM1 algorithm. For the other algorithms we do the same. Further for some of those systems we have also determined the optimal long-run average waiting time \widetilde{W} in some way. For example by solving (2.51) or (2.53) or using the fact that the fractions a_i are balanceable, etc. However, for the “larger” systems we could not do this. Then we denote with $\geq \dots$ the lower bound following from Corollary 2.6.13 for the corresponding system.

The results are given in Table 2.

Table 2

(x_1, x_2, \dots, x_n)	T	$T \cdot \bar{W}$	$T \cdot W(OS1)$	$T \cdot W(OS2)$	$T \cdot W(SWT)$	$T \cdot W(GR)$
(105,84,54)	243	189	201 (189)	189 (189)	189 (189)	207 (195)
(1000,17,5)	1022	954	963 (954)	963 (954)	968 (954)	966 (966)
(10000,66,1)	10067	9850	9851 (9850)	9851 (9850)	9885 (9850)	9899 (9899)
(10000,2,1)	10003	6666	6667 (6666)	6666 (6666)	9996 (9994)	6667 (6667)
(41,41,41,33)	156	75	80 (75)	78 (75)	78 (75)	108 (108)
(59,59,30,25)	173	$143\frac{1}{2}$	$154\frac{1}{2}$ ($143\frac{1}{2}$)	$154\frac{1}{2}$ ($143\frac{1}{2}$)	$143\frac{1}{2}$ ($143\frac{1}{2}$)	$179\frac{1}{2}$ ($179\frac{1}{2}$)
(118,95,69,35)	317	$351\frac{1}{2}$	$351\frac{1}{2}$ ($351\frac{1}{2}$)	$351\frac{1}{2}$ ($351\frac{1}{2}$)	$361\frac{1}{2}$ ($361\frac{1}{2}$)	$391\frac{1}{2}$ ($384\frac{1}{2}$)
(1000,13,1,1)	1015	$\geq 503\frac{1}{2}$	$1238\frac{1}{2}$ ($1170\frac{1}{2}$)	$1238\frac{1}{2}$ ($1170\frac{1}{2}$)	$1453\frac{1}{2}$ ($1438\frac{1}{2}$)	$903\frac{1}{2}$ ($903\frac{1}{2}$)
(553,447, 301,259)	1560	≥ 777	2022 (1983)	2022 (1983)	2043 (1989)	2094 (2018)
(1000,3,2,1)	1006	≥ 500	667 (667)	834 (834)	1487 (1480)	668 (667)
(5,4,3,2,1)	15	5	7 (7)	6 (6)	6 (6)	7 (7)
(13,12,9,5,2)	41	31	36 (36)	36 (36)	37 (35)	48 (42)
(1000,513, 102,14,3)	1632	≥ 757	2480 (2431)	2480 (2431)	2585 (2440)	2848 (2677)
(10000,3,3,1,1)	10008	≥ 4996	7498 (7498)	7498 (7498)	19976 (19964)	7498 (7498)
(7,6,5,4,3,2,1)	28	12	21 (18)	15 (15)	15 (15)	18 (17)
(4,4,4,3,3,3,3)	24	0	12 (12)	12 (12)	12 (12)	12 (12)
(33,33,33,33, 33,26,26,26,26)	269	130	244 (162)	244 (154)	262 (130)	330 (330)
(129,101,98, 87,82,61,52, 38,29,17)	694	≥ 340	1340 (1274)	1406 (1302)	1332 (1263)	1789 (1402)

Remarks on these results. First of all it follows that none of these algorithms always yields an optimal policy. It is also clear that for all these algorithms there exist systems for which this particular algorithm yields a better policy than the other algorithms. Further it seems that the various algorithms always yield a policy ψ with $W(\psi) \leq \frac{n-1}{2}$ and thus the obtained policies always satisfy the upper bound of Corollary 2.4.7. By Corollary 2.5.13 this would imply that these algorithms yield optimal policies in case that the service capacities a_1, a_2, \dots, a_n are linearly independent over \mathbb{Z} . Often the GR algorithm yields the worse policy of these algorithms. An explanation is that if the GR algorithm is applied then $k_t(\psi)$ only depends on $N_1^t(\psi)$, $N_2^t(\psi)$, \dots , $N_n^t(\psi)$, but not on the order of the routing of the foregoing jobs. So, the GR algorithm uses less information in choosing the server to which the arriving job is routed than the other algorithms. However, the GR algorithm performances rather good if a_1 is very large compared to the other fractions a_2, a_3, \dots, a_n . The reason for this is that the routing to server 1 is regular and thus optimal and most of the arriving jobs are routed to server 1 in this case. So, the routing to the other servers has not much influence on W and thus the GR algorithm yields a rather good policy in such cases. We note that on the contrary the SWT algorithm yields inferior policies if a_1 is very large compared to the other fractions (see for example the (10000, 3, 3, 1, 1) system), but rather good results if the a_i 's have similar size (sometimes even the best result). This can be explained by the fact that for systems

with large a_1 in the beginning the SWT algorithm routes too many arriving jobs to other servers than the fast server 1. Therefore the splitting sequence of server 1, $\underline{\delta}(1)$ becomes rather irregular if the SWT algorithm is applied to such systems. It seems even that in general it holds that $W(SWT) \uparrow \frac{n-1}{2}$ if $a_1 \uparrow 1$. We think that the OSSM algorithms on average have the best performance of the considered algorithms. In most cases there is not much difference in the performance of the OSSM1 and OSSM2 algorithm. However, for some systems (see the $(1000, 3, 2, 1)$ system and the same difference occurs for all the $(x, 3, 2, 1)$ systems with x large) a remarkable difference in performance occurs.

We notice that applying the SG algorithm iteratively to a solution obtained by one of these algorithm often gives a small improvement. Sometimes the improvement is rather large like in the $(33, 33, 33, 33, 33, 26, 26, 26, 26)$ system for the OSSM algorithms and especially for the SWT algorithm. Moreover it occurs regularly that from a nearly optimal policy an optimal policy can be obtained by using the SG algorithm in this way.

We now discuss an argument for the fact that in general these greedy algorithms and similar greedy algorithms do not yield optimal policies. For this we consider the $(4, 4, 4, 3, 3, 3, 3)$ system with the corresponding balanceable set of fractions $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$. In table 2 we see that for this system none of the considered algorithms yields an optimal policy. The reason for this is that a policy obtained by applying such an algorithm always starts with 1, 2, 3, i.e the first three arriving jobs are routed to the three fast servers. In fact it is clear that policies obtained by applying an OSSM algorithm or SWT algorithm have the SFQ (Shorter Faster Queue) property, i.e an arriving job is never routed to some server i if there exist a faster server j (thus $a_i < a_j$) for which the remaining workload at that moment is smaller or equal than for server i . However, optimal policies do not always have this property. For example for this system a policy is optimal if and only if it corresponds to a regular word on the alphabet $\{1, 2, 3, 4, 5, 6, 7\}$ with the above densities for the letters. In [51] the regular words with these densities are classified and it follows that there exist no regular words with these densities which have 1, 2, 3 as subword. Hence policies starting with 1, 2, 3 and policies that have the SFQ property are not optimal for this system. The most obvious regular word and corresponding optimal policy for these fractions is $(1, 4, 2, 5, 3, 6, 1, 7, 2, 4, 3, 5, 1, 6, 2, 7, 3, 4, 1, 5, 2, 6, 3, 7)^\infty$, i.e the policy that alternately routes the arriving jobs to the next fast and the next slow server. So, if we put a one for a fast server and a zero for a slow server then this policy corresponds to the regular sequence $(10)^\infty$ which has density $\frac{1}{2} = 3 \cdot \frac{1}{6}$ which is the fraction of jobs that has to be routed to one of the fast servers. This construction to obtain regular words can be used in general for systems with only

two distinct fractions. It is clear that corresponding optimal policies have in general not the SFQ property. Also for systems with fractions that are not balanceable the optimal policies do not always have the SFQ property.

For systems in which some of the fractions are equal to each other it seems a good device to first apply the various algorithms to the system obtained by taking these fractions together. From the best obtained policy for the altered system a good policy for the original system can be obtained in the same way we just obtained a regular policy for the $(4, 4, 4, 3, 3, 3, 3)$ system. For example for the $(10000, 3, 3, 1, 1)$ system this means that first a good (optimal) policy ψ for the $(5000, 3, 1)$ system is obtained and then from policy ψ follows a policy ψ' for the original $(10000, 3, 3, 1, 1)$ system.

We conclude that the considered algorithms and similar greedy algorithms yield in general policies that are not optimal. However, by applying the several algorithms and possibly using the SG algorithms iteratively we can obtain good policies quickly by taking the best one obtained. It seems that from the algorithms we considered the OSSM algorithms give in most cases the best result.

Chapter 3

Analysis of the performance of periodic routing sequences

3.1 Introduction

Load balancing in distributed multiprocessors and in communication networks is an important tool to improve their performance in mean delay of the jobs (for overviews of the literature see [16] and [31]). In many systems the load balancing protocol is an open-loop control, in which case it does not depend on the information in the system, as queue-sizes etc., but only depends on the system parameters. We focus on the performance under this type of control, which usually is called static routing control. In the publications ([43],[69], [28] and [42]) the static routing problem in a network is reduced to finding the best utilization of each service centre. Given this utilization of a particular service centre the problem then becomes to find the static routing to parallel queues with minimal average expected waiting time. For this routing problem probabilistic (also called Bernoulli) and deterministic routing policies have been analyzed. In case of probabilistic routing to homogeneous servers, the optimality of equal routing probabilities is proved in [19] and [44]. In general deterministic (generalized) round robin policies are superior to probabilistic routing for homogeneous (heterogeneous) servers. For homogeneous servers, the optimality of the round robin routing is proved in [47]. The problem of constructing the optimal generalized round robin policy for heterogeneous servers is an unsolved problem. In the papers [36] and [2] the optimal routing problem for exponential service times is

transformed to a Markov decision process (see [1] for an overview of the applications of these processes in telecommunication).

In a recent approach the optimization procedure is split into two problems. First the fractions of jobs routed to the servers are computed via a best probabilistic routing. For networks these fractions are computed via the optimal rates (see [43],[69], [28] and [42]). Approximation methods have been studied for the single server with parallel queues (see [20]). Given the fractions of jobs that have to be routed to the queues, the problem then becomes to construct the best routing pattern. In the seminal paper of Hajek ([29]) it is shown for one queue with exponential service times that the regular splitting sequence with the density d is optimal for any initial phase ϕ , if a fraction d of the arriving customers have to be routed to that queue. A regular zero-one valued splitting sequence with rational density d and initial phase $\phi \in \mathbb{R}$ is given by:

$$\{b_k^d(\phi)\} = \lfloor (k+1)d + \phi \rfloor - \lfloor kd + \phi \rfloor,$$

$k = 1, 2, \dots$. The corresponding routing rule assigns the k -th arrival to the queue if $b_k^d(\phi) = 1$. Since the zeros of a regular splitting sequence with density d form a regular splitting sequence with density $(1-d)$, it follows that the regular splitting sequence with density d gives an optimal routing sequence to two queues with exponential service times for given fractions d and $(1-d)$. It is proved in [3] and [5] that a regular splitting policy is optimal in the class of all static routing policies for two queues with general stationary interarrival and service times. An algorithm for computing the optimal routing policy for two queues with deterministic interarrival and service times has been derived in [23].

The problem of finding the best optimal routing pattern with given splitting fractions is a complex optimization problem for $N \geq 3$ queues (see [14]). There are only few sets of fractions for which the regular splitting sequences generate an exact covering of the positive integers (see [4] and [68]), and in these cases the optimal routing pattern is known. Regular sequences, which are also called Beatty, Sturmian, or bracket sequences have applications in many branches of sciences (see [67]).

For routing to $N \geq 3$ parallel queues many algorithms for finding good allocation patterns have been derived. In [57] the golden ratio policy is used for the optimal routing of jobs, in [41] for the dual problem of the static control of a multiple-access channel, and in [21] for optimal robot scheduling of web search engines. In these papers the asymptotic optimality of the golden ratio policy as the number of queues tends to infinity is studied. Other algorithms for generalized round robin routing sequences have been given in [11] and [14]. Methods to obtain good allocation

patterns can be found in [20] and [17]. Although simulations make it evident that the proposed algorithms behave well, bounds on their performance were only derived for exponential queues with no buffer space.

In this chapter we concentrate on obtaining bounds for the average expected waiting time of the jobs for a given routing pattern to parallel queues with unlimited buffer capacity and general sequences of interarrival and service times. Our analysis is mainly combinatorial. For a given pattern allocation we introduce its unbalance. All regular sequences have unbalance equal to zero, and for a splitting sequence to one queue the unbalance is roughly speaking its 'distance' to the regular sequence with the same density. For the routing pattern it is the sum of the unbalances of the splitting sequences to the queues. In [3] and [5] it is shown for routing policy ψ which routes a fraction d_i to queue i for $i = 1, 2, \dots, N$ that the average expected waiting time $\overline{W}(\psi)$ is bounded from below by \tilde{R} , where $\tilde{R} = \sum_{i=1}^N d_i \cdot \overline{W}^i(\omega(d_i))$, with $\overline{W}^i(\omega(d_i))$ the average expected waiting time for jobs routed to queue i when the routing to queue i is according to a regular sequence with density d_i . The upper bound for $\overline{W}(\psi)$ which we derive in this chapter is $\tilde{R} + \delta \cdot \overline{O}(U)$, where δ is the mean interarrival time, U is the routing sequence corresponding to policy ψ and $\overline{O}(U)$ is the unbalance of U . Hence,

$$\tilde{R} \leq \overline{W}(\psi) \leq \tilde{R} + \delta \cdot \overline{O}(U).$$

The unbalance $\overline{O}(U)$ is a combinatorial notion and it does not depend on the service or the interarrival distribution. Hence the difference between the bounds is insensitive, i.e. they are valid for any stationary sequence of interarrival and service times. However, the lower bound \tilde{R} does depend on the distribution of the interarrival and service times. The upper bound is tight for the heavy traffic situation in which the interarrival and service times are deterministic and the traffic intensity is 1 for each of the queues. So the upper bound is accurate for high traffic, whereas $\overline{W}(\psi)$ will be close to the lower bound for low traffic.

Finding the routing sequence U for given densities d_1, d_2, \dots, d_N such that the unbalance $\overline{O}(U)$ is minimal can be transformed to a combinatorial optimization problem. Greedy algorithms for obtaining good routing policies have been given in [11], [14], [60] and Chapter 2. In this chapter we consider greedy algorithms which in a special case generate billiard sequences. Suppose a billiard consisting of the N -dimensional cube $[0, 1]^N$ (see [12], [15]), with sides numbered $1, 2, \dots, N$ such that opposite sides have the same number. Give a billiard ball initial position (x_1, x_2, \dots, x_N) and initial velocity vector $(-d_1, -d_2, \dots, -d_N)$, and take the sequence $U_b = (n_1, n_2, \dots)$ indicating the order in which the sides are hit by the ball, breaking ties in an unique

way and let ψ_b be the corresponding routing policy. Then ψ_b , which routes the k -th arrival to queue i if $n_k = i$, is a routing policy with densities d_i for $i = 1, 2, \dots, N$. We show by taking a special initial position that the unbalance $\overline{O}(U_b)$ of this routing sequence is bounded by $\frac{N}{2} - 1$. Hence,

$$\tilde{R} \leq \overline{W}(\psi_b) \leq \tilde{R} + \delta\left(\frac{N}{2} - 1\right).$$

For $N = 2$ an optimal routing is constructed, and for $N \geq 3$ this billiard sequence is easily computed and has the above upper bound on its performance. In this chapter we consider only rational densities $d_i = \frac{p_i}{q_i}$ with $p_i, q_i \in \mathbb{N}$, $\gcd(p_i, q_i) = 1$. In this case the billiard sequence is periodic with period $T = \text{lcm}(q_1, q_2, \dots, q_N)$, and in fact it gives a routing pattern with period T .

The chapter is structured as follows. In Section 3.2 we start with an introduction of regular sequences together with properties which we need for our analysis. We define a partial order in the set $\mathcal{P}(T, k)$ of all sequences of zeros and ones of length T containing exactly k ones. The conjugacy class of the regular sequences in $\mathcal{P}(T, k)$ is shown to be the minimal element in an induced order. We define the notion unbalance for periodic words of zeros and ones by using the period cycle of the word, which is an element of $\mathcal{P}(T, k)$ if there are k ones in the cycle of length T . In fact we will define a primal unbalance and a dual unbalance, the dual unbalance can be used for the dual model of optimal routing of the server in polling models. The unbalance is a measure for the irregularity of the word. The unbalance has a graphical representation (see Figure 3.1), it is the area between the graph of its representation as element of $\mathcal{P}(T, k)$ and that of the regular one. This leads to another partial order which we called the graph order. The conjugacy class of the regular sequences in $\mathcal{P}(T, k)$ is shown to be the minimal element in this graph order. The results of Section 3.2 are purely combinatorial and have possibly other applications than in this routing problem.

In Section 3.3 we consider the routing to one queue and we consider all static deterministic policies which route a fraction p of all jobs to the queue. We restrict our analysis to periodic routing policies and thus to routing sequences with rational fraction p . We use the unbalance to bound the difference in average waiting times. We do this by a sample path comparison for fixed sequences of interarrival and service times. By using results on renovation and ergodicity we then obtain the above mentioned bounds for one queue. In Section 3.4 we extend the bounds to N parallel queues.

In Section 3.5 we extend the combinatorial analysis of the unbalance and derive its relation with the discrepancy function defined in Section 3.2. We introduce

greedy algorithms and the billiard sequence which is obtained by a special greedy algorithm. It is shown that there is a billiard sequence with minimal unbalance for given densities. For rational routing densities $d_i = \frac{p_i}{q_i}$ with $p_i, q_i \in \mathbb{N}$, $\gcd(p_i, q_i) = 1$ we show that the unbalance of the billiard routing pattern is bounded by $\frac{N}{2} - 1$ if an initial position (x_1, \dots, x_N) is chosen with $x_k = 0$ for k such that $q_k = \max_{i \in \{1, 2, \dots, N\}} q_i$ and $x_j = 1 - \frac{1}{q_j}$ for $j \neq k$. Many examples illustrate the notions and results in this chapter.

Notations. If $a \in \mathbb{Z}$ and $b \in \mathbb{N}$ then $a \pmod{b}$ denotes the integer $c \in \{0, 1, \dots, b-1\}$ for which there exists some $d \in \mathbb{Z}$ such that $c + d \cdot b = a$.

For any random variable X we denote by $\mathbb{E}X$ the expectation of X .

3.2 Comparing routing sequences

In this section we will compare routing sequences in a queueing system. For a particular queue the routing sequence will be a sequence of ones and zeros, where a one means that the corresponding arriving customer is routed to that queue and a zero means that the customer is not routed to that queue. We also speak of words instead of sequences. We first give some definitions.

If u and v are finite sequences, $u = (a_1, a_2, \dots, a_n)$ and $v = (b_1, b_2, \dots, b_m)$, then $uv = (a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_m)$ denotes the concatenation of the two sequences. For $k \in \mathbb{N}$ we denote by u^k the sequence $uu \dots u$ in which the number of u 's is k and we say that u^k is a power of u . A finite sequence is primitive if it is not a power of some other sequence. The following lemma is part of Proposition 1.3.1 of [48].

Lemma 3.2.1 *For every non-empty finite sequence u there exists an unique primitive sequence w such that $u = w^n$ for some $n \in \mathbb{N}$.*

Two finite sequences u, u' are said to be conjugate if there exist sequences v and w (v or w may be empty) such that $u = vw$ and $u' = wv$. This is an equivalence relation since u is conjugate to u' if and only if u' is a cyclic permutation of u . If u and u' are conjugate then we write $u \sim u'$ and the conjugacy class of all the cyclic permutations of a finite sequence u is denoted by \tilde{u} .

Lemma 3.2.2 *Let u be a finite sequence, $u = w^n$ with w primitive. Then $|\tilde{u}|$, the number of distinct conjugates of u , equals the length of sequence w .*

Proof. By Proposition 1.3.3 in [48] the number of distinct conjugates of u equals the number of distinct conjugates of w and by Proposition 1.3.2 in [48] the number of distinct conjugates of w equals the length of w . \square

If \mathbf{A} is a finite set (a so-called alphabet) and $I \subseteq \mathbb{Z}$ is an interval in \mathbb{Z} then a mapping $\Lambda : I \rightarrow \mathbf{A}$ is called an I -word or just word. Thus a word is a sequence of letters from the alphabet \mathbf{A} . A word is (in)finite if I is in(finite). A subword of Λ is the restriction of Λ to some interval $J \subseteq I$. For a finite (sub)word Λ we denote the length of Λ by $|\Lambda|$. Further for a letter $a \in A$ the number of a 's in Λ is denoted by $|\Lambda|_a$. For an I -word Λ the support of letter a is the set $S_a = \{t \in I : \Lambda(t) = a\}$. For an I -word Λ we have that letter a has density d if for every non-decreasing sequence of finite subintervals I_i with union I we have that $\frac{|S_a \cap I_i|}{|I_i|}$ tends to d . An (infinite) I -word Λ is periodic if for some $q \in \mathbb{N}$ it holds that $\Lambda(t) = \Lambda(t + q)$ for every t for which $t, t + q \in I$. Then q is said to be a period of the word and any subword of length q is said to be a period cycle of the word. The period of Λ is defined as the minimum of all the periods of Λ . If q is the period of word Λ then a subword of length q is said to be a primitive period cycle of the word. It is obvious that the following holds.

Lemma 3.2.3 *If k is a period of some word Λ then for every $m \in \mathbb{N}$ we have that $k \cdot m$ is a period of Λ .*

Note that a sequence of zeros and ones corresponds to a word on the alphabet $\{0, 1\}$. Conversely, to every letter in a word corresponds a sequence of ones and zeros via the support of the letter. In the sequel we identify a sequence (of integers) with a word and sometimes it is useful to choose an appropriate domain I for the word the sequence is identified with. So, we consider sequences of integers and words as the same thing and therefore we have the same notions and operations for words as for sequences. A sequence of zeros and ones is said to have density d if the letter 1 has density d in the corresponding word.

Definition 3.2.4 *Let u be an infinite sequence of zeros and ones and Λ be the corresponding word defined on an appropriate infinite interval I . We say that u is regular with density d if for every $n \in \mathbb{N}$ we have for every subword Λ' of Λ of length n that $|\Lambda'|_1$ equals $\lfloor nd \rfloor$ or $\lceil nd \rceil$. In that case we also say that the corresponding set S_1 is a regular set.*

The following theorem follows for $I = \mathbb{N}$ from the classification of balanced sequences in [54] and the fact that regular sequences are particular balanced sequences (see [68]).

Theorem 3.2.5 *Let u and Λ be as in Definition 3.2.4. If u is regular with density d then the support set S_1 satisfies one of the following cases.*

- (irrational case) d is irrational and there exists some $\varphi \in \mathbb{R}$ such that

$$S_1 = I \cap \{[k \cdot \frac{1}{d} + \varphi]\}_{k \in \mathbb{Z}} \text{ or } S_1 = I \cap \{\lceil k \cdot \frac{1}{d} + \varphi \rceil\}_{k \in \mathbb{Z}}.$$

- (rational case) d is rational and there exists some $\varphi \in \mathbb{Q}$ such that

$$S_1 = I \cap \{[k \cdot \frac{1}{d} + \varphi]\}_{k \in \mathbb{Z}}.$$

Theorem 3.2.5 has the following corollaries.

Corollary 3.2.6 *Let u and Λ be as above. If u is regular with irrational density d then there exists a $\theta \in \mathbb{R}$ such that $\Lambda(t) = \lfloor (t+1)d + \theta \rfloor - \lfloor td + \theta \rfloor$ or $\Lambda(t) = \lceil (t+1)d + \theta \rceil - \lceil td + \theta \rceil$ for every $t \in I$. If u is regular with rational density d then there exists a $\theta \in \mathbb{Q}$ such that $\Lambda(t) = \lfloor (t+1)d + \theta \rfloor - \lfloor td + \theta \rfloor$ for every $t \in I$.*

Corollary 3.2.7 *If u is regular with rational density $d = \frac{p}{q}$ with $p, q \in \mathbb{N}$, $\gcd(p, q) = 1$ then u is periodic with period q .*

We define a finite sequence of zeros and ones to be regular if it is a period cycle of some infinite regular periodic sequence.

In [3] and [5] it is shown that a regular sequence is optimal when one considers the routing to a single queue and a certain fraction of all the arriving customers has to be routed to that queue. They also show that routing according to some regular sequence is optimal when one has to route to two parallel queues. For $N \geq 3$ queues one often wants to route to the queues according to some (optimized) fractions, where the sum of these fractions is 1. However, for such prescribed fractions it is not always possible to route in such a way that the routing sequence for every queue is regular. Therefore we want to compare irregular sequences and for finite and periodic sequences (words) we introduce the notion of unbalance. In fact we will define a primal unbalance and a dual unbalance. The unbalance measures the irregularity of the sequence and has some useful properties. One of the obvious properties will be that a sequence has unbalance 0 if and only if it is a periodic

regular sequence or it is a period cycle of a periodic regular sequence. We first define the unbalance for finite sequences.

Notation. Let $\mathcal{P}(T, k)$ be all sequences of zeros and ones of length T containing exactly k ones and let $\mathcal{R}(T, k) \subseteq \mathcal{P}(T, k)$ be the subset of the regular sequences in $\mathcal{P}(T, k)$. Let $u \in \mathcal{P}(T, k)$ and Λ the corresponding word on the alphabet $\{0, 1\}$. We assume without loss of generality that $I = \{1, 2, \dots, T\}$ and we denote u as well as Λ by (u_1, u_2, \dots, u_T) , where $u_t = \Lambda(t)$ for $t = 1, 2, \dots, T$. Further we say for $l = 0, 1, \dots, T - 1$ that $(u_{1+l}, u_{2+l}, \dots, u_T, u_1, u_2, \dots, u_l)$ is the l -th cyclic permutation of u .

From the rational case of Theorem 3.2.5 the following lemma follows.

Lemma 3.2.8 *Let $u \in \mathcal{R}(T, k)$. Then $S_1 = \{[i \cdot \frac{T}{k} + \varphi]\}_{i=1}^k$ for some $\varphi \in \{1 - \frac{T}{k}, 1 - \frac{T - \gcd(T, k)}{k}, \dots, 1 - \frac{\gcd(T, k)}{k}\}$.*

It is clear that if $u \in \mathcal{R}(T, k)$ and $u' \sim u$ then $u' \in \mathcal{R}(T, k)$. In fact the following lemma states that $\mathcal{R}(T, k)$ forms a conjugacy class of $\mathcal{P}(T, k)$ and thus the regular sequences in $\mathcal{P}(T, k)$ are all cyclic permutations of each other.

Lemma 3.2.9 *Let $u \in \mathcal{R}(T, k)$. Then $u' \in \mathcal{R}(T, k)$ if and only if $u \sim u'$.*

Proof. We still have to prove the “only if”-part. For $u \in \mathcal{R}(T, k)$ we have that $\tilde{u} \subseteq \mathcal{R}(T, k)$. By Lemma 3.2.1 we have that $u = w^n$ for some primitive w and by Lemma 3.2.8 it follows that $n = \gcd(T, k)$. Hence the length of sequence w is $\frac{T}{\gcd(T, k)}$ and by Lemma 3.2.2 we have that $|\tilde{u}| = \frac{T}{\gcd(T, k)}$. However, by Lemma 3.2.8 we have that $|\mathcal{R}(T, k)| \leq \frac{T}{\gcd(T, k)}$. Thus $\tilde{u} = \mathcal{R}(T, k)$. \square

Notation. For sequence $u \in \mathcal{P}(T, k)$ we define the counting function $\kappa_u : \{0, 1, \dots, T\} \rightarrow \mathbb{Z}$ with $\kappa_u(n) = \sum_{t=1}^n u_t$. Thus $\kappa(n)$ counts the number of ones in the first n letters of the corresponding word Λ . We also define the discrepancy function $\chi_u : \{0, 1, \dots, T\} \rightarrow \mathbb{Q}$ by $\chi_u(n) = n \cdot \frac{k}{T} - \kappa_u(n)$ for $n = 0, 1, \dots, T$.

Now we define a partial order \preceq on $\mathcal{P}(T, k)$. For $u, v \in \mathcal{P}(T, k)$ we say that $u \preceq v$ if $\kappa_u(n) \leq \kappa_v(n)$ for $n = 1, 2, \dots, T$.

Lemma 3.2.10 *The partial order \preceq on $\mathcal{P}(T, k)$ induces a total order on $\mathcal{R}(T, k)$.*

Proof. Let $u, u' \in \mathcal{R}(T, k)$. Then by Lemma 3.2.8 there exist $\varphi, \varphi' \in \{1 - \frac{T}{k}, 1 - \frac{T - \gcd(T, k)}{k}, \dots, 1 - \frac{\gcd(T, k)}{k}\}$ such that the support set S_1 of u is

$\{[i \cdot \frac{T}{k} + \varphi]\}_{i=1}^k$, while the support set S'_1 of u' is $\{[i \cdot \frac{T}{k} + \varphi']\}_{i=1}^k$. Hence, $u \preceq u'$ if $\varphi \geq \varphi'$ and $u' \preceq u$ if $\varphi' \geq \varphi$. Thus u and u' are ordered. \square

Since $\mathcal{R}(T, k)$ is finite it follows from Lemma 3.2.10 that $\mathcal{R}(T, k)$ contains a greatest element for this order. It is easily seen that this greatest element is also the lexicographic greatest element of $\mathcal{R}(T, k)$ or equivalently, it is the sequence in which “the ones are as much to the left as possible” (under the constraint that the sequence is regular of course). This greatest element of $\mathcal{R}(T, k)$ is important for our definition of the (primal) unbalance of a sequence $u \in \mathcal{P}(T, k)$. We denote this greatest element with $\bar{\omega}(T, k)$ or just $\bar{\omega}$ if no confusion is possible. By Lemma 3.2.8 and the proof of Lemma 3.2.10 we have the following lemma which can be used to determine $\bar{\omega}(T, k)$ quickly.

Lemma 3.2.11 *For the support set S_1 of $\bar{\omega}(T, k)$ we have that*

$$S_1 = \{[i \cdot \frac{T}{k} + 1 - \frac{T}{k}]\}_{i=1}^k.$$

Corollary 3.2.12 *For $\bar{\omega}(T, k)$ we have that*

$$\kappa_{\bar{\omega}(T, k)}(n) = \lceil n \cdot \frac{k}{T} \rceil \text{ for } n = 0, 1, \dots, T.$$

We have seen in Lemma 3.2.9 that the regular sequences $\mathcal{R}(T, k)$ are a conjugacy class in $\mathcal{P}(T, k)$. We have the following theorem in which the partial order is used to give a characterising property (see Lemma 3.2.25) of the conjugacy class $\mathcal{R}(T, k)$ of regular sequences in the set of all the conjugacy classes of $\mathcal{P}(T, k)$.

Theorem 3.2.13 *Every conjugacy class \tilde{u} of $\mathcal{P}(T, k)$ contains an upper bound of $\mathcal{R}(T, k)$, i.e for every $u \in \mathcal{P}(T, k)$ there exists a $v \in \mathcal{P}(T, k)$ such that $v \sim u$ and $v \succeq w$ for every $w \in \mathcal{R}(T, k)$.*

Proof. Let $u \in \mathcal{P}(T, k)$ and let $l = \operatorname{argmax}_{n=0,1,\dots,T-1} \chi_u(n)$. We claim that $u' := (u_{1+l}, u_{2+l}, \dots, u_T, u_1, u_2, \dots, u_l)$, the l -th cyclic permutation of u , is an upper bound of $\mathcal{R}(T, k)$. Since $u' \sim u$ this proves the theorem.

To prove the claim it suffices to show that $u' \succeq \bar{\omega}$. By definition it follows that $\chi_u(l) \geq \chi_u((l+n) \pmod{T}) = \chi_u(l) + \chi_{u'}(n)$ for $n = 0, 1, \dots, T-1$ and thus $\chi_{u'}(n) \leq 0$ for $n = 0, 1, \dots, T-1$. Since $\chi_{u'}(0) = \chi_{u'}(T) = 0$ it follows that

$$\max_{n=1,2,\dots,T} \chi_{u'}(n) = \max_{n=0,1,\dots,T-1} \chi_{u'}(n) = 0.$$

Hence $\kappa_{u'}(n) \geq n \cdot \frac{k}{T}$ and thus $\kappa_{u'}(n) \geq \lceil n \cdot \frac{k}{T} \rceil$ for $n = 1, 2, \dots, T$. So, by Corollary 3.2.12 we have indeed that $u' \succeq \bar{w}$. \square

Remark. If $\gcd(k, T) = 1$ then it is easily seen for every $u \in \mathcal{P}(T, k)$ that χ_u is injective on the domain $\{0, 1, \dots, T-1\}$ and thus $\operatorname{argmax}_{n=0,1,\dots,T-1} \chi_u(n)$ is unique. A consequence is that for every $u \in \mathcal{P}(T, k)$ the upper bound of $\mathcal{R}(T, k)$ in the conjugacy class \tilde{u} is unique if $\gcd(k, T) = 1$. If $\gcd(k, T) > 1$ then the upper bound is in general not unique. For example, if $u = (1, 0, 0, 0, 1, 0, 0, 1, 0) \in \mathcal{P}(9, 3)$ then $\bar{w} = (1, 0, 0, 1, 0, 0, 1, 0, 0)$ and the conjugacy class \tilde{u} contains two upper bounds of $\mathcal{R}(9, 3)$. Namely, $(1, 0, 1, 0, 0, 0, 1, 0, 0)$ and $(1, 0, 0, 1, 0, 1, 0, 0, 0)$. Note that for $u \in \mathcal{P}(T, k)$ the proof of Theorem 3.2.13 yields an algorithm to determine the upper bound(s) of $\mathcal{R}(T, k)$ in the conjugacy class \tilde{u} .

Theorem 3.2.13 has the following “dual” theorem which is obtained in a completely analogous way. This dual theorem is needed for the definition of the dual unbalance.

Theorem 3.2.14 *Every conjugacy class \tilde{u} of $\mathcal{P}(T, k)$ contains a lower bound of $\mathcal{R}(T, k)$.*

We are now ready to define the (primal) unbalance.

Definition 3.2.15 *Let $u \in \mathcal{P}(T, k)$, $\bar{w} = \bar{w}(T, k)$ and let u' be an upper bound with respect to the partial order \preceq of $\mathcal{R}(T, k)$ in the conjugacy class \tilde{u} . Then the primal unbalance \bar{I} of u is defined by*

$$\bar{I}(u) = \frac{1}{T} \sum_{n=1}^T (\kappa_{u'}(n) - \kappa_{\bar{w}}(n)). \quad (3.1)$$

Theorem 3.2.16 *The primal unbalance \bar{I} is well defined.*

Proof. The primal unbalance \bar{I} is well defined if and only if for every pair u', u'' of upper bounds of $\mathcal{R}(T, k)$ with $u' \sim u''$ we have that $\sum_{n=1}^T \kappa_{u'}(n) = \sum_{n=1}^T \kappa_{u''}(n)$. We can assume that $u' \neq u''$. Since $u' \sim u''$ there exists an $m \in \{1, 2, \dots, T-1\}$ such that $u' = vw$, $u'' = wv$ where $v = (v_1, v_2, \dots, v_m)$ and $w = (w_1, w_2, \dots, w_{T-m})$. Since $u' \succeq \bar{w}$ and $u'' \succeq \bar{w}$ we have that $\chi_{u'}(n) \leq 0$ and $\chi_{u''}(n) \leq 0$ for $n = 0, 1, \dots, T$. Moreover $\chi_{u'}(0) = \chi_{u'}(T) = 0$ and $\chi_{u''}(0) = \chi_{u''}(T) = 0$. So,

$$\begin{aligned} \chi_{u'}(m) &= m \cdot \frac{k}{T} - \kappa_{u'}(m) = (T - (T - m)) \cdot \frac{k}{T} - \sum_{i=1}^m v_i = (T \cdot \frac{k}{T} - \kappa_{u''}(T)) \\ &- ((T - m) \cdot \frac{k}{T} - \kappa_{u''}(T - m)) = \chi_{u''}(T) - \chi_{u''}(T - m) = -\chi_{u''}(T - m) \end{aligned}$$

and thus $\chi_{u'}(m) = \chi_{u''}(T-m) = 0$. Hence $\chi_{u'}(0) = \chi_{u''}(T-m)$ and by induction it follows that $\chi_{u'}(n') = \chi_{u''}(T-m+n')$ for $n' = 0, 1, \dots, m$. Further $\chi_{u'}(m) = \chi_{u''}(0)$ and by induction it follows that $\chi_{u'}(n') = \chi_{u''}(n'-m)$ for $n' = m, m+1, \dots, T$. Hence

$$\begin{aligned} \sum_{n=1}^T \chi_{u'}(n) &= \sum_{n'=1}^m \chi_{u'}(n') + \sum_{n'=m+1}^T \chi_{u'}(n') = \sum_{n'=1}^m \chi_{u''}(T-m+n') + \\ &\sum_{n'=m+1}^T \chi_{u''}(n'-m) = \sum_{n=1}^T \chi_{u''}(n) \text{ and thus } \sum_{n=1}^T \kappa_{u'}(n) = \sum_{n=1}^T \kappa_{u''}(n). \end{aligned}$$

□

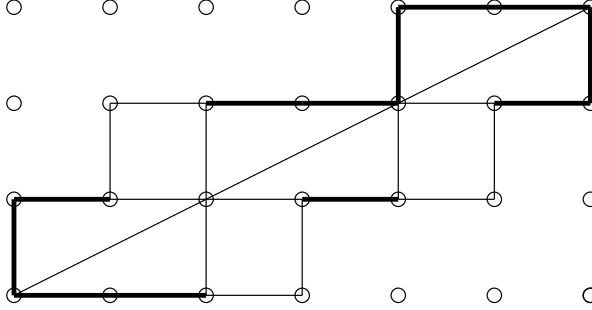
We define the dual unbalance by using the dual order \preceq' of the partial order \preceq defined by $u \preceq' v$ if and only if $v \preceq u$. With this dual order \preceq' we have in a completely analogous way the same results as for the order \preceq . Thus also for this order \preceq' we have that $\mathcal{R}(T, k)$ has a greatest element, which we denote by $\underline{\omega}(T, k)$, or just $\underline{\omega}$ if no confusion is possible. Note that $\underline{\omega}$ is the smallest element of $\mathcal{R}(T, k)$ with respect to the primal order \preceq . With this dual order we can restate Theorem 3.2.14 as “Every conjugacy class \tilde{u} of $\mathcal{P}(T, k)$ contains an upper bound (with respect to the order \preceq') of $\mathcal{R}(T, k)$.” Hence, Theorem 3.2.14 says the same as Theorem 3.2.13 except that the primal order \preceq is replaced with the dual order \preceq' .

Definition 3.2.17 *Let $u \in \mathcal{P}(T, k)$, $\underline{\omega} = \underline{\omega}(T, k)$ and let u' be an upper bound with respect to the (dual) partial order \preceq' of $\mathcal{R}(T, k)$ in the conjugacy class \tilde{u} . Then the dual unbalance \underline{I} of u is defined by*

$$\underline{I}(u) = \frac{1}{T} \sum_{n=1}^T (\kappa_{\underline{\omega}}(n) - \kappa_{u'}(n)). \quad (3.2)$$

The dual unbalance is just as the primal unbalance well defined and we can illustrate the primal unbalance $\bar{I}(u)$ and the dual unbalance $\underline{I}(u)$ of some sequence $u \in \mathcal{P}(T, k)$ graphically in the following way. We draw in the 2-dimensional plane a line segment l with slope $\frac{k}{T-k}$ from lattice point $(0, 0)$ to lattice point $(T-k, k)$. Further a sequence $v \in \mathcal{P}(T, k)$ corresponds to a connected graph from $(0, 0)$ to $(k, T-k)$ in the following way. The graph starts at $(0, 0)$ and goes one length unit to the right or above if the next element of the sequence is a zero or a one, respectively. To determine $\bar{I}(u)$ and $\underline{I}(u)$ for some $u \in \mathcal{P}(T, k)$ we have to find $u' \in \mathcal{P}(T, k)$ respectively $u'' \in \mathcal{P}(T, k)$ such that u' and u'' are in the same conjugacy class as u , $\bar{\omega} \preceq u'$ and $u'' \preceq \underline{\omega}$. The graphs of $u', u'', \bar{\omega}$ and $\underline{\omega}$ as described have the following properties.

Figure 3.1: The graphs and the unbalance



The graph of \bar{u} is never below the line segment l , but it remains as close as possible to l under this condition. Thus, the graph of \bar{u} goes through the lattice points that are on l and the lattice points that are immediately above l . Similarly the graph of $\underline{\omega}$ is never above l and goes through the lattice points on l and the lattice points immediately below l . Further the graph of u' is never below \bar{u} and thus never below l . It is easily seen that $T \cdot \bar{I}(u) = T \cdot \bar{I}(u')$ is the area below the graph of u' minus the area below the graph of \bar{u} . Similarly the graph of u'' is never above $\underline{\omega}$ and thus never above l . Moreover $T \cdot \underline{I}(u) = T \cdot \underline{I}(u'')$ is the area below the graph of $\underline{\omega}$ minus the area below the graph of u'' .

For example consider the sequence $u = (1, 0, 0, 0, 1, 0, 0, 1, 0) \in \mathcal{P}(9, 3)$ of the previous example. We take $u' = (1, 0, 1, 0, 0, 0, 1, 0, 0)$ as upper bound of $\mathcal{R}(9, 3)$ in the conjugacy class of u and we take $u'' = (0, 0, 0, 1, 0, 0, 1, 0, 1)$ as lower bound of $\mathcal{R}(9, 3)$ in the conjugacy class of u . Figure 3.1 shows the graphs of $\bar{u}, \underline{\omega}, u', u''$ and the line segment l . Note that we have drawn the lines of the graphs above l and below l thicker if the graphs of u' and \bar{u} respectively the graphs of u'' and $\underline{\omega}$ coincide. It follows that $\bar{I}(u) = \frac{1}{9}$ and $\underline{I}(u) = \frac{2}{9}$.

A direct consequence of the definitions of the primal and dual unbalance is the following lemma which states that conjugates have the same primal unbalance and the same dual unbalance.

Lemma 3.2.18 *If $u \sim u'$ then $\bar{I}(u) = \bar{I}(u')$ and $\underline{I}(u) = \underline{I}(u')$.*

For a finite word (sequence) $u = (u_1, u_2, \dots, u_T)$ we denote the reversed word by $\overleftarrow{u} := (u_T, u_{T-1}, \dots, u_1)$. Further if for some sequences $u, v \in \mathcal{P}(T, k)$ we have that $\overleftarrow{u} \sim v$ then we say that u and v are mirrors of each other. Note that if $u \in \mathcal{P}(T, k)$ is regular then u is a mirror of itself.

Proposition 3.2.19 *If $u, v \in \mathcal{P}(T, k)$ are mirrors of each other then $\bar{I}(u) = \underline{I}(v)$ and $\underline{I}(u) = \bar{I}(v)$.*

Proof. Let $u' \sim u$ be an upper bound with respect to the partial order \preceq of $\mathcal{R}(T, k)$ and put $v' = \overleftarrow{u'}$. Then

$$\bar{I}(u) = \frac{1}{T} \sum_{n=1}^T (\kappa_{u'}(n) - \kappa_{\overline{u'}}(n)) = \frac{1}{T} \sum_{n=0}^T (\kappa_{u'}(n) - \lceil n \cdot \frac{k}{T} \rceil)$$

and $v' \sim v$. Moreover for $n' = 0, 1, \dots, T$ we have that

$$\kappa_{v'}(T - n') = k - \kappa_{u'}(n') \leq k - \lceil n' \cdot \frac{k}{T} \rceil = \lfloor (T - n') \cdot \frac{k}{T} \rfloor$$

and thus $\kappa_{v'}(n) \leq \lfloor n \cdot \frac{k}{T} \rfloor$ for $n = 0, 1, 2, \dots, T$. Hence $v' \sim v$ is an upper bound with respect to the partial order \preceq' of $\mathcal{R}(T, k)$. Thus

$$\begin{aligned} \underline{I}(v) &= \frac{1}{T} \sum_{n'=1}^T (\kappa_{\underline{v'}}(n') - \kappa_{v'}(n')) = \frac{1}{T} \sum_{n'=0}^T (\lfloor n' \cdot \frac{k}{T} \rfloor - (k - \kappa_{u'}(T - n'))) = \\ &= \frac{1}{T} \cdot \sum_{n'=0}^T (\kappa_{u'}(T - n') - \lceil (T - n') \cdot \frac{k}{T} \rceil) = \frac{1}{T} \cdot \sum_{n=0}^T (\kappa_{u'}(n) - \lceil n \cdot \frac{k}{T} \rceil) = \bar{I}(u). \end{aligned}$$

Analogously it follows that $\underline{I}(u) = \bar{I}(v)$. □

For example if $u = (1, 0, 1, 1, 0, 1, 1, 0, 1, 0) \in \mathcal{P}(11, 7)$ then $v = (1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1) \in \mathcal{P}(11, 7)$ is a mirror of u . The reader can check that $\bar{I}(u) = \underline{I}(v) = \frac{1}{11}$ and $\underline{I}(u) = \bar{I}(v) = \frac{2}{11}$.

We also define the primal unbalance \bar{I} and the dual unbalance \underline{I} on the set of conjugacy classes by $\bar{I}(\tilde{u}) = \bar{I}(u)$ and $\underline{I}(\tilde{u}) = \underline{I}(u)$, where u is an arbitrary representative of \tilde{u} . By Lemma 3.2.18 this is well defined.

Lemma 3.2.20 *If $u = v^m$ where v is a primitive sequence and $m \in \mathbb{N}$ then $\bar{I}(u) = \bar{I}(v)$ and $\underline{I}(u) = \underline{I}(v)$.*

Proof. We can assume that $v \in \mathcal{P}(T, k)$. Then we have that $u \in \mathcal{P}(mT, mk)$ and by Lemma 3.2.11 we have that $\bar{\omega}(mT, mk) = \bar{\omega}(T, k)^m$. Further if $w \sim v$ and w is an upper bound of $\mathcal{R}(T, k)$ then $w^m \sim u$ and w^m is an upper bound of $\mathcal{R}(mT, mk)$. Thus we have for every $l \in \{1, 2, \dots, T\}$ and $b \in \{0, 1, \dots, m-1\}$ that

$$\kappa_w(l) - \kappa_{\bar{\omega}(T, k)}(l) = \kappa_{w^m}(l + bT) - \kappa_{\bar{\omega}(mT, mk)}(l + bT).$$

Hence

$$\bar{I}(u) = \frac{1}{mT} \sum_{n=1}^{mT} \{\kappa_{w^m}(n) - \kappa_{\bar{\omega}(mT, mk)}(n)\} = \frac{1}{T} \sum_{n=1}^T \{\kappa_w(n) - \kappa_{\bar{\omega}(T, k)}(n)\} = \bar{I}(v).$$

It follows completely analogously that $\underline{I}(u) = \underline{I}(v)$. \square

Lemma 3.2.21 *For every $u \in \mathcal{P}(T, k)$ we have that $T \cdot \bar{I}(u) \in \mathbb{Z}_{\geq 0}$ and $\bar{I}(u) = 0$ if and only if $u \in \mathcal{R}(T, k)$. The same is true for the dual unbalance $\underline{I}(u)$.*

Proof. Let u' be as in the definition of the the primal unbalance. Then $u' \succeq \bar{\omega}$ and thus $\kappa_{u'}(n) - \kappa_{\bar{\omega}}(n) \in \mathbb{Z}_{\geq 0}$ for $n = 1, 2, \dots, T$. Thus $T \cdot \bar{I}(u) = \sum_{n=1}^T (\kappa_{u'}(n) - \kappa_{\bar{\omega}}(n)) \in \mathbb{Z}_{\geq 0}$. If $u \in \mathcal{R}(T, k)$ then by Lemma 3.2.9 we have that $u' = \bar{\omega}$ and thus $\bar{I}(u) = 0$. Conversely, if $\bar{I}(u) = 0$ then it follows that $\kappa_{u'}(n) - \kappa_{\bar{\omega}}(n) = 0$ for $n = 1, 2, \dots, T$ and thus $u' = \bar{\omega} \in \mathcal{R}(T, k)$. So, by Lemma 3.2.9 $u \in \mathcal{R}(T, k)$. For the dual unbalance $\underline{I}(u)$ the proof is completely analogous. \square

Lemma 3.2.22 *Let $u, u', v, v' \in \mathcal{P}(T, k)$. Then $u \sim u'$, $v \sim v'$, $u \succeq v \succeq \bar{\omega}$ and $v' \succeq u' \succeq \bar{\omega}$ implies $u = v$ and $u' = v'$.*

Proof. By Lemma 3.2.18, the given orders and the definition of $\bar{I}(u)$ we have that $\bar{I}(u) = \bar{I}(u')$, $\bar{I}(v) = \bar{I}(v')$, $\bar{I}(u) \geq \bar{I}(v)$ and $\bar{I}(v') \geq \bar{I}(u')$. Hence $\bar{I}(u) = \bar{I}(u') = \bar{I}(v') = \bar{I}(v)$ and thus $u = v$ and $u' = v'$. \square

In the sequel we denote the set of conjugacy classes of $\mathcal{P}(T, k)$ by $\tilde{\mathcal{P}}(T, k)$ and we denote the conjugacy class of regular sequences in $\mathcal{P}(T, k)$ by $\tilde{\omega}(T, k)$ or just $\tilde{\omega}$. We now define a partial order $\leq_{\bar{g}}$ and a partial order $\leq_{\underline{g}}$ on $\tilde{\mathcal{P}}(T, k)$.

Definition 3.2.23 *Let $\tilde{u}, \tilde{v} \in \tilde{\mathcal{P}}(T, k)$. Then $\tilde{u} \leq_{\bar{g}} \tilde{v}$ if there exist $u', v' \in \mathcal{P}(T, k)$ such that $u' \in \tilde{u}$, $v' \in \tilde{v}$ and $\bar{\omega} \preceq u' \preceq v'$. Further $\tilde{u} \leq_{\underline{g}} \tilde{v}$ if there exist $u'', v'' \in \mathcal{P}(T, k)$ such that $u'' \in \tilde{u}$, $v'' \in \tilde{v}$ and $\underline{\omega} \preceq' u'' \preceq' v''$.*

Suppose that $\tilde{u} \leq_{\bar{g}} \tilde{v}$ for some $u, v \in \mathcal{P}(T, k)$ and let u', v' be as in Definition 3.2.23. Then the graph of u' is never above the graph of v' and they are both never below

the graph of $\bar{\omega}$. Therefore we call $\trianglelefteq_{\bar{g}}$ the upper graph order and $\trianglelefteq_{\underline{g}}$ the lower graph order.

Lemma 3.2.24 $\trianglelefteq_{\bar{g}}$ and $\trianglelefteq_{\underline{g}}$ are indeed partial orders on $\tilde{\mathcal{P}}(T, k)$.

Proof. The reflexivity of $\trianglelefteq_{\bar{g}}$ follows from Theorem 3.2.13 and the antisymmetry from Lemma 3.2.22. Further the transitivity of $\trianglelefteq_{\bar{g}}$ follows directly from the definition and thus $\trianglelefteq_{\bar{g}}$ is a partial order on $\tilde{\mathcal{P}}(T, k)$. The proof that $\trianglelefteq_{\underline{g}}$ is a partial order on $\tilde{\mathcal{P}}(T, k)$ is completely analogous. \square

As a direct consequence of Theorem 3.2.13 and the “dual” Theorem 3.2.14 we have the following result.

Lemma 3.2.25 The conjugacy class $\tilde{\omega}$ is the smallest element of $\tilde{\mathcal{P}}(T, k)$ with respect to the order $\trianglelefteq_{\bar{g}}$ and also with respect to the order $\trianglelefteq_{\underline{g}}$.

We say that $\tilde{u} \triangleleft_{\bar{g}} \tilde{v}$ if $\tilde{u} \trianglelefteq_{\bar{g}} \tilde{v}$ and $\tilde{u} \neq \tilde{v}$ and similar for the other partial orders. The following theorem follows directly from the definitions.

Theorem 3.2.26 Let $u, v \in \mathcal{P}(T, k)$. Then $\tilde{u} \trianglelefteq_{\bar{g}} \tilde{v}$ implies $\bar{I}(u) \leq \bar{I}(v)$ and $\tilde{u} \trianglelefteq_{\underline{g}} \tilde{v}$ implies $\underline{I}(u) \leq \underline{I}(v)$. Moreover, $\tilde{u} \triangleleft_{\bar{g}} \tilde{v}$ implies $\bar{I}(u) < \bar{I}(v)$ and $\tilde{u} \triangleleft_{\underline{g}} \tilde{v}$ implies $\underline{I}(u) < \underline{I}(v)$.

The orders $\trianglelefteq_{\bar{g}}$ and $\trianglelefteq_{\underline{g}}$ are in general not the same on $\tilde{\mathcal{P}}(T, k)$. For example if $u = (1, 0, 1, 0, 0, 0, 1, 0, 0) \in \mathcal{P}(9, 3)$ and $v = (1, 0, 1, 0, 0, 1, 0, 0, 0) \in \mathcal{P}(9, 3)$ then $\tilde{u} \triangleleft_{\bar{g}} \tilde{v}$ and $\tilde{v} \triangleleft_{\underline{g}} \tilde{u}$. Therefore it is useful to define the partial order \trianglelefteq_g on $\tilde{\mathcal{P}}(T, k)$ as “intersection” of the orders $\trianglelefteq_{\bar{g}}$ and $\trianglelefteq_{\underline{g}}$. We call \trianglelefteq_g the (strong) graph order.

Definition 3.2.27 Let $\tilde{u}, \tilde{v} \in \tilde{\mathcal{P}}(T, k)$. Then $\tilde{u} \trianglelefteq_g \tilde{v}$ if $\tilde{u} \trianglelefteq_{\bar{g}} \tilde{v}$ and $\tilde{u} \trianglelefteq_{\underline{g}} \tilde{v}$.

Remark. The partial orders $\trianglelefteq_{\bar{g}}$, $\trianglelefteq_{\underline{g}}$ and \trianglelefteq_g on $\tilde{\mathcal{P}}(T, k)$ induce in a natural way corresponding preorders on $\mathcal{P}(T, k)$. For example for $u, v \in \mathcal{P}(T, k)$ we say that $u \trianglelefteq_g v$ if $\tilde{u} \trianglelefteq_g \tilde{v}$ and similar for the other orders. However, if the orders $\trianglelefteq_{\bar{g}}$, $\trianglelefteq_{\underline{g}}$ and \trianglelefteq_g are used on $\mathcal{P}(T, k)$ in this way then they are not antisymmetric, since cyclic permutations of u are different words and therefore they are not partial orders on $\mathcal{P}(T, k)$.

We will extend the above notions to infinite periodic sequences (words). First of all we extend in a natural way the partial order \preceq and its dual order \preceq' to the

infinite sequences $u = (u_1, u_2, \dots)$ of zeros and ones corresponding to \mathbb{N} -words on the alphabet $\{0, 1\}$. We define just as for finite sequences (words) the counting function $\kappa_u : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{Z}$ by $\kappa_u(n) = \sum_{t=1}^n u_t$ and we say that $u \preceq v$ if $\kappa_u(n) \leq \kappa_v(n)$ for $n = 1, 2, \dots$. For a sequence $u = (u_1, u_2, \dots)$ of zeros and ones of density d corresponding to an \mathbb{N} -word we also define the discrepancy function $\chi_u : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}$ by $\chi_u(n) = n \cdot d - \kappa_u(n)$ for $n \in \mathbb{Z}_{\geq 0}$. Further we introduce an equivalence relation \sim for infinite periodic sequences. In fact \sim is an equivalence relation for all infinite periodic integer sequences and the corresponding words on a finite alphabet A and not only for the sequences of zeros and ones.

Definition 3.2.28 *Let u and v be infinite periodic integer sequences. Then we say that u is equivalent to v , $u \sim v$, if there exists a finite sequence w such that w is a period cycle of both u and v .*

In our application of routing sequences in a queueing system we have in general that equivalent infinite periodic sequences have the same performance. Therefore in most cases we consider infinite periodic sequences modulo this equivalence relation, i.e. if $u \sim v$ then we consider them as the same sequence. This is also necessary for the various partial orders that we define on infinite periodic sequences of zeros and ones to be antisymmetric. Further an infinite periodic sequence is now completely determined by a period cycle of the sequence. If $u \in \mathcal{P}(T, k)$ then $u^\infty = (u_1, u_2, \dots, u_T)^\infty$ is the infinite periodic sequence with period cycle u .

Let \mathcal{P} be the set of infinite periodic sequences of zeros and ones (modulo the equivalence relation), $\mathcal{R} \subseteq \mathcal{P}$ the subset of regular periodic sequences. Further for $d \in \mathbb{Q}$, $0 \leq d \leq 1$ let $\mathcal{P}(d) \subseteq \mathcal{P}$ be the subset of sequences with density d and $\mathcal{R}(d) := \mathcal{R} \cap \mathcal{P}(d)$. Then we have the following lemma.

Lemma 3.2.29 *We have that*

$$\mathcal{P} = \cup_{d \in \mathbb{Q}, 0 \leq d \leq 1} \mathcal{P}(d),$$

$$\mathcal{R} = \cup_{d \in \mathbb{Q}, 0 \leq d \leq 1} \mathcal{R}(d)$$

and $|\mathcal{R}(d)| = 1$ for every $d \in \mathbb{Q}$, $0 \leq d \leq 1$.

Proof. If $u \in \mathcal{P}$ then there exist $k \in \mathbb{Z}_{\geq 0}$, $T \in \mathbb{N}$, $k \leq T$ such that $u = v^\infty$ for some $v \in \mathcal{P}(T, k)$ and thus $u \in \mathcal{P}(\frac{k}{T}) = \mathcal{P}(d)$ for some $d \in \mathbb{Q}$, $0 \leq d \leq 1$. Hence

$$\mathcal{P} = \cup_{d \in \mathbb{Q}, 0 \leq d \leq 1} \mathcal{P}(d)$$

and thus

$$\mathcal{R} = \mathcal{R} \cap \mathcal{P} = \mathcal{R} \cap \{\cup_{d \in \mathbb{Q}, 0 \leq d \leq 1} \mathcal{P}(d)\} = \cup_{d \in \mathbb{Q}, 0 \leq d \leq 1} \mathcal{R}(d).$$

By Lemma 3.2.9 we have that $|\mathcal{R}(d)| = 1$ for every $d \in \mathbb{Q}, 0 \leq d \leq 1$. \square

We denote the unique element of $\mathcal{R}(d)$ by $\omega(d)$ for every $d \in \mathbb{Q}, 0 \leq d \leq 1$. We now define the primal and dual unbalance for infinite periodic sequences.

Definition 3.2.30 *Let $u \in \mathcal{P}$ and let $u' \in \mathcal{P}(T, k)$ be a period cycle of u . Then we define the primal unbalance of u as $\bar{I}(u) := \bar{I}(u')$ and we define the dual unbalance of u as $\underline{I}(u) := \underline{I}(u')$.*

Lemma 3.2.31 *The primal unbalance \bar{I} and dual unbalance \underline{I} are well defined on \mathcal{P} .*

Proof. Suppose u' and u'' are both period cycles of u . Then we have to show that $\bar{I}(u') = \bar{I}(u'')$. There exists a primitive period cycle w of u such that $u' \sim w^m$ and $u'' \sim w^n$ for some $m, n \in \mathbb{N}$. Then by Lemma 3.2.18 and Lemma 3.2.20 we have that $\bar{I}(u') = \bar{I}(w) = \bar{I}(u'')$. Hence \bar{I} is well defined on \mathcal{P} and completely analogously it follows that \underline{I} is well defined on \mathcal{P} . \square

Remark. When the unbalance of an infinite periodic sequence has to be computed it is of course most practical to compute the unbalance of a primitive period cycle of the sequence and not an arbitrary period cycle.

Suppose that $u, v \in \mathcal{P}(d)$ for some $d \in \mathbb{Q}, 0 \leq d \leq 1$. Then we say that u, v are mirrors of each other if there exist u', v' such that u' and v' are period cycles of u and v , respectively, and u' and v' are mirrors of each other. By Proposition 3.2.19 and the definitions of the primal and dual unbalance for infinite period sequences of zeros and ones we have the following corollary.

Corollary 3.2.32 *Suppose that $u, v \in \mathcal{P}(d)$ for some $d \in \mathbb{Q}, 0 \leq d \leq 1$ and u and v are mirrors of each other. Then $\bar{I}(u) = \underline{I}(v)$ and $\underline{I}(u) = \bar{I}(v)$.*

For every $d \in \mathbb{Q}, 0 \leq d \leq 1$ we define the partial orders $\preceq_{\bar{g}}, \preceq_{\underline{g}}$ and \preceq_g on $\mathcal{P}(d)$. Let $u, v \in \mathcal{P}(d)$ and let f_1 be the period of u and f_2 the period of v . Then by Lemma 3.2.3 u and v have both period $T := \text{lcm}(f_1, f_2)$. Moreover, if u' is a period cycle of length T of u and v' is a period cycle of length T of v then $u' \in \mathcal{P}(T, d \cdot T)$ and $v' \in \mathcal{P}(T, d \cdot T)$.

Definition 3.2.33 Let $u, v \in \mathcal{P}(d)$ for some $d \in \mathbb{Q}, 0 \leq d \leq 1$ and let $u', v' \in \mathcal{P}(T, d \cdot T)$ be corresponding period cycles as above. Then we define the partial orders $\trianglelefteq_{\bar{g}}, \trianglelefteq_{\underline{g}}$ and \trianglelefteq_g on $\mathcal{P}(d)$ by

$$\begin{aligned} u \trianglelefteq_{\bar{g}} v &\text{ if } \tilde{u}' \trianglelefteq_{\bar{g}} \tilde{v}' \text{ in } \tilde{\mathcal{P}}(T, d \cdot T), \\ u \trianglelefteq_{\underline{g}} v &\text{ if } \tilde{u}' \trianglelefteq_{\underline{g}} \tilde{v}' \text{ in } \tilde{\mathcal{P}}(T, d \cdot T) \text{ and} \\ u \trianglelefteq_g v &\text{ if } \tilde{u}' \trianglelefteq_g \tilde{v}' \text{ in } \tilde{\mathcal{P}}(T, d \cdot T). \end{aligned}$$

All the results we derived for the partial orders $\trianglelefteq_{\bar{g}}, \trianglelefteq_{\underline{g}}$ and \trianglelefteq_g on $\tilde{\mathcal{P}}(T, k)$ also hold for the partial orders $\trianglelefteq_{\bar{g}}, \trianglelefteq_{\underline{g}}$ and \trianglelefteq_g on $\mathcal{P}(d)$.

Lemma 3.2.34 For every $d \in \mathbb{Q}, 0 \leq d \leq 1$, the regular sequence $\omega(d)$ is the smallest element of $\mathcal{P}(d)$ with respect to the order $\trianglelefteq_{\bar{g}}$, the order $\trianglelefteq_{\underline{g}}$ and with respect to the order \trianglelefteq_g .

Theorem 3.2.35 Let $u, v \in \mathcal{P}(d)$. Then $u \trianglelefteq_{\bar{g}} v$ implies $\bar{I}(u) \leq \bar{I}(v)$ and $u \trianglelefteq_{\underline{g}} v$ implies $\underline{I}(u) \leq \underline{I}(v)$. Moreover, $u \triangleleft_{\bar{g}} v$ implies $\bar{I}(u) < \bar{I}(v)$ and $u \triangleleft_{\underline{g}} v$ implies $\underline{I}(u) < \underline{I}(v)$.

We give an example to illustrate this. Let $u = (1, 1, 0, 1, 0, 0)^\infty \in \mathcal{P}(\frac{1}{2})$ and $v = (1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0)^\infty \in \mathcal{P}(\frac{1}{2})$. Then $u \triangleleft_{\bar{g}} v, v \triangleleft_{\underline{g}} u$ and therefore u and v are not \trianglelefteq_g ordered. Further $\bar{I}(u) = \frac{1}{3}, \bar{I}(v) = \frac{5}{12}, \underline{I}(u) = \frac{1}{6}$ and $\underline{I}(v) = \frac{1}{12}$.

3.3 Bounding the difference in expected average waiting time between sequences

Let $\{T_i\}_{i=1,2,\dots}$ be a sequence of arrival times of customers, with the convention that $T_1 = 0$. Put $\delta_i := T_{i+1} - T_i$ for $i = 1, 2, \dots$. Then $\{\delta_i\}$ is the sequence of interarrival times. Further a fraction of these arriving customers is routed to a server according to some routing sequence $u = (u_1, u_2, \dots)$ of zeros and ones. For such a routing sequence we have the counting function $\kappa_u(n) = \sum_{t=1}^n u_t$ that we used to define a partial order \preceq for such routing sequences. Further we define the the following related function $\nu_u(i) : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{Z}$ by $\nu_u(j) = \min\{n \in \mathbb{Z}_{\geq 0} : \kappa_u(n) = j\}$ and we put $\tau_u(j) = \sum_{i=\max(\nu_u(j-1), 1)}^{\nu_u(j)-1} \delta_i$ for $j = 1, 2, \dots$. Then $\tau_u(j)$ is the time elapsed between the routing of the $j - 1$ -th and j -th customer to the server according to routing sequence u . Note that $u \preceq v$ implies $\nu_u(j) \geq \nu_v(j)$ and $\lambda_u(j) \geq \lambda_v(j)$ for

all j . If we put $\lambda_u(j) = \sum_{k=1}^j \tau_u(k)$ for $j = 1, 2, \dots$ then $\lambda_u(j)$ is the time at which the j -th customer is routed to the server according to routing sequence u . Further we have a sequence of service times $\{\sigma_j\}_{j=1,2,\dots}$, where σ_j is the service time of the j -th customer that is routed to the server according to any routing sequence. For now we assume that the interarrival times $\{\delta_i\}_{i=1,2,\dots}$ and service times $\{\sigma_j\}_{j=1,2,\dots}$ are fixed sequences of non-negative real numbers. Later we will also consider the case that they are random variables. Further we define $W_u(j, w)$ to be the workload for the server at the moment the j -th customer is routed to the server according to routing sequence u , given that the initial workload at time $T_1 = 0$ is equal to w . In other words $W_u(j, w)$ is the waiting time for the j -th customer that is routed to the server. If the initial workload w is fixed (thus the system starts for all routing sequences u with the same initial workload w) then we write for short $W_u(j)$ instead of $W_u(j, w)$. For example this is the case if it is assumed that the server is empty at $T_1 = 0$. Note that $W_u(j) + \lambda_u(j)$ is the moment that the server starts serving the j -th customer that is routed to the server. We have the following lemma.

Lemma 3.3.1 *Let $u = (u_1, u_2, \dots)$ and $v = (v_1, v_2, \dots)$ be routing sequences as above. If $u \preceq v$ and $W_v(j, w) + \lambda_v(j) \leq W_u(j, w') + \lambda_u(j)$ for some $j \in \mathbb{N}$ then for every $k \geq j$ we have that $W_v(k, w) + \lambda_v(k) \leq W_u(k, w') + \lambda_u(k)$.*

Proof. It suffices to show the assertion for $k = j + 1$, because it then follows with induction. We have that $W_v(j + 1, w) = \max(W_v(j, w) + \sigma_j - \tau_v(j + 1), 0)$ and $W_u(j + 1, w') = \max(W_u(j, w') + \sigma_j - \tau_u(j + 1), 0)$. Further $\lambda_u(j + 1) - \lambda_v(j + 1) \geq 0$ since $u \preceq v$. Hence

$$\begin{aligned} W_v(j + 1, w) &\leq \max(W_u(j, w') + \lambda_u(j) - \lambda_v(j) + \sigma_j - \tau_v(j + 1), 0) \leq \\ &\max(W_u(j, w') + \sigma_j - \tau_u(j + 1), 0) + \max(\lambda_u(j + 1) - \lambda_v(j + 1), 0) = \\ &W_u(j + 1, w') + \lambda_u(j + 1) - \lambda_v(j + 1). \end{aligned}$$

□

If $w = w'$ then the inequality holds for $j = 1$ and thus we have the following corollary.

Corollary 3.3.2 *Let u and v as above, $u \preceq v$. Then*

$$W_v(j) + \lambda_v(j) \leq W_u(j) + \lambda_u(j) \text{ for every } j \in \mathbb{N}. \quad (3.3)$$

Note that for every $j \in \mathbb{N}$ and $u \preceq v$ we have that $\lambda_u(j) - \lambda_v(j) = \sum_{i=\nu_v(j)}^{\nu_u(j)-1} \delta_i$. This is a sum of interarrival times δ_i , where the number of terms in the sum is $\nu_u(j) - \nu_v(j)$. Therefore, if we put $N_{uv}(m) = \sum_{j=1}^m (\nu_u(j) - \nu_v(j))$ for $m = 1, 2, \dots$ then we have the following lemma by Corollary 3.3.2.

Lemma 3.3.3 *Let u and v be routing sequences with $u \preceq v$. For every $m \in \mathbb{N}$*

$$\sum_{j=1}^m W_v(j) \leq \sum_{j=1}^m W_u(j) + \sum_{j=1}^m \sum_{i=\nu_v(j)}^{\nu_u(j)-1} \delta_i. \quad (3.4)$$

In the double sum the number of terms is $N_{uv}(m)$.

Since for every sequence $u = (u_1, u_2, \dots)$ the functions κ_u and ν_u are strongly related, we can also use the counting function κ instead of ν to find an expression for $N_{uv}(m)$ in view of the following lemma.

Lemma 3.3.4 *Let $u = (u_1, u_2, \dots)$ and $v = (v_1, v_2, \dots)$ be sequences of zeros and ones. Then we have for every $m \in \mathbb{N}$ that*

$$\sum_{j=1}^m (\nu_u(j) - \nu_v(j)) = \sum_{n=1}^{\max(\nu_u(m), \nu_v(m))} \{\min(m, \kappa_v(n)) - \min(m, \kappa_u(n))\}.$$

Proof. We prove it by induction on m . For $m = 1$ we may assume that $\nu_u(1) \geq \nu_v(1)$. Then

$$\begin{aligned} \nu_u(1) - \nu_v(1) &= \sum_{n=\nu_v(1)}^{\nu_u(1)-1} 1 = \sum_{n=1}^{\nu_u(1)-1} \min(1, \kappa_v(n)) = \\ &= \sum_{n=1}^{\max(\nu_u(1), \nu_v(1))} \{\min(1, \kappa_v(n)) - \min(1, \kappa_u(n))\}. \end{aligned}$$

Similarly we have for $m = 2, 3, \dots$ that

$$\nu_u(m) - \nu_v(m) = \sum_{n=\min(\nu_u(m), \nu_v(m))}^{\max(\nu_u(m), \nu_v(m))-1} \{\min(\kappa_v(n), \kappa_u(n)+1) - \min(\kappa_u(n), \kappa_v(n)+1)\}.$$

So, by induction we have for $m = 2, 3, \dots$ that

$$\begin{aligned} \sum_{j=1}^m (\nu_u(j) - \nu_v(j)) &= \sum_{j=1}^{m-1} (\nu_u(j) - \nu_v(j)) + (\nu_u(m) - \nu_v(m)) = \\ &= \sum_{n=1}^{\max(\nu_u(m-1), \nu_v(m-1))} \{\min(m-1, \kappa_v(n)) - \min(m-1, \kappa_u(n))\} + \end{aligned}$$

$$\begin{aligned} & \sum_{n=\min(\nu_u(m), \nu_v(m))}^{\max(\nu_u(m), \nu_v(m))-1} \{\min(\kappa_v(n), \kappa_u(n) + 1) - \min(\kappa_u(n), \kappa_v(n) + 1)\} = \\ & \sum_{n=1}^{\max(\nu_u(m), \nu_v(m))} \{\min(m, \kappa_v(n)) - \min(m, \kappa_u(n))\}. \end{aligned}$$

□

Corollary 3.3.5 *Let u and v be sequences as above and suppose that for some $l, m \in \mathbb{N}$ we have that $\kappa_u(l) = \kappa_v(l) = m$. Then $N_{uv}(m) = \sum_{n=1}^l (\kappa_v(n) - \kappa_u(n))$.*

For periodic sequences of the same density we have the following lemma.

Lemma 3.3.6 *Suppose that $u = (u')^\infty$ and $v = (v')^\infty$ for some $u', v' \in \mathcal{P}(T, k)$ and $\bar{\omega}(T, k) \preceq u' \preceq v'$. Suppose $m = sk$ for some $s \in \mathbb{N}$. Then $N_{uv}(m) = s \cdot T \cdot (\bar{I}(v) - \bar{I}(u))$.*

Proof. Let $m = sk$ with $s \in \mathbb{N}$. Then $\kappa_u(sT) = \kappa_v(sT) = sk$. So, by Corollary 3.3.5, $N_{uv}(m) = \sum_{n=1}^{sT} (\kappa_v(n) - \kappa_u(n))$. Since $\{\kappa_v(n) - \kappa_u(n)\}$ is periodic with period T we deduce that

$$\begin{aligned} N_{uv}(m) &= \sum_{n=1}^{sT} (\kappa_v(n) - \kappa_u(n)) = s \cdot \sum_{n=1}^T (\kappa_v(n) - \kappa_u(n)) = \\ & s \cdot \sum_{n=1}^T \{(\kappa_v(n) - \kappa_{\bar{\omega}}(n)) - (\kappa_u(n) - \kappa_{\bar{\omega}}(n))\} = s \cdot T \cdot (\bar{I}(v) - \bar{I}(u)). \end{aligned}$$

□

Corollary 3.3.7 *Suppose that $u = (u')^\infty$ and $v = (v')^\infty$ for some $u', v' \in \mathcal{P}(T, k)$ and $\bar{\omega}(T, k) \preceq u' \preceq v'$. Then for every $m \in \mathbb{N}$ we have that*

$$\lfloor \frac{m}{k} \rfloor \cdot T \cdot (\bar{I}(v) - \bar{I}(u)) \leq N_{uv}(m) \leq \lceil \frac{m}{k} \rceil \cdot T \cdot (\bar{I}(v) - \bar{I}(u))$$

and $\lim_{m \rightarrow \infty} \frac{N_{uv}(m)}{m} = \frac{T}{k} \cdot (\bar{I}(v) - \bar{I}(u))$.

Proof. Because $u' \preceq v'$ we have that $u \preceq v$ and it follows that $N_{uv}(m)$ is monotonically non-decreasing in m . Therefore the first statement follows from Lemma 3.3.6 and the second statement follows from the first. □

By Lemma 3.3.3 and Corollary 3.3.7 we have the following theorem.

Theorem 3.3.8 *Let u and v be as in Corollary 3.3.7. Suppose that for $\{\delta_i\}_{i=1,2,\dots}$, the sequence of interarrival times, we have $\limsup_{t \rightarrow \infty} \frac{1}{t} \cdot \sum_{i=0}^{t-1} \delta_{i \cdot T + l} := \gamma_l \leq \delta$ for $l = 1, 2, \dots, T$. Then*

$$\limsup_{m \rightarrow \infty} \frac{1}{m} \cdot \left(\sum_{j=1}^m W_v(j) - \sum_{j=1}^m W_u(j) \right) \leq \delta \cdot \frac{T}{k} \cdot (\bar{I}(v) - \bar{I}(u)).$$

Proof. By Lemma 3.3.3 together with some straightforward derivations,

$$\limsup_{m \rightarrow \infty} \frac{1}{m} \cdot \left(\sum_{j=1}^m W_v(j) - \sum_{j=1}^m W_u(j) \right) \leq \limsup_{m \rightarrow \infty} \frac{N_{uv}(m)}{m} \cdot \max_{l=1,\dots,T} \gamma_l \leq \delta \cdot \limsup_{m \rightarrow \infty} \frac{N_{uv}(m)}{m}.$$

Hence by Corollary 3.3.7

$$\limsup_{m \rightarrow \infty} \frac{1}{m} \cdot \left(\sum_{j=1}^m W_v(j) - \sum_{j=1}^m W_u(j) \right) \leq \delta \cdot \frac{T}{k} \cdot (\bar{I}(v) - \bar{I}(u)).$$

□

Let u be a routing sequence of zeros and ones. If the interarrival times $\{\delta_i\}$ and service times $\{\sigma_j\}$ are fixed sequences then we define the long-run average waiting time $\bar{W}(u)$ of customers routed to the server according to routing sequence u by $\lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{j=1}^m W_u(j) = \bar{W}(u)$. If $\bar{W}(u)$ exists, but the value depends on the initial workload w then $\bar{W}(u)$ denotes the value for initial workload $w = 0$ (thus the server is assumed to be empty at $T_1 = 0$).

If the interarrival times $\{\delta_i\}$ and service times $\{\sigma_j\}$ are random variables then we say that $\bar{W}(u)$ is the almost sure long-run average waiting time of customers routed to the server according to routing sequence u if $\lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{j=1}^m W_u(j) = \bar{W}(u)$ with probability one. So, if in case of random variables $\lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{j=1}^m W_u(j)$ almost surely (a.s.) exists then we say that $\bar{W}(u)$ exists (also in this case we assume for convenience that the initial workload $w = 0$).

Theorem 3.3.8 has the following corollary.

Corollary 3.3.9 *Let $\{\delta_i\}$, u and v be as in Theorem 3.3.8. If $\bar{W}(u)$ and $\bar{W}(v)$ exist then*

$$\bar{W}(v) - \bar{W}(u) \leq \delta \cdot \frac{T}{k} \cdot (\bar{I}(v) - \bar{I}(u)).$$

From ergodic theory we have the following theorem for stochastic interarrival and service times. In fact a more general result will be proved in the appendix (see Theorem 3.6.1).

Theorem 3.3.10 *Suppose that the interarrival times $\{\delta_i\}$ of customers arriving at the system are independent and identically distributed (i.i.d.) random variables with mean δ and the service times $\{\sigma_j\}$ of the considered server are i.i.d. random variables with mean σ which are independent of the interarrival times. Further let u' and u'' be routing sequences of zeros and ones that are both representatives of some $u \in \mathcal{P}(d)$ with $d \in \mathbb{Q}$ and $\frac{\sigma}{\delta} \cdot d < 1$. Then $\overline{W}(u')$ and $\overline{W}(u'')$ exist and are finite. Moreover $\overline{W}(u') = \overline{W}(u'')$ and the long-run average waiting time does not depend on the initial workload w .*

Notation. Let δ , σ and d be as in Theorem 3.3.10. Then we say that $\rho := \frac{\sigma}{\delta} \cdot d$ is the traffic intensity for the server.

By Theorem 3.3.10 all the routing sequences which are representatives of some given $u \in \mathcal{P}(d)$ have the same long-run average waiting time if the stability condition $\rho < 1$ is fulfilled. Therefore we denote this long-run average waiting time simply by $\overline{W}(u)$ for $u \in \mathcal{P}(d)$.

Theorem 3.3.11 *Let the interarrival times $\{\delta_i\}$ and the service times $\{\sigma_j\}$ be as in Theorem 3.3.10. Further let $u, v \in \mathcal{P}(d)$ for some $d \in \mathbb{Q}$ such that the traffic intensity $\rho < 1$ and $u \preceq_{\overline{g}} v$. Then*

$$\overline{W}(v) - \overline{W}(u) \leq \frac{\delta}{d}(\overline{I}(v) - \overline{I}(u)).$$

Proof. Since $u, v \in \mathcal{P}(d)$ with $d \in \mathbb{Q}$ and $u \preceq_{\overline{g}} v$ there exist $T, k \in \mathbb{N}$, $u', v' \in \mathcal{P}(T, k)$ and routing sequences u'', v'' such that $d = \frac{k}{T}$, $\overline{w}(T, k) \preceq u' \preceq v'$, $u'' = (u')^\infty$ is representative of u and $v'' = (v')^\infty$ is representative of v . By Theorem 3.3.10 we have that both $\overline{W}(u) = \overline{W}(u'')$ and $\overline{W}(v) = \overline{W}(v'')$ exist. Let the sequence $\{\delta_i(u'')\}_{i=1,2,\dots}$ be a realisation of the sequence of interarrival times when routing sequence u'' is used and let the sequence $\{\delta_i(v'')\}_{i=1,2,\dots}$ be a realisation of the interarrival times when routing sequence v'' is used. Because they are sequences of i.i.d. random variables we make the coupling that $\delta_i(u'') = \delta_i(v'') = \delta_i$ for $i = 1, 2, \dots$ where the sequence $\{\delta_i\}$ is a realisation of the sequence of interarrival times. Let the sequence $\{\sigma_j(u'')\}_{j=1,2,\dots}$ be a realisation of the sequence of service times when routing sequence u'' is used and let the sequence $\{\sigma_j(v'')\}_{j=1,2,\dots}$ be a realisation of the service times when routing sequence v'' is used. Because they are sequences of i.i.d random variables we make the coupling that $\sigma_j(u'') = \sigma_j(v'') = \sigma_j$ for $j = 1, 2, \dots$ where the sequence $\{\sigma_j\}$ is a realisation of the sequence of service times. Let $W_{u''}^c(j), W_{v''}^c(j)$ be the workload for the server at the moment the j -th customer

is routed to the server according to routing sequence u'' and v'' , respectively, with this coupling of the interarrival times and service times. Then by Theorem 3.3.10 $\lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{j=1}^m W_{u''}^c(j) = \bar{W}(u'') = \bar{W}(u)$ with probability one and $\lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{j=1}^m W_{v''}^c(j) = \bar{W}(v'') = \bar{W}(v)$ with probability one. Further for every $T \in \mathbb{N}$ and $l \in \mathbb{N}$ we have that $\lim_{t \rightarrow \infty} \frac{1}{t} \cdot \sum_{i=0}^{t-1} \delta_{i \cdot T + l} = \delta$ with probability one and thus we have analogously to Theorem 3.3.8 that

$$\limsup_{m \rightarrow \infty} \frac{1}{m} \cdot \left\{ \sum_{j=1}^m W_{v''}^c(j) - \sum_{j=1}^m W_{u''}^c(j) \right\} \leq \delta \cdot \frac{T}{k} \cdot (\bar{I}(v) - \bar{I}(u)) = \frac{\delta}{d} (\bar{I}(v) - \bar{I}(u))$$

with probability one. Hence $\bar{W}(v) - \bar{W}(u) \leq \frac{\delta}{d} (\bar{I}(v) - \bar{I}(u))$. \square

By Theorem 3.3.11 and the results in the previous section we obtain the following corollary.

Corollary 3.3.12 *Let the interarrival times $\{\delta_i\}$ and the service times $\{\sigma_j\}$ be as in Theorem 3.3.10. Further let $u \in \mathcal{P}(d)$ for some $d \in \mathbb{Q}$ such that $\rho < 1$ and let $\omega = \omega(d)$ be the regular sequence of density d . Then*

$$\bar{W}(u) \leq \bar{W}(\omega) + \frac{\delta}{d} \cdot \bar{I}(u).$$

We also have the dual results of Theorem 3.3.11 and Corollary 3.3.12. In these dual results the dual unbalance \underline{I} is used instead of the primal unbalance \bar{I} . These can be proved completely analogously using the dual order instead of the primal order. The dual results of Theorem 3.3.11 and Corollary 3.3.12 are the following.

Theorem 3.3.13 *Let the interarrival times $\{\delta_i\}$ and the service times $\{\sigma_j\}$ be as in Theorem 3.3.10. Further let $u, v \in \mathcal{P}(d)$ for some $d \in \mathbb{Q}$ such that $\rho < 1$ and $u \leq_g v$. Then*

$$\bar{W}(v) - \bar{W}(u) \geq \frac{\delta}{d} (\underline{I}(u) - \underline{I}(v)).$$

Corollary 3.3.14 *Let the interarrival times $\{\delta_i\}$ and the service times $\{\sigma_j\}$ be as in Theorem 3.3.10. Further let $u \in \mathcal{P}(d)$ for some $d \in \mathbb{Q}$ such that $\rho < 1$ and let $\omega = \omega(d)$ be the regular sequence of density d . Then*

$$\bar{W}(u) \geq \bar{W}(\omega) - \frac{\delta}{d} \cdot \underline{I}(u).$$

Remark. Contrary to Corollary 3.3.12 this dual result Corollary 3.3.14 is not useful for this type of routing system, because if u and ω are as in these corollaries then it is known that $\overline{W}(u) \geq \overline{W}(\omega)$ (See [3] and [5]).

In Theorem 3.3.10 we assume that the interarrival times $\{\delta_i\}$ are i.i.d. random variables with mean δ and the service times $\{\sigma_j\}$ are i.i.d. random variables with mean σ . So, the results hold in particular if the interarrival and service times are deterministic, i.e the interarrival times are all equal to δ and the service times are all equal to σ . Further in this case the assumption that the traffic intensity $\rho < 1$ can be relaxed to $\rho \leq 1$. Since the long-run average waiting time depends on the initial workload w if $\rho = 1$ we assume that the queue is empty at $T_1 = 0$, i.e the initial workload $w = 0$ in this case. In [46] it is shown that Theorem 3.3.10 (except for the statement that the long-run average waiting does not depend on the initial workload) in this deterministic case also holds for $\rho = 1$. Since $\rho < 1$ was only needed to prove Theorem 3.3.10 it follows that the results following after Theorem 3.3.10 also hold for $\rho = 1$ in this deterministic case. In fact for routing to a deterministic queue and deterministic interarrival times with traffic intensity $\rho = 1$ we have, instead of Theorem 3.3.11 and Corollary 3.3.12, the following stronger result.

Theorem 3.3.15 *Suppose that the interarrival times and service times are deterministic and equal to δ and σ , respectively. Further suppose that the queue is initially empty. Let $u \in \mathcal{P}(d)$ for some $d \in \mathbb{Q}$ such that $\delta = d\sigma$, hence $\rho = 1$, and let $\omega = \omega(d)$ be the regular sequence of density d . Then*

$$\overline{W}(u) = \overline{W}(\omega) + \frac{\delta}{d} \cdot \overline{I}(u).$$

Proof. Let $d = \frac{k}{T}$ with $\gcd(k, T) = 1$. Let $\overline{\omega} = (\overline{\omega}(T, k))^\infty$, which is a representative of ω . Let $u' \in \mathcal{P}(T, k)$ such that $\overline{\omega}(T, k) \preceq u'$ and $u'' := (u')^\infty$ is a representative of u . From Chapter 2 (see Theorem 2.3.2 and Lemma 2.6.2) we have that $\overline{W}(\omega) = \overline{W}(\overline{\omega}) = \lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{j=1}^m W_{\overline{\omega}}(j)$ and $\overline{W}(u) = \overline{W}(u'') = \lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{j=1}^m W_{u''}(j)$. For both routing sequence $\overline{\omega}$ and u'' the server is always busy serving customers from $T_1 = 0$ on. Indeed, let $t_0 \in \mathbb{R}_{\geq 0}$ be arbitrarily. Since customers arrive at times $0, \delta, 2\delta, \dots$ the number of customers routed to the server before time t_0 according to routing sequence $\overline{\omega}$ is $\kappa_{\overline{\omega}}(\lceil \frac{t_0}{\delta} \rceil) \geq \lceil \frac{t_0}{\delta} d \rceil = \lceil \frac{t_0}{\sigma} \rceil$. Since the server could not start earlier with the serving of these customers than at time $T_1 = 0$ and each of them has a serving time of σ , the server is not finished with serving these customers before time $\sigma \cdot \lceil \frac{t_0}{\sigma} \rceil \geq t_0$. So, for routing sequence $\overline{\omega}$ the server is always busy from $T_1 = 0$ on and this follows analogously for routing sequence u'' . From this and the fact that the server is assumed to be empty at $T_1 = 0$ it follows that

$W_{u''}(j) + \lambda_{u''}(j) = W_{\bar{w}}(j) + \lambda_{\bar{w}}(j) = (j - 1)\sigma$ for every $j \in \mathbb{N}$. Therefore, for every $m \in \mathbb{N}$, (3.4) holds with equality for u'' and \bar{w} with $\delta_i = \delta$ for $i = 1, 2, \dots$. Thus

$$\sum_{j=1}^m W_{u''}(j) - \sum_{j=1}^m W_{\bar{w}}(j) = \sum_{j=1}^m \sum_{i=\nu_{u''}(j)}^{\nu_{\bar{w}}(j)-1} \delta = \delta \cdot N_{\bar{w}u''}(m)$$

for every $m \in \mathbb{N}$. Hence by Corollary 3.3.7 we have that

$$\bar{W}(u) - \bar{W}(\omega) = \delta \cdot \lim_{m \rightarrow \infty} \frac{N_{\bar{w}u''}(m)}{m} = \delta \cdot \frac{T}{k} \cdot \bar{I}(u'') = \frac{\delta}{d} \cdot \bar{I}(u).$$

□

According to Theorem 3.3.15 the inequalities of Theorem 3.3.11 and Corollary 3.3.12 hold with equality for routing to a deterministic queue with traffic intensity $\rho = 1$ if the interarrival times are deterministic. For non-deterministic interarrival and service times the inequalities of Theorem 3.3.11 and Corollary 3.3.12 are rather tight if ρ is close to 1, or in other words if the traffic is heavy.

3.4 Routing to parallel queues

In this section we derive upper bounds on the average waiting time for routing customers to parallel queues according to some (periodic) routing policy. We assume that we have $N \geq 2$ parallel servers. Then a routing policy ψ corresponds to an \mathbb{N} -word (integer sequence) $U = (U_1, U_2, \dots)$ on the alphabet $\{1, 2, \dots, N\}$, where U_i is the server to which the i -th arriving customer is routed according to policy ψ . If the applied policy and thus also the corresponding word are both periodic we have for some $T \in \mathbb{N}$ that U_1, U_2, \dots, U_T is a primitive period cycle of the word. Then we say that T is the period of the applied policy ψ and we denote ψ and the corresponding word U as $(U_1, U_2, \dots, U_T)^\infty$.

Definition 3.4.1 *Let ψ and U be as above. Then for $t \in \mathbb{N}$ we define $W(t) = W_\psi(t)$ as the waiting time of the t -th arriving customer if policy ψ is applied, which is the remaining workload for server U_t at the moment that the t -th customer arrives. If the interarrival times and service times of all servers are fixed sequences then we say that $\bar{W}(\psi) = \bar{W}(U)$ is the long-run average waiting time of the arriving customers routed according to policy ψ if $\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=1}^{\tau} W_\psi(t) = \bar{W}(\psi)$ and the system is empty at $T_1 = 0$. If the interarrival times and service times of the several servers are random variables then we say that $\bar{W}(\psi) = \bar{W}(U)$ is the almost sure*

long-run average waiting time of the arriving customers routed according to policy ψ if $\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=1}^{\tau} W_{\psi}(t) = \overline{W}(\psi)$ with probability one and the system is empty at $T_1 = 0$.

Notation. If U is an \mathbb{N} -word (not necessarily periodic) on the alphabet $\{1, 2, \dots, N\}$ (corresponding to some routing policy) then for $i = 1, 2, \dots, N$ we denote by $u^i = (u_1^i, u_2^i, \dots)$ the corresponding routing sequence of zeros and ones for server i , i.e.

$$u_t^i = \begin{cases} 1 & \text{if } U_t = i \\ 0 & \text{if } U_t \neq i \end{cases}. \text{ For } i = 1, 2, \dots, N \text{ and } t \in \mathbb{N} \text{ we put } A^t(i) := \{t' \in$$

$\mathbb{N} : u_{t'}^i = 1\} \cap \{1, 2, \dots, t\}$. Let $d_1, d_2, \dots, d_N \in \mathbb{R}_{>0}$ with $\sum_{i=1}^N d_i = 1$. We denote by $\mathcal{S}(d_1, d_2, \dots, d_N)$ all the infinite \mathbb{N} -words $U = (U_1, U_2, \dots)$ on the alphabet $\{1, 2, \dots, N\}$ such that u^i has density d_i for $i = 1, 2, \dots, N$. If U is periodic then we have for every $i \in \{1, 2, \dots, N\}$ that $u^i \in \mathcal{P}(d)$ for some $d \in \mathbb{Q}$. Let $d_1, d_2, \dots, d_N \in \mathbb{Q}_{>0}$ such that $\sum_{i=1}^N d_i = 1$. Then we denote by $\mathcal{Q}(d_1, d_2, \dots, d_N)$ all the infinite periodic \mathbb{N} -words U on the alphabet $\{1, 2, \dots, N\}$ for which $u^i \in \mathcal{P}(d_i)$ for $i = 1, 2, \dots, N$. Further for $T, k_1, k_2, \dots, k_N \in \mathbb{N}$ with $\sum_{i=1}^N k_i = T$ we denote by $\mathcal{Q}(\{T\}, k_1, k_2, \dots, k_N)$ all the infinite \mathbb{N} -words on the alphabet $\{1, 2, \dots, N\}$ for which every subword of length T contains exactly k_i letters i for $i = 1, 2, \dots, N$.

Note that $\mathcal{Q}(\{T\}, k_1, k_2, \dots, k_N) \subseteq \mathcal{Q}(\frac{k_1}{T}, \frac{k_2}{T}, \dots, \frac{k_N}{T})$ and $\mathcal{Q}(d_1, d_2, \dots, d_N) \subseteq \mathcal{S}(d_1, d_2, \dots, d_N)$. Let ψ be a periodic routing policy with period T and U the corresponding word on the alphabet $\{1, 2, \dots, N\}$. Then there exist non-negative integers k_1, k_2, \dots, k_N with $\sum_{i=1}^N k_i = T$ such that from T consecutively arriving customers policy ψ routes exactly k_i of them to server i for $i = 1, 2, \dots, N$. Hence $U \in \mathcal{Q}(\{T\}, k_1, k_2, \dots, k_N) \subseteq \mathcal{Q}(\frac{k_1}{T}, \frac{k_2}{T}, \dots, \frac{k_N}{T}) \subseteq \mathcal{S}(\frac{k_1}{T}, \frac{k_2}{T}, \dots, \frac{k_N}{T})$. Vice versa if $U \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ with $d_i \in \mathbb{Q}_{>0}$ for $i = 1, 2, \dots, N$ and $\sum_{i=1}^N d_i = 1$ then there exist $T, k_1, k_2, \dots, k_N \in \mathbb{N}$ with $\sum_{i=1}^N k_i = T$ such that $U \in \mathcal{Q}(\{T\}, k_1, k_2, \dots, k_N)$. Further for every server $i \in \{1, 2, \dots, N\}$ we have a corresponding routing sequence $u^i = (u_1^i, u_2^i, \dots, u_T^i)^\infty \in \mathcal{P}(d_i)$ where $d_i = \frac{k_i}{T}$. For example if $U = (1, 2, 1, 2, 1, 3, 1, 2, 1, 1, 2, 3)^\infty$ then $U \in \mathcal{Q}(\{12\}, 6, 4, 2) \subseteq \mathcal{Q}(\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$. Further $u^1 = (1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0)^\infty \in \mathcal{P}(\frac{1}{2})$, $u^2 = (0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0)^\infty \in \mathcal{P}(\frac{1}{3})$ and $u^3 = (0, 0, 0, 0, 0, 1)^\infty \in \mathcal{P}(\frac{1}{6})$.

Let $\{\delta_i\}$ be the sequence of interarrival times and let $\{\sigma_j^i\}$ be the sequence of service times of server i for $i = 1, 2, \dots, N$, i.e. σ_j^i is the service time of the j -th customer that is routed to server i . We define $\overline{W}^i(\psi) = \overline{W}^i(u^i)$ as the long-run average waiting time of customers routed to server i if policy ψ is applied in the same way as in the previous section. The only differences are that routing sequence u is replaced

by routing sequence u^i and the sequence of service times $\{\sigma_j\}$ is replaced with the sequence of service times $\{\sigma_j^i\}$. We have the following theorem.

Theorem 3.4.2 *Suppose that the interarrival times $\{\delta_i\}$ of customers arriving at the system are i.i.d. random variables with mean δ and for every $i \in \{1, 2, \dots, N\}$ the service times $\{\sigma_j^i\}$ are i.i.d. random variables with mean σ_i which are independent of the interarrival times. Let ψ be a routing policy that corresponds to some word $U \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ such that $\frac{\sigma_i}{\delta} \cdot d_i < 1$ for $i = 1, 2, \dots, N$. Then $\bar{W}(\psi)$ exists and is finite. Moreover $\bar{W}(\psi) = \sum_{i=1}^N d_i \cdot \bar{W}^i(\psi)$.*

Proof. We have that $u^i \in \mathcal{P}(d_i)$ and $\frac{\sigma_i}{\delta} \cdot d_i < 1$ for $i = 1, 2, \dots, N$. Hence by Theorem 3.3.10 we have for every $i \in \{1, 2, \dots, N\}$ that $\bar{W}^i(\psi) = \bar{W}^i(u^i)$ exists and is finite. Hence

$$\begin{aligned} \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \cdot \sum_{t=1}^{\tau} W_{\psi}(t) &= \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{i=1}^N \sum_{t \in A^{\tau}(i)} W_{\psi}(t) = \\ \lim_{\tau \rightarrow \infty} \sum_{i=1}^N d_i \sum_{t \in A^{\tau}(i)} \frac{W_{\psi}(t)}{|A^{\tau}(i)|} &= \sum_{i=1}^N d_i \cdot \bar{W}^i(\psi). \end{aligned}$$

□

Notation. Let $\psi, \delta, \sigma_i, d_i$ be as in Theorem 3.4.2 for $i = 1, 2, \dots, N$. Then we say that $\rho_i := \frac{\sigma_i}{\delta} \cdot d_i$ is the traffic intensity for server i if policy ψ is applied.

Remark. If the interarrival times and the service times are deterministic then for Theorem 3.4.2 it suffices that $\rho_i \leq 1$ for $i = 1, 2, \dots, N$, since in that case $\bar{W}^i(\psi)$ exists and is finite.

Definition 3.4.3 *Let $U \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ for some $d_i \in \mathbb{Q}_{>0}$ with $\sum_{i=1}^N d_i = 1$. Then we define the total primal unbalance \bar{O} of U by*

$$\bar{O}(U) := \sum_{i=1}^N \bar{I}(u^i) \tag{3.5}$$

and the total dual unbalance \underline{O} of U by

$$\underline{O}(U) := \sum_{i=1}^N \underline{I}(u^i). \tag{3.6}$$

Further we define the partial orders $\underline{\leq}_{\bar{g}}$, $\underline{\leq}_g$ and $\underline{\leq}_g$ on $\mathcal{Q}(d_1, d_2, \dots, d_N)$ in the following way. If $U, V \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ then $U \underline{\leq}_{\bar{g}} V$ if $u^i \underline{\leq}_{\bar{g}} v^i$ for every $i \in \{1, 2, \dots, N\}$, $U \underline{\leq}_g V$ if $u^i \underline{\leq}_g v^i$ for every $i \in \{1, 2, \dots, N\}$ and $U \underline{\leq}_g V$ if $u^i \underline{\leq}_g v^i$ for every $i \in \{1, 2, \dots, N\}$.

If it is clear what is meant we just say unbalance and not total unbalance. Suppose that $U, V \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ for $d_1, d_2, \dots, d_N \in \mathbb{Q}_{>0}$ with $\sum_{i=1}^N d_i = 1$. Then we say that U, V are mirrors of each other if there exist U', V' such that U' and V' are period cycles of U respectively V and $V' = \bar{U}'$.

Theorem 3.4.4 *Suppose that $U, V \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ are mirrors of each other. Then $\bar{O}(U) = \underline{O}(V)$ and $\underline{O}(U) = \bar{O}(V)$.*

Proof. Since U and V are mirrors of each other it follows for every $i \in \{1, 2, \dots, N\}$ that $u^i, v^i \in \mathcal{P}(d_i)$ are mirrors of each other. Thus by Corollary 3.2.32 we have for $i = 1, 2, \dots, N$ that $\bar{I}(u^i) = \underline{I}(v^i)$ and

$$\bar{O}(U) = \sum_{i=1}^N \bar{I}(u^i) = \sum_{i=1}^N \underline{I}(v^i) = \underline{O}(V).$$

Analogously it follows that $\underline{O}(U) = \bar{O}(V)$. □

Example. Let $U = (1, 2, 3, 1, 4, 2, 1, 3, 1, 2, 4, 3)^\infty \in \mathcal{Q}(\{12\}, 4, 3, 3, 2) \subseteq \mathcal{Q}(\frac{1}{3}, \frac{1}{4}, \frac{1}{4}, \frac{1}{6})$. Then $\bar{I}(u^1) = \frac{1}{12}$, $\bar{I}(u^2) = 0$, $\bar{I}(u^3) = \frac{1}{12}$, $\bar{I}(u^4) = 0$ and thus $\bar{O}(U) = \frac{1}{6}$. Further $\underline{I}(u^1) = \frac{1}{4}$, $\underline{I}(u^2) = 0$, $\underline{I}(u^3) = \frac{1}{6}$, $\underline{I}(u^4) = 0$ and thus $\underline{O}(U) = \frac{5}{12}$. For $V = (1, 2, 4, 1, 3, 2, 1, 3, 4, 2, 1, 3)^\infty \in \mathcal{Q}(\{12\}, 4, 3, 3, 2) \subseteq \mathcal{Q}(\frac{1}{3}, \frac{1}{4}, \frac{1}{4}, \frac{1}{6})$ which is a mirror of U the reader can verify that $\bar{O}(V) = \frac{5}{12}$ and $\underline{O}(V) = \frac{1}{6}$.

From Theorem 3.3.11 and Theorem 3.4.2 we deduce the following theorem.

Theorem 3.4.5 *Let the interarrival times $\{\delta_i\}$ and the service times $\{\sigma_j^i\}$ for $i = 1, 2, \dots, N$ be as in Theorem 3.4.2. Let $d_i \in \mathbb{Q}_{>0}$ for $i = 1, 2, \dots, N$ with $\sum_{i=1}^N d_i = 1$. Suppose that ψ, ψ' are routing policies with corresponding words $U, V \in \mathcal{Q}(d_1, d_2, \dots, d_N)$, respectively. Suppose that $\rho_i < 1$ for $i = 1, 2, \dots, N$ and $U \underline{\leq}_{\bar{g}} V$. Then*

$$\bar{W}(\psi') - \bar{W}(\psi) \leq \delta \cdot (\bar{O}(V) - \bar{O}(U)).$$

Proof. We have that

$$\bar{W}(\psi') - \bar{W}(\psi) = \sum_{i=1}^N d_i \cdot \bar{W}^i(v_i) - \sum_{i=1}^N d_i \cdot \bar{W}^i(u_i) = \sum_{i=1}^N d_i \cdot (\bar{W}^i(v_i) - \bar{W}^i(u_i)) \leq$$

$$\sum_{i=1}^N d_i \cdot \left(\frac{\delta}{d_i}\right) \cdot (\bar{I}(v^i) - \bar{I}(u^i)) = \delta \cdot (\bar{O}(V) - \bar{O}(U)).$$

□

Analogously we have by Theorem 3.3.13 and Theorem 3.4.2 the following theorem.

Theorem 3.4.6 *Let the interarrival times $\{\delta_i\}$ and the service times $\{\sigma_j^i\}$ for $i = 1, 2, \dots, N$ be as in Theorem 3.4.2. Let $d_i \in \mathbb{Q}_{>0}$ for $i = 1, 2, \dots, N$ with $\sum_{i=1}^N d_i = 1$. Suppose that ψ, ψ' are routing policies with corresponding words $U, V \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ respectively. Suppose that $\rho_i < 1$ for $i = 1, 2, \dots, N$ and $U \leq_g V$. Then*

$$\bar{W}(\psi') - \bar{W}(\psi) \geq \delta \cdot (Q(U) - Q(V)).$$

Let the interarrival times $\{\delta_i\}$ and the service times $\{\sigma_j^i\}$ for $i = 1, 2, \dots, N$ be as in Theorem 3.4.2. Further let $d_i \in \mathbb{Q}_{>0}$ for $i = 1, 2, \dots, N$ with $\sum_{i=1}^N d_i = 1$. If $\bar{W}^i(\omega(d_i))$ exists and is finite for $i = 1, 2, \dots, N$ then we put

$$\tilde{R} = \tilde{R}(d_1, d_2, \dots, d_N) := \sum_{i=1}^N d_i \cdot \bar{W}^i(\omega(d_i)). \quad (3.7)$$

Remark. It is easily seen that if the interarrival times and service times are random variables then for the existence of \tilde{R} it suffices that $\rho_i < 1$ for $i = 1, 2, \dots, N$, while if the interarrival times and service times are deterministic then it suffices that $\rho_i \leq 1$ for $i = 1, 2, \dots, N$. Further the definition of \tilde{R} and these results about the existence of \tilde{R} can be extended to the case of irrational d_i . Namely, let $d_i \in \mathbb{R}_{>0}$ be such that $\rho_i < 1$. Then in [3] and [5] it is shown that $W^i(\omega)$ exists and is the same for every regular sequence ω of zeros and ones of density d_i . According to Theorem 2.7.11 this is also true for the case of $\rho_i = 1$ and deterministic interarrival and service times. Moreover, in Lemma 3.4.8 $W^i(\omega)$ is given explicitly for that case. So, if $d_i \in \mathbb{R}_{>0}$ for $i = 1, 2, \dots, N$ with $\sum_{i=1}^N d_i = 1$ and $\rho_i < 1$, or $\rho_i = 1$ in case of deterministic interarrival and service times, then \tilde{R} is well defined as follows:

$$\tilde{R} = \tilde{R}(d_1, d_2, \dots, d_N) := \sum_{i=1}^N d_i \cdot \bar{W}^i(\omega^i), \quad (3.8)$$

where ω^i is any regular sequence of zeros and ones of density d_i for $i = 1, 2, \dots, N$.

In general \tilde{R} depends on the distribution of the interarrival times and service times and in some cases it is possible to compute \tilde{R} explicitly. In this chapter we will do this in some examples and in Chapter 4 we give an algorithm for computing \tilde{R} explicitly for any densities in case of deterministic interarrival and service times.

By Corollary 3.3.12 and Theorem 3.4.2 we have the following theorem.

Theorem 3.4.7 *Let the interarrival times $\{\delta_i\}$ and the service times $\{\sigma_j^i\}$ for $i = 1, 2, \dots, N$ be as in Theorem 3.4.2. Let $d_i \in \mathbb{Q}_{>0}$ for $i = 1, 2, \dots, N$ with $\sum_{i=1}^N d_i = 1$. Suppose that a routing policy ψ is applied with corresponding word $U \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ and $\rho_i < 1$ for $i = 1, 2, \dots, N$. Then*

$$\overline{W}(\psi) - \tilde{R} \leq \delta \cdot \overline{O}(U).$$

From Theorem 2.7.11 we have an explicit formula for $\overline{W}^i(\omega^i)$ in case of $\rho_i = 1$, deterministic interarrival times and the server i has deterministic service times. The formula is given for interarrival times δ equal to 1, but that is no restriction since the time-unit can always be chosen such that $\delta = 1$. So, for general δ we can just multiply by δ and we get the following lemma.

Lemma 3.4.8 *Suppose that the service times $\{\sigma_j^i\}$ of some server i are deterministic equal to σ_i and the interarrival times $\{\delta_i\}$ are deterministic and equal to δ . Let ω^i be a regular sequence of zeros and ones with density $d_i \in \mathbb{R}_{>0}$ and suppose that the traffic intensity $\rho_i = d_i \cdot \frac{\sigma_i}{\delta}$ for server i is equal to 1. If d_i is rational, $d = \frac{p_i}{q_i}$ with $p_i, q_i \in \mathbb{N}$ and $\gcd(p_i, q_i) = 1$ then*

$$\overline{W}(\omega^i) = \delta \cdot \left(\frac{1}{2} - \frac{1}{2p_i} \right). \quad (3.9)$$

If d_i is irrational then $\overline{W}(\omega^i) = \frac{\delta}{2}$.

For the same case of deterministic interarrival and service times but with $\rho_i < 1$ methods to calculate $\overline{W}(\omega^i)$ for a regular sequence ω^i are given in [23] and Chapter 4 of this thesis. Thus in case of deterministic interarrival and service times $\tilde{R}(d_1, d_2, \dots, d_N)$ can be determined exactly for any $d_1, d_2, \dots, d_N \in \mathbb{R}_{>0}$ with $\sum_{i=1}^N d_i = 1$ if $\rho_i \leq 1$ for $i = 1, 2, \dots, N$. In the example at the end of this section \tilde{R} is determined in a special case with Poisson arrivals and exponentially distributed service times. For general bounds on \tilde{R} see [62].

We have the following theorem which emphasizes the sharpness of the bound of Theorem 3.4.7.

Theorem 3.4.9 *Let the interarrival times $\{\delta_i\}$ be deterministic equal to δ and the service times $\{\sigma_j^i\}$ be deterministic equal to σ_i for $i = 1, 2, \dots, N$. Let $d_i \in \mathbb{Q}_{>0}$ for $i = 1, 2, \dots, N$ with $\sum_{i=1}^N d_i = 1$. Let $p_i, q_i \in \mathbb{N}$ be such that $d_i = \frac{p_i}{q_i}$ with $\gcd(p_i, q_i) = 1$ for $i = 1, 2, \dots, N$. Suppose that a routing policy ψ is applied with corresponding word $U \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ and $\rho_i = 1$ for $i = 1, 2, \dots, N$. Then*

$$\bar{W}(\psi) = \tilde{R} + \delta \cdot \bar{O}(U) = \delta \cdot \left(\frac{1}{2} - \sum_{i=1}^N \frac{1}{2q_i} + \bar{O}(U) \right).$$

Proof. According to the remark following Theorem 3.4.2 we can apply Theorem 3.4.2 and by Theorem 3.3.15 it follows that

$$\bar{W}(\psi) = \tilde{R} + \delta \cdot \bar{O}(U).$$

By (3.9) we have that

$$\tilde{R} = \sum_{i=1}^N d_i \cdot \bar{W}(\omega(d_i)) = \sum_{i=1}^N \frac{p_i}{q_i} \cdot \delta \cdot \left(\frac{1}{2} - \frac{1}{2p_i} \right) = \delta \cdot \left(\frac{1}{2} - \sum_{i=1}^N \frac{1}{2q_i} \right). \quad (3.10)$$

Example. We consider a queueing system with 3 parallel servers where the interarrival times are deterministic and equal to $\delta = 3$. The arriving jobs are routed to the servers according to the policy ψ that corresponds to the word

$$U = (1, 2, 1, 2, 1, 3, 1, 2, 1, 3)^\infty \in \mathcal{Q}(\{10\}, 5, 3, 2) \subseteq \mathcal{Q}\left(\frac{1}{2}, \frac{3}{10}, \frac{1}{5}\right).$$

All the service times are deterministic and given by $\sigma_1 = 6$, $\sigma_2 = 10$ and $\sigma_3 = 15$. Hence $\rho_i = 1$ for $i = 1, 2, 3$. Further $\bar{I}(u^1) = 0$, $\bar{I}(u^2) = \frac{1}{10}$, $\bar{I}(u^3) = \frac{1}{10}$ and thus $\bar{O}(U) = \frac{1}{5}$. So, according to Theorem 3.4.9 we have that

$$\bar{W}(\psi) = 3 \cdot \left(\frac{1}{2} - \left(\frac{1}{4} + \frac{1}{20} + \frac{1}{10} \right) + \frac{1}{5} \right) = \frac{9}{10}$$

which can be checked by direct calculation.

Theorem 3.4.10 *Let the interarrival times $\{\delta_i\}$ and the service times $\{\sigma_j^i\}$ for $i = 1, 2, \dots, N$ be as in Theorem 3.4.2. Let $d_i \in \mathbb{Q}_{>0}$ for $i = 1, 2, \dots, N$ with $\sum_{i=1}^N d_i = 1$. Suppose that a routing policy ψ is applied with corresponding word $U \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ and $\rho_i < 1$ for $i = 1, 2, \dots, N$. Then*

$$\tilde{R} \leq \bar{W}(\psi) \leq \tilde{R} + \delta \cdot \bar{O}(U).$$

Proof. From [3] and [5] we have that $\tilde{R} \leq \overline{W}(\psi)$ and according to Theorem 3.4.7 $\overline{W}(\psi) \leq \tilde{R} + \delta \cdot \overline{O}(U)$. \square

Note. Suppose that we have a queueing system where the arrivals are according to a Poisson process with parameter λ . Suppose that the service times of server i are exponentially distributed with parameter μ_i . If d_i , the fraction of jobs that is routed to server i , equals $\frac{1}{q_i}$ for some $q_i \in \mathbb{N}$, then $\overline{W}^i(\omega(d_i))$ can be calculated in the following way. For the routing sequence $\omega(d_i)$ among every q_i arriving jobs exactly one job is routed to server i . So, the interarrival times at server i consist of q_i Poisson arrivals with parameter λ . Hence the interarrival times for the queue of server i are Erlang distributed, that is, according to an $E_\lambda^{q_i}/M/1$ queue. Thus $\overline{W}^i(\omega(d_i))$ is the same as the average waiting time for a $E_\lambda^{q_i}/M/1$ queue, where the parameter of the service times is μ_i . So (see [27]) if $x_i \in (0, 1)$ is a solution of the equation

$$x = \left(\frac{\lambda}{\lambda + \mu_i - \mu_i \cdot x} \right)^{q_i}, \quad (3.11)$$

then $\overline{W}^i(\omega(d_i)) = \frac{x_i}{\mu_i \cdot (1 - x_i)}$.

In the following example we have calculated $\overline{W}^i(\omega(d_i))$ for $i = 1, 2, 3, 4$ in this way. Further we have explicitly calculated the lower bound \tilde{R} and upper bound $\tilde{R} + \delta \cdot \overline{O}(U)$ of Theorem 3.4.10 for $\overline{W}(\psi)$, where U is the word corresponding to the applied routing policy ψ .

Example. We consider a queueing system with 4 parallel servers where the arrivals are according to a Poisson process with parameter $\lambda = 11$. Hence for the mean interarrival time δ we have that $\delta = \frac{1}{11}$. The arriving jobs are routed to the servers according to the policy ψ that corresponds to the word

$$U = (1, 2, 3, 1, 4, 2, 1, 3, 1, 2, 4, 3)^\infty \in \mathcal{Q}(\{12\}, 4, 3, 3, 2) \subseteq \mathcal{Q}\left(\frac{1}{3}, \frac{1}{4}, \frac{1}{4}, \frac{1}{6}\right)$$

which we considered in a previous example. For every $i \in \{1, 2, 3, 4\}$ the service times are exponentially distributed with parameter μ_i and $\mu_1 = 4$, $\mu_2 = \mu_3 = 3$ and $\mu_4 = 2$. Then we find that $\overline{W}^1(\omega(\frac{1}{3})) = 1.7792$ (rounded to 4 decimals) and thus by Corollary 3.3.12 $\overline{W}^1(u^1) \leq 1.7792 + 3 \cdot \frac{1}{11} \cdot \frac{1}{12} = 1.8019$. Similarly $\overline{W}^2(\omega(\frac{1}{4})) = \overline{W}^2(u^2) = 2.2105$, $\overline{W}^3(\omega(\frac{1}{4})) = 2.2105$, $\overline{W}^3(u^3) \leq 2.2408$ and $\overline{W}^4(\omega(\frac{1}{6})) = \overline{W}^4(u^4) = 3.0732$. Hence

$$\tilde{R} = \frac{1}{3} \cdot \overline{W}^1(\omega(\frac{1}{3})) + \frac{1}{4} \cdot \overline{W}^2(\omega(\frac{1}{4})) + \frac{1}{4} \cdot \overline{W}^3(\omega(\frac{1}{4})) + \frac{1}{6} \cdot \overline{W}^4(\omega(\frac{1}{6})) = 2.2105.$$

and $\tilde{R} + \delta \cdot \overline{O}(U) = 2.2105 + \frac{1}{11} \cdot \frac{1}{6} = 2.2257$. So, by Theorem 3.4.10, $2.2105 \leq \overline{W}(\psi) \leq 2.2257$.

3.5 Billiard sequences and routing sequences

In this section we give some properties of sequences (words) and their implications for the corresponding routing policies. In particular we have some properties for so-called billiard sequences (see [12] and [15]). For every set of rational densities we obtain a billiard sequence with unbalance smaller or equal than $\frac{N}{2} - 1$, where N is the number of parallel queues. We derive a relation between the discrepancy function and the unbalance and several properties can be obtained from that. We show that for given rational densities there exists a periodic billiard sequence which has minimal unbalance among all sequences with those densities.

Theorem 3.5.1 *Let $d_1, d_2, \dots, d_N \in \mathbb{Q}_{>0}$ with $\sum_{i=1}^N d_i = 1$ for some $N \in \mathbb{Z}_{\geq 2}$. Then there exists a word $U \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ such that $\overline{O}(U) \leq \frac{N}{2} - 1$.*

Theorem 3.5.1 follows from Theorem 3.5.5 which we formulate and prove later. The dual of Theorem 3.5.1, which is a consequence of Theorem 3.5.1 and Theorem 3.4.4, is the following theorem.

Theorem 3.5.2 *Let $d_1, d_2, \dots, d_N \in \mathbb{Q}_{>0}$ with $\sum_{i=1}^N d_i = 1$ for some $N \in \mathbb{Z}_{\geq 2}$. Then there exists a word $U \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ such that $\underline{O}(U) \leq \frac{N}{2} - 1$.*

Theorem 3.5.1 together with Theorem 3.4.7 give the following corollary.

Corollary 3.5.3 *Let the interarrival times $\{\delta_i\}$, the service times $\{\sigma_j^i\}$ and d_i for $i = 1, 2, \dots, N$ be as in Theorem 3.4.7. Suppose that $\rho_i < 1$ for $i = 1, 2, \dots, N$. Then there exists a routing policy ψ corresponding to a word $U \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ such that*

$$\overline{W}(\psi) \leq \tilde{R}(d_1, d_2, \dots, d_N) + \delta \cdot \left(\frac{N}{2} - 1\right).$$

Theorem 3.5.5 shows how to construct a word $U \in \mathcal{Q}(d_1, d_2, \dots, d_N)$ such that $\overline{O}(U) \leq \frac{N}{2} - 1$. In Section 2.5 we introduced the SG (Special Greedy) algorithm. For the construction of U we apply the following SG algorithm.

SG algorithm. Let $d_1, d_2, \dots, d_N \in \mathbb{R}_{>0}$ with $\sum_{i=1}^N d_i = 1$ and $x_1, x_2, \dots, x_N \in \mathbb{R}_{\geq 0}$ be given. Then $U = (U_1, U_2, \dots)$ is determined inductively in the following way. Suppose that U_1, U_2, \dots, U_{n-1} have been determined and thus $\kappa_{u^i}(n-1)$ is known for $i = 1, 2, \dots, N$. Choose U_n such that $U_n = i \in \{1, 2, \dots, N\}$ for which $\frac{x_i + \kappa_{u^i}(n-1)}{d_i}$ is minimal.

If $0 \leq x_i \leq 1$ for $i = 1, 2, \dots, N$ then a word U constructed by the above SG algorithm is known in the literature as a billiard word or sequence (see for example [12]). Indeed, if a billiard ball travels in an N -dimensional cube with elastic reflection on the sides of the cube and a sequence of integers from $\{1, 2, \dots, N\}$ (the integers correspond to the sides of the cube and opposite sides correspond to the same integer) is constructed by writing down an integer if the ball reflects against a side of the cube corresponding to that integer then such a sequence is called a billiard sequence. It is easily seen that the word constructed by the SG algorithm is a billiard sequence obtained by starting in position (x_1, x_2, \dots, x_N) and having initial velocity vector $(-d_1, -d_2, \dots, -d_N)$ in the unit cube $[0, 1]^N$. In the SG algorithm the choice of U_n is not determined if $\frac{x_i + \kappa_{u^i}(n-1)}{d_i}$ is minimal for several $i \in \{1, 2, \dots, N\}$. We assume that in such cases a consistent choice is made. For example always the smallest i is chosen for which $\frac{x_i + \kappa_{u^i}(n-1)}{d_i}$ is minimal. If this SG algorithm is used then the word obtained by the algorithm is determined by $x_1, x_2, \dots, x_N \in \mathbb{R}_{\geq 0}$. For billiard sequences this consistent choice means that if the billiard ball hits multiple sides at the same time then the order in which the letters corresponding to those sides appear in the billiard sequence is prescribed. We will call such a billiard sequence a consistent billiard sequence. More precisely, we define a consistent billiard sequence on the alphabet $\{0, 1, \dots, N\}$ to be a sequence (\mathbb{N} -word) that can be obtained by applying the SG algorithm in a consistent way to some $0 \leq x_i < 1$ for $i = 1, 2, \dots, N$. Note that we take $x_i < 1$ instead of $x_i \leq 1$ for $i = 1, 2, \dots, N$. The reason for this becomes clear in the proof of Lemma 3.5.4. A billiard sequence that is not consistent can be aperiodic even if the densities d_1, d_2, \dots, d_N are all positive rational numbers. However, for consistent billiard sequences we have the following lemma.

Lemma 3.5.4 *Let $d_1, d_2, \dots, d_N \in \mathbb{Q}_{>0}$ with $\sum_{i=1}^N d_i = 1$. Let $p_i, q_i, k_i \in \mathbb{N}$ for $i = 1, 2, \dots, N$ and $T \in \mathbb{N}$ be such that $d_i = \frac{p_i}{q_i} = \frac{k_i}{T}$ with $\gcd(p_i, q_i) = 1$ for $i = 1, 2, \dots, N$ and $\gcd(k_1, k_2, \dots, k_N) = 1$. Let U be a consistent billiard sequence with these densities. Then $U \in \mathcal{Q}(\{T\}, k_1, k_2, \dots, k_N) \subseteq \mathcal{Q}(d_1, d_2, \dots, d_N)$.*

Proof. The consistent billiard sequence U can be constructed by starting in (x_1, x_2, \dots, x_N) with $0 \leq x_i < 1$ for $i = 1, 2, \dots, N$ and initial velocity vector

$(-d_1, -d_2, \dots, -d_N)$. We assume that the velocity of the billiard ball is such that for every $i \in \{1, 2, \dots, N\}$ it takes $1/d_i$ time units to travel between the opposite sides of the cube corresponding to i . Then for every $t \geq 0$ the billiard ball has reflected $\lceil t \cdot d_i - x_i \rceil$ times on sides corresponding to letter i in the time interval $[0, t]$. Hence in the time interval $[0, T]$ the billiard ball has reflected $\lceil T \cdot d_i - x_i \rceil = k_i$ times on sides corresponding to i for $i = 1, 2, \dots, N$. Since $\sum_{i=1}^N k_i = T$ it follows that $\kappa_{u^i}(T) = k_i$ for every $i \in \{1, 2, \dots, N\}$ and $\frac{x_i + \kappa_{u^i}(T)}{d_i} = \frac{x_i}{d_i} + T$ for every $i \in \{1, 2, \dots, N\}$. Since U is obtained by applying the SG algorithm in a consistent way to x_1, x_2, \dots, x_N it follows that $U_{T+t} = U_t$ for every $t \in \{1, 2, \dots, T\}$ and by induction we have that U is periodic with period T . Hence $U \in \mathcal{Q}(\{T\}, k_1, k_2, \dots, k_N) \subseteq \mathcal{Q}(d_1, d_2, \dots, d_N)$. \square

Theorem 3.5.5 *Let d_i, p_i, q_i, k_i and T be as in Lemma 3.5.4, $N \geq 2$ and let $k \in \{1, 2, \dots, N\}$ be such that $q_k = \max_{i \in \{1, 2, \dots, N\}} q_i$. Put $x_k^* = 0$ and $x_j^* = 1 - \frac{1}{q_j}$ for every $j \in \{1, 2, \dots, N\}$ for which $j \neq k$. Construct a consistent billiard sequence U starting in $(x_1^*, x_2^*, \dots, x_N^*)$ with initial velocity vector $(-d_1, -d_2, \dots, -d_N)$ in the unit cube $[0, 1]^N$. Then $U \in \mathcal{Q}(\{T\}, k_1, k_2, \dots, k_N) \subseteq \mathcal{Q}(d_1, d_2, \dots, d_N)$. Moreover $\overline{O}(U) = 0$ if $N = 2$ and $\overline{O}(U) < \frac{N}{2} - 1$ if $N \geq 3$.*

We prove this theorem later. In the theorem we have chosen a particular starting point for the billiard sequence, but we think that this is not necessary. We have the following conjecture.

Conjecture 3.5.6 *For every consistent periodic billiard sequence U on N letters both $\overline{O}(U) \leq \frac{N}{2} - 1$ and $\underline{O}(U) \leq \frac{N}{2} - 1$ for $N \geq 2$.*

Billiard sequences have more nice properties. For example, in [60] there is a notion for (routing) sequences to be called m -balanced. Regular sequences are 1-balanced for that notion. From the results in that paper it follows that billiard sequences on an alphabet of N letters are $N - 1$ -balanced.

We need a number of lemmas and theorems in order to prove Theorem 3.5.5. These results may be of independent interest. A relation between the discrepancy function and the unbalance is derived in Theorem 3.5.10.

Lemma 3.5.7 *Let $p, q \in \mathbb{N}$. Then*

$$\sum_{n=1}^q \lceil n \cdot \frac{p}{q} \rceil = \frac{1}{2}(pq + p + q - \gcd(p, q)).$$

Proof. Put $p' = \frac{p}{\gcd(p,q)}$ and $q' = \frac{q}{\gcd(p,q)}$. Recall that for $a \in \mathbb{Z}$ and $b \in \mathbb{N}$ we denote by $a \pmod{b}$ the integer $a' \in \{0, 1, \dots, b-1\}$ such that $a = a' + kb$ for some $k \in \mathbb{Z}$. Then we have that

$$\begin{aligned} \sum_{n=1}^q (\lceil n \cdot \frac{p}{q} \rceil - n \cdot \frac{p}{q}) &= \frac{1}{q} \cdot \sum_{n=1}^q (-np) \pmod{q} = \\ \frac{\gcd(p,q)}{q} \cdot \sum_{n=1}^{q'} \gcd(p,q) \cdot (-np') &\pmod{q'} = \\ \frac{(\gcd(p,q))^2}{q} \sum_{i=0}^{q'-1} i &= \frac{(\gcd(p,q))^2}{q} \cdot \frac{q'}{2} \cdot (q' - 1). \end{aligned}$$

Hence

$$\begin{aligned} \sum_{n=1}^q \lceil n \cdot \frac{p}{q} \rceil &= \sum_{n=1}^q n \cdot \frac{p}{q} + \sum_{n=1}^q (\lceil n \cdot \frac{p}{q} \rceil - n \cdot \frac{p}{q}) = \\ \frac{p}{2} \cdot (q+1) + \frac{1}{2} \gcd(p,q) \cdot \left(\frac{q}{\gcd(p,q)} - 1 \right) &= \frac{1}{2} (pq + p + q - \gcd(p,q)). \end{aligned}$$

□

Recall the notation introduced in the previous section and the definition of the discrepancy function χ from Section 3.2.

Definition 3.5.8 Let $U \in \mathcal{S}(d_1, d_2, \dots, d_N)$ for some $d_1, d_2, \dots, d_N \in \mathbb{R}_{>0}$ with $\sum_{i=1}^N d_i = 1$. Then we define

$$c_i = c_i(U) = \sup_{n \in \mathbb{Z}_{\geq 0}} \chi_{u^i}(n)$$

for $i = 1, 2, \dots, N$.

Lemma 3.5.9 If $U \in \mathcal{Q}(\{T\}, k_1, k_2, \dots, k_N)$ for some $T, k_1, k_2, \dots, k_N \in \mathbb{N}$ then

$$c_i = \max_{n \in \{0, 1, \dots, T-1\}} \chi_{u^i}(n) \text{ for } i = 1, 2, \dots, N.$$

Proof. For $i \in \{1, 2, \dots, N\}$ we have that $d_i = \frac{k_i}{T}$ is the density of sequence u^i . So, for every $n \in \mathbb{Z}_{\geq 0}$

$$\chi_{u^i}(n+T) = \chi_{u^i}(n) + d_i \cdot T - (\kappa_{u^i}(n+T) - \kappa_{u^i}(n)) = \chi_{u^i}(n).$$

□

Theorem 3.5.10 Let $d_i \in \mathbb{Q}_{>0}$ for $i = 1, 2, \dots, N$ such that $\sum_{i=1}^N d_i = 1$. Let $p_i, q_i \in \mathbb{N}$ such that $d_i = \frac{p_i}{q_i}$ with $\gcd(p_i, q_i) = 1$ for $i = 1, 2, \dots, N$ and let $U \in \mathcal{Q}(d_1, d_2, \dots, d_N)$. Then

$$\bar{O}(U) = \sum_{i=1}^N c_i(U) - \frac{N}{2} + \sum_{i=1}^N \frac{1}{2q_i}.$$

Proof. There exist $T, k_1, k_2, \dots, k_N \in \mathbb{N}$ such that $U \in \mathcal{Q}(\{T\}, k_1, k_2, \dots, k_N) \subseteq \mathcal{Q}(d_1, d_2, \dots, d_N)$. For $i \in \{1, 2, \dots, N\}$ put $v_i = (u^i(1), u^i(2), \dots, u^i(T))$. Then $v_i \in \mathcal{P}(T, k_i)$ is a period cycle of u^i . Let $v'_i \sim v_i$ be an upper bound with respect to the partial order \preceq of $\mathcal{R}(T, k_i)$. Since v'_i is the l -th cyclic permutation of v_i for some $l \in \{0, 1, \dots, T-1\}$, it follows that (see the proof of Theorem 3.2.13) $\chi_{v_i}(l) = c_i = c_i + \chi_{v'_i}(0)$ and thus $\chi_{v_i}((l+n) \pmod{T}) = \chi_{v'_i}(n) + c_i$ for $n = 0, 1, \dots, T$. Hence by the definition of the primal unbalance

$$c_i = \frac{1}{T} \sum_{n=1}^T (\chi_{v_i}(n) - \chi_{v'_i}(n)) = \frac{1}{T} \sum_{n=1}^T (\kappa_{v'_i}(n) - \kappa_{v_i}(n)) = \quad (3.12)$$

$$\bar{I}(u^i) + \frac{1}{T} \sum_{n=1}^T \kappa_{\bar{w}(T, k_i)}(n) - \frac{1}{T} \sum_{n=1}^T \kappa_{u^i}(n).$$

By Corollary 3.2.12 and Lemma 3.5.7 we have that

$$\begin{aligned} \sum_{i=1}^N \frac{1}{T} \sum_{n=1}^T \kappa_{\bar{w}(T, k_i)}(n) &= \sum_{i=1}^N \left(\frac{1}{T} \cdot \sum_{n=1}^T \lceil n \cdot \frac{k_i}{T} \rceil \right) = \\ \frac{1}{2} \cdot \sum_{i=1}^N \left(k_i + \frac{k_i}{T} + 1 - \frac{\gcd(k_i, T)}{T} \right) &= \frac{1}{2} \cdot (T+1+N - \sum_{i=1}^N \frac{1}{q_i}). \end{aligned}$$

So, by (3.12) we have that

$$\sum_{i=1}^N c_i = \bar{O}(U) - \frac{1}{T} \sum_{n=1}^T n + \frac{1}{2} \cdot (T+1+N - \sum_{i=1}^N \frac{1}{q_i}) = \bar{O}(U) + \frac{N}{2} - \sum_{i=1}^N \frac{1}{2q_i}.$$

□

We need the following definitions. Let $d_1, d_2, \dots, d_N \in \mathbb{R}_{>0}$ with $\sum_{i=1}^N d_i = 1$. Put

$$\tilde{S} = \tilde{S}(d_1, d_2, \dots, d_N) := \inf_{U \in \mathcal{S}(d_1, d_2, \dots, d_N)} \sum_{i=1}^N c_i(U).$$

Moreover, define

$$D = D(d_1, d_2, \dots, d_N) = \{(x_1, x_2, \dots, x_N) \in \mathbb{R}^N : n \geq \sum_{i=1}^N \max(\lceil n \cdot d_i - x_i \rceil, 0) \text{ for every } n \in \mathbb{Z}_{\geq 0}\}$$

and

$$\tilde{D} = \tilde{D}(d_1, d_2, \dots, d_N) := \inf_{(x_1, x_2, \dots, x_N) \in D} \sum_{i=1}^N x_i.$$

For every $U \in \mathcal{S}(d_1, d_2, \dots, d_N)$ we have

$$(c_1(U), c_2(U), \dots, c_N(U)) \in D(d_1, d_2, \dots, d_N). \quad (3.13)$$

Consider the following algorithm to construct a sequence U with given densities. It is similar to the SG algorithm, but the choice of U_n is slightly less restricted. Therefore it is called the GG (General Greedy) algorithm.

GG algorithm. Let $d_1, d_2, \dots, d_N \in \mathbb{R}_{>0}$ with $\sum_{i=1}^N d_i = 1$ and $x_1, x_2, \dots, x_N \in \mathbb{R}_{\geq 0}$ be given. Then $U = (U_1, U_2, \dots) \in \mathcal{S}(d_1, d_2, \dots, d_N)$ is determined inductively in the following way. Suppose that U_1, U_2, \dots, U_{n-1} have been determined and thus $\kappa_{n,i}(n-1)$ is known for $i = 1, 2, \dots, N$. Choose $U_n = i \in \{1, 2, \dots, N\}$ for which $\lfloor \frac{x_i + \kappa_{n,i}(n-1)}{d_i} \rfloor$ is minimal.

Note that $U \in \mathcal{S}(d_1, d_2, \dots, d_N)$ for every word U constructed by the GG algorithm and that the SG algorithm is a special GG algorithm. The following results (until Theorem 3.5.19) follow from Chapter 2 in which a slightly different notation is used. These results are used for proving Theorem 3.5.5. Therefore, we reformulate these results from Chapter 2 in the notation of Chapter 3 and we have some more comments. First we note that some of the results are valid for arbitrary real densities and not only for rational densities.

Lemma 3.5.11 *Let $d_1, d_2, \dots, d_N \in \mathbb{R}_{>0}$ with $\sum_{i=1}^N d_i = 1$. Let $U \in \mathcal{S}(d_1, d_2, \dots, d_N)$ be constructed by applying the GG algorithm with $x_1, x_2, \dots, x_N \in \mathbb{R}_{\geq 0}$ such that $(x_1, x_2, \dots, x_N) \in D(d_1, d_2, \dots, d_N)$. Then $c_i(U) \leq x_i$ for $i = 1, 2, \dots, N$.*

A direct consequence of (3.13) and Lemma 3.5.11 is that

$$\tilde{S} = \tilde{D} \tag{3.14}$$

Theorem 3.5.12 *For every set of fractions $d_1, d_2, \dots, d_N \in \mathbb{R}_{>0}$ with $\sum_{i=1}^N d_i = 1$ there exist $U^* \in \mathcal{S}(d_1, d_2, \dots, d_N)$ and corresponding $(x_1^*, x_2^*, \dots, x_N^*) = (c_1(U^*), c_2(U^*), \dots, c_N(U^*)) \in D$ such that*

$$\sum_{i=1}^N c_i(U^*) = \tilde{S} \text{ and } \sum_{i=1}^N x_i^* = \tilde{D}.$$

Thus according to Theorem 3.5.12 there are U and x_i which attain the respective infima.

Lemma 3.5.13 *Let $d_1, d_2, \dots, d_N \in \mathbb{R}_{>0}$ with $\sum_{i=1}^N d_i = 1$ and suppose that*

$$\sum_{i=1}^N x_i = N - 1 \text{ and } 0 \leq x_i \leq 1 + d_i \text{ for } i = 1, 2, \dots, N. \tag{3.15}$$

Then $(x_1, x_2, \dots, x_n) \in D$.

Particular x_1, x_2, \dots, x_N that fulfill (3.15) are for example $x_i = 0$ for some $i \in \{1, 2, \dots, N\}$ and $x_j = 1$ for all $j \neq i$. Another example is $x_i = 1 - d_i$ for $i = 1, 2, \dots, N$. Combining (3.14), Theorem 3.5.12 and Lemma 3.5.13 gives the following corollary.

Corollary 3.5.14 *For every set of fractions $d_1, d_2, \dots, d_N \in \mathbb{R}_{>0}$ with $\sum_{i=1}^N d_i = 1$ we have that*

$$\tilde{S} = \min_{U \in \mathcal{S}(d_1, d_2, \dots, d_N)} \sum_{i=1}^N c_i \leq N - 1.$$

We think that for any billiard sequence U it holds that $\sum_{i=1}^N c_i(U) \leq N - 1$. The following theorem asserts that it is possible that $\tilde{S} = N - 1$ and thus Corollary 3.5.14 is sharp.

Theorem 3.5.15 *Let $d_1, d_2, \dots, d_N \in \mathbb{R}_{>0}$ with $\sum_{i=1}^N d_i = 1$. If d_1, d_2, \dots, d_N are linearly independent over \mathbb{Z} then $\tilde{S}(d_1, d_2, \dots, d_N) = N - 1$.*

Put

$$\mathcal{S}_0(d_1, d_2, \dots, d_N) = \{U \in \mathcal{S}(d_1, d_2, \dots, d_N) \text{ for which } \sum_{i=1}^N c_i = \tilde{S}\}$$

and

$$D_0 = D_0(d_1, d_2, \dots, d_N) = \{(x_1, x_2, \dots, x_N) \in D(d_1, d_2, \dots, d_N) \text{ for which } \sum_{i=1}^N x_i = \tilde{D}\}.$$

By Theorem 3.5.12 $\mathcal{S}_0(d_1, d_2, \dots, d_N) \neq \emptyset$ and $D_0(d_1, d_2, \dots, d_N) \neq \emptyset$. Moreover, if $U \in \mathcal{S}_0(d_1, d_2, \dots, d_N)$ then $(c_1(U), c_2(U), \dots, c_N(U)) \in D_0(d_1, d_2, \dots, d_N)$ and vice versa if

$(x_1, x_2, \dots, x_N) \in D_0(d_1, d_2, \dots, d_N)$ then $U \in \mathcal{S}_0(d_1, d_2, \dots, d_N)$ for any word U constructed by applying the GG algorithm with these x_1, x_2, \dots, x_N .

Theorem 3.5.16 *Let $d_1, d_2, \dots, d_N \in \mathbb{R}_{>0}$ with $\sum_{i=1}^N d_i = 1$. Then for every $\varepsilon > 0$ and $i \in \{1, 2, \dots, N\}$ there exist $(x_1, x_2, \dots, x_N) \in D_0(d_1, d_2, \dots, d_N)$ with $x_i < \varepsilon$ and $x_j < 1$ for all $j \neq i$.*

In the foregoing results it was not required for the densities d_i to be rational. In the remaining of this section the densities d_i are rational with $\sum_{i=1}^N d_i = 1$. Further $p_i, q_i, k_i \in \mathbb{N}$ for $i = 1, 2, \dots, N$ and $T \in \mathbb{N}$ are such that $d_i = \frac{p_i}{q_i} = \frac{k_i}{T}$ with $\gcd(p_i, q_i) = 1$ and $\gcd(k_1, k_2, \dots, k_N) = 1$ for $i = 1, 2, \dots, N$.

Lemma 3.5.17 *Let $U \in \mathcal{S}(d_1, d_2, \dots, d_N)$ be such that $c_i(U) < \infty$ for $i \in \{1, 2, \dots, N\}$. Then $q_i \cdot c_i(U) \in \mathbb{Z}_{\geq 0}$.*

Proof. For every $n \in \mathbb{Z}_{\geq 0}$ we have that $q_i \cdot \chi_{u^i}(n) = p_i \cdot n - q_i \cdot \kappa_{u^i}(n) \in \mathbb{Z}$. □

Corollary 3.5.18 *Suppose that $(x_1, x_2, \dots, x_N) \in D_0(x_1, x_2, \dots, x_N)$. Then $q_i \cdot x_i \in \mathbb{Z}_{\geq 0}$ for $i = 1, 2, \dots, N$.*

From this corollary we have that $x_j < 1$ implies $x_j \leq 1 - \frac{1}{q_j}$ for every j . Combining this with Theorem 3.5.16, we obtain the following result.

Theorem 3.5.19 *For every $i \in \{1, 2, \dots, N\}$ there exists some $(x_1, x_2, \dots, x_N) \in D_0(d_1, d_2, \dots, d_N)$ with $x_i = 0$ and $x_j \leq 1 - \frac{1}{q_j}$ for every $j \neq i$. Moreover*

$$\tilde{S} \leq N - 1 - \sum_{i=1}^N \frac{1}{q_i} + \min_{i \in \{1, 2, \dots, N\}} \frac{1}{q_i}.$$

Proof of Theorem 3.5.5. By Lemma 3.5.4 we have that $U \in \mathcal{Q}(\{T\}, k_1, k_2, \dots, k_N) \subseteq \mathcal{Q}(d_1, d_2, \dots, d_N)$. According to Theorem 3.5.19 there exist $(x_1, x_2, \dots, x_N) \in D_0(d_1, d_2, \dots, d_N) \subseteq D(d_1, d_2, \dots, d_N)$ with $x_i \leq x_i^*$ for $i = 1, 2, \dots, N$. It follows directly from the definition of $D(d_1, d_2, \dots, d_N)$ that if $(x_1, x_2, \dots, x_N) \in D(d_1, d_2, \dots, d_N)$ and $x_i \leq x_i'$ for $i = 1, 2, \dots, N$ then $(x_1', x_2', \dots, x_N') \in D(d_1, d_2, \dots, d_N)$. Hence $(x_1^*, x_2^*, \dots, x_N^*) \in D(d_1, d_2, \dots, d_N)$. Since constructing a consistent billiard sequence is a special form of applying a GG algorithm we have by Lemma 3.5.11 that $c_i(U) \leq x_i^*$ for $i = 1, 2, \dots, N$. Thus

$$\sum_{i=1}^N c_i(U) \leq \sum_{i=1}^N x_i^* = N - 1 - \sum_{i=1}^N \frac{1}{q_i} + \min_{i \in \{1, 2, \dots, N\}} \frac{1}{q_i}.$$

So, by Theorem 3.5.10

$$\bar{O}(U) = \sum_{i=1}^N c_i(U) - \frac{N}{2} + \sum_{i=1}^N \frac{1}{2q_i} \leq \frac{N}{2} - 1 + \min_{i \in \{1, 2, \dots, N\}} \frac{1}{q_i} - \sum_{i=1}^N \frac{1}{2q_i}. \quad (3.16)$$

If $N = 2$ then $q_1 = q_2$ and thus $\sum_{i=1}^N \frac{1}{2q_i} = \min_{i \in \{1, 2, \dots, N\}} \frac{1}{q_i}$. Combining this with (3.16) gives $\bar{O}(U) \leq 0$ and thus $\bar{O}(U) = 0$ if $N = 2$. If $N > 2$ then

$$\sum_{i=1}^N \frac{1}{2q_i} \geq \sum_{i=1}^N \min_{i \in \{1, 2, \dots, N\}} \frac{1}{2q_i} > \min_{i \in \{1, 2, \dots, N\}} \frac{1}{q_i}$$

and thus by (3.16), $\bar{O}(U) < \frac{N}{2} - 1$ if $N > 2$. □

We now show that for every set of rational densities there exists a consistent and thus periodic billiard sequence which has minimal unbalance among all (periodic) sequences with those densities.

Theorem 3.5.20 *For every set of rational densities $\{d_1, d_2, \dots, d_N\}$ there exists some consistent billiard sequence $U^* \in \mathcal{Q}(\{T\}, k_1, k_2, \dots, k_N) \subseteq \mathcal{Q}(d_1, d_2, \dots, d_N)$ such that*

$$\sum_{i=1}^N c_i(U^*) = \tilde{S} \text{ and } \bar{O}(U^*) = \min_{U \in \mathcal{Q}(d_1, d_2, \dots, d_N)} \bar{O}(U).$$

Proof. By Theorem 3.5.16 there exist $(x_1, x_2, \dots, x_N) \in D_0(d_1, d_2, \dots, d_N)$ with $0 \leq x_i < 1$ for $i = 1, 2, \dots, N$. Let U^* be obtained by applying the SG algorithm in a consistent way to such x_1, x_2, \dots, x_N . Then U^* is a consistent billiard sequence and by Lemma 3.5.4 we have that $U^* \in \mathcal{Q}(\{T\}, k_1, k_2, \dots, k_N) \subseteq$

$\mathcal{Q}(d_1, d_2, \dots, d_N)$. Moreover, since $(x_1, x_2, \dots, x_N) \in D_0(d_1, d_2, \dots, d_N)$ we have that $U^* \in \mathcal{S}_0(d_1, d_2, \dots, d_N)$ and thus $\sum_{i=1}^N c_i(U^*) = \tilde{S}$. By Theorem 3.5.10 it follows that $\overline{O}(U^*) = \min_{U \in \mathcal{Q}(d_1, d_2, \dots, d_N)} \overline{O}(U)$. \square

Such a billiard sequence with minimal unbalance can be found in an algorithmic way. Indeed, in case of rational densities d_i we showed in Chapter 2 that an $(x_1, x_2, \dots, x_N) \in D_0(d_1, d_2, \dots, d_N)$ with $0 \leq x_i < 1$ for $i = 1, 2, \dots, N$ can be found by solving an integer linear programming problem (ILP). Thus for given rational densities it is possible to obtain a consistent billiard sequence U with minimal unbalance \overline{O} by first solving an ILP and then applying the SG algorithm in a consistent way. Then the routing policy ψ corresponding to U is periodic and the upper bound from Theorem 3.4.7 for $\overline{W}(\psi)$ is minimized by ψ for those densities. Note that from Theorem 3.4.4 it follows that if V is a mirror of such a routing sequence U then V has minimal dual unbalance \underline{O} . The following example illustrates some of the results of this section.

Example. Let $N = 5$, $d_1 = \frac{3}{8}$, $d_2 = \frac{7}{24}$, $d_3 = \frac{1}{6}$, $d_4 = \frac{1}{8}$ and $d_5 = \frac{1}{24}$. We apply the SG algorithm in a consistent way (in case of a tie the letter of smallest index is chosen) with $x_1 = \frac{7}{8}$, $x_2 = 0$, $x_3 = \frac{5}{6}$, $x_4 = \frac{7}{8}$ and $x_5 = \frac{23}{24}$. Then we obtain the word $U = (2, 1, 2, 1, 3, 2, 4, 1, 2, 1, 3, 1, 2, 4, 1, 3, 2, 1, 2, 1, 3, 4, 5, 1)^\infty$ and thus $U \in \mathcal{Q}(\{24\}, 9, 7, 4, 3, 1) \subseteq \mathcal{Q}(\frac{3}{8}, \frac{7}{24}, \frac{1}{6}, \frac{1}{8}, \frac{1}{24})$. Using Lemma 3.5.9 it can be checked that $c_1(U) = \frac{5}{8}$, $c_2(U) = 0$, $c_3(U) = \frac{2}{3}$, $c_4(U) = \frac{3}{4}$, $c_5(U) = \frac{11}{12}$ and thus $\sum_{i=1}^N c_i(U) = \frac{71}{24}$. Note that $c_i(U) \leq x_i$ for every $i \in \{1, 2, 3, 4, 5\}$ as expected by Lemma 3.5.11. By Theorem 3.5.10 or direct computation it follows that $\overline{O}(U) = \frac{17}{24}$ and thus $\overline{O}(U) < \frac{N}{2} - 1$ as expected by Theorem 3.5.5. For the word $V = (1, 5, 4, 3, 1, 2, 1, 2, 3, 1, 4, 2, 1, 3, 1, 2, 1, 4, 2, 3, 1, 2, 1, 2)^\infty$ which is a mirror of U we have by Theorem 3.4.4 that $\underline{O}(V) = \overline{O}(U) = \frac{17}{24} < \frac{N}{2} - 1$.

By Lemma 3.5.11 we have that if the SG algorithm is applied with $x_i = c_i(U)$ for $i = 1, 2, \dots, 5$ then a word $U' \in \mathcal{Q}(\frac{3}{8}, \frac{7}{24}, \frac{1}{6}, \frac{1}{8}, \frac{1}{24})$ is obtained with $c_i(U') \leq c_i(U)$ for $i = 1, 2, \dots, 5$ and thus $\overline{O}(U') \leq \overline{O}(U)$. Hence by using the SG algorithm iteratively in this way a sequence of periodic words with the given densities is found with non-increasing total primal unbalance. We do this for this example starting with U and we get the following sequence of words. First we obtain $U' = (2, 1, 2, 3, 1, 4, 2, 1, 1, 3, 2, 1, 2, 4, 1, 3, 2, 1, 1, 2, 3, 4, 5, 1)^\infty$ with $c_1(U') = \frac{5}{8}$, $c_2(U') = 0$, $c_3(U') = \frac{1}{2}$, $c_4(U') = \frac{5}{8}$, $c_5(U') = \frac{11}{12}$ and $\overline{O}(U') = \frac{5}{12}$. By applying the SG algorithm to $x_i = c_i(U')$ we obtain the word

$U'' = (2, 1, 3, 2, 1, 4, 2, 1, 3, 1, 2, 1, 4, 2, 1, 3, 2, 1, 1, 2, 3, 4, 5, 1)^\infty$ with $c_1(U'') = \frac{5}{8}$, $c_2(U'') = 0$, $c_3(U'') = \frac{1}{2}$, $c_4(U'') = \frac{5}{8}$, $c_5(U'') = \frac{11}{12}$ and $\overline{O}(U'') = \frac{5}{12}$. Since $c_i(U'') = c_i(U')$ for $i = 1, 2, \dots, 5$ applying the SG algorithm to $x_i = c_i(U'')$ gives the word U'' again. Thus the iterative process stops with U'' (recall the remarks in Sec-

tion 2.6 on this iterative process), but the word U'' does not have minimal unbalance for these densities. Namely, we have found that $(0, \frac{7}{24}, \frac{1}{2}, \frac{3}{4}, \frac{23}{24}) \in D_0(\frac{3}{8}, \frac{7}{24}, \frac{1}{6}, \frac{1}{8}, \frac{1}{24})$, and thus the word $W = (1, 2, 1, 3, 2, 1, 4, 2, 1, 3, 1, 2, 1, 4, 2, 3, 1, 2, 1, 3, 1, 2, 4, 5)^\infty$ obtained by applying the SG algorithm to this solution with $\bar{O}(W) = \frac{1}{4}$ has minimal primal unbalance for these densities.

Conclusion

We investigated in this chapter the static routing to parallel queues with different service times. We introduced a combinatorial notion which we called unbalance. For a periodic routing pattern we derived an upper bound on the average waiting time as function of its unbalance. From a combinatorial point of view it would be of interest to extend the notion of unbalance to non-periodic routing sequences which have irrational routing fractions. For applications it would be important to extend the bounds derived in this chapter to the following models:

1. optimal routing of the server in polling systems;
2. parallel queues with two or more servers for each queue;
3. sojourn times in queueing networks.

3.6 Appendix

In Theorem 3.3.10 and subsequent results we assumed that $\{\delta_i\}$ ($\{\sigma_j\}$) were i.i.d. random variables. Below we show that instead of i.i.d. random variables we may assume that they are stationary ergodic sequences. This greater generality is very useful for applications to traffic processes in telecommunication networks. To the authors' knowledge there does not exist a proof for the i.i.d. case in the literature. But, it is a special case of the more general result of Theorem 3.6.1 (see also the remark following the theorem). Recall that $W_u(j, w)$ denotes the workload for the server at the moment that the j -th customer is routed to the server (according to routing sequence u), given that the initial workload at time $T_1 = 0$ is equal to w .

Theorem 3.6.1 *Let u' and u'' be routing sequences of zeros and ones that are both representatives of some $u \in \mathcal{P}(d)$ where $d = \frac{k}{T} \in \mathbb{Q}$ and the period cycle of u is an element of $\mathcal{P}(T, k)$. Suppose that the interarrival times $\{\delta_i\}$ of customers arriving at the system are a stationary ergodic sequence with $\delta = \mathbb{E}\delta_1$. Moreover, assume that the sequence $\{\delta_{nT+l}\}_{n=0}^\infty$ is ergodic for each $l \in \{1, 2, \dots, T\}$. Suppose that*

the service times σ_j of the customers routed to the server form a stationary ergodic sequence with $\sigma = \mathbb{E}\sigma_1$, and are stochastically independent of the interarrival times. Moreover, assume that the sequence $\{\sigma_{nk+l}\}_{n=0}^{\infty}$ is ergodic for each $l \in \{1, 2, \dots, k\}$. Suppose that $\frac{\sigma}{\delta} \cdot d < 1$. Then $\lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{n=0}^{m-1} \delta_{nT+l} = \delta$ almost surely (a.s.) for $l = 1, 2, \dots, T$ and there exists a deterministic value $\overline{W}(u)$ such that

$$\lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{j=0}^{m-1} \overline{W}_{w'}(j, w') = \lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{j=0}^{m-1} W_{w''}(j, w'') = \overline{W}(u) \text{ a.s.}$$

for any initial workloads w' and w'' .

Proof. Since $\{\delta_{nT+l}\}_{n=0}^{\infty}$ is a stationary ergodic sequence it follows directly from Birkhoff's theorem ([18] Theorem 6.28) that $\lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{n=0}^{m-1} \delta_{nT+l} = \mathbb{E}\delta_l$ a.s. Since δ_i is a stationary sequence we have that $\mathbb{E}\delta_l = \delta$ for $l = 1, 2, \dots, T$. It is shown in [10] that there exist k sequences $\{W_u^*(nk + j)\}_{n=0}^{\infty}$, $j = 1, 2, \dots, k$, which are all stationary ergodic, such that these sequences couple in finite time with the workload sequences. Indeed, for any w there exists a finite random variable M such that almost surely,

$$W_u(nk + j, w) = W_u^*(nk + j) \text{ for all } n \geq M. \quad (3.17)$$

Define,

$$\overline{W}(u) = \frac{1}{k} \cdot \sum_{j=1}^k \mathbb{E}W_u^*(j).$$

From Birkhoff's ergodic theorem it follows that almost surely

$$\lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{n=0}^{m-1} W_u^*(nk + j) = \mathbb{E}W_u^*(j) \text{ for } j = 1, 2, \dots, k.$$

Hence,

$$\lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{j=1}^m W_u^*(j) = \frac{1}{k} \cdot \sum_{j=1}^k \mathbb{E}W_u^*(j) = \overline{W}(u) \text{ a.s.}$$

With the coupling result (3.17) it then follows that

$$\lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{j=1}^m \overline{W}_u(j, w) = \overline{W}(u) \text{ a.s.}$$

For a representative u' of u there is an $l \in \{1, 2, \dots, T\}$ such that $u_n = u'_{l+n}$, $n = 1, 2, \dots$. If we couple the service times in the routing sequences u and u' such that $\sigma_j = \sigma'_{l+j}$ then $W_u(n, w) = W_{u'}(l + n, w')$ a.s. for $n = 1, 2, \dots$ where $w := W_{u'}(l, w')$. Hence

$$\lim_{m \rightarrow \infty} \frac{1}{m} \cdot \sum_{j=1}^M W_{u'}(j, w') = \overline{W}(u) \text{ a.s.}$$

□

Remark. If $\{\delta_i\}$ and $\{\sigma_j\}$ are independent sequences of i.i.d. random variables with mean δ respectively σ , then all the assumptions of Theorem 3.6.1 are satisfied. Indeed, every subsequence of i.i.d. random variables is stationary ergodic (see [18] Corollary 6.33). In modern telecommunication systems the arrival process is composed of heterogeneous traffic streams and in general more complex than a renewal process. Often a Markov Modulated Poisson process is considered as the appropriate process (see [22]). This process also satisfies the assumptions of Theorem 3.6.1.

Chapter 4

Deterministic parallel queueing systems: on the average waiting time for regular routing and the corresponding lower bound

4.1 Introduction

In this chapter we consider the deterministic optimal routing to N parallel inhomogeneous queues. We assume that the arrival epochs are a deterministic sequence, $T = 0, 1, 2, \dots$ say, and the service times at the queues are deterministic, S_i at queue i say. In stochastic models there is a difference between open-loop control in which there is no state information and closed-loop control in which the controller knows the loads at the queues. In the deterministic model there is no such difference in state information. But, as in the open-loop control we describe the routing policy as a sequence of the numbers of the queues to which the successive arriving customers are routed. See Chapter 3 for an overview of papers on open-loop control. An overview of results on closed-loop control can be found in [35] and [45].

The optimal routing for homogeneous queues, i.e. queues with $S_i = S$, for $i = 1, \dots, N$, is well-known to be the round robin policy (see [47]). For inhomogeneous queues the computation of the optimal policy may be difficult. In the seminal paper [29] Hajek showed for an exponential queue that the optimal routing of a given fraction of the customers to that queue is a regular or bracket sequence. For $N = 2$ it is proved in [3] and [5] that the optimal routing to each of the queues is a regular or bracket sequence for general stochastic interarrival and service time sequences. However, if $N \geq 3$ the optimal routing fails to be the composition of regular sequences, because in general regular sequences can not be combined to a feasible routing policy. Finding the optimal routing becomes a combinatorial hard problem for $N \geq 3$. In Chapter 2 we studied the deterministic model where the system is fully loaded, i.e. $\sum_{i=1}^N \frac{1}{S_i} = 1$, and a mathematical programming problem was designed to compute the optimal routing. In this chapter we consider the case $\sum_{i=1}^N \frac{1}{S_i} \geq 1$. Now finding the optimal routing is split into two parts:

- Find best routing densities d_i to the queues $i = 1, \dots, N$
- Construct a best allocation pattern for the given routing densities d_i

For $N = 2$ the second part is completely solved, since the allocation uses regular sequences as mentioned above. In [23] the problem of finding the best densities is studied, and a very interesting connection is made there to combinatorics on words. Moreover, a recursive formula for the average waiting time for a given density is obtained by using a (backward) continued fraction decomposition of the inverse of the service time, and an algorithm is derived to compute an optimal routing sequence. Our analysis is in the same spirit but we consider the more general case of $N \geq 2$ queues. We provide an algorithm to compute a lower bound for the total average waiting time for the optimal routing by finding best routing densities given that the routing to each of the queues is regular. In Chapter 3 we derived an upper bound for the total waiting time for a given allocation pattern as a function of its unbalance (which is a measure of the deviation from the regular routing). Combining these results we use a billiard sequence to provide an upper bound for the average waiting time of the optimal routing policy too. The upper and lower bounds coincide in case $N = 2$. In this chapter the analysis of deterministic queues is based on results from number theory, involving Farey intervals, best approximation points, convergents and continued fraction expansions.

The chapter is organized as follows. Section 4.2 gives a description of the queueing system together with the definitions of routing policy and regular sequences. In Section 4.3 the Lindley recursion together with the extension of Little's relation between

the average number of customers and the waiting time is given for the routing model. Its proof is given in the appendix 4.8. In Section 4.4 Farey intervals are introduced. The period sequence of the upper bracket sequence of which the rational density belongs to some Farey interval is determined as a finite composition of the period sequences of the upper bracket sequences of the end-points densities of that Farey interval. This result gives the key to the proof that for upper bracket sequences the average number of customers $L(d)$ is a linear function of the density d on a Farey interval. This result is proved in Section 4.5, where it is also shown that $L(d)$ is a convex function of the density. In Section 4.6 best lower approximation points of $\frac{1}{S_i}$, where S_i is the service time for queue i , are introduced. They correspond to the so-called jump points in [23], where recursive formulae for computing the average waiting time $W(d_i)$ were obtained. The main results of Section 4.6 are the explicit formulae for the average number of customers $L_{S_i}(d_i)$ (and of the average waiting time $W(d_i)$) at best lower approximation points d_i of $\frac{1}{S_i}$. Together with the fact that consecutive best lower approximation points constitute a Farey interval on which $L_{S_i}(d_i)$ is linear, this provides an efficient algorithm for computing $L_{S_i}(d_i)$ given that the best lower approximation points are known. Section 4.6 also contains an exposition, based on well-known results from number theory, how the ordered set of best approximation points can be computed by the use of the continued fraction expansion and the induced convergents. The efficient algorithm to compute this ordered set is illustrated by an example. In Section 4.7 the above mentioned results are applied to find the minimal regular routing value. More precisely we derive an algorithm for computing

$$\min_{d_1 + \dots + d_N = 1} \sum_{i=1}^N L_{S_i}(d_i),$$

using the above mentioned convexity and linearity properties on Farey intervals. The algorithm is again illustrated by an example. We characterize the set of densities for which the minimum is achieved. Also, as a consequence of the characterization we find that there exist minimizing rational densities if $\sum_{i=1}^N \frac{1}{S_i} > 1$. Indeed, the algorithm computes such a set of rational densities. For $N = 2$ this implies that the regular policy with these densities is periodic, and hence an optimal periodic routing policy is found by the algorithm. For $N \geq 3$ the algorithm computes a lower bound for the optimal routing policy. Together with results of Chapter 3 which we summarize in Section 4.7, we find a billiard sequence with rational densities (hence a periodic policy) which has an average waiting time not larger than the lower bound plus $\frac{N}{2} - 1$. The performance of this billiard sequence should be close to optimal if N is not too large. For $N = 2$ the lower bound equals the upper bound.

4.2 Description of the queueing system and notation

We consider a queueing system with $N \geq 2$ parallel FIFO (First In First Out) queues. Customers arrive at moments T_1, T_2, \dots with $T_1 = 0$ and $T_1 \leq T_2 \leq \dots$ and the n -th arriving customer has to be routed to one of the queues at the moment T_n of its arrival. We consider deterministic routing policies using no state information. Then, as in the previous chapter, the routing policy is described by an infinite word $U = (U_1, U_2, \dots)$ on the alphabet $\{1, 2, \dots, N\}$ where U_n is the queue to which the n -th arriving customer is routed. The following notations (and definitions) are also similar to what we used in the previous chapter, but we briefly recall it to make this chapter self-contained.

For every queue $i \in \{1, 2, \dots, N\}$ we have a sequence $u^i = (u_1^i, u_2^i, \dots)$ of zeros and ones where $u_n^i = 1$ if $U_n = i$ and $u_n^i = 0$ if $U_n \neq i$. Then u^i is called the routing, splitting or admission sequence for queue i (see for example [29], [62], [4], [37] and [46]).

For a sequence of nonnegative integers $u = (u_1, u_2, \dots)$ we define the counting function $\kappa_u : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{Z}$ by $\kappa_u(n) := \sum_{i=1}^n u_i$. Thus $\kappa_u(n)$ is the partial sum of the first n terms. Note that if u^i is a splitting sequence then $\kappa_{u^i}(n)$ is the number of customers that is routed to queue i among the first n arriving customers.

Put $\delta_n := T_{n+1} - T_n$ for $n = 1, 2, \dots$. Then $\{\delta_n\}$ is the sequence of interarrival times. Further for queue $i \in \{1, 2, \dots, N\}$ we have a sequence of service times $\{\sigma_j^i\}_{j=1,2,\dots}$, where σ_j^i is the service time of the j -th customer that is routed to queue i .

In a deterministic queueing system the interarrival times are constant and the service times are constant for every queue. Thus there exist positive numbers $\delta, \sigma_1, \sigma_2, \dots, \sigma_N$ such that $\delta = \delta_1 = \delta_2 = \dots$ and $\sigma_i = \sigma_1^i = \sigma_2^i = \sigma_3^i = \dots$ for $i = 1, 2, \dots, N$. By time scaling with a factor δ we assume without loss of generality that $\delta = 1$ and then $S_i := \frac{\sigma_i}{\delta}$ is the service time of server i for the new time scale. Thus such a deterministic queueing system with N parallel FIFO queues can be described by the vector (S_1, S_2, \dots, S_N) of service times when the interarrival times are assumed to be equal to the time unit. In the sequel we use the description (S_1, S_2, \dots, S_N) system when we mean a queueing system with deterministic interarrival and service times as described above.

4.2.1 Routing policies and words

A routing policy U is described by a sequence of symbols (letters) (U_1, U_2, \dots) and can be seen as an (infinite) word (sequence) on a finite alphabet. If we have N parallel queues then U is a so-called \mathbb{N} -word on the alphabet $\{1, 2, \dots, N\}$. We refer to the beginning of Section 3.2 for the definitions of the density of a letter, period of a word, subword, prefix, suffix, concatenation of sequences, etc. See [49] and [33] for more on words and combinatorics.

We recall that the domain of a word I is a subinterval of \mathbb{Z} and the length of a finite (i.e. I is finite) word W is denoted by $|W|$. Moreover, a sequence (word) of zeros and ones is said to have density d if the letter 1 has density d . It follows that the density of letter 0 is $1 - d$ in that case. Regular sequences (see Definition 3.2.4) are also called bracket sequences, since the symbols are determined by a bracket form (see Corollary 3.2.6). Moreover, the variable θ in this bracket form is called the phase of the bracket sequence. In the sequel all infinite regular sequences we consider are \mathbb{N} -words. Hence if I is infinite, then $I = \mathbb{N}$. The regular support set of the ones of such a regular sequence is denoted by $H_1 \subseteq \mathbb{N}$.

Definition 4.2.1 *The upper bracket sequence of density d is the \mathbb{N} -word $u = (u_1, u_2, \dots)$ on the alphabet $\{0, 1\}$ for which $u_t = \lceil td \rceil - \lceil (t-1)d \rceil$ for every $t \in \mathbb{N}$. The lower bracket sequence of density d is the \mathbb{N} -word $u = (u_1, u_2, \dots)$ on the alphabet $\{0, 1\}$ for which $u_t = \lfloor td \rfloor - \lfloor (t-1)d \rfloor$ for every $t \in \mathbb{N}$.*

If we make the assumption that letter 0 is smaller than letter 1 in the alphabet $\{0, 1\}$ then the upper bracket sequence of density d is the lexicographic greatest regular sequence of density d and the lower bracket sequence is the lexicographic smallest regular sequence of density d . We have the following properties of these sequences (see Chapter 3).

Lemma 4.2.2 *For the upper bracket sequence u of density d we have that $\kappa_u(n) = \lceil nd \rceil$ for every $n \in \mathbb{Z}_{\geq 0}$ and for the lower bracket sequence v of density density d we have that $\kappa_v(n) = \lfloor nd \rfloor$ for every $n \in \mathbb{Z}_{\geq 0}$.*

Lemma 4.2.3 *Let u and v be as in Lemma 4.2.2. Then for the corresponding support sets $H_1(u)$ and $H_1(v)$ of the ones of these bracket sequences we have that*

$$H_1(u) = \{\lfloor \frac{t}{d} \rfloor + 1\}_{t=0}^{\infty} \text{ and } H_1(v) = \{\lceil \frac{t}{d} \rceil\}_{t=1}^{\infty}.$$

4.3 Definitions and preliminary results

For $t \geq 0$ and $i \in \{1, 2, \dots, N\}$ let $V_i(t)$ be the remaining workload measured in time units for server i at time t and let $B_i(t)$ be the number of customers waiting in the buffer of queue i at moment t . Then $\underline{V}(t) := (V_1(t), V_2(t), \dots, V_N(t))$ is the vector of remaining workloads in the system at time t , $\underline{B}(t) := (B_1(t), B_2(t), \dots, B_N(t))$ is the vector of buffer contents at time t and $B(t) := \sum_{i=1}^N B_i(t)$ is the total number of customers waiting in one of the buffers at time t . Let $G_i(t) = \{t' \in [0, t] : V_i(t') = 0\}$ and let $m_i(t)$ be the Lebesgue measure of $G_i(t)$. Then $m_i(t)$ is the time that server i was idle up to time t . We say that a server i is non-idling in a time interval $[t_1, t_2]$ if $m_i(t_1) = m_i(t_2)$.

For an (S_1, S_2, \dots, S_N) system we assume that $V_i(0) = 0$ for $i = 1, 2, \dots, N$ and for a given routing policy $U = (U_1, U_2, \dots)$ and time $t > 0$ we can compute $\underline{V}(t)$ recursively by Lindley's formula which gives that

$$V_i(t) = \max(V_i(t') + u_{t'}^i \cdot S_i + (t' - t), 0) \text{ for } i = 1, 2, \dots, N \text{ where } t' = \lceil t \rceil - 1. \quad (4.1)$$

Note that we define $V_i(t)$ such that the function $t \rightarrow V_i(t)$ is left continuous, but not right continuous in general. However, the only right discontinuity of this function can be at integer t . Moreover, it follows inductively that for every $i \in \{1, 2, \dots, N\}$ and $t, t_0 \in \mathbb{R}$ with $t \geq t_0 \geq 0$ we have that

$$V_i(t) - V_i(t_0) = (\kappa_{u^i}(\lceil t \rceil) - \kappa_{u^i}(\lceil t_0 \rceil)) \cdot S_i + (m_i(t) - m_i(t_0)) - (t - t_0). \quad (4.2)$$

Note that in an (S_1, S_2, \dots, S_N) system it holds that $T_t = t - 1$ for $t = 1, 2, \dots$. Thus $\kappa_{u^i}(\lceil t \rceil)$ is the number of customers routed to server i before time t . Further for an (S_1, S_2, \dots, S_N) system we have for every $t \geq 0$ that $B_i(t) = \lfloor \frac{V_i(t)}{S_i} \rfloor$ for $i = 1, 2, \dots, N$ and thus $\underline{B}(t)$ and $B(t)$ can be easily obtained from $\underline{V}(t)$.

Definition 4.3.1 Let $U = (U_1, U_2, \dots)$ be a routing policy. Then we define

$$L(U) := \limsup_{t' \rightarrow \infty} \frac{1}{t'} \cdot \int_{t=0}^{t'} B(t) dt$$

as the average number of customers waiting in one of the buffers for policy U . We define

$$W(U) := \limsup_{t' \rightarrow \infty} \frac{1}{t'} \sum_{t=1}^{t'} V_{U_t}(T_t)$$

as the average waiting time for the arriving customers for policy U .

We show in the appendix as an extension of Little's law for routing policies that

$$L(U) = \lambda \cdot W(U) \text{ if } W(U) < \infty, \quad (4.3)$$

where λ is the arrival rate of customers in the queueing system. In an (S_1, S_2, \dots, S_N) system we have by the time scaling that the arrival rate of customers λ is equal to 1 and thus it follows by (4.3) that

$$L(U) = W(U) \text{ if } W(U) < \infty. \quad (4.4)$$

The objective in this chapter is optimal routing with respect to minimizing the average waiting time. Thus we look for routing policies U such that $L(U)$ and $W(U)$ are minimal.

For this problem it is useful to consider for a particular server $i \in \{1, 2, \dots, N\}$ the average number of customers waiting in buffer i and the average waiting time of the customers that are routed to this server i . So, for $i = 1, 2, \dots, N$ we define

$$L^i(U) := \limsup_{t' \rightarrow \infty} \frac{1}{t'} \cdot \int_{t=0}^{t'} B_i(t) d(t)$$

as the average number of customers waiting in buffer i for policy U . Recall that $H_i(U)$ is the support of letter i of the \mathbb{N} -word U and for $t \in \mathbb{N}$ let $I_t = \{1, 2, \dots, t\}$. Moreover, for $i = 1, 2, \dots, N$ we define

$$W^i(U) := \limsup_{t' \rightarrow \infty} \frac{1}{|H_i(U) \cap I_{t'}|} \cdot \sum_{t \in H_i(U) \cap I_{t'}} V_i(I_t) = \limsup_{t' \rightarrow \infty} \frac{1}{\kappa_{u^i}(t')} \cdot \sum_{t=1}^{t'} u_t^i \cdot V_i(t-1) \quad (4.5)$$

as the average waiting time of customers routed to server i for policy U .

Note that $L^i(U)$ and $W^i(U)$ depend on u^i , the routing sequence for server i , but they do not depend on u^j if $j \neq i$. Therefore, if the routing to server i is according to some infinite sequence w of zeros and ones then we also use the notations $L^i(w)$ and $W^i(w)$ instead of $L^i(U)$ and $W^i(U)$, respectively. Moreover, if we consider a single queue of the system and it is not necessary to refer to the number of the server i then we just write $L(w)$ and $W(w)$, respectively. We define $\overline{d}_i := \limsup_{t \rightarrow \infty} \frac{\kappa_{u^i}(t)}{t}$ which we call the upper density of routing sequence u^i and we define $\underline{d}_i := \liminf_{t \rightarrow \infty} \frac{\kappa_{u^i}(t)}{t}$ which we call the lower density of u^i . If $\overline{d}_i = \underline{d}_i$ then the routing sequence u^i has

a density $d_i := \lim_{t \rightarrow \infty} \frac{\kappa_{u^i}(t)}{t}$ and d_i is the arrival rate for server i in that case. In the appendix 4.8 we show that

$$L^i(u^i) = d_i \cdot W^i(u^i) \text{ if } d_i \text{ exists and } W^i(u^i) < \infty. \quad (4.6)$$

From (4.4) and (4.6) we obtain the following proposition.

Proposition 4.3.2 *Let U be a routing policy applied in an (S_1, S_2, \dots, S_N) system such that routing sequence u^i has density d_i and $W^i(u^i) < \infty$ for $i = 1, 2, \dots, N$. Then*

$$W(U) = L(U) = \sum_{i=1}^N L^i(u^i) = \sum_{i=1}^N d_i \cdot W^i(u^i).$$

Note that $\sum_{i=1}^N d_i = 1$ if the conditions of Proposition 4.3.2 are satisfied.

4.4 The upper bracket sequence and Farey intervals

In the sequel we denote by $w(d)$ the upper bracket sequence of density d . Thus for the counting function κ we have that $\kappa_{w(d)}(n) = \lceil n \cdot d \rceil$ for every $n \in \mathbb{Z}_{\geq 0}$. If d is rational, $d = \frac{p}{q}$ with $\gcd(p, q) = 1$ let $\omega(d) := (w(d)_1, w(d)_2, \dots, w(d)_q)$ be the period word of $w(d)$. Note that we use the slightly different notations $w(d)$ and $\omega(d)$ for the upper bracket sequence and the period word of this sequence, respectively.

Let d_1, d_2 be rational fractions with $0 \leq d_1 \leq d_2 \leq 1$, $d_i = \frac{p_i}{q_i}$ and $\gcd(p_i, q_i) = 1$ for $i = 1, 2$. Then $I = [d_1, d_2]$ is called a Farey interval if and only if $q_1 \cdot p_2 - p_1 \cdot q_2 = 1$. From elementary number theory (see for example [30], chapter 3, page 23) we have the following facts about Farey intervals. If $[d_1, d_2]$ is a Farey interval and $d_0 = \frac{p_1 + p_2}{q_1 + q_2}$. Then $I' = [d_1, d_0]$ and $I'' = [d_0, d_2]$ are also Farey intervals. Moreover, all rational numbers in (d_1, d_2) have denominator greater than or equal to $q_1 + q_2$.

4.4.1 Factorisation of the upper bracket sequence

Recall that a factor of a word is a finite subsequence of consecutive letters of the word. Let X be a set of nonempty finite words over some alphabet \mathbf{A} . A finite

X - factorisation of a finite word w is a finite sequence (x_1, x_2, \dots, x_n) such that $w = x_1 x_2 \dots x_n$ and $x_i \in X$ for $i = 1, 2, \dots, n$. An infinite X -factorisation of an \mathbb{N} - word w is an infinite sequence (x_1, x_2, \dots) such that $w = x_1 x_2 \dots$ and $x_i \in X$ for $i = 1, 2, \dots$. Sometimes we just say factorisation instead of (in)finite X -factorisation. The set X is called a code (\mathbb{N} -code) if every finite (infinite) word over the alphabet \mathbf{A} has at most one X - factorisation. It is easily seen that every \mathbb{N} -code is also a code, but the converse is not true (see [49], chapter 6, page 177).

Lemma 4.4.1 *Let $d_1, d_2 \in \mathbb{Q} \cap [0, 1]$ with $d_1 \neq d_2$. Then $\omega(d_1)$ and $\omega(d_2)$ are not powers of the same word.*

Proof. Suppose that there exist $k, l \in \mathbb{N}$ and a sequence of zeros and ones u such that $u^k = \omega(d_1)$ and $u^l = \omega(d_2)$. Then it follows that d_1 , the density of $\omega(d_1)$, equals the density of u and also d_2 , the density of $\omega(d_2)$, equals the density of u . Hence $d_1 = d_2$, which gives a contradiction. \square

Theorem 4.4.2 *Let $d_1, d_2 \in \mathbb{Q} \cap [0, 1]$ with $d_1 \neq d_2$. Then $X := \{\omega(d_1), \omega(d_2)\}$ is an \mathbb{N} -code for words over the alphabet $\{0, 1\}$.*

Proof. By Lemma 4.4.1 and Corollary 6.2.5 in [49] (which follows immediately from the so-called Defect theorem) there exists neither a finite nor an infinite nontrivial relation between $\omega(d_1)$ and $\omega(d_2)$. Hence X is a code and also an \mathbb{N} -code.

Lemma 4.4.3 *Let $I = [d_1, d_2]$ be a Farey interval and $d_0 = \frac{p_1+p_2}{q_1+q_2}$ as above. Then $\omega(d_0) = \omega(d_2)\omega(d_1)$.*

Proof. We put $u := \omega(d_2)\omega(d_1)$. We have $\gcd(p_1 + p_2, q_1 + q_2) = 1$ because $q_1 + q_2$ is the smallest denominator in (d_1, d_2) . Hence $|u| = |\omega(d_0)| = q_1 + q_2$ and it suffices to prove that $\kappa_u(n) = \kappa_{\omega(d_0)}(n)$ for $1 \leq n \leq q_1 + q_2$. For $1 \leq n \leq q_2$ we have that

$$\kappa_{\omega(d_0)}(n) = \lceil d_0 \cdot n \rceil \leq \lceil d_2 \cdot n \rceil = \kappa_u(n).$$

Suppose $\lceil d_0 \cdot n \rceil < \lceil d_2 \cdot n \rceil$ for some $1 \leq n \leq q_2$. Then there exist a positive integer k such that $\frac{n \cdot (p_1+p_2)}{q_1+q_2} \leq k < \frac{n \cdot p_2}{q_2}$. Hence $n \cdot q_2 \cdot (p_1+p_2) \leq k \cdot q_2 \cdot (q_1+q_2) < n \cdot p_2 \cdot (q_1+q_2)$ and thus $0 \leq q_2 \cdot (k \cdot (q_1 + q_2) - n \cdot (p_1 + p_2)) < n$. Since $n \leq q_2$ it follows that $k \cdot (q_1 + q_2) - n \cdot (p_1 + p_2) = 0$ and thus $d_0 = \frac{p_1+p_2}{q_1+q_2} = \frac{k}{n}$ with $1 \leq n \leq q_2$, which gives a contradiction. So, $\kappa_u(n) = \kappa_{\omega(d_0)}(n)$ for $1 \leq n \leq q_2$. Analogously it follows that $\kappa_u(n) = \kappa_{\omega(d_0)}(n)$ for $q_2 + 1 \leq n \leq q_1 + q_2$. \square

Theorem 4.4.4 *Let $I = [d_1, d_2]$ be a Farey interval and put $X := \{\omega(d_1), \omega(d_2)\}$. Then for every $d \in (d_1, d_2)$ there exists a unique infinite X -factorisation of the upper bracket sequence $w(d)$ of density d . Moreover, if d is rational then there exists a unique finite X -factorisation of the period word $\omega(d)$ of $w(d)$.*

Proof. The factorization is obtained by the following algorithm. Initialize by putting $x_0 = d_1 = \frac{p_0}{q_0}$, $y_0 = d_2 = \frac{r_0}{s_0}$, $u_0 = \omega(x_0)$, $v_0 = \omega(y_0)$ and $i := 0$. At step i we have that $x_i = \frac{p_i}{q_i}$ with $\gcd(p_i, q_i) = 1$ and $y_i = \frac{r_i}{s_i}$ with $\gcd(r_i, s_i) = 1$ and we do the following iteratively: $z_i := \frac{p_i + r_i}{q_i + s_i}$, $w_i := v_i u_i$. If $d = z_i$ then stop. Else if $d < z_i$ then $x_{i+1} := x_i$, $u_{i+1} := u_i$, $y_{i+1} := z_i$, $v_{i+1} := w_i$. Else if $d > z_i$ then $x_{i+1} := z_i$, $u_{i+1} := w_i$, $y_{i+1} := y_i$, $v_{i+1} := v_i$. If the algorithm has not stopped then let $i := i + 1$ and go to the next step.

From the algorithm it follows by induction that w_i is factored in $u_0 = \omega(d_1)$ and $v_0 = \omega(d_2)$ for every $i \in \mathbb{Z}_{\geq 0}$ and by Lemma 4.4.3 we have that $w_i = \omega(z_i)$. If d is rational then the algorithm stops for some $i = N$ and then $d = z_N$ and $\omega(d) = \omega(z_N) = w_N$ is factored in $\omega(d_1)$ and $\omega(d_2)$. If d is irrational then $\lim_{i \rightarrow \infty} z_i = d$ and $\lim_{i \rightarrow \infty} w_i = w(d)$. \square

4.5 The average number of customers in a single queue

We consider a single queue $i \in \{1, 2, \dots, N\}$ of an (S_1, S_2, \dots, S_N) system. Let $S = S_i$ be the service time in this queue. We suppose that the routing sequence for this queue has density $d \in [0, 1]$. Then d is the arrival rate of customers for this queue and the traffic intensity for this queue is $\rho := d \cdot S$. The stability condition for the queue is $\rho \leq 1$, or equivalently $d \leq \frac{1}{S}$. Suppose that the routing sequence u^i is a regular sequence of density $d \in [0, 1]$. Then it is proved in [39] that $W^i(u^i)$ and $L^i(u^i)$ exist for every ρ and are independent of the initial phase θ . It follows from Loynes' stability theorem (see [13]) that the limit is finite for $\rho < 1$. Moreover, for $\rho = 1$ we showed this in Chapter 2. If $\rho > 1$ then the queue is not stable and $L^i(u^i) = W^i(u^i) = \infty$ in that case. Combining these results it follows that if u is a regular sequence of density $d \in [0, 1]$ then

$$L^i(u) = L^i(w(d)) \text{ and } W^i(u) = W^i(w(d)). \quad (4.7)$$

Thus for the average number of customers waiting in a single queue all regular

sequences of density d have the same performance as the upper bracket sequence $w(d)$ of density d . Moreover, this performance does only depend on the service time S of the queue and the density d . Therefore to simplify notation we write $L_S(d)$, or $L(d)$, if the value of S is clear from the context, instead of the more involved $L^i(w(d))$. Similarly we write $W_S(d)$ or $W(d)$ instead of $W^i(w(d))$. So, for a single queue with constant service time S we have that $L(d) = L_S(d)$ is the average number of customers waiting in the queue when the routing to the queue is according to a regular sequence of density d and $W(d) = W_S(d)$ is the average waiting time of these customers. Moreover, in [5] it is proved for rather general settings that routing according to a regular sequence of density d is optimal among all routing sequences of (lower) density at least d . This result implies the following lemma.

Lemma 4.5.1 *For every service time S and any routing sequence u^i of (lower) density at least d we have that $W^i(u^i) \geq W(d)$ and $L^i(u^i) \geq L(d)$.*

In the sequel if we consider a single queue then the remaining workload for the server (measured in time units) at time t is simply denoted by $V(t)$ instead of $V_i(t)$. Remember that the interarrival times of customers at the routing point are constant and equal to 1 by time scaling and that we assume that $V(0) = 0$. One of the objectives in this chapter is to give an efficient algorithm for computing $L_S(d)$ and $W_S(d)$ for arbitrary $S > 0$ and $d \in [0, 1]$. For this it is useful to consider for a fixed value of S the functions $d \rightarrow L(d)$ and $d \rightarrow W(d)$, where both these functions have domain $[0, 1]$ and codomain $[0, \infty]$. In [23] an algorithm is given to compute $W_S(d)$ and some properties of the function $W(d)$ are obtained. The following lemma follows directly from the results in [23].

Lemma 4.5.2 *If $d \leq \frac{1}{S}$ where S is the service time then $W(d) = 0$*

From Lemma 4.5.2 it follows that $W(d) = 0$ for every $d \in [0, 1]$ if the service time $S \leq 1$, which is a trivial case. Therefore we assume in the sequel that $S > 1$. If $S > 1$ then the queue is stable if $d \in [0, \frac{1}{S}]$ and unstable if $d \in (\frac{1}{S}, 1]$. Thus by (4.6) we have the following lemma.

Lemma 4.5.3

$$L(d) = d \cdot W(d) \text{ if } d \in [0, \frac{1}{S}].$$

Moreover, if $d \in (\frac{1}{S}, 1]$ then $L(d) = W(d) = \infty$.

So, there is a direct relation between $L(d)$ and $W(d)$ and if you have computed one of them then the other follows from Lemma 4.5.3. In Chapter 2 (see Theorem 2.7.11) the value of $W(d)$ was computed for $d = \frac{1}{S}$, the most right point of the interval of stability and traffic intensity $\rho = 1$. The following result was obtained.

Lemma 4.5.4 *For service time $S > 1$ we have that $W(\frac{1}{S}) = \frac{1}{2}$ if S is irrational. If $S = \frac{p}{q}$ with $p, q \in \mathbb{N}$, $\gcd(p, q) = 1$ then $W(\frac{1}{S}) = \frac{1}{2} - \frac{1}{2q}$.*

The following properties of the function $W(d)$ are proved in [23].

Theorem 4.5.5 *For given service time $S > 1$ the function $W(d)$ is continuous and monotonically increasing on the interval $[0, \frac{1}{S}]$. Moreover, on the interval $[\frac{1}{\lceil S \rceil}, \frac{1}{S}]$ the function $W(d)$ is strictly increasing.*

It also follows from the more general results in [39] that the function $L(d)$ is continuous and monotonically increasing on the stability interval. We prove some properties of the function $L(d)$ which do not hold for $W(d)$. These properties of the function $L(d)$ will be used in our algorithm for computing $L_S(d)$ and $W_S(d)$.

Let u be a finite sequence of zeros and ones of length k and suppose that k consecutively arriving customers are routed to the queue according to u . Suppose that the first of these k customers arrives at time t_0 and thus the last at time $t_0 + k - 1$. Then we say that u lasts from t_0 to $t_0 + k$ and thus $V(t_0)$ is the workload at the beginning of u and $V(t_0 + k)$ is the workload at the end of u . We say that u is workload non-increasing if for every initial workload $V(t_0)$ it holds that $V(t_0 + k) \leq V(t_0)$. Note that from Lindley's equation (4.1) it follows that for this it is sufficient that $V(t_0) = 0$ implies that $V(t_0 + k) = 0$. The following lemma is useful for proving properties of $L(d)$.

Lemma 4.5.6 *Let $d \in \mathbb{Q}$, $0 \leq d \leq 1$. Then $\omega(d)$, the period word of the upper bracket sequence of density d , is workload non-increasing if and only if $d \leq \frac{1}{S}$.*

Proof. Let $d = \frac{p}{q}$ where $p, q \in \mathbb{N}$ with $\gcd(p, q) = 1$. Then $q = |\omega(d)|$ and we may assume that $\omega(d)$ lasts from some $t_0 \in \mathbb{Z}_{\geq 0}$ to $t_1 = t_0 + q \in \mathbb{Z}_{\geq 0}$. Suppose that $d > \frac{1}{S}$. Then by (4.2) we have for $S = S_i$

$$V(t_1) = V(t_0) + \kappa_{\omega(d)}(q) \cdot S_i + (m_i(t_1) - m_i(t_0)) - q \geq$$

$$V(t_0) + p \cdot S_i - q = V(t_0) + q(d \cdot S_i - 1) > V(t_0).$$

Hence $\omega(d)$ is not workload non-increasing if $d > \frac{1}{S}$. Conversely, suppose that $d \leq \frac{1}{S}$ and $V(t_1) > V(t_0)$. Let $t^* = \max : t \in [t_0, t_1]$ for which $V(t^*) \leq V(t_0)$. Since $V(t)$ is monotonically non-increasing in any interval $(a, a + 1]$ for which $a \in \mathbb{Z}_{\geq 0}$ it follows that $t^* \in \mathbb{Z}_{\geq 0}$. Moreover, it follows that the server is non-idling in the interval $[t^*, t_1]$. So, by (4.2) and Lemma 4.2.2 we have that

$$\begin{aligned} V(t_1) &= V(t^*) + (\kappa_{\omega(d)}(q) - \kappa_{\omega(d)}(t^* - t_0)) \cdot S - (t_1 - t^*) \leq \\ V(t_0) + (p - \lceil d(t^* - t_0) \rceil) \cdot S - (t_0 + q - t^*) &\leq V(t_0) + (p - d(t^* - t_0)) \cdot S - (t_0 + q - t^*) \leq \\ V(t_0) + (p - d(t^* - t_0)) \cdot \frac{1}{d} - (t_0 + q - t^*) &= V(t_0), \end{aligned}$$

which contradicts $V(t_1) > V(t_0)$. Hence $\omega(d)$ is workload non-increasing if $d \leq \frac{1}{S}$. \square

Suppose that $d \in \mathbb{Q}$, $d \leq \frac{1}{S}$ and that the customers are routed to the queue according to the upper bracket sequence $w(d) = \omega(d)^\infty$. Then it follows from Lemma 4.5.6 that if the workload is 0 at the beginning of an $\omega(d)$ factor then it is also 0 at the end of the factor and thus at the beginning of the next factor. So, since the workload is 0 at $t = 0$, the beginning of the first $\omega(d)$ factor, it follows that the workload is 0 at the end of every $\omega(d)$ factor. Thus the workload process is renewed after every $\omega(d)$ factor. Hence we have the following corollary.

Corollary 4.5.7 *If $d \in \mathbb{Q}$, $d \leq \frac{1}{S}$ then $W(d)$ is equal to the average waiting time of customers routed to the queue during the first period $\omega(d)$. Similarly $L(d)$ is equal to the average number of customers in the queue during the first period $\omega(d)$.*

In [39] the following theorem is shown in a more general form. In fact, the convexity of $L(d)$ for any stationary sequence of generally distributed interarrival times and stationary generally distributed service times is proved, which implies the special deterministic case. We give a simple proof for the deterministic queueing system of this chapter.

Theorem 4.5.8 *For given service time $S > 1$ the function $L(d)$ is convex on the interval of stability $[0, \frac{1}{S}]$.*

Proof. Since $L(d)$, the average number of customers in the queue under the bracket policy with density d , is a continuous function of d on the interval of stability $[0, \frac{1}{S}]$, it is sufficient to show the following mid-point convexity of $L(d)$ (see [50] Theorem 1.7). Let $d_1 = \frac{p_1}{q_1}$ and $d_2 = \frac{p_2}{q_2}$ where $p_i, q_i \in \mathbb{N}$ with $\gcd(p_i, q_i) = 1$ and period

words $w_i := \omega(d_i)$, $i = 1, 2$. Define the mid-point $d := \frac{1}{2}(d_1 + d_2)$ and show that $L(d) \leq \frac{1}{2}(L(d_1) + L(d_2))$. In order to prove this we consider the admission sequence

$$w = (w_1^{q_2} w_2^{q_1})^\infty \tag{4.8}$$

A simple calculation gives that the density of w equals

$$\frac{p_1 q_2 + p_2 q_1}{2 q_1 q_2} = d.$$

It follows from Lemma 4.5.6 that the queue is empty after each w_i factor. Since $|w_1^{q_2}| = |w_2^{q_1}| = q_1 q_2$ it follows that

$$L(w) = \frac{1}{2}(L(d_1) + L(d_2)).$$

Hence, since the bracket policy is optimal (see [3]) we have

$$L(d) \leq L(w) = \frac{1}{2}(L(d_1) + L(d_2)).$$

□

The following theorem shows that the function $L(d)$ is not only convex, but also piecewise linear. This property will be of great use in our algorithm for computing $L_S(d)$ and $W_S(d)$.

Theorem 4.5.9 *Let $d_1, d_2 \in [0, 1]$ be rational numbers such that $I = [d_1, d_2]$ is a Farey interval and $d_1 < d_2 \leq \frac{1}{S}$, where S is the service time. Let $d \in I$, $d = \lambda \cdot d_1 + (1 - \lambda) \cdot d_2$, where $\lambda \in [0, 1]$. Then $L(d) = \lambda \cdot L(d_1) + (1 - \lambda) \cdot L(d_2)$.*

Proof. Arriving customers are routed to the queue according to routing sequence $w(d)$ and to compute $L(d)$ we assume that the workload is 0 at starting moment $t = 0$. According to Theorem 4.4.4 $w(d)$ can be factored in factors $\omega(d_1)$ and $\omega(d_2)$. From the assumption $d_1 < d_2 \leq \frac{1}{S}$ and Lemma 4.5.6 it follows that both the factors $\omega(d_1)$ and $\omega(d_2)$ are workload non-increasing. Thus if the workload in the queue is 0 at the beginning of such a factor then it is also 0 at the end of the factor and thus it is 0 at the beginning of the next factor. So, the queue is always empty at the beginning of an $\omega(d_1)$ or $\omega(d_2)$ factor. Hence the average number of customers in the queue during any $\omega(d_1)$ or $\omega(d_2)$ factor is $L(d_1)$ respectively $L(d_2)$. Hence, if p is the fraction of time the system is in an $\omega(d_1)$ factor and thus $(1 - p)$ is the fraction of time the system is in an $\omega(d_2)$ factor then

$$L(d) = p \cdot L(d_1) + (1 - p) \cdot L(d_2). \tag{4.9}$$

It is easily seen that $p = \lambda$ and thus (4.9) is valid with $p = \lambda$. This implies the theorem. \square

Remark. Thus we have that the function $L(d)$ is linear on Farey intervals contained in the stability interval. The linearity of the function $L(d)$ on Farey intervals can be generalised to some more generally distributed service times if the following condition holds. If $[d_1, d_2]$ with $0 \leq d_1 < d_2 \leq 1$ is the Farey interval then both the $\omega(d_1)$ and $\omega(d_2)$ factor should be workload non-increasing with probability one. If this condition holds then the proof of Theorem 4.5.9 remains valid.

4.6 Computation of the average waiting time

4.6.1 The value of $W(d)$ in best lower approximation points

Definition 4.6.1 *Let $x > 0$ be a given real number and let $r = \frac{p}{q}$ with $p \in \mathbb{Z}_{\geq 0}$, $q \in \mathbb{N}$ with $\gcd(p, q) = 1$ be a rational number in the interval $(0, x]$. Then r is a best lower approximation of x if there does not exist a rational number in the interval $(r, x]$ with denominator smaller than or equal to q .*

Note that x is a best lower approximation of x itself if and only if x is a rational number.

Lemma 4.6.2 *Let $S > 1$ be the service time and d be a best lower approximation of $\frac{1}{S}$. Then during an $\omega(d)$ factor the server is non-idling before the last letter of the factor. Moreover, if $d = \frac{p}{q}$ where $p, q \in \mathbb{N}$ with $\gcd(p, q) = 1$ and the arriving customers are routed to the queue according to the upper bracket sequence $w(d)$ then*

$$V(t) = \lceil \lceil t \rceil d \rceil S - t \text{ for every } t \in [0, q - 1]. \quad (4.10)$$

Proof. Let $d = \frac{p}{q}$ where $p, q \in \mathbb{N}$ with $\gcd(p, q) = 1$. Then $q = |\omega(d)|$ and we may assume that $\omega(d)$ lasts from some $t_0 \in \mathbb{Z}_{\geq 0}$ to $t_1 = t_0 + q \in \mathbb{Z}_{\geq 0}$. We have that $0 < d \leq \frac{1}{S} < 1$ and thus $q \geq 2$. Recall that $V(t)$ denotes the workload in the queue at moment $t \geq 0$. Let $k \in \{1, 2, \dots, q - 1\}$. Suppose that $\lceil d \cdot k \rceil S - k \leq 0$. Then it follows that $d = \frac{p}{q} \leq \frac{\lceil d \cdot k \rceil}{k} \leq \frac{1}{S}$. Since $\frac{\lceil d \cdot k \rceil}{k}$ is a rational number with denominator smaller than q it follows that $d < \frac{\lceil d \cdot k \rceil}{k} \leq \frac{1}{S}$, which contradicts that d is a best lower approximation of $\frac{1}{S}$. Thus $\lceil d \cdot k \rceil S - k > 0$. Hence by (4.2) and Lemma 4.2.2 we have that

$$V(t_0 + k) \geq V(t_0) + \kappa_{\omega(d)}(k) \cdot S - k = V(t_0) + \lceil d \cdot k \rceil S - k > V(t_0).$$

So, we have that $V(t_0 + k) > V(t_0)$ for every $k \in \{1, 2, \dots, q-1\}$ and thus $V(t) > 0$ for every $t \in (t_0, t_0 + q - 1]$. Thus the server is non-idling in the time interval $[0, q - 1]$. So, by (4.2) and Lemma 4.2.2 we have for every $t \in [0, q - 1]$ that

$$V(t) = V(0) + \kappa_{w(d)}(\lceil t \rceil) \cdot S - t = \lceil \lceil t \rceil d \rceil S - t.$$

□

Theorem 4.6.3 *Let $S > 1$ be the service time and d be a best lower approximation of $\frac{1}{S}$. Let $d = \frac{p}{q}$ where $p, q \in \mathbb{N}$ with $\gcd(p, q) = 1$. Then*

$$W(d) = \frac{p-1}{2p}(pS + 1 - q) \text{ and} \quad (4.11)$$

$$L(d) = \frac{p-1}{2q}(pS + 1 - q). \quad (4.12)$$

Proof. We have that $d \leq \frac{1}{S}$. Thus according to Lemma 4.5.7 $W(d)$ is equal to the average waiting time in the first period $\omega(d)$. From $|\omega(d)| = q$ and (4.5) it follows that

$$W(d) = \frac{1}{\kappa_{\omega(d)}(q)} \cdot \sum_{t=1}^q \omega(d)_t \cdot V(t-1) = \frac{1}{p} \cdot \sum_{t=0}^{q-1} \omega(d)_{t+1} \cdot V(t). \quad (4.13)$$

By Definition 4.2.1 and (4.10) we have for $t = 0, 1, \dots, q-1$ that

$$\omega(d)_{t+1} \cdot V(t) = (\lceil (t+1)d \rceil - \lceil t \cdot d \rceil)(\lceil td \rceil S - t). \quad (4.14)$$

Let $A = \{0, \lfloor \frac{1}{d} \rfloor, \lfloor \frac{2}{d} \rfloor, \dots\}$. Then it follows by Lemma 4.2.3 for every $t \in \mathbb{Z}_{\geq 0}$ that

$$w(d)_{t+1} = \lceil (t+1)d \rceil - \lceil t \cdot d \rceil = \begin{cases} 1 & \text{if } t \in A \\ 0 & \text{if } t \notin A \end{cases} \quad (4.15)$$

On combining (4.13), (4.14) and (4.15) we obtain

$$\begin{aligned} W(d) &= \frac{1}{p} \cdot \sum_{t \in A \cap \{0, 1, \dots, q-1\}} (\lceil td \rceil S - t) = \frac{1}{p} \cdot \sum_{t'=0}^{p-1} (\lceil \lfloor \frac{t'}{d} \rfloor d \rceil S - \lfloor \frac{t'}{d} \rfloor) = \\ &= \frac{S}{p} \cdot \sum_{t'=0}^{p-1} (\lceil \lfloor \frac{t'}{d} \rfloor d \rceil) - \frac{1}{p} \cdot \sum_{t'=0}^{p-1} (\lfloor t' \cdot \frac{q}{p} \rfloor). \end{aligned} \quad (4.16)$$

It is easily seen that if $0 < x < 1$ then $\lceil \lfloor \frac{k}{x} \rfloor x \rceil = k$ for $k = 0, 1, \dots$. Thus

$$\frac{S}{p} \cdot \sum_{t'=0}^{p-1} (\lceil \lfloor \frac{t'}{d} \rfloor d \rceil) = \frac{S}{p} \cdot \sum_{t'=0}^{p-1} t' = \frac{S(p-1)}{2}.$$

Moreover, analogously to Theorem 100 in [30] (cf. Lemma 3.5.7) it follows that

$$\sum_{t'=0}^{p-1} (\lfloor t' \cdot \frac{q}{p} \rfloor) = \frac{1}{2}(p-1)(q-1).$$

Hence by (4.16) we have that

$$W(d) = \frac{S(p-1)}{2} - \frac{1}{2p}(p-1)(q-1) = \frac{p-1}{2p}(pS+1-q).$$

By Lemma 4.5.3 it follows that

$$L(d) = d \cdot W(d) = \frac{p-1}{2q}(pS+1-q).$$

□

So, for given service time $S > 1$ the values of $W(d)$ and $L(d)$ are easily obtained for all best lower approximations of $\frac{1}{S}$. Before we give a procedure to obtain iteratively the best lower approximations of $\frac{1}{S}$, we first deduce the following lemma.

Lemma 4.6.4 *If I is a subinterval of positive Lebesgue measure of the open interval $(0, 1)$, then there exists a unique rational number of lowest denominator in I .*

Proof. Since I has positive Lebesgue measure and the rational numbers are dense in the real numbers, I contains rational numbers. Suppose there exist $a, b \in \mathbb{N}$ such that $[\frac{a}{b}, \frac{a+1}{b}] \subseteq I$ and there exist no rational numbers in the interval $(\frac{a}{b}, \frac{a+1}{b})$ with denominator smaller or equal than b . Then $\frac{a}{b}, \frac{a+1}{b}$ are consecutive elements of the Farey series of order b (see [30], page 23) and thus we have by Theorem 28 in [30] that $(a+1)b - ba = b = 1$. Hence $\frac{a}{b}$ is an integer which contradicts that $I \subseteq (0, 1)$. □

Let r_1 be the rational number of lowest denominator in the interval $(0, \frac{1}{S}]$ and put $i := 1$. Do the following iteratively. If $r_i = \frac{1}{S}$ then stop. Else let r_{i+1} be the rational number of lowest denominator in the interval $(r_i, \frac{1}{S}]$ and let $i := i + 1$. According to Lemma 4.6.4 the r_i are well defined.

Lemma 4.6.5 *The number d is a best lower approximation of $\frac{1}{S}$ if and only if $d = r_i$ for some $i \in \mathbb{N}$. Moreover, if S is rational then there exists some $k \in \mathbb{N}$ such that*

$$\frac{1}{\lceil S \rceil} = r_1 < r_2 < \dots < r_k = \frac{1}{S}.$$

If S is irrational then

$$\frac{1}{\lceil S \rceil} = r_1 < r_2 < \dots \text{ and } \lim_{i \rightarrow \infty} r_i = \frac{1}{S}.$$

Proof. Let $r_i = \frac{p_i}{q_i}$ with $p_i, q_i \in \mathbb{N}$ and $\gcd(p_i, q_i) = 1$. We first prove that $r_1 = \frac{1}{\lceil S \rceil}$. Suppose that $q_1 < \lceil S \rceil$. Then $r_1 = \frac{p_1}{q_1} \geq \frac{1}{q_1} > \frac{1}{S}$, which gives a contradiction. Since $\frac{1}{\lceil S \rceil} \leq \frac{1}{S}$ we have that $q_1 = \lceil S \rceil$ and by Lemma 4.6.4 it follows that $r_1 = \frac{1}{\lceil S \rceil}$. By definition and Lemma 4.6.4 we have that the r_i are best lower approximations. Moreover, both the sequence r_1, r_2, \dots of best lower approximations and the sequence q_1, q_2, \dots of denominators are strictly increasing. Let $d = \frac{a_1}{a_2}$ be a best lower approximation of $\frac{1}{S}$ with $a_1, a_2 \in \mathbb{N}$ and $\gcd(a_1, a_2) = 1$. Then there exist some $l \in \mathbb{N}$ such that $q_i \geq a_2$ if and only if $i \geq l$. From the definition it follows that $q_l = a_2$ and thus we have by Lemma 4.6.4 that $r_l = \frac{a_1}{a_2} = d$. So, d is a best lower approximation of $\frac{1}{S}$ if and only if $d = r_i$ for some $i \in \mathbb{N}$. If S is rational then $\frac{1}{S}$ is a best lower approximation of itself and thus we have that $r_k = \frac{1}{S}$ for some $k \in \mathbb{N}$. If S is irrational then $\frac{1}{S}$ is irrational and it follows from the definition that $r_i < \frac{1}{S}$ for every $i \in \mathbb{N}$. Thus we have that $r := \lim_{i \rightarrow \infty} r_i \leq \frac{1}{S}$. Suppose that $r < \frac{1}{S}$. Then by definition there exist no rational numbers in the interval $(r, \frac{1}{S})$, which contradicts the fact that the rational numbers are dense in the real numbers. \square

Lemma 4.6.6 *If r_i, r_{i+1} are consecutive best lower approximations of $\frac{1}{S}$ then $[r_i, r_{i+1}]$ is a Farey interval.*

Proof. Let $r_i = \frac{a_1}{b_1}$ with $a_1, b_1 \in \mathbb{N}$, $\gcd(a_1, b_1) = 1$ and let $r_{i+1} = \frac{a_2}{b_2}$ with $a_2, b_2 \in \mathbb{N}$, $\gcd(a_2, b_2) = 1$. By definition and Lemma 4.6.4 we have that $b_1 < b_2$ and that there exist no rational numbers in the interval (r_i, r_{i+1}) with denominator smaller than or equal to b_2 . Hence $\frac{a_1}{b_1}, \frac{a_2}{b_2}$ are consecutive members of the Farey series of order b_2 and thus by Theorem 28 in [30] we have that $b_1 \cdot a_2 - a_1 \cdot b_2 = 1$. Thus $[r_i, r_{i+1}]$ is a Farey interval. \square

For every service time $S > 1$ we have by Theorem 4.5.8 that the function $L(d)$ is convex on the interval $[0, \frac{1}{S}]$. Moreover, it follows from Lemma 4.5.2 that $L(d)$ is linear on $[0, r_1]$ and from Theorem 4.5.9 and Lemma 4.6.6 that $L(d)$ is linear on every interval $[r_i, r_{i+1}]$. Thus it follows from Lemma 4.6.5 that $L(d)$ is piecewise linear on the interval $[0, \frac{1}{S}]$. Therefore the slope of $L(d)$ can only change in the best lower approximation points r_i . In [23] these points r_i are called jump points. According to Lemma 4.5.3 $L(d) = W(d) = \infty$ for $d > \frac{1}{S}$.

The value of $L_S(d)$ and $W_S(d)$ can be exactly computed for every given service time $S > 1$ and $d \in [0, 1]$. The following algorithm can be used to compute $L(d)$ and $W(d)$ for any d .

Algorithm. If $0 \leq d \leq \frac{1}{\lceil S \rceil}$ then $L(d) = W(d) = 0$. Else if $\frac{1}{S} < d \leq 1$ then $L(d) = W(d) = \infty$. Else if $d = \frac{1}{S}$ then $W(d)$ is given by Lemma 4.5.4 and $L(d) =$

$d \cdot W(d)$. Else we have that $\frac{1}{\lceil S \rceil} < d < \frac{1}{S}$. Then there exist consecutive best lower approximations points r_i and r_{i+1} such that $d \in [r_i, r_{i+1}]$ (in the following subsection we give a method based on the continued fraction expansion of $\frac{1}{S}$ to determine these points r_i, r_{i+1} in an efficient way). If $d = r_i$ or $d = r_{i+1}$ then $W(d)$ and $L(d)$ are given by Theorem 4.6.3. Else $r_i < d < r_{i+1}$. Then compute $L(r_i)$ and $L(r_{i+1})$ by applying Theorem 4.6.3 and put $\lambda = \frac{r_{i+1}-d}{r_{i+1}-r_i}$. Then $L(d) = \lambda \cdot L(r_i) + (1-\lambda) \cdot L(r_{i+1})$ and $W(d) = \frac{L(d)}{d}$.

We illustrate the algorithm by the following example.

Example. We calculate $W_S(d)$ for $S = \frac{31}{19}$ and $d = \frac{\sqrt{6}}{4}$. The best lower approximations of $\frac{1}{S} = \frac{19}{31}$ are $r_1 = \frac{1}{2}$, $r_2 = \frac{3}{5}$, $r_3 = \frac{11}{18}$ and $r_4 = \frac{19}{31}$. Since $d \in [\frac{11}{18}, \frac{19}{31}]$ we obtain from Theorem 4.6.3 that $L(\frac{11}{18}) = \frac{5}{19}$ and $L(\frac{19}{31}) = \frac{9}{31}$. We have $\lambda = \frac{\frac{19}{31}-d}{\frac{19}{31}-\frac{11}{18}} = 342 - \frac{279\sqrt{6}}{2}$ and $L(d) = \lambda \cdot L(\frac{11}{18}) + (1-\lambda) \cdot L(\frac{19}{31}) = \frac{72\sqrt{6}}{19} - 9$. Hence $W_S(d) = \frac{L(d)}{d} = \frac{288}{19} - 6\sqrt{6}$.

4.6.2 Best lower approximations and the continued fraction expansion

In this subsection we show that the best lower approximations r_1, r_2, \dots of $\frac{1}{S}$ can be obtained by applying the continued fraction expansion of $\frac{1}{S}$. For given $d \in (\frac{1}{\lceil S \rceil}, \frac{1}{S})$ this gives an efficient algorithm for computing consecutive best lower approximation r_i, r_{i+1} of $\frac{1}{S}$ such that $d \in [r_i, r_{i+1}]$.

See for example [30] for definitions and properties of the continued fraction expansion of a real number. Put $\alpha := \frac{1}{S}$. The partial quotients a_0, a_1, \dots of the (simple) continued fraction expansion of α are recursively defined by:

$$\left\{ \begin{array}{l} a_0 = \lfloor \alpha \rfloor; \quad \alpha_1 = \frac{1}{\alpha - a_0} \\ a_n = \lfloor \alpha_n \rfloor; \quad \alpha_{n+1} = \frac{1}{\alpha_n - a_n} \text{ for } n = 1, 2, \dots \end{array} \right\}.$$

Then

$$\alpha = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \dots \frac{1}{a_n + \dots}}}} := [a_0, a_1, \dots, a_n, \dots].$$

Since $S > 1$ we have that $0 < \alpha < 1$ and thus $a_0 = 0$. Moreover, a_1, a_2, \dots are positive integers. If α is rational then $\alpha_m - a_m = 0$ for some $m \in \mathbb{N}$ and the process of computing the partial quotients stops for $n = m$. Then the continued fraction expansion of α is finite and $\alpha = [0, a_1, a_2, \dots, a_m]$. If α is irrational then the continued fraction expansion of α is infinite.

We define p_n, q_n recursively by

$$\begin{aligned} p_0 &= a_0, & p_1 &= a_0 a_0 + 1, & p_n &= a_n p_{n-1} + p_{n-2} (n \geq 2), \\ q_0 &= 1, & q_1 &= a_1, & q_n &= a_n q_{n-1} + q_{n-2} (n \geq 2) \end{aligned}$$

Then $\frac{p_n}{q_n} = [a_0, a_1, \dots, a_n]$ and $\frac{p_n}{q_n}$ is called the n th convergent of $\alpha = [a_0, a_1, \dots]$. If $\alpha = [a_0, a_1, \dots, a_m]$ is rational then the m th convergent $\frac{p_m}{q_m} = \alpha$ is the last convergent of α and thus the number of convergents of α is finite. If α is irrational then the number of convergents of α is infinite. We put $x_n := \frac{p_n}{q_n}$. Then the convergents of α have the following properties (see [30]).

Lemma 4.6.7 (i) *The even convergents x_{2n} increase strictly with n , while the odd convergents x_{2n+1} decrease strictly.*

(ii) *Every even convergent is smaller than α and every odd convergent is greater than α (except that in case of rational α the last convergent is equal to α , whether this be even or odd).*

(iii) *If α is irrational then $\lim_{n \rightarrow \infty} x_n = \alpha$.*

Lemma 4.6.8 *If $n > 1$, $0 < q \leq q_n$, and $\frac{p}{q} \neq \frac{p_n}{q_n}$, then*

$$\left| \frac{p_n}{q_n} - \alpha \right| < \left| \frac{p}{q} - \alpha \right|.$$

Lemma 4.6.7 and Lemma 4.6.8 have the following corollary.

Corollary 4.6.9 *The even convergents $\frac{p_{2n}}{q_{2n}}$, $n = 1, 2, \dots$ are best lower approximations of α .*

So, the even convergents obtained by the continued fraction expansion of α are best lower approximations of α . However, in general the converse is not true. First of all if α is equal to the n -th convergent $\frac{p_n}{q_n}$ for some odd positive integer n then this odd convergent is also a best approximation of α . Moreover, to obtain all best lower approximations of α from the continued fraction of α we also have to consider the so-called even intermediate convergents, which we define now.

Definition 4.6.10 *Let $\alpha = [a_0, a_1, \dots]$ if α is irrational or $\alpha = [a_0, a_1, \dots, a_m]$ for some $m \in \mathbb{N}$ if α is rational. Then a rational number $\frac{p}{q}$ is an even intermediate*

convergent of α if and only if $p = p_{n-2} + c \cdot p_{n-1}$ and $q = q_{n-2} + c \cdot q_{n-1}$ for some positive even integer n (with n smaller or equal than m if α is rational) and $c \in \{1, 2, \dots, a_n - 1\}$, where $\frac{p_n}{q_n}$ is the n -th convergent of α .

Note that for every positive even integer $n \leq m$ we have that

$$\begin{aligned} \frac{p_{n-2}}{q_{n-2}} &< \frac{p_{n-2} + p_{n-1}}{q_{n-2} + q_{n-1}} < \frac{p_{n-2} + 2p_{n-1}}{q_{n-2} + 2q_{n-1}} < \dots < \\ \frac{p_{n-2} + (a_n - 1)p_{n-1}}{q_{n-2} + (a_n - 1)q_{n-1}} &< \frac{p_{n-2} + a_n p_{n-1}}{q_{n-2} + a_n q_{n-1}} = \frac{p_n}{q_n} \leq \alpha. \end{aligned} \quad (4.17)$$

So, all even intermediate convergents of α are smaller than α . Moreover, it can be shown (see for example [55]) that every even intermediate convergent $\frac{p}{q}$ of α is a best lower approximation of α . Conversely, we have the following lemma, which follows directly from results in [55].

Lemma 4.6.11 *Let $r_i = \frac{p}{q}$ with $\gcd(p, q) = 1$ be a best lower approximation of $\alpha \in (0, 1)$. Then one of the following three possibilities holds for r_i .*

- r_i is an even convergent $\frac{p_{2n}}{q_{2n}}$ with $n \geq 1$ of α .
- r_i is an even intermediate convergent of α .
- r_i is equal to α , which is an odd convergent of itself.

Recall that we ordered the best lower approximations r_1, r_2, \dots of $\alpha \in (0, 1)$ such that $r_1 < r_2 < \dots$ and that $\frac{p_n}{q_n}$ is the n -th convergent of α . Using the preceding results it is possible to identify the ordered set $B := (r_1, r_2, \dots)$ of best lower approximations of α in terms of the convergents and the partial quotients of α . Doing this we have three slightly different cases. First note that $p_0 = 0$ and $q_0 = 1$.

1. If $\alpha = [0, a_1, a_2, \dots]$ is irrational then the ordered set B is infinite,

$$B = \left(\frac{p_1}{q_1 + 1}, \frac{2p_1}{2q_1 + 1}, \dots, \frac{a_2 p_1}{a_2 q_1 + 1} = \frac{p_2}{q_2}, \frac{p_3 + p_2}{q_3 + q_2}, \frac{2p_3 + p_2}{2q_3 + q_2}, \dots, \frac{p_4}{q_4}, \dots \right). \quad (4.18)$$

2. If α is rational and an even convergent of itself, $\alpha = [0, a_1, a_2, \dots, a_k] = \frac{p_k}{q_k}$ where k is even, then B is finite,

$$B = \left(\frac{p_1}{q_1 + 1}, \frac{2p_1}{2q_1 + 1}, \dots, \frac{a_2 p_1}{a_2 q_1 + 1} = \frac{p_2}{q_2}, \frac{p_3 + p_2}{q_3 + q_2}, \frac{2p_3 + p_2}{2q_3 + q_2}, \dots, \right. \\ \left. \frac{p_{k-2}}{q_{k-2}}, \frac{p_{k-1} + p_{k-2}}{q_{k-1} + q_{k-2}}, \frac{2p_{k-1} + p_{k-2}}{2q_{k-1} + q_{k-2}}, \dots, \frac{a_k p_{k-1} + p_{k-2}}{a_k q_{k-1} + q_{k-2}} = \frac{p_k}{q_k} = \alpha \right). \quad (4.19)$$

3. If α is rational and an odd convergent of itself, $\alpha = [0, a_1, a_2, \dots, a_k] = \frac{p_k}{q_k}$ where k is odd, then B is finite,

$$B = \left(\frac{p_1}{q_1 + 1}, \frac{2p_1}{2q_1 + 1}, \dots, \frac{a_2 p_1}{a_2 q_1 + 1} = \frac{p_2}{q_2}, \frac{p_3 + p_2}{q_3 + q_2}, \frac{2p_3 + p_2}{2q_3 + q_2}, \dots, \frac{p_{k-1}}{q_{k-1}}, \frac{p_k}{q_k} = \alpha \right). \quad (4.20)$$

Example. Suppose that $\alpha = \frac{339}{410}$. Then the continued fraction algorithm gives that $\alpha = [0, 1, 4, 1, 3, 2, 3, 2]$ and the convergents $\{\frac{p_n}{q_n}\}_{n=0}^7$ of α are $\frac{0}{1}, \frac{1}{1}, \frac{4}{5}, \frac{5}{6}, \frac{19}{23}, \frac{43}{52}, \frac{148}{179}, \frac{339}{410}$. By (4.20) it follows that the ordered set of best lower approximations of α is $B = (\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{9}{11}, \frac{14}{17}, \frac{19}{23}, \frac{62}{75}, \frac{105}{127}, \frac{148}{179}, \frac{339}{410})$.

We now give the algorithm based on continued fraction expansion for computing for any $d \in (\frac{1}{\lceil S \rceil}, \frac{1}{S})$ two consecutive best lower approximations r_i, r_{i+1} of $\alpha = \frac{1}{S}$ such that $d \in [r_i, r_{i+1}]$.

Algorithm. Apply the continued fraction algorithm to find consecutively the partial quotients a_1, a_2, \dots and the corresponding convergents $\frac{p_1}{q_1}, \frac{p_2}{q_2}, \dots$ of α until we have found an even convergent $\frac{p_{2n}}{q_{2n}}$, that is greater or equal than d or we have that $\frac{p_{N-1}}{q_{N-1}} < d < \frac{p_N}{q_N} = \alpha$ for some odd integer $N > 1$. Note that according to Lemma 4.6.7 this first part of the algorithm stops after finite time. In the latter case we put $r_i = \frac{p_{N-1}}{q_{N-1}}$ and $r_{i+1} = \frac{p_N}{q_N}$. Then $d \in (r_i, r_{i+1})$ and we are finished. So, suppose we have the former case. Then by Lemma 4.6.7 we have that $\frac{p_{2n-2}}{q_{2n-2}} < d \leq \frac{p_{2n}}{q_{2n}}$ and by (4.17) there exists some unique integer k , $0 \leq k < a_{2n}$, such that $\frac{k \cdot p_{2n-1} + p_{2n-2}}{k \cdot q_{2n-1} + q_{2n-2}} < d \leq \frac{(k+1) \cdot p_{2n-1} + p_{2n-2}}{(k+1) \cdot q_{2n-1} + q_{2n-2}}$. Obviously this holds for $k = \lfloor \frac{d \cdot q_{2n-1} - p_{2n-2}}{p_{2n-1} - d \cdot q_{2n-1}} \rfloor$. By putting $r_i = \frac{k \cdot p_{2n-1} + p_{2n-2}}{k \cdot q_{2n-1} + q_{2n-2}}$ and $r_{i+1} = \frac{(k+1) \cdot p_{2n-1} + p_{2n-2}}{(k+1) \cdot q_{2n-1} + q_{2n-2}}$ we have that $d \in (r_i, r_{i+1}]$ and we are finished.

We give an example to illustrate this algorithm.

Example.

Suppose $\alpha := \frac{1}{5} = \frac{7433}{9734} + \frac{\sqrt{18773}}{9734}$ and $d = 0.77761$. Then applying the continued fraction algorithm to α we consecutively find $a_0 = 0$, $\frac{p_0}{q_0} = \frac{0}{1}$, $a_1 = 1$, $\frac{p_1}{q_1} = 1$, $a_2 = 3$, $\frac{p_2}{q_2} = \frac{3}{4}$, $a_3 = 2$, $\frac{p_3}{q_3} = \frac{7}{9}$, $a_4 = 137$, $\frac{p_4}{q_4} = \frac{962}{1237}$ and we have that $d \leq \frac{p_4}{q_4}$. Then we compute that $k = \lfloor \frac{d \cdot q_2 - p_2}{p_3 - d \cdot q_3} \rfloor = 73$ and it follows that $r_i = \frac{k \cdot p_3 + p_2}{k \cdot q_3 + q_2} = \frac{514}{661}$ and $r_{i+1} = \frac{(k+1) \cdot p_3 + p_2}{(k+1) \cdot q_3 + q_2} = \frac{521}{670}$.

Remark. The number of operations needed for this algorithm based on the continued fraction expansion is linear in the number of digits of the denominator of r_{i+1} . So, this algorithm is very efficient in practice. Moreover, as illustrated in the last example of the previous subsection the exact value of $W_S(d)$ can be easily calculated after we have obtained the consecutive best lower approximations r_i, r_{i+1} of $\frac{1}{5}$ such that $d \in (r_i, r_{i+1}]$. So, this is an efficient method for calculating the exact value of $W_S(d)$ and in most cases it is quicker than the recursive algorithm used in [23]. Indeed, if this recursive algorithm is applied then, when $d \in (r_i, r_{i+1}]$, all the best lower approximations $r_1 < r_2 < \dots < r_i < r_{i+1}$ of $\frac{1}{5}$ and the corresponding value of $W_S(r_n)$ for $n = 1, 2, \dots, i+1$ are explicitly calculated. So, if q is the denominator of r_{i+1} then this recursive algorithm for computing $W_S(d)$ has to do a number of order q calculations in worst case. In our algorithm not all the best lower approximations of $\frac{1}{5}$ are explicitly calculated. Indeed, only the convergents of $\frac{1}{5}$ and the last two best lower approximations r_i, r_{i+1} with corresponding values $W_S(r_i)$ and $W_S(r_{i+1})$ are calculated. Therefore our algorithm only has to do a number of order $\log(q)$ calculations.

4.7 Minimization over multiple queues

4.7.1 Properties of a minimal point

In this section we obtain results on the minimal average waiting time in case of N parallel queues as introduced in Section 4.2.

For a given $\bar{S} := (S_1, S_2, \dots, S_N)$ queueing system with N parallel FIFO queues and $(d_1, d_2, \dots, d_n) \in [0, 1]^N$ we put

$$L_{\bar{S}}(d_1, d_2, \dots, d_n) = L(d_1, d_2, \dots, d_n) := \sum_{i=1}^N L_{S_i}(d_i).$$

Let U be a routing policy applied in an $\bar{S} = (S_1, S_2, \dots, S_N)$ system such that routing sequence u^i has density d_i for $i = 1, 2, \dots, N$. Then by Proposition 4.3.2 and Lemma 4.5.1 we have that

$$W(U) = \sum_{i=1}^N L^i(u^i) \geq \sum_{i=1}^N L_{S_i}(d_i) = L_{\bar{S}}(d_1, d_2, \dots, d_n). \quad (4.21)$$

Put

$$H^N := \{(x_1, x_2, \dots, x_N) \in \mathbb{R}^N : x_i \geq 0, i = 1, 2, \dots, N \text{ and } \sum_{i=1}^N x_i = 1\}.$$

It is easily seen that if some routing policy U has densities d_1, d_2, \dots, d_N then $(d_1, d_2, \dots, d_N) \in H^N$. Moreover, since H^N is a convex and compact subset of \mathbb{R}^N and $L_{\bar{S}}(d_1, d_2, \dots, d_n)$ is convex in (d_1, d_2, \dots, d_n) (see Theorem 4.5.8) we have the following lemma.

Lemma 4.7.1 *For an $\bar{S} := (S_1, S_2, \dots, S_N)$ queueing system put*

$$M(\bar{S}) := \{(d_1, d_2, \dots, d_N) \in H : L_{\bar{S}}(d_1, d_2, \dots, d_n) = \min_{H^N} L_{\bar{S}}(x_1, x_2, \dots, x_N)\}. \quad (4.22)$$

Then $M(\bar{S})$ is an nonempty convex and compact subset of H^N .

We put

$$R(\bar{S}) = R((S_1, S_2, \dots, S_N)) := L_{\bar{S}}(d_1, d_2, \dots, d_n), \text{ where } (d_1, d_2, \dots, d_n) \in M(\bar{S}).$$

Observe $R(\bar{S})$ is a lower bound for the average waiting time in the \bar{S} system, since Theorem 4.1 and Theorem 4.2 of [5] give the following result, in which we do not require that the routing sequences u^i have densities d_i for $i = 1, 2, \dots, N$.

Proposition 4.7.2 *For any routing policy U applied to the $\bar{S} := (S_1, S_2, \dots, S_N)$ queueing system we have that*

$$W(U) \geq R(\bar{S}).$$

We also have upper bounds for the average waiting time in the $\bar{S} = (S_1, S_2, \dots, S_N)$ system. By Theorem 3.4.7 we have the following proposition.

Proposition 4.7.3 *Let U be a periodic routing policy such that the routing sequence u^i has density d_i for $i = 1, 2, \dots, N$. Then*

$$W(U) \leq L_{\bar{S}}(d_1, d_2, \dots, d_N) + \bar{O}(U),$$

where $\bar{O}(U) \geq 0$ is the total primal unbalance of U .

The total primal unbalance can always be bounded. Indeed, by Theorem 3.5.1 we have the following proposition.

Proposition 4.7.4 *Let rational densities $d_1, d_2, \dots, d_N \in \mathbb{Q}_{>0}$ with $\sum_{i=1}^N d_i = 1$ be given. Then there exists a periodic routing policy U such that routing sequence u^i has density d_i for $i = 1, 2, \dots, N$ and $\bar{O}(U) \leq \frac{N}{2} - 1$.*

We recall from Section 3.5 that such a routing sequence policy U with given densities d_i and $\bar{O}(U) \leq \frac{N}{2} - 1$ can easily be obtained by constructing a billiard sequence. Moreover, the upper bound of Proposition 4.7.3 is tight. Note that by the definition of $L_{\bar{S}}(d_1, d_2, \dots, d_N)$ we have that $W(U) = L_{\bar{S}}(d_1, d_2, \dots, d_N)$ if for every $i \in \{1, 2, \dots, N\}$ the routing sequence u^i is a regular sequence of density d_i . Indeed, by the definition of the total primal unbalance, $\bar{O}(U) = 0$ if and only if u^i is a regular sequence for $i = 1, 2, \dots, N$.

Suppose that $\sum_{i=1}^N \frac{1}{S_i} < 1$. Then for every $(x_1, x_2, \dots, x_N) \in H^N$ there exists some i for which $x_i > \frac{1}{S_i}$ and thus Lemma 4.5.3 implies $R(\bar{S}) = \infty$. So, by Proposition 4.7.2, $W(U) = \infty$ for every routing policy U . Thus the queueing system is unstable in this case and optimal policies do not exist. Therefore we assume in the sequel that $\sum_{i=1}^N \frac{1}{S_i} \geq 1$. We shall show in Proposition 4.7.14 that the lower bound $R(\bar{S})$ is finite and in Theorem 4.7.15 that there exist policies U with $W(U) \leq \frac{N-1}{2}$. Moreover, we characterize the set $M(\bar{S})$ and we show that this set contains some special rational point if $\sum_{i=1}^N \frac{1}{S_i} > 1$. Hence there exists a periodic policy for which the average waiting time is bounded by $R((S_1, S_2, \dots, S_N)) + \frac{N}{2} - 1$. A corollary of this result is that there exists a periodic policy which is optimal if $N = 2$, which was proved in [23] and is generalized to the routing of customers in two parallel networks of queues in tandem with deterministic service times in [24]. Finally we give an algorithm to obtain a rational point $x \in M(\bar{S})$ and the corresponding lower bound $R(\bar{S})$. If $N = 2$ then the rational point obtained by the algorithm yields an optimal routing policy as in [23]. If $N > 2$ then in general it is not possible to obtain an optimal policy from the rational point, but it can be used to construct periodic policies with performance close to the lower bound $R(\bar{S})$.

For given service time $S > 1$ and $d \in [0, 1]$ we define

Definition 4.7.5

$$l_S(d) = \begin{cases} 0 & \text{if } d \in [0, \frac{1}{\lceil \bar{S} \rceil}] \\ \lim_{\varepsilon \downarrow 0} \frac{L_S(d) - L_S(d-\varepsilon)}{\varepsilon} & \text{if } d \in (\frac{1}{\lceil \bar{S} \rceil}, \frac{1}{\bar{S}}] \\ \infty & \text{if } d \in (\frac{1}{\bar{S}}, 1] \end{cases} \quad (4.23)$$

and

$$r_S(d) = \begin{cases} 0 & \text{if } d \in [0, \frac{1}{\lceil \bar{S} \rceil}) \\ \lim_{\varepsilon \downarrow 0} \frac{L_S(d+\varepsilon) - L_S(d)}{\varepsilon} & \text{if } d \in [\frac{1}{\lceil \bar{S} \rceil}, \frac{1}{\bar{S}}) \\ \infty & \text{if } d \in [\frac{1}{\bar{S}}, 1] \end{cases} . \quad (4.24)$$

From Theorem 4.5.8 it follows that $l_S(d)$ and $r_S(d)$ are monotonically non-decreasing in d . Moreover, they assume only countably many values, since they are constant between consecutive best lower approximations of $\frac{1}{\bar{S}}$.

The following lemma follows directly from Theorem 4.6.3 and Lemma 4.6.5.

Lemma 4.7.6 *If $S > 1$ is rational then $l_S(\frac{1}{\bar{S}}) < \infty$. If $S > 1$ is irrational then $l_S(\frac{1}{\bar{S}}) = \infty$.*

The following proposition gives a characterization of elements of $M(\bar{S})$.

Proposition 4.7.7 *Let $\bar{S} = (S_1, S_2, \dots, S_N)$ and $x = (x_1, x_2, \dots, x_N) \in H^N$. Then $x \in M(\bar{S})$ if and only if for every pair $i, j \in \{1, 2, \dots, N\}$ it holds that either $l_{S_i}(x_i) \leq r_{S_j}(x_j) < \infty$ or $r_{S_j}(x_j) = \infty$.*

Proof. We first prove the “only if” part. Suppose that $x \in M(\bar{S})$ and there exist $i, j \in \{1, 2, \dots, N\}$ such that $r_{S_j}(x_j) < \infty$ and $l_{S_i}(x_i) > r_{S_j}(x_j)$. Since $\sum_{i=1}^N \frac{1}{S_i} \geq 1$ and thus $L_{\bar{S}}(x_1, x_2, \dots, x_N) = R(\bar{S}) < \infty$, it follows that $0 \leq x_n \leq \frac{1}{S_n}$ for $n = 1, 2, \dots, N$. Moreover, it follows from (4.23) and (4.24) that $x_i > \frac{1}{\lceil \bar{S}_i \rceil} > 0$ and $x_j < \frac{1}{S_j}$. Put $y(\varepsilon) := (y_1, y_2, \dots, y_N)$, where $y_i := x_i - \varepsilon$, $y_j := x_j + \varepsilon$ and $y_l := x_l$ for $l \neq i, j$. Then $L_{\bar{S}}(x) - L_{\bar{S}}(y(\varepsilon)) = \sum_{n=1}^N (L_{S_n}(x_n) - L_{S_n}(y_n)) = (L_{S_i}(x_i) - L_{S_i}(x_i - \varepsilon)) - (L_{S_j}(x_j + \varepsilon) - L_{S_j}(x_j))$ and thus

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \cdot (L_{\bar{S}}(x) - L_{\bar{S}}(y(\varepsilon))) = l_{S_i}(x_i) - r_{S_j}(x_j) > 0.$$

So, if $\varepsilon' > 0$ is an arbitrarily small positive number then $y(\varepsilon') \in H^N$ and $L_{\bar{S}}(x) - L_{\bar{S}}(y(\varepsilon')) > 0$, which contradicts the fact that $x \in M(\bar{S})$.

The “if” part follows from the convexity of the function. \square

By Lemma 4.7.6 and Proposition 4.7.7 we have the following corollary.

Corollary 4.7.8 *Let $x = (x_1, x_2, \dots, x_N) \in M(\bar{S})$ where $\bar{S} = (S_1, S_2, \dots, S_N)$ with $\sum_{i=1}^N \frac{1}{S_i} > 1$. Then for every $j \in \{1, 2, \dots, N\}$ for which S_j is irrational we have that $x_j < \frac{1}{S_j}$.*

Proof. Since $\sum_{i=1}^N \frac{1}{S_i} \geq 1$ we have that $R(\bar{S}) < \infty$ and thus $x_i \leq \frac{1}{S_i}$ for $i = 1, 2, \dots, N$. Suppose that S_j is irrational and $x_j = \frac{1}{S_j}$. Then by Lemma 4.7.6 we have that $l_{S_j}(x_j) = \infty$. Since $\sum_{i=1}^N \frac{1}{S_i} > 1$ and $\sum_{i=1}^N x_i = 1$, there exists some $k \neq j$ for which $x_k < \frac{1}{S_k}$. Then $r_{S_k}(x_k) < \infty = l_{S_j}(x_j)$, which according to Proposition 4.7.7 contradicts the fact that $x = (x_1, x_2, \dots, x_N) \in M(\bar{S})$. \square

Theorem 4.7.9 *Let $\bar{S} = (S_1, S_2, \dots, S_N)$ be such that $\sum_{i=1}^N \frac{1}{S_i} > 1$. Then there exists some $x = (x_1, x_2, \dots, x_N) \in M(\bar{S})$ and some $j \in \{1, 2, \dots, N\}$ such that for every $i \neq j$ it holds that x_i is either a best lower approximation of $\frac{1}{S_i}$ or $x_i = 0$.*

Proof. Suppose that $x = (x_1, x_2, \dots, x_N) \in M(\bar{S})$ and there exist $k, l \in \{1, 2, \dots, N\}$, $k < l$, such that $x_k, x_l \neq 0$ and neither x_k is a best lower approximation of $\frac{1}{S_k}$ nor x_l is a best lower approximation of $\frac{1}{S_l}$. Put $r_0 = 0$ and let $\frac{1}{\lceil S_k \rceil} = r_1 < r_2 < \dots$ be the best lower approximations of $\frac{1}{S_k}$. Similarly, put $s_0 = 0$ and let $\frac{1}{\lceil S_l \rceil} = s_1 < s_2 < \dots$ be the best lower approximations of $\frac{1}{S_l}$. By Corollary 4.7.8 we have that $0 < x_k < \frac{1}{S_k}$ and $0 < x_l < \frac{1}{S_l}$. Hence there exist $n, m \in \mathbb{Z}_{\geq 0}$ such that $r_n < x_k < r_{n+1}$ and $s_m < x_l < s_{m+1}$. By Theorem 4.5.9 we have that the functions $L_{S_k}(d)$ and $L_{S_l}(d)$ are linear in d on the intervals $[r_n, r_{n+1}]$ and $[s_m, s_{m+1}]$ respectively. Let $x(\varepsilon) = (x_1, x_2, \dots, x_{k-1}, x_k + \varepsilon, x_{k+1}, \dots, x_{l-1}, x_l + \varepsilon, x_{l+1}, \dots, x_N)$ and put $A = \min(r_{n+1} - x_k, x_l - s_m)$ and $B = \max(r_n - x_k, x_l - x_{m+1})$. Then there exists some $C \in \mathbb{R}$ such that

$$L_{\bar{S}}(x(\varepsilon)) = R(\bar{S}) + C \cdot \varepsilon$$

for every $\varepsilon \in [B, A]$. Since $B < 0$, $A > 0$ and $L_{\bar{S}}(x(\varepsilon)) \geq R(\bar{S})$ for every $\varepsilon \in \mathbb{R}$ it follows that $C = 0$. Hence $x(\varepsilon) \in M(\bar{S})$ for every $\varepsilon \in [B, A]$. In particular we have that $x(A) \in M(\bar{S})$ and thus either $(x_1, x_2, \dots, x_{k-1}, r_{n+1}, x_{k+1}, \dots, x_{l-1}, x_l - A, x_{l+1}, \dots, x_N) \in M(\bar{S})$ or $(x_1, x_2, \dots, x_{k-1}, x_k + A, x_{k+1}, \dots, x_{l-1}, s_m, x_{l+1}, \dots, x_N) \in M(\bar{S})$. The theorem follows by induction. \square

Corollary 4.7.10 *Let $\bar{S} = (S_1, S_2, \dots, S_N)$ be such that $\sum_{i=1}^N \frac{1}{S_i} > 1$. Then the set $M(\bar{S})$ contains a point with rational coordinates.*

Remark. In fact, if $R(\bar{S}) > 0$ then in most cases the set $M(\bar{S})$ consists of only one point, which according to Corollary 4.7.10 and Theorem 4.7.9 has rational coordinates and all coordinates except at most one are best lower approximations of the inverse of the service time of the corresponding server.

By Proposition 4.7.3, Proposition 4.7.4 and Corollary 4.7.10 we have the following theorem.

Theorem 4.7.11 *For every $\bar{S} = (S_1, S_2, \dots, S_N)$ system with $\sum_{i=1}^N \frac{1}{S_i} > 1$ there exists a periodic policy U such that $W(U) \leq R(\bar{S}) + \frac{N}{2} - 1$.*

Proof. By Corollary 4.7.10 there exists a point $(d_1, d_2, \dots, d_N) \in M(\bar{S})$ with $d_i \in \mathbb{Q}_{\geq 0}$ for $i = 1, 2, \dots, N$. By Proposition 4.7.4 there exists a periodic policy U such that for $i = 1, 2, \dots, N$ the routing sequence u^i has density d_i for $i = 1, 2, \dots, N$ and $\bar{O}(U) \leq \frac{N}{2} - 1$. Then by Proposition 4.7.3 we have that

$$W(U) \leq L_{\bar{S}}(d_1, d_2, \dots, d_N) + \bar{O}(U) \leq R(\bar{S}) + \frac{N}{2} - 1.$$

□

From the sequel it becomes clear how such a policy U can be obtained. By Proposition 4.7.2 and Theorem 4.7.11 we obtain that for $N = 2$ there exist optimal periodic policies if $\frac{1}{S_1} + \frac{1}{S_2} > 1$ (see [23]).

Corollary 4.7.12 *For an $\bar{S} = (S_1, S_2)$ system with $\frac{1}{S_1} + \frac{1}{S_2} > 1$ there exists an optimal periodic policy U with $W(U) = R(\bar{S})$.*

Such an optimal periodic policy can be obtained as follows. Find a rational point $(d_1, d_2) \in M(\bar{S})$. Let u^1 be the upper bracket sequence of density d_1 (or any other regular sequence of zeros and ones of density d_1). Then we construct the routing policy U by taking u^1 and replacing every zero by a two. Since d_1 is rational we have that u^1 and thus also U is periodic. Moreover, u^1 is a regular sequence of density d_1 and u^2 is a regular sequence of density $1 - d_1 = d_2$, since u^2 is the complement (every zero is a one and every one is a zero) of u^1 . Hence $\bar{O}(U) = 0$ and $W(U) = L_{\bar{S}}(d_1, d_2) = R(\bar{S})$.

In general if we have a routing policy U such that u^i is a regular sequence of density d_i for $i = 1, 2, \dots, N$ and $(d_1, d_2, \dots, d_N) \in M(\bar{S})$, then $W(U) = L_{\bar{S}}(d_1, d_2, \dots, d_N) = R(\bar{S})$ and it follows that this routing policy U is optimal. However, if $N > 2$ then generally there does not exist a routing policy U such that u^i is a regular

sequence of density d_i for $i = 1, 2, \dots, N$ (see [4]). So, in that case it is possible that $W(U) > R(\bar{S})$ for every policy U .

If $\sum_{i=1}^N \frac{1}{S_i} = 1$ then $M(\bar{S})$ consists of the single point $(\frac{1}{S_1}, \frac{1}{S_2}, \dots, \frac{1}{S_N})$, which in general is not a rational point. However, from Chapter 2 we have the following result.

Proposition 4.7.13 *For every $\bar{S} = (S_1, S_2, \dots, S_N)$ system with $\sum_{i=1}^N \frac{1}{S_i} = 1$ there exists a policy U such that $W(U) \leq R(\bar{S}) + \frac{N}{2} - 1$.*

We have the following proposition.

Proposition 4.7.14 *For every $\bar{S} = (S_1, S_2, \dots, S_N)$ system with $\sum_{i=1}^N \frac{1}{S_i} \geq 1$ we have that $R(\bar{S}) \leq \frac{1}{2}$.*

Proof. Let $x = (x_1, x_2, \dots, x_N) \in M(\bar{S})$. Since $\sum_{i=1}^N \frac{1}{S_i} \geq 1$ we have that $0 \leq x_i \leq \frac{1}{S_i}$ for $i = 1, 2, \dots, N$. So, by Lemma 4.5.4 and Theorem 4.5.5 we have that $W_{S_i}(x_i) \leq \frac{1}{2}$ for $i = 1, 2, \dots, N$. Hence

$$R(\bar{S}) = \sum_{i=1}^N L_{S_i}(x_i) = \sum_{i=1}^N x_i \cdot W_{S_i}(x_i) \leq \sum_{i=1}^N x_i \cdot \frac{1}{2} = \frac{1}{2}.$$

□

By Theorem 4.7.11, Proposition 4.7.13 and Proposition 4.7.14 we have the following theorem.

Theorem 4.7.15 *For every $\bar{S} = (S_1, S_2, \dots, S_N)$ system with $\sum_{i=1}^N \frac{1}{S_i} \geq 1$ there exists a policy U such that $W(U) \leq \frac{N-1}{2}$. Moreover, if $\sum_{i=1}^N \frac{1}{S_i} > 1$ then there exists a periodic policy U such that $W(U) \leq \frac{N-1}{2}$.*

4.7.2 Algorithms for determining a minimal point

In this subsection we give an algorithm for determining for a given $\bar{S} = (S_1, S_2, \dots, S_N)$ system with $\sum_{i=1}^N \frac{1}{S_i} > 1$ a rational point $(x_1, x_2, \dots, x_N) \in M(\bar{S})$ and the corresponding value of $R(\bar{S})$. Recall that in most cases the set $M(\bar{S})$ consists of only one point and then the rational point is obviously unique.

Put

$$F(\bar{S}) := \{(x_1, x_2, \dots, x_N) \in H^N : L_{\bar{S}}(x_1, x_2, \dots, x_N) < \infty\}$$

Then $M(\bar{S}) \subseteq F(\bar{S})$. Suppose that $x = (x_1, x_2, \dots, x_N) \in F(\bar{S}) \setminus M(\bar{S})$. Then according to the characterization of $M(\bar{S})$ given by Theorem 4.7.7 we have that there exist $i, j \in \{1, 2, \dots, N\}$, $i \neq j$ such that $l_{S_i}(x_i) > r_{S_j}(x_j)$ and $r_{S_j}(x_j) < \infty$. Hence there exists some $\varepsilon > 0$ such that $L_{\bar{S}}(x_1, x_2, \dots, x_{i-1}, x_i - \varepsilon, x_{i+1}, \dots, x_{j-1}, x_j + \varepsilon, x_{j+1}, \dots, x_N) < L_{\bar{S}}(x_1, x_2, \dots, x_N) < \infty$. So, if we have a point $x = (x_1, x_2, \dots, x_N) \in H^N$ and we have for all $i, j \in \{1, 2, \dots, N\}$ with $i \neq j$ that $L_{\bar{S}}(x_1, x_2, \dots, x_i - h, \dots, x_j + h, \dots, x_N) \geq L_{\bar{S}}(x_1, x_2, \dots, x_N)$, where h is a small positive number, then the (Euclidean) distance between $x = (x_1, x_2, \dots, x_N)$ and the set $M(\bar{S})$ is small, since this distance is of order h . For a given $\bar{S} = (S_1, S_2, \dots, S_N)$ system with $\sum_{i=1}^N \frac{1}{S_i} > 1$ we can use this property to obtain a point (arbitrarily) close to the set $M(\bar{S})$. We now give the main steps of the algorithm for obtaining this point. Later we give more details on the implementation of every step.

Sketch of Algorithm.

step 1. Pick an initial point $x := (x_1, x_2, \dots, x_N) \in F(\bar{S})$ and choose an appropriate initial step size $h > 0$ and an “acceptable error” $\varepsilon > 0$.

step 2. Find $i, j \in \{1, 2, \dots, N\}$ with $i \neq j$ such that $L_{\bar{S}}(x_1, x_2, \dots, x_i - h, \dots, x_j + h, \dots, x_N) < L_{\bar{S}}(x_1, x_2, \dots, x_N)$ or determine that there do not exist such i, j . In the former case go to step 3, in the latter case go to step 4.

step 3. Put $x := (x_1, x_2, \dots, x_i - h, \dots, x_j + h, \dots, x_N)$ and return to step 2.

step 4. Reduce the value of the step size h by a reasonable amount (for example $h := \frac{h}{2}$) and determine whether the new value of h is smaller than ε or not. In the former case go to step 5, in the latter case return to step 2.

step 5. The algorithm terminates by giving as output the point x .

We will show that the point x obtained in this way can be used to obtain a rational point $x' \in M(\bar{S})$ if the distance between the point x and the set $M(\bar{S})$ is small enough. We first give some more details on our implementation of several steps of the algorithm above. In step 1 we suggest to put $x_i = \frac{\frac{1}{S_i}}{\sum_{j=1}^N \frac{1}{S_j}}$ for $i = 1, 2, \dots, N$.

It is easily seen that then $x \in F(\bar{S})$. Then for the initial step size h we choose $h = 2^{-(k+1)}$, where $k = \lceil \frac{\log(\max_{i=1,2,\dots,N}(\frac{1}{S_i} - x_i))}{\log(\frac{1}{2})} \rceil$. We suggest putting $\varepsilon := \frac{h}{100}$, since

then the distance between the obtained point x and the set $M(\bar{S})$ will usually be small enough to obtain a rational point in $M(\bar{S})$. However, a smaller value for ϵ should be chosen if it turns out that this distance is not small enough to find the rational point. In many other cases it is possible to use a greater value for ϵ .

For step 2 the most obvious way to find such i, j is just systematically checking for $i = 1, 2, \dots, N, j = 1, 2, \dots, N, i \neq j$ whether $L_{\bar{S}}(x_1, x_2, \dots, x_i - h, \dots, x_j + h, \dots, x_N) < L_{\bar{S}}(x_1, x_2, \dots, x_N)$ until you find such i, j or you have tried all pairs i, j with $i \neq j$ and none suffices. In this way the value of $L_{\bar{S}}(x_1, x_2, \dots, x_i - h, \dots, x_j + h, \dots, x_N)$ has to be computed $O(N^2)$ times. In practice this is fast enough, since N is usually a rather small number and $L_{\bar{S}}(x_1, x_2, \dots, x_i - h, \dots, x_j + h, \dots, x_N)$ can be computed quickly by the algorithm described in the previous section. However, it is not really necessary to check all pairs i, j since for example you could decide to go to step 4 and decrease the step size if you have determined that with high probability there do not exist such i, j . Moreover, some time can be gained if the pairs i, j are checked in a smart order. Therefore, we suggest a slightly more sophisticated method for implementing step 2 of the algorithm in case that N is not a small number. Start with computing the left derivatives $l_{S_i}(x_i)$ and the right derivatives $r_{S_i}(x_i)$ for $i = 1, 2, \dots, N$. Then calculate the difference $l_{S_i}(x_i) - r_{S_j}(x_j)$ for all the pairs $i, j \in \{1, 2, \dots, N\}$ with $i \neq j$. Check whether $L_{\bar{S}}(x_1, x_2, \dots, x_i - h, \dots, x_j + h, \dots, x_N) < L_{\bar{S}}(x_1, x_2, \dots, x_N)$ for the pairs $i, j \in \{1, 2, \dots, N\}$ for which $l_{S_i}(x_i) - r_{S_j}(x_j) > 0$ and run through these pairs in order of decreasing value of $l_{S_i}(x_i) - r_{S_j}(x_j)$. In the computation of $L_{\bar{S}}(x_1, x_2, \dots, x_i - h, \dots, x_j + h, \dots, x_N)$ the (intermediate) results for computing the left and right derivatives $l_{S_i}(x_i)$ and $r_{S_i}(x_i)$ should be used to increase the efficiency of this method.

In step 4 we suggest to reduce the step size by a factor 2.

We illustrate the algorithm to obtain a point x close to the set $M(\bar{S})$ in the following example.

Example. Let $N = 3$ and $\bar{S} = (\frac{93}{57}, \frac{106}{39}, \frac{183}{20})$. Putting $x_i = \frac{1}{\sum_{j=1}^N \frac{1}{\bar{S}_j}}$ for $i = 1, 2, 3$ we obtain the initial point $x = (\frac{368562}{655529}, \frac{221247}{655529}, \frac{65270}{655529}) = (0.562236, 0.337509, 0.100255) \in F(\bar{S})$ (rounded to 6 decimals) and corresponding value $L(x_1, x_2, x_3) = 0.128134$. The initial step size $h := 2^{-6} = 0.015625$ and the error ϵ is chosen to be $\frac{h}{100} = \frac{1}{6400} = 0.00015625$. For this step size h we do not find a pair i, j with $i \neq j$ that suffices. So, $h := 2^{-7}$ and we find that $i = 2, j = 1$ suffices. Then $x_1 := x_1 + h = 0.570048, x_2 := x_2 - h = 0.329697$, and we obtain the new point $(x_1, x_2, x_3) = (0.570048, 0.329697, 0.100255) \in F(\bar{S})$ and corresponding value $L(x_1, x_2, x_3) = 0.125732$. We do not find a new pair i, j which suffices and thus $h := 2^{-8}$. Next we find that $i = 1, j = 2$ suffices and we obtain the new

point $(x_1, x_2, x_3) = (0.566142, 0.333603, 0.100255) \in F(\bar{S})$ and corresponding value $L(x_1, x_2, x_3) = 0.119799$. Continuing the procedure in this way we obtain consecutively $(x_1, x_2, x_3) := (0.562236, 0.333603, 0.104161)$, $L(x_1, x_2, x_3) = 0.118669$, $h := 2^{-9}$, $h := 2^{-10}$, $(x_1, x_2, x_3) := (0.561259, 0.333603, 0.105138)$, $L(x_1, x_2, x_3) = 0.118386$, $h := 2^{-11}$, $(x_1, x_2, x_3) := (0.561748, 0.3331155, 0.105138)$, $L(x_1, x_2, x_3) = 0.118202$, $h := 2^{-12}$, $(x_1, x_2, x_3) := (0.561504, 0.333359, 0.103138)$, $L(x_1, x_2, x_3) = 0.117865$, $h := 2^{-13}$. Since the new step size $h := 2^{-13}$ is smaller than the chosen error $\varepsilon = \frac{1}{6400}$, the algorithm stops. The distance between the last obtained point $x = (x_1, x_2, x_3) = (0.561504, 0.333359, 0.103138) \in F(\bar{S})$ and the set $M(\bar{S})$ is of order ε and thus $L(x_1, x_2, x_3) = 0.117865$ is an approximation of $R(\bar{S})$, the lower bound for the average waiting time in the \bar{S} system.

So, as illustrated above we have an algorithm to obtain a point (arbitrarily) close to the set $M(\bar{S})$ and thus an approximation of $R(\bar{S})$. We now discuss a method to obtain a rational point in the set $M(\bar{S})$ from this and if we have found this rational point then we can calculate the exact value of $R(\bar{S})$ by the algorithm of the previous section.

Recall that by Theorem 4.7.9 there exists a point $(x_1, x_2, \dots, x_N) \in M(\bar{S})$ and some $j \in \{1, 2, \dots, N\}$ such that x_i is 0 or a best lower approximation of $\frac{1}{S_i}$ for every $i \neq j$. Hence x_i is a rational number of relatively low denominator for every $i \neq j$ and $x_j = 1 - \sum_{i \neq j} x_i$ is also a rational number. Let $(x_1^*, x_2^*, \dots, x_N^*)$ be a point close to the set $M(\bar{S})$ obtained by the algorithm demonstrated above and let x_1^i, x_2^i, \dots be the convergents of x_i^* . Then for every $i \neq j$ it is very likely that $x_i = x_{k_i}^i$, where k_i is a relatively small positive integer since the denominator of x_i is relatively small. Let ε be the error used in the algorithm to determine $(x_1^*, x_2^*, \dots, x_N^*)$. Then for every $i \in \{1, 2, \dots, N\}$ we use the continued fraction algorithm to determine the convergents x_1^i, x_2^i, \dots of x_i^* until we have that $|x_{k_i}^i - x_i^*| < \varepsilon$ for some $k_i \in \mathbb{N}$. We put $y_i := x_{k_i}^i$ for $i = 1, 2, \dots, N$. Then y_i is a rational number for $i = 1, 2, \dots, N$. Let y_j be the one with the greatest denominator among them. Put $y_j = 1 - \sum_{i \neq j} y_i$. Then $\sum_{i=1}^N y_i = 1$ and if ε was chosen small enough then (y_1, y_2, \dots, y_N) is likely to be a rational point in the set $M(\bar{S})$. In fact, if the set $M(\bar{S})$ consists of the single point (x_1, x_2, \dots, x_N) as is usually the case, then we should have that $y_i = x_i$ for $i = 1, 2, \dots, N$. In that case we have for every $i \neq j$ that y_i is either 0 or a best lower approximation of $\frac{1}{S_i}$. In general we can verify whether the obtained rational point (y_1, y_2, \dots, y_N) is an element of $M(\bar{S})$ or not by the condition of Proposition 4.7.7. If $(y_1, y_2, \dots, y_N) \in M(\bar{S})$ then the exact value of $R(\bar{S})$ can be calculated by $R(\bar{S}) = L(y_1, y_2, \dots, y_N)$. If it turns out that $(y_1, y_2, \dots, y_N) \notin M(\bar{S})$ then the algorithm could be modified.

First of all it is possible to choose a smaller value for the error ε and try again. It is also possible to exhaustively search for a rational point $(x_1, x_2, \dots, x_N) \in M(\bar{S})$ for which x_i is either 0 or a best lower approximation of $\frac{1}{S_i}$ for every $i \neq j$ by using the enumerations (see (4.18), (4.19) and (4.20)) of the best lower approximations of $\frac{1}{S_i}$ for $i = 1, 2, \dots, N$. For $N = 2$ this enumeration of possible values for the x_i and the fact that the function $L(x_1, x_2)$ is convex in x_1 and x_2 is used in [23] to find a rational point in $M(\bar{S})$. However, it is more complicated if N gets larger and then this approach would take much more time. On the other hand, if you first apply the above algorithm to obtain a point $(x_1^*, x_2^*, \dots, x_N^*)$ close to the set $M(\bar{S})$ then you approximately know where you should start the search. This should reduce the time it takes to find the rational point by an exhaustive search.

We illustrate the algorithm to find a rational point $(y_1, y_2, \dots, y_N) \in M(\bar{S})$ from a point $(x_1^*, x_2^*, \dots, x_N^*)$ close to the set $M(\bar{S})$ by applying the continued fraction algorithm.

Example. We continue the previous example in which we obtained the point $(x_1^*, x_2^*, x_3^*) = (0.561504, 0.333359, 0.103138)$ for the $\bar{S} = (\frac{93}{57}, \frac{106}{39}, \frac{183}{20})$ system by taking $\varepsilon = \frac{1}{6400}$. We apply the continued fraction algorithm to $x_1^* = 0.561504$ (rounded to 6 decimals) and we find that $x_1^* = [0, 1, 1, 3, 1, 1, 3, \dots]$ with convergents $\frac{0}{1}, \frac{1}{1}, \frac{1}{2}, \frac{4}{7}, \frac{5}{9}, \frac{9}{16}, \frac{32}{57}, \dots$. Since for the 6-th convergent $\frac{32}{57}$ we have that $|\frac{32}{57} - x_1^*| < \varepsilon$ we stop the continued fraction expansion of x_1^* and put $y_1 = \frac{32}{57}$. Similarly, we obtain $y_2 = \frac{1}{3}$ and $y_3 = \frac{2}{19}$. According to the algorithm we should put $y_1 := 1 - y_2 - y_3$, but the value of y_1 does not change by this, since we already have $y_1 + y_2 + y_3 = 1$. So, we have obtained the rational point $(y_1, y_2, y_3) = (\frac{32}{57}, \frac{1}{3}, \frac{2}{19})$ and next we verify that $(y_1, y_2, y_3) \in M(\bar{S})$ by the condition of Proposition 4.7.7. So, we calculate the left derivatives $l_{S_i}(y_i)$ and the right derivatives $r_{S_i}(y_i)$ for $i = 1, 2, 3$. We obtain $l_{S_1}(y_1) = \frac{34}{19}$, $l_{S_2}(y_2) = 0$, $l_{S_3}(y_3) = \frac{3}{2}$, $r_{S_1}(y_1) = \frac{34}{19}$, $r_{S_2}(y_2) = \frac{51}{13}$ and $r_{S_3}(y_3) = \frac{87}{20}$. Since $l_{S_i}(y_i) \leq r_{S_j}(y_j)$ for every $i \neq j$ we have indeed that $(y_1, y_2, y_3) \in M(\bar{S})$. Moreover, y_2 and y_3 are best lower approximations of $\frac{1}{S_2}$ and $\frac{1}{S_3}$ respectively, as expected. This follows directly from the fact that $l_{S_2}(y_2) < r_{S_2}(y_2)$ and $l_{S_3}(y_3) < r_{S_3}(y_3)$. We have $R(\bar{S}) = L(y_1, y_2, y_3) = \frac{2551}{21660}$. So, for the $\bar{S} = (\frac{93}{57}, \frac{106}{39}, \frac{183}{20})$ system we have that $W(U) \geq \frac{2551}{21660}$ for every routing policy U . Moreover, it follows from Proposition 4.7.4 and Theorem 4.7.11 that it is easy to obtain (by constructing a billiard sequence) a periodic policy U' with densities $(d_1, d_2, d_3) = (y_1, y_2, y_3) = (\frac{32}{57}, \frac{1}{3}, \frac{2}{19})$ (and thus a period of 57) for which $W(U') \leq R(\bar{S}) + \frac{N}{2} - 1 = \frac{13381}{21660}$.

4.8 Appendix: An extension of Little's Law for routing policies

Stidham proved in [65] the Little relation under the assumption that W , the limiting average of the waiting times, and L , the limiting average of the number of customers, do exist. In our optimal routing problem we do not want to restrict the set of policies to those for which these limits exist. Even if the limiting density does exist then the limiting averages L and W may fail to exist (see Example 4.8.3). Here we show that Little's relation can be replaced with the inequalities of Theorem 4.8.2 if L and W are defined in terms of limsups as in Section 4.3.

The results in this appendix are valid for any input-output queueing system and thus we have a more general setting than before. However, we recall that the objective is to show the validity of (4.3), (4.4) and (4.6) for the particular deterministic queueing system considered in this chapter. For convenience we use notation similar to that used before.

We suppose that customers arrive at (or are admitted by some routing policy to) the queueing system at moments $t_n \in \mathbb{R}$, $n = 1, 2, \dots$ where $0 \leq t_1 \leq t_2 \leq \dots$. Let W_n be the waiting time in the system of the n -th arriving customer and let $B(t)$ be the number of customers waiting in the system at time t . Then we define $L := \limsup_{T \rightarrow \infty} \frac{1}{T} \cdot \int_{t=0}^T B(t) d(t)$ as the time average number of customers in the system and $W := \limsup_{n \rightarrow \infty} \frac{1}{T} \cdot \sum_{n=1}^T W_n$ as the average waiting time of the arriving customers. Moreover, we put $A(t) := |n \in \mathbb{N} : t_n \leq t|$, the number of arrivals before time t and $D(t) := |n \in \mathbb{N} : t_n + W_n \leq t|$, the number of departures before time t . From the definition it is clear that

$$A(t) = B(t) + D(t) \text{ for every } t \in \mathbb{Z}_{\geq 0}.$$

We put $\bar{\lambda} := \limsup_{t \rightarrow \infty} \frac{A(t)}{t}$ as the upper arrival rate and $\underline{\lambda} := \liminf_{t \rightarrow \infty} \frac{A(t)}{t}$ as the lower arrival rate. We assume that $\bar{\lambda} < \infty$. If $\underline{\lambda} = \bar{\lambda}$ then $\lambda := \lim_{t \rightarrow \infty} \frac{A(t)}{t}$ is the arrival rate.

Lemma 4.8.1 *If $W < \infty$ then*

$$\liminf_{t \rightarrow \infty} \frac{D(t)}{t} = \liminf_{t \rightarrow \infty} \frac{A(t)}{t} \text{ and } \limsup_{t \rightarrow \infty} \frac{D(t)}{t} = \limsup_{t \rightarrow \infty} \frac{A(t)}{t}. \quad (4.25)$$

Moreover,

$$\limsup_{t \rightarrow \infty} \frac{1}{A(t)} \sum_{n=1}^{A(t)} W_n = \limsup_{t \rightarrow \infty} \frac{1}{A(t)} \sum_{n: t_n + W_n \leq t} W_n. \quad (4.26)$$

Proof. Since $W < \infty$ we have $\lim_{n \rightarrow \infty} \frac{W_n}{n} = 0$. Thus $\lim_{n \rightarrow \infty} \frac{W_n}{t_n} = 0$, since $\bar{\lambda} < \infty$. Hence, for every $\varepsilon > 0$ there exists an $n_0 \in \mathbb{N}$ such that $W_n < t_n \varepsilon$ for every $n > n_0$. Let $t_0 \in R_{\geq 0}$ be chosen so large that $A(t_0) > n_0$. Then

$$\begin{aligned} A(t_0) &\geq D(t_0) = |n : t_n + W_n \leq t_0| \geq |n \leq n_0 : t_n + W_n \leq t_0| + \\ &|n > n_0 : t_n(1 + \varepsilon) \leq t_0| = |n \leq n_0 : t_n + W_n \leq t_0| - \\ &|n \leq n_0 : t_n \cdot (1 + \varepsilon) \leq t_0| + |n : t_n \leq t_0/(1 + \varepsilon)|. \end{aligned} \quad (4.27)$$

Dividing (4.27) by t_0 and taking liminfs for $t_0 \rightarrow \infty$ we obtain

$$\liminf_{t_0 \rightarrow \infty} \frac{A(t_0)}{t_0} \geq \liminf_{t_0 \rightarrow \infty} \frac{D(t_0)}{t_0} \geq \liminf_{t_0 \rightarrow \infty} \frac{A(t_0)/(1 + \varepsilon)}{t_0}.$$

Since ε was arbitrary, it follows that $\liminf_{t \rightarrow \infty} \frac{D(t)}{t} = \liminf_{t \rightarrow \infty} \frac{A(t)}{t}$. Analogously it follows that $\limsup_{t \rightarrow \infty} \frac{D(t)}{t} = \limsup_{t \rightarrow \infty} \frac{A(t)}{t}$. Moreover, since $A(t) \rightarrow \infty$ as $t \rightarrow \infty$, it also follows from (4.27) that

$$\limsup_{t \rightarrow \infty} \frac{1}{A(t)} \sum_{n=1}^{A(t)} W_n \geq \limsup_{t \rightarrow \infty} \frac{1}{A(t)} \sum_{n: t_n + W_n \leq t} W_n \geq \frac{1}{1 + \varepsilon} \limsup_{t \rightarrow \infty} \frac{1}{A(t)} \sum_{n=1}^{A(t)} W_n.$$

This implies (4.26) since ε was arbitrary. \square

Remark. The proof of this lemma is similar to the proof of Little's Law by Stidham in [65]. The main difference is that the existence of certain limits is not assumed, but that instead of that liminfs and limsups are used.

By Lemma 4.8.1 we can prove that the following extension of Little's Law holds.

Theorem 4.8.2 *Suppose that $W < \infty$ for some input-output queueing system as described above. Then, for the upper arrival rate $\bar{\lambda}$ we have that*

$$L \leq \bar{\lambda} \cdot W. \quad (4.28)$$

Moreover, for the lower arrival rate $\underline{\lambda}$ we have that

$$L \geq \underline{\lambda} \cdot W \quad (4.29)$$

and if the arrival rate λ exists then

$$L = \lambda \cdot W \quad (4.30)$$

Proof. Using a standard argument in the derivation of Little's law, more precisely by considering the total number of waiting time units up to time t (see [58] page 102 or [65]), we find the inequalities

$$\sum_{n \in \mathbb{N}: t_n + W_n \leq t} W_n \leq \int_{s=0}^t B(s) ds \leq \sum_{n=1}^{A(t)} W_n. \quad (4.31)$$

By the second inequality we have

$$\begin{aligned} \bar{\lambda} \cdot W &= \limsup_{t \rightarrow \infty} \frac{A(t)}{t} \cdot \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n W_k = \limsup_{t \rightarrow \infty} \frac{A(t)}{t} \cdot \limsup_{t \rightarrow \infty} \frac{1}{A(t)} \sum_{k=1}^{A(t)} W_k \geq \\ &\limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{k=1}^{A(t)} W_k \geq \limsup_{t \rightarrow \infty} \frac{1}{t} \int_{s=0}^t B(s) ds = L. \end{aligned}$$

Similarly, (4.26) and the first inequality imply

$$\begin{aligned} \lambda \cdot W &= \liminf_{t \rightarrow \infty} \frac{A(t)}{t} \cdot \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n W_k = \liminf_{t \rightarrow \infty} \frac{A(t)}{t} \cdot \limsup_{t \rightarrow \infty} \frac{1}{A(t)} \sum_{k=1}^{A(t)} W_k = \\ &\liminf_{t \rightarrow \infty} \frac{A(t)}{t} \cdot \limsup_{t \rightarrow \infty} \frac{1}{A(t)} \sum_{n: t_n + W_n \leq t} W_n \leq \limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{n: t_n + W_n \leq t} W_n \leq \\ &\limsup_{t \rightarrow \infty} \frac{1}{t} \int_{s=0}^t B(s) ds = L. \end{aligned}$$

By combining the first and second assertion we have $L = \lambda \cdot W$ if λ exists. \square

We now apply Theorem 4.8.2 for routing to a deterministic queueing system as described in Section 4.3. If we consider the (S_1, S_2, \dots, S_N) system in total, then λ is equal to 1 and independent of the routing policy U . Hence (4.4) follows immediately from (4.30). Moreover, if we consider a single queue i of the system then the arrival rate λ is equal to the density d_i (if it exists) of the routing sequence u^i corresponding to this queue. Thus (4.30) also implies (4.6).

In Section 4.3 we defined the average waiting time W and the average buffer content L by limsups instead of limits. The following example shows that this is necessary, since in general the limits do not exist (even not under the assumption that the routing sequence has a density and thus the arrival rate exists).

Example 4.8.3 Consider a single deterministic queue and assume that the density d of the routing sequence exists and is rational, $d = \frac{p}{q}$ with $\gcd(p, q) = 1$ say. Suppose that the service time S is such that

$$\frac{1}{\lceil S \rceil} < d < \frac{1}{S}.$$

Let w be the upperbracket sequence with period word $\omega = (\omega_1, \dots, \omega_q)$ and with average workload $L(d)$. We construct a policy $u = (u_1, u_2, \dots)$ which consists of blocks of type ω^k alternated with blocks of type ϖ^l where $\varpi = (\varpi_1, \dots, \varpi_q)$, with $\varpi_k = 1$ ($1 \leq k \leq p$) and $\varpi_k = 0$ ($p+1 \leq k \leq q$).

Choose (p, q) and S such that

$$L(\varpi^\infty) > L(d).$$

Choose a sequence $\{\varepsilon_k\}_{k=1}^\infty$ of positive numbers with $\lim_{k \rightarrow \infty} \varepsilon_k = 0$ and define inductively unbounded sequences t_k, t'_k . Let $t_k = t'_{k-1} + m_k q$ and define on the interval $t_k < t < t'_k$ the u_t as ω^{m_k} . Take m_k so large that

$$\left| \frac{1}{t_k} \int_{s=0}^{t_k} B(s) ds - L(d) \right| < \varepsilon_k.$$

Let $t'_k = t_k + n_k q$ and define on the interval $t_k < t < t'_k$ the u_t as ϖ^{n_k} and take n_k so large that

$$\left| \frac{1}{t'_k} \int_{s=0}^{t'_k} B(s) ds - L(\varpi^\infty) \right| < \varepsilon_k.$$

Then, since the fraction of ones in each ω and ϖ is equal to d , we have

$$\lim_{t \rightarrow \infty} \frac{\kappa_u(t)}{t} = d.$$

Moreover, $\liminf_{t \rightarrow \infty} \frac{1}{t} \int_{s=0}^t B(s) ds = L(d) < L(\varpi^\infty) = \limsup_{t \rightarrow \infty} \frac{1}{t} \int_{s=0}^t B(s) ds$.

The conclusion is that the existence of the limiting density does not imply the existence of the limiting average waiting time nor the existence of the limiting average buffer content.

Chapter 5

On the optimality of a stationary policy for deterministic parallel queueing systems

5.1 Introduction

We consider a deterministic parallel queueing system as in the previous chapter and we use the same notations as much as possible. So, if N is the number of queues then the system is denoted by (S_1, S_2, \dots, S_N) where S_i is the service time for queue i for $i = 1, 2, \dots, N$. Customers arrive with a constant interarrival time of 1 at the integer moments $t = 0, 1, 2, \dots$. Arriving customers are routed to one of the parallel queues at the moment of their arrival, with as objective to minimize the long-run average waiting time of the arriving customers. So, we look for a routing policy U such that $W(U) = \limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{n=1}^t W_n$ is minimal, where W_n is the waiting time of the n -th arriving customer.

In this chapter we will describe this problem as a Markov Decision Chain (MDC) with average cost optimization (see for example [56],[59] and [61] for theory on MDC's). The decision epochs are at the moments that the customers arrive, i.e at the integer points $t = 0, 1, \dots$. We describe the state space, the action space, the

transition rules and the one-step costs to obtain a Markov decision chain. Next we prove that there exists a stationary optimal policy for this MDC if we assume that $\sum_{i=1}^N \frac{1}{S_i} > 1$. As a corollary of this result we prove the existence of a periodic optimal policy if all the service times S_i are rational numbers.

5.2 The optimality of a stationary policy

5.2.1 The description of the Markov Decision Chain

The state space. We first define a state space X for the MDC. Recall from Section 4.3 that $V_i(t)$ denotes the remaining workload measured in time units for server i at time t and $\underline{V}(t) := (V_1(t), V_2(t), \dots, V_N(t))$ is the vector of remaining workloads in the system at time t . We use this vector to describe the state of the process at the decision epochs $t = 0, 1, 2, \dots$. We assume as in the previous chapter that the initial state at $t = 0$ is the zero state $(0, 0, \dots, 0)$. Then we have for any $t \in \mathbb{Z}_{\geq 0}$ and $i \in \{1, 2, \dots, N\}$ that $V_i(t)$ is of the form $\max(k_i \cdot S_i - l_i, 0)$ where $k_i, l_i \in \mathbb{Z}_{\geq 0}$. Therefore we define for the (S_1, S_2, \dots, S_N) system the state space X as the set

$$X := \{(x_1, x_2, \dots, x_N) \in \mathbb{R}^N : x_i = \max(k_i \cdot S_i - l_i, 0), k_i, l_i \in \mathbb{Z}_{\geq 0} \text{ for } i = 1, 2, \dots, N\}.$$

Note that the state space X is an infinite countable set.

The action space. For every state $s \in X$ we define the action space $A = A(s)$ to be the finite set $\{1, 2, \dots, N\}$. Indeed, if at some decision epoch the system is in state $s \in X$ and the policy chooses action $a \in A$ then the arriving customer is routed to server a .

The transition rules. To describe the transition rules of the MDC we give for every pair $s, t \in X$ the probability $P_{st}(a)$ that, if the system is in state s and action $a \in A$ is chosen, the system is in state t at the next decision epoch. However, since we have a deterministic queueing system there exists for every $s \in X$ and $a \in A$ exactly one state $k \in X$ such that $P_{sk}(a) = 1$ and $P_{st}(a) = 0$ for every $t \neq k$. For the (S_1, S_2, \dots, S_N) system let $s \in X$ be the vector (x_1, x_2, \dots, x_N) and $a \in A = \{1, 2, \dots, N\}$. Then the state $k \in X$ for which $P_{sk}(a) = 1$ is given by the vector (y_1, y_2, \dots, y_N) where $y_a = \max(x_a + S_a - 1, 0)$ and $y_n = \max(x_n - 1, 0)$ for every $n \neq a$.

The one-step costs. Given a state $s = (x_1, x_2, \dots, x_N) \in X$ and an action $a \in A$

the one-step cost $C(s, a)$ is the waiting time for the arriving customer, which is routed to server a . Thus $C(s, a) = x_a$, the remaining workload for server a at the decision epoch.

5.2.2 Definitions and theory on the MDC

We use a notation similar to the one used in [61]. Assume that the initial state is s and that the process is operated under an arbitrary policy θ , i.e a sequence of decision rules which prescribes the way an action is chosen in every situation (that is the current state and the total history of the process) at every decision epoch t , $t = 0, 1, \dots$. More precisely, the decision rule at time epoch t may depend on all information of the process until t , i.e on the states at time epochs $0, 1, \dots, t$ and the actions chosen at time epochs $0, 1, \dots, t - 1$. The state at decision epoch t is denoted by the random variable X_t and the action chosen at t is denoted by the random variable A_t . Moreover, the joint probability distribution (X_t, A_t) is given by $P_\theta(X_t = j, A_t = A | X_0 = s)$. Associated to the random pair (X_t, A_t) is the cost $C(X_t, A_t)$, which is the cost incurred at decision epoch t under policy θ . This is also a random variable and the expected average cost at time t is given by

$$E_\theta[C(X_t, A_t) | X_0 = s] = \sum_{j \in X} \sum_{a \in A} C(j, a) P_\theta(X_t = j, A_t = a | X_0 = s).$$

Suppose that a so-called discount factor α , $0 < \alpha < 1$, an initial state s , and policy θ are given. Then the expected (total) discounted cost under θ is denoted by $V_{\theta, \alpha}$ and defined as

$$V_{\theta, \alpha}(s) = E_\theta\left[\sum_{t=0}^{\infty} \alpha^t C(X_t, A_t) | X_0 = s\right] = \sum_{t=0}^{\infty} \alpha^t E_\theta[C(X_t, A_t) | X_0 = s]. \quad (5.1)$$

The expected discounted value function is defined for every $s \in X$ as

$$V_\alpha(s) = \inf_{\theta} V_{\theta, \alpha}(s), \quad (5.2)$$

where the infimum is taken over all possible policies.

Remark. The notation $V_\alpha(s)$ for the expected discounted value function should not be confused with the notation $V_i(t)$, which we use for the remaining workload in server i at moment t . We stress out that in case of a subscript α always the expected discounted value function is meant.

Besides the expected discounted cost under a policy θ we also define the long-run expected average cost under policy θ . This corresponds to the long-run average waiting time of arriving customers in the (S_1, S_2, \dots, S_N) system, which we want to minimize.

Suppose that an initial state s and policy θ are given. Then the long-run expected average cost under policy θ is denoted by $J_\theta(s)$ and defined by

$$J_\theta(s) = \limsup_{n \rightarrow \infty} \frac{1}{n} E_\theta \left[\sum_{t=0}^{n-1} C(X_t, A_t) \mid X_0 = s \right]. \quad (5.3)$$

Note that for some policies the long-run expected average cost may be infinite.

The long-run expected average cost function is defined for every $s \in X$ as

$$J(s) = \inf_{\theta} J_\theta(s), \quad (5.4)$$

where the infimum is taken over all possible policies.

Definition 5.2.1 *A policy θ is optimal for the (expected) discounted cost criterion with discount factor α , if $V_{\theta, \alpha}(s) = V_\alpha(s)$ for every $s \in X$. A policy θ is optimal for the average cost criterion if $J_\theta(s) = J(s)$ for every $s \in X$.*

If θ is a deterministic policy (see page 21 of [61] for the definition) then the random variables X_t, A_t for $t = 0, 1, \dots$ are in fact deterministic if the initial state X_0 is given. This is easily seen by considering the fact that the transition rules are deterministic for the MDC corresponding to an (S_1, S_2, \dots, S_N) system. Recall that for an (S_1, S_2, \dots, S_N) system it is assumed that X_0 is the zero state $s = (0, 0, \dots, 0)$. Hence to such a policy θ we can associate the infinite sequence (A_0, A_1, A_2, \dots) of integers, which is a routing sequence (policy) for the (S_1, S_2, \dots, S_N) system as we considered in the previous chapter. We also note that in this case we have that $E_\theta[C(X_t, A_t) \mid X_0 = s] = C(X_t, A_t)$, which simplifies the right hand sides of (5.1) and (5.3).

In this section we prove that for an MDC corresponding to an (S_1, S_2, \dots, S_N) system with $\sum_{i=1}^N \frac{1}{S_i} > 1$ there exists an optimal deterministic stationary policy. We show in Section 5.3 that the routing sequence associated with such an optimal deterministic stationary policy is ultimately periodic if we assume that the service times S_i are rational. We show in that case that there exists an optimal periodic routing sequence if the initial state s is the zero state $(0, 0, \dots, 0)$.

5.2.3 Sufficient conditions for the existence of an optimal stationary policy

In this subsection we describe the properties of a (deterministic) stationary policy and we give sufficient conditions (assumptions) for the existence of such a policy for the average cost criterion. In the next subsection we show that this assumptions are satisfied for the MDC corresponding to an (S_1, S_2, \dots, S_N) system.

For the definition of a (deterministic) stationary policy see for example [59] or [61]. In the sequel we just say stationary policy instead of deterministic stationary policy. To summarize a stationary policy is deterministic, memoryless (Markov) and the decision rules are independent of the time epoch t . So, a stationary policy is (described by) a function $f : X \rightarrow A$, which implies that if the system is in state $s \in X$ at any decision epoch t then always action $f(s) \in A$ is chosen. Thus a stationary policy depends only on the current state of the process and not on any other history of the process.

It is well known (see for example Theorem 4.1.4 in [61]) that for a MDC with a countable state space, one-step costs which are bounded from below and a finite action set (as is here the case) that for any discount factor $0 < \alpha < 1$ there always exist an optimal stationary policy for the discounted cost criterion. However, for the average cost criterion this is not always the case and some extra assumptions are needed for the existence of a stationary optimal policy. Sufficient assumptions are given in Section 7.2 of [61]. These assumptions are as follows.

There exists some state $z \in X$ such that

- (A1). The quantity $(1-\alpha)V_\alpha(z)$ is bounded for $\alpha \in (0, 1)$. (This implies that $V_\alpha(z) < \infty$ and hence we may define the function $h_\alpha(s) =: V_\alpha(s) - V_\alpha(z)$, $s \in X$.)
- (A2). There exists a function $M : X \rightarrow \mathbb{R}_{\geq 0}$ such that $h_\alpha(s) \leq M(s)$ for $s \in X$ and $\alpha \in (0, 1)$.
- (A3). There exists a nonnegative real number K such that $-K \leq h_\alpha(s)$ for $s \in X$ and $\alpha \in (0, 1)$.

We call z the distinguished state. Note that $h_\alpha(z) = 0$. Moreover, the first assumption is related to the finiteness of the minimum average cost. According to Section 7.2 and particularly Theorem 7.2.3 in [61] we have the following result (see also Theorem 8.10.7 in [61]).

Theorem 5.2.2 *Suppose the above assumptions hold for some MDC with a countable state space X and a finite action set $A(s)$ for every $s \in X$. Then $J(s)$ is finite and independent of s for every $s \in X$ and thus the minimum average cost is a finite constant J , independent of the initial state of the process. Moreover, $J = \lim_{\alpha \uparrow 1} (1 - \alpha) \cdot V_\alpha(s)$ for $s \in X$ and there exists an optimal stationary policy f with long-run average cost $J_f = J$.*

5.2.4 Verification of the assumptions

In this subsection we show that the assumptions (A1) - (A3) are satisfied for an MDC corresponding to an (S_1, S_2, \dots, S_N) system if $\sum_{i=1}^N \frac{1}{S_i} > 1$. The existence of an optimal stationary policy then follows from Theorem 5.2.2.

To verify the assumptions we choose the zero state $(0, 0, \dots, 0)$ as distinguished state z . We first show that the first assumption (A1) holds.

Verification of (A1) Let ψ be the policy for which the actions at the decision epochs are chosen according to the following rule. Let $X_t = (x_1^t, x_2^t, \dots, x_N^t)$ be the state at decision epoch t . Then A_t , the action chosen at time t , is defined as the minimal $i \in A$ for which $x_i^t = \min_{j \in A} x_j^t$. So, under policy ψ an arriving customer is always routed to a server with minimal remaining workload and in case of a tie the one of lowest index is chosen. So, under policy ψ the customers are routed according to a shortest waiting time algorithm. Note that ψ is a stationary policy. Moreover, we have for every time epoch t that $C(X_t, A_t) = \min_{a \in A} C(X_t, a)$ and thus ψ is a so-called myopic policy.

Lemma 5.2.3 *If the MDC is operated under the stationary myopic policy ψ and the initial state X_0 equals $z = (0, 0, \dots, 0)$, then for every $t \in \mathbb{Z}_{\geq 0}$ we have that*

$$\sum_{i=1}^N \frac{x_i^t}{S_i} < N - 1. \quad (5.5)$$

Proof. Since $N > 1$, it is clear that (5.5) holds for $t = 0$. Suppose the assertion is not true. Let $T \in \mathbb{N}$ be the first decision epoch for which $\sum_{i=1}^N \frac{x_i^T}{S_i} \geq N - 1$. Put $k := A_{T-1}$, the last action before T . Suppose that $x_k^T \geq S_k$. Since $x_k^T = \max(x_k^{T-1} + S_k - 1, 0)$, it follows that $x_k^{T-1} \geq 1$ and thus $x_i^{T-1} \geq 1$ for every $i \in A$. Hence $x_j^T = x_j^{T-1} - 1$ for every $j \neq k$ and $x_k^T = x_k^{T-1} + S_k - 1$. Thus

$$\sum_{i=1}^N \frac{x_i^T}{S_i} = \sum_{i=1}^N \frac{x_i^{T-1}}{S_i} - \sum_{i=1}^N \frac{1}{S_i} + 1 < \sum_{i=1}^N \frac{x_i^{T-1}}{S_i},$$

since $\sum_{i=1}^N \frac{1}{S_i} > 1$. This contradicts the fact that $\sum_{i=1}^N \frac{x_i^{T-1}}{S_i} < N - 1 \leq \sum_{i=1}^N \frac{x_i^T}{S_i}$. So, we have that $x_k^T < S_k$ and thus $x_k^{T-1} < 1$. Suppose there exists some $l \in A$, $l \neq k$ for which $x_l^{T-1} \leq 1$. Then $x_l^T = 0$, $x_k^T < S_k$. Hence $\sum_{j \in A, j \neq k, l} \frac{x_j^T}{S_j} > N - 2$ and thus there exists some $m \in A$, $m \neq k, l$ for which $x_m^T > S_m > 0$. Therefore there exists some decision epoch $t < T$ for which $A_t = m$. Let t^* be the maximal $t < T$ such that $A_{t^*} = m$. It follows by induction that $x_m^{t^*+1} = x_m^T + (T - t^* - 1) > S_m + (T - t^* - 1)$ and thus $x_m^{t^*} > T - t^*$. Moreover, since $x_l^T = 0$ it also follows by induction that $x_l^{t^*} \leq T - t^*$. Hence $x_l^{t^*} < x_m^{t^*}$ which contradicts the fact that $A_{t^*} = m$. So, we have for every $j \neq k$ that $x_j^{T-1} > 1$ and thus $x_j^T = x_j^{T-1} - 1$. Hence

$$\sum_{i=1}^N \frac{x_i^T}{S_i} \leq \sum_{i=1}^N \frac{x_i^{T-1}}{S_i} - \sum_{i=1}^N \frac{1}{S_i} + 1 < \sum_{i=1}^N \frac{x_i^{T-1}}{S_i},$$

which again gives a contradiction. Thus we have proved the lemma \square

Suppose that under policy ψ there exists some decision epoch t for which

$$\min_{i \in A} x_i^t \geq 1 + (N - 2) \cdot \max_{i \in A} S_i =: D.$$

Then we have that

$$\sum_{i=1}^N \frac{x_i^T}{S_i} \geq \sum_{i=1}^N \frac{D}{S_i} \geq \sum_{i=1}^N \frac{1}{S_i} + (N - 2) > N - 1,$$

which contradicts Lemma 5.2.3. Hence we have under policy ψ that $C(X_t, A_t) < D$ for $t = 0, 1, \dots$. So, by (5.1) we have for $0 < \alpha < 1$ that

$$V_{\psi, \alpha}(z) < \sum_{t=0}^{\infty} \alpha^t D = \frac{D}{1 - \alpha}.$$

Hence $V_{\alpha}(z) < \frac{D}{1 - \alpha}$. We conclude that (A1) holds for this MDC.

Lemma 5.2.4 *Let ψ_1, ψ_2 be deterministic policies for the MDC corresponding to an (S_1, S_2, \dots, S_N) system. Let $X_t := (x_1^t, x_2^t, \dots, x_N^t)$, $t = 0, 1, \dots$ be the state at decision epoch t under policy ψ_1 given initial state X_0 . Let A_t , $t = 0, 1, \dots$ be the chosen action at decision epoch t under policy ψ_1 given initial state X_0 . Let $Y_t := (y_1^t, y_2^t, \dots, y_N^t)$, $t = 0, 1, \dots$ be the state at decision epoch t under policy ψ_2 given initial state Y_0 . Let B_t , $t = 0, 1, \dots$ be the chosen action at decision epoch t under policy ψ_2 given initial state Y_0 . Suppose that for some $t_1, t_2 \in \mathbb{Z}_{\geq 0}$, $n \in \mathbb{N}$, $i \in \{1, 2, \dots, N\}$ we have that $x_i^{t_1} \leq y_i^{t_2}$ and $A_{t_1+k} = B_{t_2+k}$ for $k = 0, 1, \dots, n - 1$. Then $x_i^{t_1+k} \leq y_i^{t_2+k}$ for $k = 0, 1, 2, \dots, n$*

Proof. We prove it by induction. We have $x_i^{t_1+k} \leq y_i^{t_2+k}$ for $k = 0$. Suppose that $x_i^{t_1+k} \leq y_i^{t_2+k}$ for some $0 \leq k < n$. If $A_{t_1+k} = B_{t_2+k} \neq i$, then

$$x_i^{t_1+k+1} = \max(x_i^{t_1+k} - 1, 0) \leq \max(y_i^{t_2+k} - 1, 0) = y_i^{t_2+k+1}.$$

If $A_{t_1+k} = B_{t_2+k} = i$ then

$$x_i^{t_1+k+1} = \max(x_i^{t_1+k} + S_i - 1, 0) \leq \max(y_i^{t_2+k} + S_i - 1, 0) = y_i^{t_2+k+1}.$$

Hence $x_i^{t_1+k} \leq y_i^{t_2+k}$ implies $x_i^{t_1+k+1} \leq y_i^{t_2+k+1}$ and the lemma follows by induction. \square

Verification of (A2). Let ψ_α be a stationary optimal policy for the discounted cost criterion with discount factor α . Let $X_t := (x_1^t, x_2^t, \dots, x_N^t)$, $t = 0, 1, \dots$ be the state at decision epoch t under policy ψ_α given initial state $X_0 = z = (0, 0, \dots, 0)$ and let A_t , $t = 0, 1, \dots$ be the action at decision epoch t .

Let $s \in X$ be an arbitrary state. To show that (A2) holds it suffices to show that there exists some policy θ and a finite number $M := M(s)$ such that $V_{\theta, \alpha}(s) - V_{\psi_\alpha, \alpha}(z) \leq M$, since this implies $h_\alpha(s) \leq M$. Let $Y_t := (y_1^t, y_2^t, \dots, y_N^t)$, $t = 0, 1, \dots$ be the state at decision epoch t under some deterministic policy θ given initial state $Y_0 = s = (y_1^0, y_2^0, \dots, y_N^0)$ and let B_t be the action at time t for $t = 0, 1, \dots$.

Since $\sum_{i=1}^N \frac{1}{S_i} > 1$, there exists some $\beta > 0$ such that $\sum_{i=1}^N \frac{1}{S_i} = 1 + \beta$. Put $C = \lceil \max_{i \in A} y_i^0 \rceil$ and put $T = \lceil \frac{C \cdot N}{\beta} \rceil$. The idea is to construct a deterministic policy θ such that $Y_{T+C} = X_T$, i.e. to show that the state reached at decision epoch T under policy ψ_α given initial state X_0 can be reached from any initial state $Y_0 = s \in X$ at decision epoch $T + C$ under policy θ .

For $t = 0, 1, \dots$ and $i \in A$ let $N_i^t = |\{n < t : A_n = i\}|$, the number of times action i is chosen among the first t decision epochs under policy ψ_α and initial state $X_0 = (0, 0, \dots, 0)$. Suppose $(C + N_i^T) \cdot S_i > T$ for $i = 1, 2, \dots, N$. Then

$$T = \sum_{i=1}^N N_i^T > \sum_{i=1}^N \left(\frac{T}{S_i} - C \right) = T \cdot (1 + \beta) - N \cdot C.$$

Since this implies that $T < \frac{C \cdot N}{\beta}$, we have a contradiction. Hence there exists some $k \in \{1, 2, \dots, N\}$ for which $(C + N_k^T) \cdot S_k \leq T$. For policy θ we choose $B_t = k$ for $t = 0, 1, \dots, C - 1$. Then for Y_C , the state reached at decision epoch C , we have that $y_i^C = 0$ for every $i \neq k$ and

$$y_k^C = \max(y_k^0 + C \cdot (S_k - 1), 0) \leq \max(C + C \cdot (S_k - 1), 0) = C \cdot S_k.$$

Next we choose further actions for policy θ by putting $B_{C+t} = A_t$ for $t = 0, 1, \dots$. Then we have by Lemma 5.2.4 for every $i \neq k$ that $y_i^{C+t} = x_i^t$ for $t = 0, 1, \dots$, since

$y_i^C = x_i^0 = 0$. Similarly, it follows that $y_k^{C+t} \geq x_k^t$ for $t = 0, 1, \dots$. Suppose that $y_k^{C+t} > 0$ for $t = 0, 1, \dots, T$. Then it follows by induction that

$$y_k^{C+T} = y_k^C + N_k^T \cdot S_k - T \leq (C + N_k^T) \cdot S_k - T \leq 0,$$

which gives a contradiction. Thus there exists some $t' \in \{0, 1, \dots, T\}$ for which $y_k^{C+t'} = 0 = x_k^{t'}$. Then, it follows by Lemma 5.2.4 that $y_k^{C+t} = x_k^t$ for every $t \geq t'$ and thus $y_k^{C+T} = x_k^T$. Hence $Y_{C+T} = X_T$. Moreover, since $B_{C+t} = A_t$ for $t = 0, 1, \dots$, it follows that $Y_{C+T+t} = X_{T+t}$ and thus for the costs we have that $C(Y_{C+T+t}, B_{C+T+t}) = C(X_{T+t}, A_{T+t})$ for $t = 0, 1, \dots$.

Put $K := \max_{t \in \{0, 1, \dots, T-1\}} C(Y_t, B_t)$. Then we have that

$$V_\alpha(s) \leq V_{\theta, \alpha}(s) = \sum_{t=0}^{\infty} \alpha^t C(Y_t, B_t) \leq K \cdot (C + T) + \alpha^C \cdot \sum_{t=T}^{\infty} \alpha^t C(X_t, A_t) \leq$$

$$K \cdot (C + T) + V_\alpha(z).$$

Hence $h_\alpha(s) = V_\alpha(s) - V_\alpha(z) \leq M(s)$, where $M(s) := K \cdot (C + T)$. We conclude that (A2) holds for this MDC.

Verification of (A3) To show that (A3) holds it suffices to prove that $h_\alpha(s) \geq 0$ for every $s \in X$ and $\alpha \in (0, 1)$. Let ψ_α be an optimal stationary policy for the discounted cost criterion with discount factor α . Let $X_t := (x_1^t, x_2^t, \dots, x_N^t)$, $t = 0, 1, \dots$ be the state at decision epoch t under ψ_α , given initial state $X_0 = s = (x_1^0, x_2^0, \dots, x_N^0)$ and let A_t , $t = 0, 1, \dots$ be the action at decision epoch t . We define a deterministic policy θ by putting $B_t = A_t$ for $t = 0, 1, \dots$, where B_t is the action at decision epoch t for policy θ . Let $Y_t := (y_1^t, y_2^t, \dots, y_N^t)$, $t = 0, 1, \dots$ be the state at decision epoch t under θ , given initial state $Y_0 = z = (0, 0, \dots, 0)$. By Lemma 5.2.4 we have that $y_j^t \leq x_j^t$ for every $j \in \{1, 2, \dots, N\}$ and $t \in \mathbb{Z}_{\geq 0}$. So, $C(Y_t, B_t) = C(Y_t, A_t) \leq C(X_t, A_t)$ for $t = 0, 1, \dots$ and thus we have that

$$V_\alpha(z) \leq V_{\theta, \alpha}(z) = \sum_{t=0}^{\infty} \alpha^t C(Y_t, B_t) \leq \sum_{t=0}^{\infty} \alpha^t C(X_t, A_t) = V_{\psi_\alpha, \alpha}(s) = V_\alpha(s).$$

Hence $h_\alpha(s) \geq 0$ and we conclude that (A3) holds.

We have shown that the assumptions (A1), (A2) and (A3) are satisfied for an MDC corresponding to an (S_1, S_2, \dots, S_N) system if $\sum_{i=1}^N \frac{1}{S_i} > 1$. By Theorem 5.2.2 we have the following theorem.

Theorem 5.2.5 *Suppose we have an (S_1, S_2, \dots, S_N) system with $\sum_{i=1}^N \frac{1}{S_i} > 1$. Then, for the MDC corresponding to this system there exists an optimal stationary*

policy f with minimal long-run average waiting time of the arriving customers equal to

$$J(S_1, S_2, \dots, S_N) := J_f = J < \infty.$$

5.3 Properties of routing sequences corresponding to optimal stationary policies

5.3.1 Ultimately periodic routing sequences

Suppose we have an (S_1, S_2, \dots, S_N) system with $\sum_{i=1}^N \frac{1}{S_i} > 1$ and let f be an optimal stationary policy for the MDC corresponding to this system. Let $X_t := (x_1^t, x_2^t, \dots, x_n^t)$, $t = 0, 1, \dots$ be the state at time t and A_t , $t = 0, 1, \dots$ be the action taken at time t given that the initial state X_0 is the zero state $(0, 0, \dots, 0)$. Recall that the variables X_t and A_t are deterministic since f is a deterministic policy and that the routing sequence $U = (U_1, U_2, \dots)$ corresponding to f is determined by $U_n = A_{n-1}$ for $n = 1, 2, \dots$. So, U is an one-sided infinite sequence of integers in the domain $\{1, 2, \dots, N\}$, which as before can be seen as an \mathbb{N} -word on the alphabet $\{1, 2, \dots, N\}$. In the next theorem we show that if we assume that the service times S_1, S_2, \dots, S_N are rational numbers, then there exists some positive integer m and a positive integer T (the period) such that $U_n = U_{n+T}$ for every $n \in \mathbb{N}$, $n \geq m$. In other words U is ultimately periodic.

Theorem 5.3.1 *Suppose we have an (S_1, S_2, \dots, S_N) system with $\sum_{i=1}^N \frac{1}{S_i} > 1$ and the service times (S_1, S_2, \dots, S_N) are rational numbers. Let f be an optimal stationary policy for this system and $U = (U_1, U_2, \dots)$ the corresponding routing sequence. Then U is ultimately periodic.*

Proof. Let $W(U)$, the long-run average waiting time for routing sequence U , and $L(U)$, the long-run average number of customers in the system, be defined as in Section 4.3. By Theorem 5.2.5 we have that $W(U) = J_j = J < \infty$ and thus it follows by (4.4) that

$$L(U) = W(U) < \infty. \tag{5.6}$$

Since f is a stationary policy it suffices to show that there exist $t \in \mathbb{Z}_{\geq 0}$ and $T \in \mathbb{N}$ such that $X_t = X_{t+T}$, i.e. the state reached at time $t + T$ is the same as the state

reached at time t . Indeed, in that case it follows that $U_n = U_{n+T}$ for every $n > t$. Suppose that there do not exist such t, T . Then it is easily seen that the reached states X_0, X_1, X_2, \dots are distinct. Recall that $B(t)$ denotes the number of customers in the system at time t . Note that for $t = 0, 1, \dots$ we have $B(t) = \sum_{i=1}^N \lceil \frac{x_i^t}{S_i} \rceil$. Since $B(\lceil t \rceil) \leq B(t)$ for every $t \geq 0$ we have that

$$L(U) \geq \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n B(t) = \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n \sum_{i=1}^N \lceil \frac{x_i^t}{S_i} \rceil. \quad (5.7)$$

Since all the service times S_i are rational numbers, it is easily seen that for every real number $\varepsilon > 0$ there exist only finitely many states $(x_1, x_2, \dots, x_N) \in X$ for which $\sum_{i=1}^N \lceil \frac{x_i^t}{S_i} \rceil \leq L(U) + \varepsilon$. So, X_0, X_1, \dots cannot be distinct. This contradiction proves the theorem. \square

5.3.2 The existence of an optimal periodic routing sequence in case of rational service times

In the previous subsection we have shown that for an (S_1, S_2, \dots, S_N) system for which $\sum_{i=1}^N \frac{1}{S_i} > 1$ and the service times S_i are rational there exists an optimal ultimately periodic routing sequence $U = (U_1, U_2, \dots)$. In this subsection we derive the slightly stronger result that for such systems there always exists an optimal periodic routing sequence.

A sequence $V = (V_1, V_2, \dots)$ is called a shift of a sequence $U = (U_1, U_2, \dots)$ if $V_n = U_{n+m}$ for some $m \in \mathbb{Z}_{\geq 0}$ and all $n \in \mathbb{N}$.

Lemma 5.3.2 *Let $U = (U_1, U_2, \dots)$ and $V = (V_1, V_2, \dots)$ be routing sequences for an (S_1, S_2, \dots, S_N) system such that $W(U) < \infty$ and V is a shift of U . Then $W(V) \leq W(U)$.*

Proof. Let $X_t = (x_1^t, x_2^t, \dots, x_N^t)$, $Y_t = (y_1^t, y_2^t, \dots, y_N^t)$ be the state at decision epoch t under respectively routing sequence U and routing sequence V given that in both cases the initial state is the zero state. Let $m \in \mathbb{Z}_{\geq 0}$ be such that $U_{n+m} = V_n$ for $n = 1, 2, \dots$. For $j = 1, 2, \dots, N$ we have that $x_j^m \geq y_j^0 = 0$. So, by Lemma 5.2.4 it follows that $x_j^{m+t} \geq y_j^t$ for every $t \in \mathbb{Z}_{\geq 0}$ and $C(X_{t+m}, U_{t+m+1}) \geq C(Y_t, V_{t+1})$ for $t = 0, 1, \dots$. Hence

$$W(V) = \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} C(Y_t, V_{t+1}) \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} C(X_{t+m}, U_{t+m+1}) =$$

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} C(X_t, U_{t+1}) = W(U). \quad \square$$

Theorem 5.3.3 *Suppose we have an (S_1, S_2, \dots, S_N) system for which $\sum_{i=1}^N \frac{1}{S_i} > 1$ and the service times S_i are rational numbers. Then there exists an optimal periodic routing sequence.*

Proof. By Theorem 5.3.1 there exists an ultimately periodic routing sequence $U = (U_1, U_2, \dots)$ which is optimal and thus $W(U) = J(S_1, S_2, \dots, S_N) = J < \infty$. Let m, T be the minimal positive integers such that $U_n = U_{n+T}$ for every $n \geq m$. Let $V = (V_1, V_2, \dots)$ where $V_n = U_{m+n-1}$ for $n = 1, 2, \dots$. It is easily seen that V is periodic with period T . Moreover, since V is a shift of U we have by Lemma 5.3.2 that $W(V) = W(U) = J < \infty$ and thus V is an optimal routing sequence. \square

5.4 Optimal routing sequences in case of irrational service times

In the previous section we showed that there exists an optimal periodic routing sequence in case of rational service times if $\sum_{i=1}^N \frac{1}{S_i} > 1$. In this section we consider the case that not all the service times S_i are rational. An obvious question is whether there still exists an optimal periodic routing sequence in that case. We have a partial answer to that question. Namely, if $N = 2$ then there exists also in case of irrational service times an optimal periodic routing sequence (see Theorem 4.7.12 and [23]). However, if $N > 2$ then this problem is open.

We conjecture that if $N > 2$ there exist (S_1, S_2, \dots, S_N) systems with $\sum_{i=1}^N \frac{1}{S_i} > 1$ for which there does not exist an optimal periodic routing sequence if $N > 2$. For example consider an (S_1, S_2, S_3) system where $S_1 = 2 - \varepsilon$ with ε an arbitrarily small positive number, $S_2 = 2\sqrt{2}$ and $S_3 = 2\sqrt{2} + 4$. Note that $\frac{1}{S_2} + \frac{1}{S_3} = \frac{1}{2}$ and thus $\sum_{i=1}^3 \frac{1}{S_i} > 1$. Let $U = (U_1, U_2, \dots)$ be an optimal routing sequence for this system and recall that $N_i^t = |\{n \leq t : U_n = i\}|$ for $i = 1, 2, 3$ is the number of times that a customer is routed to queue i among the first t decision epochs. We think that $p_i := \lim_{t \rightarrow \infty} \frac{N_i^t}{t}$, the fraction of customers routed to queue i , exists for $i = 1, 2, 3$ and that $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{S_2}$ and $p_3 = \frac{1}{S_3}$. If this is valid, then p_2 and p_3 are irrational, which implies that U is not periodic.

So, we believe that if $N > 2$ then there exist (S_1, S_2, \dots, S_N) systems with $\sum_{i=1}^N \frac{1}{S_i} > 1$ for which there does not exist an optimal periodic routing sequence. On the

other hand this is only possible if we have for routing sequences U corresponding to optimal stationary policies that there exists some $i \in \{1, 2, \dots, N\}$ for which $\limsup_{t \rightarrow \infty} \frac{N_i^t}{t} = \frac{1}{S_i}$, where S_i is irrational. Indeed, if for $i = 1, 2, \dots, N$ it holds that $\limsup_{t \rightarrow \infty} \frac{N_i^t}{t} < \frac{1}{S_i}$ or S_i is rational then the Markov decision process reaches only finitely many states. Analogously to the proof of Theorem 5.3.1 this would imply that U is ultimately periodic and thus there exists an optimal periodic routing sequence. So, we believe that there exist (S_1, S_2, \dots, S_N) systems with $\sum_{i=1}^N \frac{1}{S_i} > 1$ for which there does not exist an optimal periodic routing sequence, but that such systems are quite exceptional.

Since we assume that there exist (S_1, S_2, \dots, S_N) systems with $\sum_{i=1}^N \frac{1}{S_i} > 1$ for which there does not exist an optimal periodic routing sequence, we are interested whether optimal routing sequences satisfy in general some other (weaker) properties than periodicity. From Lemma 2.3.6 we have for (S_1, S_2, \dots, S_N) systems with $\sum_{i=1}^N \frac{1}{S_i} = 1$ that for every optimal routing sequence $U = (U_1, U_2, \dots)$ it holds that $p_i := \lim_{t \rightarrow \infty} \frac{N_i^t}{t}$ exists for $i = 1, 2, \dots, N$. In other words every letter in an optimal routing sequence has a density for such systems. We think that (S_1, S_2, \dots, S_N) systems with $\sum_{i=1}^N \frac{1}{S_i} > 1$ behave similarly. More precisely, we have the following conjecture.

Conjecture 5.4.1 *Suppose we have an (S_1, S_2, \dots, S_N) system with $\sum_{i=1}^N \frac{1}{S_i} > 1$. Let f be an optimal stationary policy for this system and $U = (U_1, U_2, \dots)$ the corresponding optimal routing sequence. Then $\lim_{t \rightarrow \infty} \frac{N_i^t}{t}$ exists for $i = 1, 2, \dots, N$.*

It is obvious that the existence of the densities of every letter in a routing sequence is a weaker property than periodicity. Another property that is weaker than periodicity is uniform recurrence, which means that every finite subsequence occurs infinitely often with bounded gaps between two consecutive occurrences.

Definition 5.4.2 *Let $U = (U_1, U_2, \dots)$ be an infinite word on the alphabet $\mathcal{A} = \{1, 2, \dots, N\}$ and let $w = (w_1, w_2, \dots, w_k)$ be a finite word of length k on \mathcal{A} . Let*

$$G = \{n \in \mathbb{N} : U_{n+i-1} = w_i \text{ for } i = 1, 2, \dots, k\}.$$

If $\lim_{t \rightarrow \infty} \frac{|G \cap \{1, 2, \dots, t\}|}{t}$ exists, then the value of this limit is the density of w in U .

We think that for every (S_1, S_2, \dots, S_N) system with $\sum_{i=1}^N \frac{1}{S_i} > 1$ there exists an optimal routing sequence U such that U is uniformly recurrent and every finite word on the alphabet $\{1, 2, \dots, N\}$ has a density in U , but the problem is still open.

We conclude that for (S_1, S_2, \dots, S_N) systems with $\sum_{i=1}^N \frac{1}{S_i} > 1$ there exist optimal stationary policies for the corresponding MDC. From this it follows that there exists an optimal periodic routing sequence if the service times S_i are rational numbers. However, in case of irrational service times it is in general not even known whether there exist an optimal routing sequence for which every letter has a density. The existence of optimal periodic routing sequences or the existence of optimal routing sequences in which every letter has a density are also interesting problems for other parallel queueing systems with stochastic interarrival and service times.

Chapter 6

Multimodular functions and partial orders on routing sequences

6.1 Introduction

In several papers ([7, 8]), various attempts have been made to study the question which of two deterministic periodic admission sequences (periodic sequences of non-negative integers) gives the smaller average expected waiting time. A partial order, called the cone order is introduced in [7], and it is shown that the average waiting time and more generally any multimodular function is monotone with respect to the cone order. It is natural to define a multimodular order by requiring that any multimodular function is monotone. In contrast to [34], where the convex order for stochastic admission sequences is used, we consider only deterministic admission sequences. Note that deterministic admission sequences can only be ordered for all multimodular functions if they are equal. Therefore we consider multimodular functions with a fixed minimal point as was done in [7]. In Section 6.2 we introduce the multimodular order and we show that the cone order is equivalent to the multimodular order. In Subsection 6.2.3 the shift invariant counterparts of these orders are studied and it is shown that the regular admission sequence is the minimal point.

For the optimal routing problem to N queues a lower bound is obtained by using

regular admission sequences (or what is the same, bracket sequences) with 'minimizing' routing fractions (densities) as admission sequences to all queues (see [5]). But generally only in the routing to $N = 2$ queues, the routing fractions will be balanceable (see [4]) and only in that case the admission sequences can be glued together and be made to a feasible routing policy.

After we have generalized the definition of (both the primal and dual) unbalance from Section 3.2 to (periodic) sequences of nonnegative integers we show that the unbalance is a shift invariant and multimodular function on such sequences. The unbalance is a combinatorial notion and the analysis in this chapter is mainly combinatorial. It turns out that the notion of unbalance is also useful in other models than routing to queues. Indeed, the bounds involving the unbalance (see Chapter 3) can be generalized from waiting times to similar bounds for sequences of multimodular functions as in [5] (see the remark at the end of Section 6.3).

In Section 6.3 we also generalize the definition of the graph orders to periodic sequences of nonnegative integers. This gives the distance of a given periodic sequence of nonnegative integers to the bracket sequence and thereby defines its unbalance. The relations between the shift invariant orders and the graph orders are studied in Section 6.4. We show through counterexamples that they are not equivalent. More details can be found in the monograph ([9]), which contains the results of this chapter.

6.2 The multimodular order and the cone order

Let D be a multimodular matrix with rowvectors $d_0, d_1, \dots, d_n \in \mathbb{R}^n$ verifying $d_0 + \dots + d_n = 0$ where d_1, d_2, \dots, d_n are linearly independent over \mathbb{R} and let $M_D \subset \mathbb{R}^n$ be the corresponding mesh, that is the set of all the points $\{a_0 d_0 + a_1 d_1 + \dots + a_n d_n, \quad a_i \in \mathbb{Z}, \quad i = 0, \dots, n\}$.

Definition 6.2.1 *A function $f : M_D \rightarrow \mathbb{R}$ is D -multimodular if and only if for all $a \in M_D$, and for all i, j with $0 \leq i < j \leq n$,*

$$f(a + d_i) + f(a + d_j) \geq f(a) + f(a + d_i + d_j). \quad (6.1)$$

For $r \in M_D$ let

$$\mathcal{F}(D, r) = \left\{ f : M_D \rightarrow \mathbb{R} \text{ such that } f \text{ is } D\text{-multimodular and } f(r) = \min_{x \in M_D} f(x) \right\}$$

be the set of D -multimodular functions with global minimum point at r .

6.2.1 The cone order

This section is devoted to the definition of the cone order. More on this can be found in [7].

Definition 6.2.2 *A D-atom is a simplex in \mathbb{R}^n , made of the $n + 1$ points*

$$\begin{aligned}
 p_0 &= a, \\
 p_1 &= a + d_{\xi(0)}, \\
 p_2 &= a + d_{\xi(0)} + d_{\xi(1)} \\
 &\vdots \\
 p_n &= a + d_{\xi(0)} + d_{\xi(1)} + \cdots + d_{\xi(n-1)}
 \end{aligned} \tag{6.2}$$

where $a \in M_D$ (called the root) and ξ is a permutation of $\{0, \dots, n\}$. This atom will be denoted as $\mathcal{S}(p_0, \dots, p_n)$.

Theorem 6.2.3 *The set of all the D-atoms forms a triangulation of \mathbb{R}^n , called a multimodular triangulation.*

Proof. Let x be a point in \mathbb{R}^n . We can write $x = \alpha_0 d_0 + \cdots + \alpha_n d_n$, with non-negative coordinates, one of which is 0 since $d_0 + \cdots + d_n = 0$. We construct $a = \lfloor \alpha_0 \rfloor d_0 + \cdots + \lfloor \alpha_n \rfloor d_n$ and ξ such that $\alpha_{\xi(i)} - \lfloor \alpha_{\xi(i)} \rfloor \geq \alpha_{\xi(i+1)} - \lfloor \alpha_{\xi(i+1)} \rfloor$. We define $\beta_n = 0$, $\beta_{n-1} = \alpha_{\xi(n-1)} - \lfloor \alpha_{\xi(n-1)} \rfloor$, $\beta_i = \alpha_{\xi(i)} - \lfloor \alpha_{\xi(i)} \rfloor - (\alpha_{\xi(i+1)} - \lfloor \alpha_{\xi(i+1)} \rfloor)$. All of them verify $0 \leq \beta_i \leq 1$ and $\sum_{i=0}^{n-1} \beta_i \leq 1$. We have $x = a + \beta_0 d_{\xi(0)} + \cdots + \beta_{n-1} (d_{\xi(0)} + \cdots + d_{\xi(n-1)})$. Therefore, x belongs to the atom with root a and permutation ξ .

Now, assume that a point x belongs to the interior of two different atoms with respective roots a and b and permutations ξ and τ . Since everything is shift invariant, we may assume that $b = 0$ and τ is the identity. We may write

$$\begin{aligned}
 x &= a + \alpha_0 d_{\xi(0)} + \cdots + \alpha_{n-1} (d_{\xi(0)} + \cdots + d_{\xi(n-1)}) \\
 &= \beta_0 d_0 + \cdots + \beta_{n-1} (d_0 + \cdots + d_{n-1})
 \end{aligned}$$

with $\sum_{i=0}^{n-1} \beta_i \leq 1$ and $\sum_{i=0}^{n-1} \alpha_i \leq 1$. Since x is in the interior of both atoms, we also have $\beta_i > 0$ and $\alpha_i > 0$ for $i = 0, \dots, n - 1$. Therefore, by uniqueness of the

decomposition of x , and writing $a = a_0d_0 + \cdots + a_nd_n$,

$$\begin{aligned}
a_0 + \sum_{j=\xi^{-1}(0)}^{n-1} \alpha_{\xi(j)} &= \beta_0 + \cdots + \beta_{n-1} \\
a_1 + \sum_{j=\xi^{-1}(1)}^{n-1} \alpha_{\xi(j)} &= \beta_1 + \cdots + \beta_{n-1} \\
&\vdots \\
a_{n-1} + \sum_{j=\xi^{-1}(n-1)}^{n-1} \alpha_{\xi(j)} &= \beta_{n-1} \\
a_n + \sum_{j=\xi^{-1}(n)}^{n-1} \alpha_{\xi(j)} &= 0.
\end{aligned}$$

Since the partial sums of the α_i or of the β_i are smaller than one and since a_i are integer numbers, we have $a_i = 0$ for all $i = 0, \dots, n$. Therefore both atoms have the same root. Now, the equality of the partial sums taken one by one implies first that $\sum_{j=\xi^{-1}(n)}^{n-1} \alpha_{\xi(j)} = 0$. Since $\alpha_i > 0$ for all i , the only possibility is $\xi^{-1}(n) = n$. Considering vectors d_k and d_{k+1} , we have

$$\sum_{j=\xi^{-1}(k)}^{n-1} \alpha_{\xi(j)} - \sum_{j=\xi^{-1}(k+1)}^{n-1} \alpha_{\xi(j)} = \beta_k > 0.$$

This implies that $\xi^{-1}(k) > \xi^{-1}(k+1)$. Thus ξ is the identical permutation. Therefore, both atoms are equal. \square

Consider one atom $S = \mathcal{S}(p_0, p_1, \dots, p_n)$ containing $r = p_0$ as a vertex (the global minimum point of the multimodular functions considered above). Let ξ and τ be the permutations on $\{0, \dots, n\}$ such that $\tau(0) = 0$ and

$$p_{\tau(1)} = r + d_{\xi(0)} \tag{6.3}$$

$$p_{\tau(2)} = p_{\tau(1)} + d_{\xi(1)} \tag{6.4}$$

$$\vdots = \vdots \tag{6.5}$$

$$p_{\tau(n)} = p_{\tau(n-1)} + d_{\xi(n-1)} \tag{6.6}$$

$$r = p_{\tau(n)} + s_{\xi(n)}. \tag{6.7}$$

For $1 \leq i \leq n$ we define $b_i = \sum_{j=1}^i d_{\xi(j)}$. Let $b = (b_1, \dots, b_n)$ be the $n \times n$ matrix for which the i -th column is vector b_i . Note that $p_{\tau(i)} = r + b_i$ for $i = 1, 2, \dots, n$. The cone associated with such an atom S , denoted by $C(S)$, contains all the points x of M_D such that $x = r + c^t(x)b$ for some non-negative column vector $c = c(x)$ in \mathbb{N}_0^n .

Conversely, for any point x in M_D , there exists such a cone containing x and we will have $x = r + c^t(x)b$, where the columns of b corresponding to the support of $c(x)$ are uniquely defined. We shall denote $d(r, x) = c_1(x) + \dots + c_n(x)$ and call it the cone-distance from x to r .

It is easily seen that there are $(n+1)!$ atoms with root in r and thus there are exactly $(n+1)!$ cones corresponding to the given multimodular matrix D and root r . Let $\mathcal{C} = \{C_i\}_{i=1}^{(n+1)!}$ be the set of cones corresponding to D and root $r = (0, 0, \dots, 0)$ (the origin). Note that from the preceding argument and Theorem 6.2.3 it follows that $\{\cup C_i : C_i \in \mathcal{C}\} = \mathbb{R}^n$. For $C_i \in \mathcal{C}$ let C'_i be the intersection of C_i and M_D . In the following proposition which follows directly from the definitions and preceding results we elucidate the properties of the objects we have associated to a multimodular matrix D with rowvectors $d_0, d_1, \dots, d_n \in \mathbb{R}^n$.

Proposition 6.2.4 *Let $C_i \in \mathcal{C}$. Then there exist $b_1, b_2, \dots, b_n \in \mathbb{R}^n$ and a bijection $\sigma : \{1, 2, \dots, n+1\} \rightarrow \{0, 1, \dots, n\}$ such that*

1. b_1, b_2, \dots, b_n are linearly independent.
2. $b_j = \sum_{i=1}^j d_{\sigma(i)}$ for $j = 1, 2, \dots, n$.
3. $C_i = \{\sum_{i=1}^n \lambda_i \cdot b_i : \lambda_i \geq 0 \text{ for } i = 1, 2, \dots, n\}$.
4. $M_D = \{\sum_{i=1}^n \lambda_i \cdot b_i : \lambda_i \in \mathbb{Z} \text{ for } i = 1, 2, \dots, n\}$.
5. $C'_i = \{\sum_{i=1}^n \lambda_i \cdot b_i : \lambda_i \in \mathbb{Z}_{\geq 0} \text{ for } i = 1, 2, \dots, n\}$.

For a convex cone $C \subset \mathbb{R}^n$ we say that a set of vectors v_1, v_2, \dots, v_k are generators of C if v_1, v_2, \dots, v_k are linearly independent and $C = \{\sum_{i=1}^k \lambda_i \cdot v_i : \lambda_i \geq 0 \text{ for } i = 1, 2, \dots, k\}$. In that case we say that C is a k -dimensional cone. In particular we have for b_1, b_2, \dots, b_n as in Proposition 6.2.4 that they are generators of the n -dimensional convex cone C_i .

Definition 6.2.5 *For $x, y \in M_D$ we denote $x \leq_C y$ if there exists some cone $C_i \in \mathcal{C}$, with generators b_1, b_2, \dots, b_n , such that $x, y \in C_i$ and $c(x) \leq c(y)$ component-wise in cone C_i .*

If $x \leq_C y$ then we say that x and y are cone ordered and \leq_C is called the cone order.

6.2.2 The multimodular order and the cone order

Definition 6.2.6 *Let the multimodular matrix D and global minimum point $r \in M_D$ be given. Then for $x, y \in M_D$ we say that $x \leq_{mm} y$ if $f(x) \leq f(y)$ for every $f \in \mathcal{F}(D, r)$.*

It is easy to verify that both the multimodular order \leq_{mm} and the cone order \leq_C are reflexive, antisymmetric and transitive and thus they are partial orders on M_D . The following theorem says that these partial orders on M_D are equivalent.

Theorem 6.2.7 *For given multimodular matrix D and global minimum point $r \in M_D$ we have for $x, y \in M_D$ that $x \leq_{mm} y$ if and only if $x \leq_C y$.*

First we prove that $x \leq_C y$ implies $x \leq_{mm} y$. Assume that x and y are in the same cone C with $x \leq_C y$ and let b_1, \dots, b_n be the generators of this cone. Note that we can assume that $d(x, y) := |d(r, y) - d(r, x)| = 1$. If not, then we prove step by step along the path from x to y along the direction of the generators, say $x = x_1 \leq_C \dots \leq_C x_m = y$ that $f(x) = f(x_1) \leq \dots \leq f(x_m) = f(y)$ for $f \in \mathcal{F}(D, r)$. The proof will now proceed by induction on $d(r, x)$. The property is true if $d(x, r) = 0$, since r is the argmin of f on S . Now, assume $d(x, r) \geq 1$. Pick a point z in cone C such that $x = z + b_i$. (equivalently, we have $c(y) + e_i = c(x)$ and $c(y) \leq 0$). Note that $d(z, r) = d(x, r) - 1$ and $z \leq_C x$. By induction, this means $f(z) \leq f(x)$. Since $d(x, y) = 1$, there exists some j such that $y = x + b_j$. Now we distinguish two cases.

- If $i = j$, then by convexity of f ,

$$f(z + b_i) - f(z) \leq f(z + b_i + b_i) - f(z + b_i).$$

We also know by induction that $f(z + b_i) - f(z) \geq 0$. Thus $f(x) \leq f(y)$.

- If $i \neq j$, then we put $w = z + b_j$. We may assume that $i > j$ (the case $j < i$ is similar by inverting the role played by b_i and b_j in the following).

Since b_i is a sum of base vectors and all base vectors add up to 0, it is also a sum of opposites of base vectors. Note that all these base vectors are distinct from the base vectors involved in b_j .

We have

$$b_j = d_{\xi(1)} + \cdots + d_{\xi(j)}, \quad (6.8)$$

$$b_i = -d_{\xi(i+1)} - d_{\xi(i+2)} - \cdots - d_{\xi(n+1)}. \quad (6.9)$$

Therefore, by the multimodularity of f ,

$$f(w) - f(z) = f(x + d_{\xi(1)} + \cdots + d_{\xi(j)} + d_{\xi(i+1)} + \cdots + d_{\xi(n+1)}) -$$

$$f(x + d_{\xi(i+1)} + \cdots + d_{\xi(n+1)}) \leq f(x + d_{\xi(1)} + \cdots + d_{\xi(j)}) - f(x) = f(y) - f(x).$$

Since $d(r, w) = d(r, z) + 1$, we have $f(w) - f(z) \geq 0$ by induction. Thus $f(y) - f(x) \geq 0$.

To finish the proof of Theorem 6.2.7 we have to show that $x \leq_{mm} y$ implies that $x \leq_C y$. By translation we can assume without loss of generality that the global minimum point $r = (0, 0, \dots, 0) \in M_D$. Let b_1, b_2, \dots, b_n be the generators of the cone $C_i \in \mathcal{C}$ as in Proposition 6.2.4. If $x \in M_D$ then we have by Proposition 6.2.4 that there exist unique $x_j \in \mathbb{Z}$ such that $x = \sum_{j=1}^n x_j \cdot b_j$. Moreover, for $x = \sum_{j=1}^n x_j \cdot b_j \in \mathbb{R}^n$ we define $\text{supp}_i(x) = \{k \in \{1, 2, \dots, n\} \text{ for which } x_k > 0\}$. Note that $x_k > 0$ is stronger than in the standard definition of support. For $k = 1, 2, \dots, n$ we define the function $f_k : M_D \rightarrow \mathbb{Z}_{\geq 0}$ by

$$f_k(x) = \max(x_k, 0). \quad (6.10)$$

Proposition 6.2.8 *For every $k \in \{1, 2, \dots, n\}$ the function f_k is D -multimodular with global minimum point in $r = \mathbf{0}$.*

Proof. Put $f = f_k$. Since $f(x) \geq 0$ for every $x \in M_D$ and $f(r) = 0$, it follows that r is a global minimum point of f . It remains to prove that f is a D -multimodular function. So, we have to show that

$$f(x + d_i) + f(x + d_j) \geq f(x) + f(x + d_i + d_j) \quad (6.11)$$

for every $x \in M_D$ and $i, j \in \{0, 1, \dots, n\}$ with $i \neq j$.

According to property 2 of Proposition 6.2.4 there exists a bijection $\sigma : \{1, 2, \dots, n + 1\} \rightarrow \{0, 1, 2, \dots, n\}$ such that $b_j = \sum_{i=1}^j d_{\sigma(i)}$ for $j = 1, 2, \dots, n$. From this it follows that $d_{\sigma(1)} = b_1$, $d_{\sigma(j)} = b_j - b_{j-1}$ for $2 \leq j \leq n$ and $d_{\sigma(n+1)} = -\sum_{i=1}^n d_{\sigma(i)} =$

$-b_n$. Put $l = \sigma(k)$, $m = \sigma(k + 1)$. If $i \notin \{l, m\}$ and $j \notin \{l, m\}$ then we have for every $x \in M_D$ that $f(x + d_i) = f(x + d_j) = f(x) = f(x + d_i + d_j)$. Hence (6.11) holds. Suppose that $i \in \{l, m\}$ and $j \notin \{l, m\}$. Then we have for every $x \in M_D$ that $f(x + d_i) = f(x + d_i + d_j)$ and $f(x + d_j) = f(x)$ and thus (6.11) holds. Suppose that $\{i, j\} = \{l, m\}$ and assume without loss of generality that $i = l$ and $j = m$. If $f(x) > 0$ then $f(x + d_i) = f(x) + 1$, $f(x + d_j) = f(x) - 1$, $f(x + d_i + d_j) = f(x)$ and thus (6.11) holds. If $f(x) = 0$ then $f(x + d_i) \geq f(x)$, $f(x + d_j) = f(x + d_i + d_j) = f(x) = 0$ and thus (6.11) holds too. So, we can conclude that (6.11) holds in every case and thus $f \in \mathcal{F}(D, r)$. \square

For $i = 1, 2, \dots, (n+1)!$ let $\mathcal{G}(C_i)$ be the set of functions $\{f_1, f_2, \dots, f_n\}$ for cone C_i and let $\mathcal{G} = \bigcup_{i=1}^{(n+1)!} \mathcal{G}(C_i)$. We state some corollaries of Proposition 6.2.8.

Corollary 6.2.9 *For given multimodular matrix D and global minimum point $r = (0, 0, \dots, 0) \in \mathbb{R}^n$ we have that $\mathcal{G} \subseteq \mathcal{F}(D, r)$.*

Corollary 6.2.10 *If $x, y \in M_D$, $x \leq_{mm} y$ and there exists some cone $C_i \in \mathcal{C}$ such that $x, y \in C_i$ then $x \leq_C y$.*

Proof. Let b_1, b_2, \dots, b_n be the generators of cone C_i as in Proposition 6.2.4 and put $x = \sum_{i=1}^n x_i \cdot b_i$ and $y = \sum_{i=1}^n y_i \cdot b_i$. Then, by Proposition 6.2.4, $x_i, y_i \in \mathbb{Z}_{\geq 0}$ for $i = 1, 2, \dots, n$. So, for $f_k \in \mathcal{G}(C_i)$ we have that $f_k(x) = x_k$ and $f_k(y) = y_k$ for $k = 1, 2, \dots, n$. Hence by Proposition 6.2.8 and $x \leq_{mm} y$ we derive that $0 \leq x_k = f_k(x) \leq f_k(y) = y_k$ for $k = 1, 2, \dots, n$. Thus $x \leq_C y$. \square

Corollary 6.2.11 *If $x, y \in M_D$, $x \leq_{mm} y$, then $\text{supp}_i(x) \subset \text{supp}_i(y)$ for $i = 1, 2, \dots, (n+1)!$.*

Corollary 6.2.11 can be proved in the same way as Corollary 6.2.10. The next corollary follows immediately from Corollary 6.2.10 and Corollary 6.2.11.

Corollary 6.2.12 *If $x, y \in M_D$, $x \leq_{mm} y$ and there exists some cone $C_i \in \mathcal{C}$ such that x is an interior point of C_i then $x \leq_C y$.*

To prove Theorem 6.2.7 it suffices by Corollary 6.2.10 to prove that there exists some cone $C_i \in \mathcal{C}$ such that $x, y \in C_i$ if $x \leq_{mm} y$. If $x = r = \mathbf{0}$, $y = r = \mathbf{0}$ or $y = \lambda \cdot x$ with $\lambda > 0$ then it follows directly that $x, y \in C_i$ for some $C_i \in \mathcal{C}$. Suppose that $y = \lambda \cdot x$ with $\lambda < 0$ and $x \neq r$. Let $C_i \in \mathcal{C}$ be a cone containing x with generators b_1, b_2, \dots, b_n . Then $x = \sum_{i=1}^n x_i \cdot b_i$ and $y = \sum_{i=1}^n \lambda \cdot x_i \cdot b_i$ with $x_i \geq 0$ for $i = 1, 2, \dots, n$. Moreover we have that $x_k > 0$ for some $k \in \{1, 2, \dots, n\}$ by

$x \neq r$. Then for $f_k \in \mathcal{G}(C_i)$ we deduce that $f_k(x) = x_k > 0$ and $f_k(y) = 0$, which contradicts $x \leq_{mm} y$. So, we can assume that x and y are linearly independent vectors in \mathbb{R}^n . Let $H := \text{span}(x, y)$ be the two dimensional subspace that contains x and y . For $C_j \in \mathcal{C}$ let C_j^* be the intersection of C_j and H . Then the C_j^* are convex cones of dimension smaller or equal than two in the subspace H . To prove Theorem 6.2.7 we derive the following lemmas.

Lemma 6.2.13 *Suppose that $C_i^* = C_i \cap H$ is a two dimensional cone generated by u and v . Then $\text{supp}_i(u)$ is not a subset of $\text{supp}_i(v)$ and $\text{supp}_i(v)$ is not a subset of $\text{supp}_i(u)$.*

Proof. Let b_1, b_2, \dots, b_n be the generators of C_i as in Proposition 6.2.4. Then $u = \sum_{k \in \text{supp}_i(u)} u_k \cdot b_k$ with $u_k > 0$ and $v = \sum_{k \in \text{supp}_i(v)} v_k \cdot b_k$ with $v_k > 0$. For a small positive number ε we have that $w := v - \varepsilon \cdot u \in H$. If $\text{supp}_i(u) \subseteq \text{supp}_i(v)$ then $w = \sum_{k \in \text{supp}_i(v)} w_k \cdot b_k$ with $w_k > 0$ and thus $w \in C_i^*$. However, $w \in C_i^*$ contradicts the fact that u and v are generators of C_i^* . The case $\text{supp}_i(v) \subseteq \text{supp}_i(u)$ is similar. \square

Lemma 6.2.14 *Let $u, v, w \in H$ with u, v linearly independent and u, w linearly independent. Let A_1 be the two-dimensional cone generated by u and v , A_2 the two-dimensional cone generated by u and w and A_3 the one-dimensional cone generated by u . Let $w = a_1 \cdot u + a_2 \cdot v$. Then $A_1 \cap A_2 = A_3$ if and only if $a_2 < 0$.*

Proof. It is obvious that $A_3 \subseteq A_1 \cap A_2$. Let $x = \nu \cdot u + w$ where $\nu \geq \max(-a_1, 0)$. Then $x \in A_2 \setminus A_3$. It follows that $x = (\nu + a_1) \cdot u + a_2 \cdot v$ and thus $x \in A_1 \cap A_2$ if $a_2 \geq 0$. So, $A_1 \cap A_2 = A_3$ implies that $a_2 < 0$. Conversely, suppose that

$$x' = \lambda \cdot u + \lambda' \cdot w \in A_1 \cap A_2.$$

Since $x' \in A_2$ we have that $\lambda \geq 0$ and $\lambda' \geq 0$. Moreover, from

$$x' = \lambda \cdot u + \lambda' \cdot (a_1 \cdot u + a_2 \cdot v) = (\lambda + \lambda' \cdot a_1) \cdot u + \lambda' \cdot a_2 \cdot v$$

and $x' \in A_1$, it follows that $\lambda' \cdot a_2 \geq 0$. Thus if $a_2 < 0$ then $\lambda' = 0$ and $x' \in A_3$. \square

We are now ready to finish the proof of Theorem 6.2.7

Proof. Since $\cup_{i=1}^{(n+1)!} C_i = \mathbb{R}^n$ we have that $\cup_{i=1}^{(n+1)!} C_i^* = H$. Therefore one of the following two cases holds:

1. there exists some two dimensional cone C_i^* such that in the subspace H we have that x is in the interior of cone C_i^* .

2. there exist two dimensional cones C_i^*, C_j^* such that in the subspace H we have that x is on the boundary of C_i^* , x is on the boundary of C_j^* and x is in the interior of $C_i^* \cup C_j^*$.

Case 1. Let a and b be generators of the cone C_i^* . Then $a, b \in C_i$ and $x = \lambda \cdot a + \mu \cdot b$ with $\lambda, \mu > 0$. Hence $x \in C_i$ and $\text{supp}_i(x) = \text{supp}_i(a) \cup \text{supp}_i(b)$. Suppose that $y \notin C_i^*$. Then $y = y_1 \cdot a + y_2 \cdot b$ with $\min(y_1, y_2) < 0$ and we may assume without loss of generality that $y_1 < 0$. According to Lemma 6.2.13 there exists some $k \in \text{supp}_i(a) \setminus \text{supp}_i(b)$. Then $k \notin \text{supp}_i(y)$, $k \in \text{supp}_i(x)$. Thus $\text{supp}_i(x)$ is not a subset of $\text{supp}_i(y)$, but this contradicts Corollary 6.2.11. Hence $y \in C_i^*$ and thus $x, y \in C_i$. According to Corollary 6.2.10 it follows that $x \leq_C y$.

Case 2. There exist $a, b \in H$ such that x and a are generators of cone C_i^* and x and b are generators of cone C_j^* . Let u_1, u_2, v_1, v_2 be such that $y = u_1 \cdot x + u_2 \cdot a = v_1 \cdot x + v_2 \cdot b$. If $u_1 < 0$ then it follows analogously to the proof in case 1 that $\text{supp}_i(x)$ is not a subset of $\text{supp}_i(y)$, which yields a contradiction again. Thus $u_1 \geq 0$ and analogously $v_1 \geq 0$. We will prove that $y \in C_i^* \cup C_j^*$. Suppose that $y \notin C_i^*$. Then $u_2 < 0$. Let a_1, a_2 be such that $a = a_1 \cdot x + a_2 \cdot b$. By Lemma 6.2.14 $a_2 < 0$. Then $y = (u_1 + u_2 \cdot a_1)x + (u_2 \cdot a_2)b$. Hence $v_2 = u_2 \cdot a_2 > 0$. So, if $y \notin C_i^*$ then $y \in C_j^*$. Hence $y \in C_i^* \cup C_j^*$. Since $x \in C_i^* \cap C_j^*$ it follows that $x, y \in C_i$ or $x, y \in C_j$. Hence $x \leq_C y$ by Corollary 6.2.10. \square

Remark. For given multimodular matrix D and $x, y \in M_D$ it is a priori not easy to determine whether $x \leq_{mm} y$. However, according to Theorem 6.2.7 this problem can be solved by checking whether x and y are cone ordered. Indeed it is not hard to check whether $x \leq_C y$ or not. This explains the advantage of introducing additional (partial) orders like the cone order. We introduce and discuss some more (partial) orders for various problems in the remaining of this chapter.

6.2.3 Shift invariant counterparts

We consider sequences (of nonnegative integers) of a given length $T \in \mathbb{N}$ and given sum $S \in \mathbb{N}$. The set of such sequences is denoted by $P(T, S)$. So, $P(T, S) = \{(x_1, x_2, \dots, x_T) : x_i \in \mathbb{Z}_{\geq 0} \text{ for every } i, \sum_{i=1}^T x_i = S\}$ and this is a submesh of dimension $T - 1$ of \mathbb{Z}^T . Let D' be a multimodular matrix of size $T \times T$ induced by the submesh $P(T, S)$.

Let $x, x' \in P(T, S)$. Then we say that x and x' are conjugate if they are cyclic

permutations of each other. If there exist finite (possibly empty) sequences v and w such that $x = vw$ and $x' = wv$ then x and x' are conjugate. It is easily seen that this conjugacy is an equivalence relation on $P(T, S)$. We write $x \sim x'$ if x and x' are conjugate. We denote by \tilde{x} the conjugacy class of $x \in P(T, S)$, which is the set of all cyclic permutations of x . By $\tilde{P}(T, S)$ we denote the set of conjugacy classes of $P(T, S)$. If $x \sim x'$ then we also say that x and x' are shifts of each other. If $\tilde{x} = \tilde{y}$ then we say that x is a representative of \tilde{y} .

Let $f : P(T, S) \rightarrow \mathbb{R}$ be a function such that $f(x) = f(x')$ if $x \sim x'$. Then we say that f is a shift invariant function. A shift invariant function $f : P(T, S) \rightarrow \mathbb{R}$ induces a function $\tilde{f} : \tilde{P}(T, S) \rightarrow \mathbb{R}$ by $\tilde{f}(y) = f(x)$ where $x \in P(T, S)$ is a representative of $y \in \tilde{P}(T, S)$. We denote by $\mathcal{F}^{\text{shift}}(D')$ the set of functions mapping $\tilde{P}(T, S)$ to \mathbb{R} which are induced by D' -multimodular functions that are shift invariant. We use $\mathcal{F}^{\text{shift}}(D')$ to define a partial order $\leq_{\text{mm}s}$ on $\tilde{P}(T, S)$ in the same way as $\mathcal{F}(D, r)$ was used to define the partial order \leq_{mm} on M_D . The partial order $\leq_{\text{mm}s}$ is called the shift invariant multimodular order.

Definition 6.2.15 *Let a multimodular matrix D' induced by the submesh $P(T, S)$ be given. Then for $x, y \in \tilde{P}(T, S)$ we say that $x \leq_{\text{mm}s} y$ if $f(x) \leq f(y)$ for every $f \in \mathcal{F}^{\text{shift}}(D')$.*

Let $x \in P(T, S)$. Then we say that $x = (x_1, x_2, \dots, x_T)$ is regular if the induced infinite sequence $x^\infty := (x_1, \dots, x_T, x_1, \dots, x_T, \dots)$ is a bracket (or regular) sequence and in that case the conjugacy class $\tilde{x} \in \tilde{P}(T, S)$ is also called regular. We denote by $R(T, S) \subseteq P(T, S)$ the subset of regular sequences. The following lemma can be proved analogously to the proof for sequences of zeros and ones that was given in Chapter 3 (see Lemma 3.2.9).

Lemma 6.2.16 *Let $T, S \in \mathbb{N}$ be given. Then there exists exactly one element in $\tilde{P}(T, S)$ which is regular.*

We denote by $\tilde{\omega} = \tilde{\omega}(T, S)$ the unique element of $\tilde{P}(T, S)$ that is regular. Then the following theorem (see [7]) holds.

Theorem 6.2.17 *Let a multimodular matrix D' induced by the submesh $P(T, S)$ be given and let $f \in \mathcal{F}^{\text{shift}}(D')$. Then a global minimum of f is attained at $\tilde{\omega}(T, S)$.*

Note that the global minimum $\tilde{\omega}(T, S)$ does not depend on D' .

Corollary 6.2.18 *For every D' we have that $\tilde{\omega}(T, S)$ is the smallest element for the partial order \leq_{mms} on $\tilde{P}(T, S)$.*

Both the multimodular order \leq_{mm} and the shift invariant multimodular order \leq_{mms} are defined on $\tilde{P}(T, S)$. We also define a shift invariant cone order \leq_{Cs} on $\tilde{P}(T, S)$ which is the counterpart of the cone order \leq_C .

Definition 6.2.19 *Let D' be as in Theorem 6.2.17. Then for $u, v \in \tilde{P}(T, S)$ we say that $u \leq_{Cs} v$ if and only if there exist representatives u' of u and v' of v such that $u' \leq_C v'$, where \leq_C is the cone order for the multimodular matrix D' and some root r which is a representative of $\tilde{\omega}(T, S)$.*

According to Theorem 6.2.7 $u' \leq_C v'$ if and only if $f(u') \leq f(v')$ for every $f \in \mathcal{F}(D', r)$. Thus Theorem 6.2.7 and the definitions imply the following corollary.

Corollary 6.2.20 *Let D' be as in Theorem 6.2.17. Then for $u, v \in \tilde{P}(T, S)$ we have that $u \leq_{mms} v$ if $u \leq_{Cs} v$.*

According to Corollary 6.2.20 the shift invariant cone order implies the shift invariant multimodular order.

6.3 The graph order and the unbalance

In this section we first generalize the definitions of the graph orders and the primal and dual unbalance from Section 3.2 to (periodic) sequences of nonnegative integers. Then we prove that the unbalance (both the primal and the dual) is a shift invariant multimodular function.

Let $u = (u_1, u_2, \dots, u_T) \in P(T, S)$ and for $l = 0, 1, \dots, T - 1$ let $u^{(l)} := (u_{l+1}, u_{l+2}, \dots, u_T, u_1, u_2, \dots, u_l) \in P(T, S)$ be the l -th cyclic permutation of u . For a sequence $u \in P(T, S)$ we define a counting function $\kappa_u : \{0, 1, \dots, T\} \rightarrow \mathbb{Z}$ by $\kappa_u(n) = \sum_{t=1}^n u_t$ and a discrepancy function $\phi_u : \{0, 1, \dots, T\} \rightarrow \mathbb{Q}$ by $\phi_u(n) = \kappa_u(n) - n \cdot \frac{S}{T}$ for $n = 0, 1, \dots, T$.

Lemma 6.3.1 *For $u \in P(T, S)$ and $l \in \{0, 1, \dots, T - 1\}$ we have that $\phi_{u^{(l)}}(n) \geq 0$ for $n = 0, 1, \dots, T$ if and only if $\phi_u(l) = \min_{n=0,1,\dots,T} \phi_u(n)$. Moreover, $\phi_{u^{(l)}}(n) \leq 0$ for $n = 0, 1, \dots, T$ if and only if $\phi_u(l) = \max_{n=0,1,\dots,T} \phi_u(n)$.*

Proof. Suppose that $\phi_{u^{(v)}}(n) < 0$ for some $n \in \{0, 1, \dots, T-1\}$. Then $\phi_u((l+n) \pmod T) = \phi_u(l) + \phi_{u^{(v)}}(n) < \phi_u(l)$. Thus $\phi_u(l) = \min_{n=0,1,\dots,T} \phi_u(n)$ implies that $\phi_{u^{(v)}}(n) \geq 0$ for $n = 0, 1, \dots, T$. Suppose that $\phi_u(l') < \phi_u(l)$ for some $l' \in \{0, 1, \dots, T-1\}$ with $l' \neq l$. Then $\phi_{u^{(v)}}((l'-l) \pmod T) = \phi_u(l') - \phi_u(l) < 0$. Thus $\phi_{u^{(v)}}(n) \geq 0$ for $n = 0, 1, \dots, T$ implies that $\phi_u(l) = \min_{n=0,1,\dots,T} \phi_u(n)$. Analogously it follows that $\phi_{u^{(v)}}(n) \leq 0$ for $n = 0, 1, \dots, T$ if and only if $\phi_u(l) = \max_{n=0,1,\dots,T} \phi_u(n)$. \square

We define a partial order \preceq on $P(T, S)$. This induces partial orders on $\tilde{P}(T, S)$ from which we derive bounds for the expected average waiting time.

Definition 6.3.2 For $u, v \in P(T, S)$ we say that $u \preceq v$ if $\kappa_u(n) \leq \kappa_v(n)$ for $n = 1, 2, \dots, T$.

We have the following (see Chapter 3) where it is proved for sequences of zeros and ones. The following proposition is similar to Lemma 3.2.9 and Lemma 3.2.10 which hold for sequences of zeros and ones.

Proposition 6.3.3 Let $u \in R(T, S)$. Then $u' \in R(T, S)$ if and only if $u \sim u'$. Moreover the partial order \preceq on $P(T, S)$ induces a total order on $R(T, S)$.

Since $R(T, S)$ is finite it follows from Proposition 6.3.3 that $R(T, S)$ contains a greatest element for this order. We denote this greatest element by $\omega(T, S)$, or just ω if no confusion is possible. For this element ω the partial sums $\kappa_\omega(n)$ for $n = 1, 2, \dots, T$ are as great as possible under the restriction that ω is regular. For example $\omega(7, 4)$ is the sequence $(1, 1, 0, 1, 0, 1, 0)$. We can determine $\omega(T, S)$ quickly by using the fact (see Corollary 3.2.12) that

$$\kappa_{\omega(T,S)}(n) = \lceil n \cdot \frac{S}{T} \rceil \text{ for } n = 0, 1, \dots, T. \quad (6.12)$$

We have seen in Proposition 6.3.3 that the regular sequences $R(T, S)$ form a conjugacy class in $P(T, S)$. It follows that $u \in R(T, S)$ if and only if u is a representative of $\tilde{\omega}(T, S) \in \tilde{P}(T, S)$. Combining Lemma 6.3.1, Proposition 6.3.3 and (6.12) we obtain the following theorem in which the partial order \preceq is used to give a characterising property of the conjugacy class $R(T, k) = \tilde{\omega}(T, S)$ of regular sequences in $\tilde{P}(T, S)$.

Theorem 6.3.4 Every conjugacy class \tilde{u} of $P(T, S)$ contains an upper bound of $R(T, S)$, i.e. for every $u \in P(T, S)$ there exists a $v \in P(T, S)$ such that $v \sim u$ and $v \succeq w$ for every $w \in R(T, S)$.

Proof. For $u \in P(T, S)$ let $l \in \{0, 1, \dots, T-1\}$ be such that $\phi_u(l) = \min_{n=0,1,\dots,T} \phi_u(n)$. Let $v = u^{(l)}$. Then $v \sim u$ and by Lemma 6.3.1 we have that $\phi_v(n) \geq 0$ for $n = 0, 1, \dots, T$. Hence $\kappa_v(n) \geq \lceil n \cdot \frac{S}{T} \rceil$ for $n = 0, 1, \dots, T$. So, by (6.12), $v \succeq \omega(T, S)$. By Proposition 6.3.3 $\omega(T, S) \succeq w$ and thus $v \succeq w$ for every $w \in R(T, S)$. \square

The following preorders $\preceq_{\bar{g}}$, $\preceq_{\underline{g}}$ and \preceq_g on $P(T, S)$ are called the primal or upper graph order, the dual or lower graph order, and the strong graph order, respectively.

Definition 6.3.5 *Let $u, v \in P(T, S)$. Then $u \preceq_{\bar{g}} v$ if there exist $u', v' \in P(T, S)$ such that $u' \in \tilde{u}$, $v' \in \tilde{v}$ and $0 \leq \phi_{u'}(n) \leq \phi_{v'}(n)$ for $n = 0, 1, \dots, T$. Furthermore, $u \preceq_{\underline{g}} v$ if there exist $u'', v'' \in P(T, S)$ such that $u'' \in \tilde{u}$, $v'' \in \tilde{v}$ and $0 \geq \phi_{u''}(n) \geq \phi_{v''}(n)$ for $n = 0, 1, \dots, T$. Finally $u \preceq_g v$ if $u \preceq_{\bar{g}} v$ and $u \preceq_{\underline{g}} v$.*

If $u, v \in P(T, S)$ are cyclic permutations of each other then $u \preceq_g v$ and $v \preceq_g u$. Thus these orders are not antisymmetric. Therefore the preorders $\preceq_{\bar{g}}$, $\preceq_{\underline{g}}$ and \preceq_g on $P(T, S)$ are not partial orders. However, they induce partial orders on $\tilde{P}(T, S)$:

Definition 6.3.6 *Let $u, v \in \tilde{P}(T, S)$, let $u' \in P(T, S)$ a representative of u and $v' \in P(T, S)$ a representative of v . Then we say that $u \preceq_g v$ if and only if $u' \preceq_g v'$ and similarly for the orders $\preceq_{\bar{g}}$ and $\preceq_{\underline{g}}$.*

As in Section 3.2 (see Lemma 3.2.24) it follows that these induced graph orders on $\tilde{P}(T, S)$ are partial orders. For a sequence $u \in P(T, S)$ we have a primal unbalance $\bar{I}(u)$ and a dual unbalance $\underline{I}(u)$.

Definition 6.3.7 *Let $u \in P(T, S)$. Then the primal unbalance of u is*

$$\bar{I}(u) := \frac{1}{T} \cdot \sum_{n=1}^T (\kappa_{u'}(n) - \lceil n \cdot \frac{S}{T} \rceil),$$

where $u' \in \tilde{u}$ such that $\phi_{u'}(n) \geq 0$ for $n = 1, 2, \dots, T$. The dual unbalance of u is

$$\underline{I}(u) := \frac{1}{T} \cdot \sum_{n=1}^T (\lfloor n \cdot \frac{S}{T} \rfloor - \kappa_{u'}(n)),$$

where $u' \in \tilde{u}$ such that $\phi_{u'}(n) \leq 0$ for $n = 1, 2, \dots, T$.

Remark. For $u \in P(T, S)$ the primal unbalance is well defined. Namely, by Lemma 6.3.1 there exist $u' \in \tilde{u}$ such that $\phi_{u'}(n) \geq 0$ for $n = 1, 2, \dots, T$. Moreover, if

$u', u'' \in \tilde{u}$ are such that $\phi_{u'}(n) \geq 0$ and $\phi_{u''}(n) \geq 0$ for $n = 1, 2, \dots, T$ then $\sum_{n=1}^T \kappa_{u'}(n) = \sum_{n=1}^T \kappa_{u''}(n)$ (see Theorem 6.3.8). Analogously it follows that the dual unbalance is well defined. It is easily seen that for every sequence $u \in P(T, S)$ both $\bar{I}(u) \geq 0$ and $\underline{I}(u) \geq 0$. Namely, for $u' \in P(T, S)$ we have that $\phi_{u'}(n) \geq 0$ if and only if $\kappa_{u'}(n) \geq \lceil n \cdot \frac{S}{T} \rceil$ and that $\phi_{u'}(n) \leq 0$ if and only if $\kappa_{u'}(n) \leq \lfloor n \cdot \frac{S}{T} \rfloor$. Moreover, a sequence $u \in P(T, S)$ is regular if and only if $\bar{I}(u) = 0$ if and only if $\underline{I}(u) = 0$ (see Chapter 3).

Theorem 6.3.8 *Let $u \in P(T, S)$. Then the following statements are equivalent for $l \in \{0, 1, \dots, T-1\}$.*

- (i) $\bar{I}(u) = \frac{1}{T} \cdot \sum_{n=1}^T (\kappa_{u^{(l)}}(n) - \lceil n \cdot \frac{S}{T} \rceil)$.
- (ii) $\sum_{n=1}^T \kappa_{u^{(l)}}(n) = \max_{i=0,1,\dots,T-1} \sum_{n=1}^T \kappa_{u^{(i)}}(n)$.
- (iii) $\min_{n=0,1,\dots,T-1} \phi_{u^{(l)}}(n) = \phi_{u^{(l)}}(0) = 0$.
- (iv) $\phi_u(l) = \min_{i=0,1,\dots,T-1} \phi_u(i)$.

Proof. By Lemma 6.3.1 (iii) and (iv) are equivalent. We now prove that (ii) implies (iii). Suppose there exists some $t \in \{1, 2, \dots, T-1\}$ such that $\phi_{u^{(t)}}(t) = -\mu$ with $\mu > 0$. Let x be the prefix of length t of $u^{(t)}$ and let y be the suffix of length $T-t$ of $u^{(t)}$. Then $u^{(t)} = xy$. Let $z := yx = u^{((l+t) \pmod T)} \in \tilde{u}$. Then $\phi_{u^{(t)}}(n) = \phi_z(n-t) - \mu$ for $n = t, t+1, \dots, T$ and $\phi_{u^{(t)}}(n) = \phi_z(n+T-t) - \mu$ for $n = 1, 2, \dots, t-1$. Hence $\sum_{n=1}^T \phi_{u^{(t)}}(n) = \sum_{n=1}^T \phi_z(n) - T \cdot \mu$ and thus

$$\sum_{n=1}^T \kappa_{u^{(t)}}(n) = \sum_{n=1}^T \kappa_z(n) - T \cdot \mu < \sum_{n=1}^T \kappa_z(n) \leq \max_{i=0,1,\dots,T-1} \sum_{n=1}^T \kappa_{u^{(i)}}(n),$$

which contradicts (ii). From the definition of the primal unbalance it follows directly that (iii) implies (i). To finish the proof we show that (i) implies (ii). Suppose that $\sum_{n=1}^T \kappa_{u^{(l)}}(n) < \max_{i=0,1,\dots,T-1} \sum_{n=1}^T \kappa_{u^{(i)}}(n)$. Let $l' \in \{0, 1, \dots, T-1\}$ with $l' \neq l$ such that

$\sum_{n=1}^T \kappa_{u^{(l')}}(n) = \max_{i=0,1,\dots,T-1} \sum_{n=1}^T \kappa_{u^{(i)}}(n)$. Since we have proved that (ii) implies (i) it follows that

$$\bar{I}(u) = \frac{1}{T} \cdot \sum_{n=1}^T (\kappa_{u^{(l')}}(n) - \lceil n \cdot \frac{S}{T} \rceil) > \frac{1}{T} \cdot \sum_{n=1}^T (\kappa_{u^{(l)}}(n) - \lceil n \cdot \frac{S}{T} \rceil),$$

which contradicts (i). □

The following result follows immediately from the definitions.

Lemma 6.3.9 *Let $u, v \in P(T, S)$. If $u \preceq_{\bar{g}} v$ then $\bar{I}(u) \leq \bar{I}(v)$. If $u \preceq_g v$ then $\underline{I}(u) \leq \underline{I}(v)$. If $u \preceq_g v$ then $\bar{I}(u) \leq \bar{I}(v)$ and $\underline{I}(u) \leq \underline{I}(v)$.*

The following theorem shows that the primal and dual unbalance are multimodular with respect to the L -triangulation.

Theorem 6.3.10 *The primal unbalance and the dual unbalance are shift invariant and multimodular on $P(T, S)$ with respect to the base d_1, d_2, \dots, d_T where $d_i = e_i - e_{i+1}$ for $i = 1, 2, \dots, T-1$, $d_T = e_T - e_1$ and e_i is the i -th unit vector of length T for $i = 1, 2, \dots, T$.*

Proof. For $u \in P(T, S)$ let $f(u) = \bar{I}(u)$ and $g(u) = \underline{I}(u)$. Thus $f : P(T, S) \rightarrow \mathbb{Q}$ is the primal unbalance function and $g : P(T, S) \rightarrow \mathbb{Q}$ is the dual unbalance function. From the definition it follows that if $u', u'' \in P(T, S)$ are cyclic permutations of each other then $f(u') = f(u'')$ and $g(u') = g(u'')$. Thus the primal unbalance and dual unbalance are shift invariant. To prove the multimodularity of the primal unbalance function f we have to show for every $u \in P(T, S)$ and $i, j \in \{1, 2, \dots, T\}$, $i \neq j$ that

$$f(u + d_i) + f(u + d_j) \geq f(u) + f(u + d_i + d_j). \quad (6.13)$$

We first show for every $x \in P(T, S)$ and $i \in \{1, 2, \dots, T\}$ that

$$f(x + d_i) \leq f(x) + \frac{1}{T}. \quad (6.14)$$

Put $y = x + d_i$. Without loss of generality we can assume that

$$\sum_{n=1}^T \kappa_y(n) = \max_{i=0,1,\dots,T-1} \sum_{n=1}^T \kappa_{y^{(i)}}(n). \text{ Then } f(y) = \frac{1}{T} \cdot \sum_{n=1}^T (\kappa_y(n) - \lceil n \cdot \frac{S}{T} \rceil)$$

and

$$f(x) = f(y - d_i) \geq \frac{1}{T} \cdot \sum_{n=1}^T (\kappa_{y-d_i}(n) - \lceil n \cdot \frac{S}{T} \rceil) \geq \frac{1}{T} \cdot \left(\sum_{n=1}^T (\kappa_y(n) - \lceil n \cdot \frac{S}{T} \rceil) - 1 \right) = f(y) - \frac{1}{T}$$

by Theorem 6.3.8. So, (6.14) holds. Next we show that (6.13) holds. Without loss of generality we assume that $\min_{n=0,1,\dots,T-1} \phi_u(n) = \phi_u(0) = 0$. Then $f(u) = \frac{1}{T} \cdot \sum_{n=1}^T (\kappa_u(n) - \lceil n \cdot \frac{S}{T} \rceil)$ by Theorem 6.3.8. Suppose that $i \neq T$ and $j \neq T$. Then $\kappa_{u+d_i}(n) \geq \kappa_u(n)$, $\kappa_{u+d_j}(n) \geq \kappa_u(n)$ and $\kappa_{u+d_i+d_j}(n) \geq \kappa_u(n)$ for $n = 0, 1, \dots, T$. Hence

$$\min_{n=0,1,\dots,T-1} \phi_{u+d_i}(n) = \min_{n=0,1,\dots,T-1} \phi_{u+d_j}(n) = \min_{n=0,1,\dots,T-1} \phi_{u+d_i+d_j}(n) =$$

$$\min_{n=0,1,\dots,T-1} \phi_u(n) = 0.$$

From the definition of the primal unbalance it follows that $f(u + d_i) = f(u + d_j) = f(u) + \frac{1}{T}$ and $f(u + d_i + d_j) = f(u) + \frac{2}{T}$ and thus $f(u + d_i) + f(u + d_j) = f(u) + f(u + d_i + d_j)$. It remains to show that (6.13) holds in case $i = T$ or $j = T$. We may assume that $j = T$ and $i \neq T$. Then from the preceding argument $f(u + d_i) = f(u) + \frac{1}{T}$. Moreover by (6.14) $f(u + d_i + d_j) \leq f(u + d_j) + \frac{1}{T}$. Hence

$$f(u + d_i) + f(u + d_j) \geq (f(u) + \frac{1}{T}) + (f(u + d_i + d_j) - \frac{1}{T}) = f(u) + f(u + d_i + d_j).$$

Thus the primal unbalance function f is multimodular and it follows analogously that the dual unbalance function g is multimodular. \square

Note. Since the primal and dual unbalance functions are shift invariant on $P(T, S)$, they induce functions on $\tilde{P}(T, S)$ by $\tilde{I}(v) = \bar{I}(u)$ and $\underline{I}(v) = \underline{I}(u)$ if $u \in P(T, S)$ is a representative of $v \in \tilde{P}(T, S)$. By Theorem 6.3.10 these induced primal and dual unbalance functions are elements of $\mathcal{F}^{\text{shift}}(D)$, where D is the multimodular matrix having the multimodular base d_1, d_2, \dots, d_T as row vectors.

Corollary 6.3.11 *For this multimodular base d_1, d_2, \dots, d_T and $u, v \in \tilde{P}(T, S)$ we have that $u \leq_{mms} v$ implies that $\tilde{I}(u) \leq \tilde{I}(v)$ and $\underline{I}(u) \leq \underline{I}(v)$.*

Remark. The upper bound of Corollary 3.3.12 for the expected average waiting time can be generalized to a sequence of multimodular functions satisfying the relations of section 3 of [3]. Indeed, it follows from the results there that the regular or bracket sequence is the minimal admission sequence. If the density is rational then the corresponding regular admission sequence is periodic. In order to obtain upper bounds as in Theorem 3.3.12 and Theorem 3.4.7 we need a generalization of Lemma 3.3.1, especially the relation (3.3). For multimodular functions f_k corresponding to $W_u(j)$ it suffices that

$$f_k(a_1, \dots, a_l + 1, a_{l+1} - 1, \dots, a_k) \leq f_k(a_1, \dots, a_l, a_{l+1}, \dots, a_k) + \delta,$$

for $1 \leq l \leq k - 1$ and some constant δ . Other multimodular models satisfy this relation too, e.g. the models of [6].

6.4 Relations and counterexamples

6.4.1 The shift invariant cone order and the graph order

Consider the following example. Let $u = (1, 3, 2, 2, 3) \in \mathcal{P}(5, 11)$ and $v = (2, 1, 3, 2, 3) \in \mathcal{P}(5, 11)$. It is easily seen that u and v are not ordered for the dual lower graph order (but they are ordered for the upper graph order). So, u and v are not graph ordered. However, we show that u and v are ordered for shift invariant multimodular functions with respect to the standard base d_0, d_1, d_2, d_3, d_4 where $d_i = e_i - e_{i+1}$ for $i = 1, 2, 3, 4$ and $d_0 = (-1, 0, 0, 0, 1)$. Namely, for a shift invariant multimodular function f we have $f(u + d_4) = f((1, 3, 2, 3, 2)) = f(v)$ and $f(u + d_1) = f((2, 2, 2, 2, 3)) = f((2, 2, 2, 3, 2)) = f(u + d_1 + d_4)$ by shift invariance. By multimodularity $f(u + d_1) + f(u + d_4) \geq f(u) + f(u + d_1 + d_4)$. Thus $f(u) \leq f(u + d_4) = f(v)$ for every multimodular shift invariant function f .

Conclusion. The shift invariant multimodular order does not imply the graph order. Thus the shift invariant multimodular order does not imply the shift invariant cone order or the shift invariant cone order does not imply the graph order (or both do not hold).

We now investigate whether $u \leq_{C_s} v$. To do this we take the regular sequence $r = (3, 2, 2, 2, 2)$ as the root of all the cones we consider and we let u', v' run through all the shifts of u and v , respectively. Then we check whether $u' \leq v'$ for some cone. Doing this we find that u and v are indeed ordered for the shift invariant cone order. Namely, for $u' = (3, 1, 3, 2, 2)$ and $v' = (2, 1, 3, 2, 3)$ we have that $u' = r + d_0 + d_1 + d_3 + d_4$ and $v' = r + 2d_0 + d_1 + d_3 + d_4$. Hence, if we consider the cone generated by $b_1 = d_0, b_2 = d_0 + d_1, b_3 = d_0 + d_1 + d_3$ and $b_4 = d_0 + d_1 + d_3 + d_4$ then $u' = r + b_4$ and $v' = r + b_1 + b_4$. By the way, this u' and v' is the only pair of shifts that are cone ordered for some cone with root in r . Since $u' \leq_C v'$ and thus $u \leq_{C_s} v$ we reach the following conclusion.

Conclusion. The shift invariant cone order does not imply the graph order.

Remarks. An explanation for the fact that the ordering of u' and v' in this cone does not imply the lower graph order is the following. For u' you have to start with the second coordinate to get the graph for the lower graph order and for v' you have

to start with the first coordinate. This is possible, because for r you have the start at the second coordinate for the lower graph order, while the generating base vectors of this cone “suggest” the lower graph order for starting at the first coordinate. One of the consequences is that this cone contains other regular sequences than r . For example $r + b_1 = (2, 2, 2, 2, 3)$ is another regular sequence in this cone. It turns out that the cone order in such cones does not imply the graph order.

If you consider the mirrored sequences of u and v then you a similar problem with the upper graph order instead of the lower graph order occurs.

Note that u and v are ordered for the unbalance (despite the fact that they are not ordered for the graph order). Namely, for the primal (upper) unbalance \bar{I} we have that $\bar{I}(u) = 1 \leq 2 = \bar{I}(v)$ and for the dual (lower) unbalance \underline{I} we have that $\underline{I}(u) = 1 = \underline{I}(v)$. Of course this order for the unbalance follows immediately from $\tilde{u} \leq_{mms} \tilde{v}$ and Corollary 6.3.11.

6.4.2 The shift invariant multimodular order and the shift invariant cone order

A counterexample is given here to show that the shift invariant multimodular order does not imply the shift invariant cone order. In fact it is shown that the shift invariant cone order is not even a partial order, because it is not transitive.

Counterexample. Again we consider sequences in $P(5, 11)$ and the same multimodular standard base $\{d_i\}_{i=0,1,\dots,5}$ as before. Let $u = (1, 2, 3, 2, 3)$, $v = (2, 2, 3, 3, 1)$ and $w = (2, 1, 0, 4, 4)$ and as regular sequence again $r = (3, 2, 2, 2, 2)$. The shift invariant cone order satisfies $u \leq_{C_s} v$ and $v \leq_{C_s} w$. However, we do not have that $u \leq_{C_s} w$. Namely, $u' = (3, 2, 3, 1, 2) = r + d_3 = r + b_1$ and $v' = (2, 3, 3, 1, 2) = r + d_2 + 2d_3 + d_4 + d_0 = r + b_1 + b_4$, where $b_1 = d_3$, $b_2 = d_3 + d_0$, $b_3 = d_3 + d_0 + d_2$ and $b_4 = d_3 + d_0 + d_2 + d_4$. Hence $u \leq_{C_s} v$. Further $v'' = (3, 1, 2, 2, 3) = r + d_0 + d_1 = r + b_2$ and $w'' = (2, 1, 0, 4, 4) = r + 4d_0 + 3d_1 + 2d_2 + 2d_4 = r + b_1 + b_2 + 2b_4$, where $b_1 = d_0$, $b_2 = d_0 + d_1$, $b_3 = d_0 + d_1 + d_2$ and $b_4 = d_0 + d_1 + d_2 + d_4$. Hence $v \leq_{C_s} w$. However, there exist no cone with root in r such that some shift of u is smaller in that cone than some shift of w . Hence, u and w are not shift invariant cone ordered, while from $u \leq_{C_s} v$ and $v \leq_{C_s} w$ it follows that u and w are ordered for the shift invariant multimodular order.

Remark. Since the shift invariant cone order is not transitive, it is not a partial order. It is natural to consider the smallest preorder which contains the shift invari-

ant cone order and is transitive. It is easily seen that this order implies the shift invariant multimodular order, but we do not know whether the converse is also true.

6.4.3 The graph order and the shift invariant multimodular order

In this counterexample we show that the graph order does not imply the shift invariant multimodular order. Consider the sequences $u = (0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1) \in P(32, 15)$ and $v = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1) \in P(32, 15)$. Then $u \preceq_g v$. Thus u and v are ordered for the graph order. Moreover, $\bar{I}(u) < \bar{I}(v)$ and $\underline{I}(u) < \underline{I}(v)$. However, it is not true that $f(u) \leq f(v)$ for every shift invariant multimodular function $f : P(32, 15) \rightarrow \mathbb{R}$. Namely, consider u and v as period cycles for the splitting sequences corresponding to the admission of arriving customers to some server. Suppose that the interarrival times are deterministic and equal to 1 and the service times are deterministic and equal to $(1 + \epsilon)$ where ϵ is a small positive number. For $x \in P(32, 15)$ let $W(x)$ be the average waiting time of customers admitted according to x . The average waiting time is multimodular with respect to the standard base of the L - triangulation (see [7]). Since the traffic intensity for the server is smaller than one, the average waiting time function W is also shift invariant (see Chapter 3). Thus W is a shift invariant multimodular function. However, for the given interarrival and service times $W(u) = \frac{11}{15} \cdot \epsilon$ and $W(v) = \frac{10}{15} \cdot \epsilon$. Thus $W(v) < W(u)$ for these service times and u and v are not ordered for the shift invariant multimodular order.

6.4.4 The shift invariant orders: conclusion

Let $u, v \in P(T, S)$ be given. Then u has a performance at least as good as v if $u \leq_{mms} v$, where \leq_{mms} is the shift invariant multimodular order. Therefore, given $u, v \in P(T, S)$, we would like to be able to determine whether $u \leq_{mms} v$. However, this is not easy in general. Therefore we have also investigated the graph order \preceq_g and the shift invariant cone order \leq_{Cs} . These orders have (at least) in common with the shift invariant multimodular order \leq_{mms} that for all these orders the regular sequence is the minimal element.

The advantage of the graph order is that we have given a straightforward algorithm to check whether $u \preceq_g v$ or not. However, we have also shown that the graph order is quite different from the shift invariant multimodular order. Indeed, $u \preceq_g$

v does not imply that $u \leq_{mms} v$. Vice versa, $u \leq_{mms} v$ does not imply that $v \preceq_g u$. So, although the graph order gives some information (in particular bounds on the difference in performance), it can not be used to show that some sequence has a better performance than some other sequence. The advantage of the shift invariant cone order is that it can be used for showing that some sequence has a better performance than another, since $u \leq_{C_s} v$ implies that $u \leq_{mms} v$. Moreover, the problem of determining whether $u \leq_{C_s} v$ is quite tractable, although it takes generally more time than for the graph order. However, the main disadvantage of the shift invariant cone order is that it is more restrictive than the shift invariant multimodular order. Namely, we have shown that the shift invariant cone order is not transitive and thus $u \leq_{mms} v$ does not imply that $u \leq_{C_s} v$. So, this is not as good as in the normal (not shift invariant) case for which we have shown (see Theorem 6.2.7) that the multimodular order and the cone order are equivalent. This problem that the shift invariant cone order is more restrictive than the shift invariant multimodular order could (partially) be resolved by considering the smallest preorder which contains the shift invariant cone order and is transitive. However, for this order it could just as for the shift invariant multimodular order be hard to determine whether two given sequences are ordered. So, in the shift invariant case we have no order that satisfies our demands in every aspect.

Bibliography

- [1] E. Altman (2000). Applications of Markov decision processes in telecommunication: a survey. In *Markov Decision Processes*. Eds. A. Shwartz and E. Feinberg.
- [2] E. Altman, S. Bhulai, B. Gaujal and A. Hordijk (2000). Open-loop routing to M parallel servers with no buffers. *J. Appl. Probab.*, **37**, 668-684.
- [3] E. Altman, B. Gaujal and A. Hordijk (2000). Admission control in stochastic event graphs. *IEEE Trans. Automat. Control*, **45**, 854-867.
- [4] E. Altman, B. Gaujal and A. Hordijk (2000). Balanced sequences and optimal routing. *J. Assoc. Comput. Mach.*, **47**, 752-775.
- [5] E. Altman, B. Gaujal and A. Hordijk (2000). Multimodularity, convexity and optimization properties. *Math. Oper. Res.*, **25**, 324-347.
- [6] E. Altman, B. Gaujal and A. Hordijk (2000). Optimal open-loop control of vacations, polling and service assignment. *Queueing Systems*, **36**, 303-325.
- [7] E. Altman, B. Gaujal and A. Hordijk (2000). Simplex convexity with application to open-loop stochastic control in networks. In *IEEE 39th Conf. on Decision and Control*, 1852-1857.
- [8] E. Altman, B. Gaujal, and A. Hordijk (2002). Regular ordering and applications in control policies. *Discrete Event Dyn. Syst.*, **12**, 187-210.
- [9] E. Altman, B. Gaujal, and A. Hordijk. *Discrete-event control of stochastic networks: Multimodularity and Regularity*. In preparation.
- [10] E. Altman and A. Hordijk (1997). Applications of Borovkov's renovation theory to non-stationary stochastic recursive sequences and their control. *Adv. in Appl. Probab.*, **29**, 388-413.

- [11] Y. Arian and Y. Levy (1992). Algorithms for generalized round robin routing. *Oper. Res. Lett.*, **12**, 313-319.
- [12] P. Arnoux, C. Mauduit, I. Shiokawa and J. Tamura (1994). Complexity of sequences defined by billiards in the cube. *Bull. Soc. Math. France*, **122**, 1-12.
- [13] F. Baccelli and P. Brémaud (1994). *Elements of queueing theory*, Springer-Verlag, Berlin.
- [14] A. Bar-Noy, R. Bhatia, J. Naor and B. Schieber (1998). Minimizing service and operation costs of periodic scheduling. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM, New York, 11-20.
- [15] Y. Baryshnikov (1995). Complexity of trajectories in rectangular billiards. *Comm. Math. Phys.*, **175**, 43-56.
- [16] R. K. Boel and J. H. van Schuppen (1989). Distributed routing for load balancing. *Proceedings of the IEEE*, **77**, 210-221.
- [17] S. Borst and K. G. Kamakrishnan (1999). Optimization of template-driven scheduling mechanisms. *J. Sched.*, **2**, 19-33.
- [18] J. Breiman (1968). *Probability*, Addison-Wesley, London.
- [19] C. S. Chang, X. Chao and M. Pinedo (1990). A note on queues with Bernoulli routing. In *29th Conference on Decision and Control*.
- [20] M. B. Combé and O. J. Boxma (1994). Optimization of static traffic allocation policies. *Theoret. Comput. Sci.*, **125**, 17-43.
- [21] E. G. Coffman, Jr., Z. Liu and R. R. Weber (1998). Optimal robot scheduling for web search engines. *J. Sched.*, **1**, 15-29.
- [22] W. Fischer and K. S. Meier-Hellstern (1992). The Markov-modulated Poisson process (MMPP) cook book. *Performance Evaluation*, **18**, 149-171.
- [23] B. Gaujal and E. Hyon (2000). Optimal routing policy in two deterministic queues. *Calculateurs Parallèles*, **13**, 601-634.
- [24] B. Gaujal and E. Hyon (2002). Optimal routing in deterministic queues in tandem. Technical Report INRIA RR-4393.
- [25] B. Gaujal, A. Hordijk and D. A. van der Laan (2001). On orders and bounds for multimodular functions. Technical Report MI 2001-29, Leiden University.

- [26] R. Graham (1973). Covering the positive integers by disjoint sets of the form $\{[n\alpha + \beta] : n = 1, 2, \dots\}$, *J. Combin. Theory Ser. A*, **15**, 354-358.
- [27] D. Gross and C. M. Harris (1974). *Fundamentals of queueing theory*, Wiley Series in Probability and Mathematical Statistics, New York.
- [28] B. Hajek (1983). The proof of a folk theorem on queueing delay with applications to routing in networks. *J. Assoc. Comput. Mach.*, **30**, 834-851.
- [29] B. Hajek (1985). Extremal splittings of point processes. *Math. Oper. Res.*, **10**, 543-556.
- [30] G. H. Hardy and E. M. Wright (1960). *An introduction to the theory of numbers*, Fourth edition, The Clarendon Press, Oxford.
- [31] R. Hariharan, V. G. Kulkarni and S. Stidham (1990). A survey of research relevant to virtual circuit routing in telecommunication networks. Technical Report WC/OC/TR 90-13, University of North Carolina at Chapel Hill.
- [32] J. M. Harrison (1985). *Brownian motion and stochastic flow systems*, Wiley, New York.
- [33] A. Heinis (2001). *Arithmetics and combinatorics of words of low complexity*, Ph.D.-thesis, Leiden University.
- [34] A. Hordijk (2001). Comparison of queues with different discrete-time arrival processes. *Probab. Engrg. Inform. Sci.*, **15**, 1-14.
- [35] A. Hordijk and G. Koole (1992). On the assignment of customers to parallel queues. *Probab. Engrg. Inform. Sci.*, **6**, 495-511.
- [36] A. Hordijk, G. M. Koole and J. A. Loeve (1994). Analysis of a customer assignment model with no state information. *Probab. Engrg. Inform. Sci.*, **8**, 419-429.
- [37] A. Hordijk and D.A. van der Laan (2000). Periodic routing to parallel queues with bounds on the average waiting time. Technical Report MI 2000-44, Leiden University.
- [38] A. Hordijk and D. A. van der Laan (2001). Bounds for deterministic periodic routing sequences. *Integer programming and combinatorial optimization*, Editors K. Aardal and B. Gerards, Proceedings of the 8th international IPCO conference, Springer-Verlag, Utrecht, 236-250.

- [39] A. Hordijk and D. A. van der Laan (2002). Note on the convexity of the stationary waiting time as function of the density. To appear in *Probab. Engrg. Inform. Sci.*
- [40] A. Hordijk and D. A. van der Laan (2002). On the average waiting time for regular routing to deterministic queues. Technical Report MI 2002-24, Leiden University.
- [41] A. Itai and Z. Rosberg (1984). A golden ratio control policy for a multiple-access channel. *IEEE Trans. Automat. Control*, **29**, 712-718.
- [42] H. Kameda and Y. Zhang (1995). Uniqueness of the solution for optimal static routing in open BCMP queueing networks. *Math. Comput. Modelling*, **22**, 119-130.
- [43] L. Kleinrock (1976). *Queueing systems. Volume II: Computer applications*, Wiley, New York.
- [44] G. M. Koole (1999). On the static assignment to parallel servers. *IEEE Trans. Automat. Control*, **44**, 1588-1592.
- [45] G. Koole, P. Sparaggis and D. Towsley (1999). Minimizing response times and queue lengths in systems of parallel queues, *J. Appl. Probab.*, **36**, 1185-1193.
- [46] D. A. van der Laan (2000). Routing jobs to servers with deterministic service times. Report MI no. 2000-20, Leiden University. Submitted to *Math. Oper. Res.* Available on www.math.leidenuniv.nl/reports/2000-20.shtml.
- [47] Z. Liu and R. Richter (1998). Optimal load balancing on distributed homogeneous unreliable processors. *Oper. Res.*, **46**, 563-573.
- [48] M. Lothaire (1983). *Combinatorics on Words*, Cambridge University Press, Cambridge, Reissued in 1997.
- [49] M. Lothaire (2002). *Algebraic Combinatorics on Words*, Cambridge University Press, Cambridge.
- [50] P. A. P. Moran (1968). *An Introduction to Probability Theory*, The Clarendon Press, Oxford.
- [51] R. Morikawa. On eventually covering families generated by the bracket function $i-v$, *Bull. Fac. Liberal Arts Nagasaki Univ.* **23** no. 1 (1982), 17-22, **24** no. 1 (1983), 1-9, **25** no. 1 (1984), 1-11, **25** no. 2 (1985), 1-8, **26** no. 1 (1985), 15-17, **36** no. 1 (1995), 1-17.

- [52] R. Morikawa. Disjoint sequences generated by the bracket function i-vi, Number theory and combinatorics. Japan, 1984. World Sci. Publishing, Singapore (1985), 305-321 and *Bull. Fac. Liberal Arts Nagasaki Univ.* **26** no. 1 (1985), 1-13, **28** no. 2 (1988), 1-24, **30** no. 1 (1989), 1-10, **32** no. 2 (1992), 181-185, **34** no. 1 (1993), 1-23.
- [53] M. Morse and G.A. Hedlund (1938). Symbolic dynamics, *Amer. J. Math.*, **60**, 815-866.
- [54] M. Morse and G.A. Hedlund (1940). Symbolic dynamics II. *Amer. J. Math.*, **62**, 1-42.
- [55] O. Perron (1954). *Die Lehre von den Kettenbrüchen*, Third edition, Stuttgart.
- [56] M.I. Puterman, (1994). *Markov Decision Processes*, Wiley, New York.
- [57] Z. Rosberg and D. Towsley (1985). Customer routing to parallel servers with different rates. *IEEE Trans. Automat. Control*, **30**, 1140-1143.
- [58] S. M. Ross (1970). *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco.
- [59] S. M. Ross (1983). *Introduction to Stochastic Dynamic Programming*, Academic Press, New York.
- [60] S. Sano and N. Miyoshi (2000). Applications of m -balanced sequences to some network scheduling problems. *Discrete Event system: Analysis and Control, Proceedings of the 5th Workshop on Discrete Event Systems (WODES 2000)*, Kluwer Academic Publisher, 317-325.
- [61] L. I. Sennott (1999). *Stochastic Dynamic Programming and the Control of Queueing Systems*, Wiley, New York.
- [62] H. Shirakawa, M. Mori, and M. Kijima (1989). Evaluation of Regular Splitting Queues. *Comm. Statist. Stochastic Models*, **5**, 219-234.
- [63] J. Simpson (1991). Disjoint covering systems of rational Beatty sequences. *Discrete Math.*, **92**, 361-369.
- [64] Y. G. Sinai (1976). *Introduction to Ergodic Theory*, Princeton University Press, Princeton, N.J.
- [65] S. Stidham, Jr. (1974). A last Word on $L = \lambda W$, *Oper. Res.*, **22**, 417-421.

- [66] R. Tijdeman (1996). On complementary triples of Sturmian bisequences. *Indag. Math. (N.S)* **7**, 419-424.
- [67] R. Tijdeman (2000). Exact covers of balanced sequences and Fraenkel's conjecture. *Algebraic number theory and diophantine analysis*, Eds. F. Halter-Koch and R.F. Tichy, Walter de Gruyter, 467-483.
- [68] R. Tijdeman (2000). Fraenkel's conjecture for six sequences. *Discrete Math.*, **222**, 223-234.
- [69] T. P. Yum (1981). The design and analysis of a semidynamic deterministic routing rule. *IEEE Trans. Comm.*, **29**, 498-504.

Index

- action space, 158
- admission sequence, 122
 - deterministic, 171
 - stochastic, 171
- allocation pattern, 18
- alphabet, 57
- atom, 173
- average cost optimization, 157
- average waiting time, 10

- balanceable, 13, 57
- balanced, 56
- balanced sequence, 19
- Beatty sequence, 55
- Bernoulli routing, 10
- best lower approximation, 133
- billiard sequence, 40, 75, 107
 - consistent, 107
- bracket sequence, 12
 - lower, 123
 - upper, 123

- closed-loop control, 9
- cone, 175
 - convex, 175
- cone order, 171
 - shift invariant, 182
- conjugate, 77
- continued fraction expansion, 137
- convergent, 138
 - even, 138
 - intermediate, 138
 - odd, 138
- cost, 158
 - discounted, 159
 - one-step, 158
- counting function, 80, 182

- cyclic permutation, 182

- decision epoch, 9, 157
- density, 14
 - lower, 125
 - upper, 125
- discrepancy function, 80, 182

- factorisation, 126
- Farey interval, 126
- FIFO, 9, 122

- General Greedy (GG), 31
- generalized round robin routing, 10
- generator, 175
- GG algorithm, 111
- GR algorithm, 66
- graph order, 87
 - lower, 87, 184
 - strong, 87, 184
 - upper, 87, 184

- integer linear program (ILP), 14

- letter, 57
- linear programming (LP), 14
- Little's law, 28, 125

- Markov Decision Chain (MDC), 15, 157
- mathematical prog. problem (MPP), 14
- mirror, 85, 101
- multimodular, 172
 - function, 171
 - matrix, 172
 - triangulation, 173
- multimodular order, 171

- shift invariant, 181
- open-loop control, 9
- OSSM algorithm, 64
- parallel queueing system, 9
 - deterministic, 119
- partial order, 80
- partial quotient, 137
- period word, 126
- periodic, 78
 - ultimately, 166
- policy
 - myopic, 162
 - periodic, 19
 - stationary
 - deterministic, 161
- proportional periodic (p.p.), 43
- recurrent
 - uniformly, 169
- regular, 78
- regular sequence, 12, 55
- regular set, 55
- routing, 10
 - deterministic, 10
 - semi-dynamic, 18
 - probabilistic, 10
- routing policy, 9
- routing sequence, 10, 122
 - periodic, 11
- SG algorithm, 107
- shift, 167
 - cyclic, 54
- shift invariant, 172
- shift invariant function, 181
- Shorter Faster Queue (SFQ), 70
- sojourn time, 20
- Special Greedy (SG), 39
- splitting sequence, 55, 122
- stability interval, 130
- state space, 158
- stationary ergodic sequence, 116
- subword, 57
- SWT algorithm, 66
- total unbalance, 101
 - dual, 100
 - primal, 100
- traffic intensity, 95, 128
- transition rules, 158
- unbalance, 75, 172, 182
 - dual, 83, 184
 - primal, 82, 184
- waiting time, 20
- word, 57

Index of notation.

(S_1, S_2, \dots, S_N) , 122	$\mathbb{R}_{>t}$, 19
$A(i)$, 55	$\mathbb{Z}_{>t}$, 19
$B(t)$, 124	χ_u , 80
$B_i(t)$, 124	δ , 94
$C(s, a)$, 159	δ_i , 90
$F(\bar{S})$, 147	κ_u , 80, 182
$G_i(t)$, 124	λ , 125
I_t , 22	$\lceil \rceil$, 19
$J(s)$, 160	\leq_C , 175
$J_\theta(s)$, 160	\leq_{Cs} , 182
$L(U)$, 124	\leq_{mms} , 181
$L(d)$, 129	\leq_{mm} , 176
$L^i(U)$, 125	$\lfloor \rfloor$, 19
$M(\bar{S})$, 142	$\mathbb{E}X$, 77
M_D , 172	\mathcal{C} , 175
N_i^t , 22	$\mathcal{F}(D, r)$, 172
$P(T, S)$, 180	\mathcal{G} , 178
$P_{st}(a)$, 158	$\mathcal{P}(T, k)$, 80
Q_i^t , 22	$\mathcal{P}(d)$, 88
$R(T, S)$, 181	$\mathcal{R}(T, k)$, 80
$R(\bar{S})$, 142	$\mathcal{R}(d)$, 88
S , 23	$\omega(T, S)$, 183
S^t , 22	$\omega(d)$, 126
S_a , 57	$\bar{I}(u)$, 82, 184
S_i , 122	$\bar{W}(\psi)$, 99
T_i , 90	$\bar{W}(u)$, 94
V , 20	ϕ_u , 182
$V_i(t)$, 124	\preceq , 80, 183
$V_\alpha(s)$, 159	\preceq_g , 184
$V_{\theta, \alpha}(s)$, 159	$\preceq_{\bar{g}}$, 184
W , 20	$\preceq_{\underline{g}}$, 184
$W(U)$, 125	ψ^s , 35
$W(d)$, 129	ρ , 95
$W^i(U)$, 125	σ_j^i , 99
W_n , 10	σ_j , 90
$\mathbb{Q}_{>t}$, 19	$\underline{I}(u)$, 83, 184

$\underline{\delta}_i$, 55
 $\widetilde{P}(T, S)$, 181
 \widetilde{R} , 102
 \widetilde{S} , 23
 \widetilde{V} , 22
 \widetilde{W} , 22
 $a \pmod{b}$, 77
 a_i , 20
 $d_i(\psi)$, 30
 h_i^t , 36
 $k_s(\psi)$, 20
 $l_S(d)$, 143
 $m_i(t)$, 124
 p_i , 27
 $r_S(d)$, 143
 r_i , 135
 s_i , 43
 u^i , 99
 u_i^s , 20
 v_i^s , 20
 $w(d)$, 126
 w_i^t , 44
 z_i , 58
 $\mathcal{Q}(d_1, d_2, \dots, d_N)$, 99
 $\mathcal{S}(d_1, d_2, \dots, d_N)$, 99
 $\overline{O}(U)$, 100
 $\overline{\omega}(T, k)$, 81
 \trianglelefteq_g , 87, 100
 $\trianglelefteq_{\overline{g}}$, 87, 100
 $\underline{\trianglelefteq}_g$, 87, 100
 $\underline{Q}(U)$, 100

Samenvatting.

In dit proefschrift bestuderen we wachtrijsystemen met parallelle wachtrijen, waarbij iedere wachtrij zijn eigen karakteristieke bediende (server) heeft. Klanten (opdrachten) die in het systeem arriveren worden naar één van de parallelle wachtrijen gestuurd, waarin de klant moet wachten totdat de bediende behorende bij deze wachtrij alle voorgaande klanten bediend heeft.

Zo een systeem met parallelle wachtrijen lijkt op de situatie in een supermarkt met wachtrijen voor de kassa's, maar een belangrijk verschil is dat in een supermarkt een klant zelf de (wacht)rij kiest in welke hij gaat staan. Hier gaan we er echter vanuit dat het sturen van de klanten centraal wordt geregeld. Hierbij zou gedacht kunnen worden aan het sturen van opdrachten in een computersysteem met verschillende servers of aan het versturen van pakketten van informatie in een telecommunicatienetwerk.

Het sturen van opdrachten in een systeem met parallelle wachtrijen wordt ook wel routeren genoemd. Een zogenaamde routeringsstrategie geeft ruwweg gesproken voor elke arriverende klant een voorschrift dat bepaalt naar welke server de klant gezonden zal worden. Gewoonlijk willen we een routeringsstrategie vinden zodanig dat de (verwachte) gemiddelde wachttijd over alle klanten minimaal is. Hierbij is de wachttijd van een klant de tijd tussen het arriveren van een klant en het moment dat hij daadwerkelijk bediend gaat worden.

In dit proefschrift concentreren we ons op routeren volgens een statische routeringsstrategie. Dit betekent dat de beslissing om een klant naar een bepaalde wachtrij te sturen onafhankelijk is van tijdsafhankelijke informatie in het systeem, zoals bijvoorbeeld het aantal wachtenden in elke wachtrij op het moment dat de klant arriveert of de wachttijden in elke wachtrij. Een statische routeringsstrategie hangt namelijk alleen maar af van de basisgegevens van het systeem. Dat wil zeggen het aankomstproces en gegevens over de bedieningsduren van alle bedienden. In tegenstelling tot statische routeringsstrategieën zijn er ook dynamische routeringsstrategieën, waarbij wel tijdsafhankelijke informatie wordt gebruikt. Een goede dynamische routeringsstrategie geeft in het algemeen een beter resultaat dan een statische, maar een praktisch bezwaar van een dynamische strategie is dat de beslissingen tijdens het proces zelf gemaakt moeten worden. Daarentegen kunnen de beslissingen bij een statische strategie van tevoren vastgelegd worden, waardoor een statische strategie

veel makkelijker te implementeren is.

In dit proefschrift beschouwen we dus statische routeringsstrategieën en in het bijzonder deterministische statische routeringsstrategieën. Zo'n strategie kan worden gekarakteriseerd door een oneindige routeringsrij $U = (U_1, U_2, \dots)$. Hierbij geeft U_n de wachtrij weer waar de n -de arriverende klant naar toe wordt gestuurd volgens deze strategie. Indien we aannemen dat er $N \geq 2$ parallelle wachtrijen zijn en we indiceren ze als 1 tot en met N , dan kan zo een routeringsrij $U = (U_1, U_2, \dots)$ gezien worden als een oneindig lang woord op het alfabet $\{1, 2, \dots, N\}$. We zijn met name geïnteresseerd in de structuur van optimale routeringsrijen. Dus de structuur van de woorden (patronen) behorend bij optimale deterministische statische routeringsstrategieën, waarbij een strategie optimaal wordt genoemd als er geen andere is die een beter resultaat geeft. Hier komt dat er meestal op neer dat er geen strategie is die een lagere verwachte gemiddelde wachttijd over alle arriverende klanten geeft.

Eén van de eerste vragen die we ons stellen is of er optimale strategieën met bijbehorende optimale routeringsrijen bestaan. Indien dit bevestigend beantwoord kan worden dan kunnen er ook andere problemen aangaande de structuur onderzocht worden. Bijvoorbeeld of er een optimale strategie bestaat waarvoor de bijbehorende optimale routeringsrij $U = (U_1, U_2, \dots)$ periodiek is. Dat wil zeggen dat er een natuurlijk getal T (de periode) bestaat zodanig dat $U_n = U_{n+T}$ voor $n = 1, 2, \dots$

We analyseren ook de routeringsrijen en bijbehorende verwachte gemiddelde wachttijden voor een enkele wachtrij met bijbehorende server. De routeringsrij behorend bij een enkele wachtrij wordt gegeven door een oneindige rij $u = (u_1, u_2, \dots)$ van nullen en enen. Hierbij geldt $u_n = 1$ als de n -de arriverende klant naar deze wachtrij wordt gestuurd en $u_n = 0$ indien hij niet naar deze wachtrij wordt gestuurd. Een dergelijke rij van enen en nullen wordt ook wel een toelatingsrij genoemd. Het is bekend dat als een zekere fractie d van alle arriverende klanten naar een bepaalde wachtrij gestuurd wordt, dat dan de verwachte gemiddelde wachttijd in die wachtrij minimaal is als de bijbehorende toelatingsrij $u = (u_1, u_2, \dots)$ een regelmatige rij met dichtheid d is. Een rij $u = (u_1, u_2, \dots)$ van nullen en enen is regelmatig met dichtheid d als voor alle natuurlijke getallen k, n elke deelrij $(u_k, u_{k+1}, \dots, u_{k+n-1})$ van lengte n precies $\lfloor nd \rfloor$ of $\lceil nd \rceil$ enen bevat. Hierbij staat $\lfloor \cdot \rfloor$ voor het naar beneden afronden tot een geheel getal en $\lceil \cdot \rceil$ voor het naar boven afronden tot een geheel getal. Dat wil dus zeggen dat de rij u regelmatig met dichtheid d is dan en slechts dan als $-1 < nd - \sum_{i=k}^{k+n-1} u_i < 1$ voor alle natuurlijke getallen k, n .

Voor bepaalde systemen zullen we de verwachte gemiddelde wachttijd in een enkele wachtrij exact berekenen in geval dat de bijbehorende toelatingsrij een regelmatige rij met dichtheid d is. Met behulp daarvan kan een ondergrens voor de totale gemiddelde wachttijd over alle arriverende klanten afgeleid worden. Bovendien leiden we voor willekeurige periodieke toelatingsrijen een bovengrens af voor het verschil in verwachte gemiddelde wachttijd met een regelmatige toelatingsrij met dezelfde dichtheid. Hierdoor kan ook een bovengrens voor de verwachte totale gemiddelde wachttijd voor een bepaalde routeringsstrategie gegeven worden.

Hier volgt een korte beschrijving van de inhoud van de hoofdstukken van dit proefschrift.

Hoofdstuk 1 is een inleiding. In hoofdstuk 2 analyseren we wachtrijsystemen met constant tijdsverschil tussen opeenvolgende arriverende klanten en ook constante bedieningsduren voor elke wachtrij. Door het schalen van de tijd kunnen we dan aannemen dat het tijdsverschil tussen opeenvolgende arriverende klanten altijd precies één tijdseenheid is. Bovendien nemen we aan dat de zogenaamde verkeersintensiteit gelijk aan 1 is, wat er op neer komt dat het systeem maximaal belast is. We laten voor zulke systemen een direct verband zien tussen de gemiddelde wachttijd van alle arriverende klanten en de zogenaamde totale ongebruikte capaciteit van alle servers. We leiden hieruit af dat de minimale gemiddelde wachttijd kleiner dan of gelijk aan $\frac{N-1}{2}$ is, waarbij $N \geq 2$ het aantal wachtrijen is. Aangaande de structuur van optimale routeringsrijen tonen we voor zulke systemen aan dat er een optimale routeringsrij bestaat, die tevens een biljartrij is. Dat wil zeggen dat de rij geconstrueerd kan worden volgens een zogenaamd biljartalgoritme. In geval dat de bedieningsduren rationaal zijn volgt hieruit het bestaan van een periodieke optimale routeringsrij. Tenslotte vergelijken we de prestaties van enkele routeringsalgoritmes.

In hoofdstuk 3 nemen we niet zoals in Hoofdstuk 2 aan dat er een constant tijdsverschil is tussen opeenvolgende arriverende klanten. We nemen wel aan dat deze tijdsverschillen onafhankelijk van elkaar zijn en bovendien identiek verdeeld volgens een bepaalde kansverdeling. Hetzelfde nemen we aan voor de bedieningsduren voor elk van de servers. Allereerst analyseren we de toelatingsrijen van enen en nullen en bijbehorende verwachte gemiddelde wachttijden voor een enkele wachtrij met bijbehorende server. Voor periodieke rijen van enen en nullen definiëren we het combinatorische begrip van onbalans. Een regelmatige periodieke rij heeft bijvoorbeeld een onbalans gelijk aan nul. Vervolgens leiden we een bovengrens af voor het verschil in verwachte gemiddelde wachttijd tussen routing volgens een bepaalde periodieke toelatingsrij en een regelmatige rij met dezelfde dichtheid. Deze bovengrens is evenredig met de onbalans van de rij. Vervolgens breiden we de definitie van onbalans uit tot periodieke routeringsrijen op een alfabet van $N \geq 2$ letters. Dit noemen

we de totale onbalans en we verkrijgen een bovengrens voor de verwachte totale gemiddelde wachttijd van een routeringsstrategie $U = (U_1, U_2, \dots)$ die afhangt van de totale onbalans van U . Tenslotte tonen we aan dat er voor gegeven dichtheden altijd een biljartrij is met minimale totale onbalans.

In hoofdstuk 4 nemen we net als in hoofdstuk 2 aan dat er een constant tijdsverschil is tussen opeenvolgende arriverende klanten en dat de bedieningsduren voor elke wachtrij constant zijn. De tijdseenheid wordt weer zo gekozen dat het tijdsverschil tussen opeenvolgende arriverende klanten gelijk aan één is. In tegenstelling tot hoofdstuk 2 beschouwen we nu echter ook systemen met verkeersintensiteit kleiner dan 1. Voor een enkele wachtrij geven we een efficiënt algoritme om in geval van een regelmatige toelatingsrij met een bepaalde dichtheid de gemiddelde wachttijd te berekenen. Voor het routeren naar N parallelle wachtrijen kan vervolgens gezocht worden naar routeringsdichtheden d_1, d_2, \dots, d_N met $\sum_{i=1}^N d_i = 1$, met een bijbehorende (best mogelijke) ondergrens voor de minimale totale gemiddelde wachttijd. We leiden af dat als de verkeersintensiteit kleiner dan 1 is, er altijd rationale routeringsdichtheden zijn die de beste ondergrens geven. In geval van maar twee parallelle wachtrijen is het bekend dat er een optimale routeringsrij bestaat zodanig dat beide bijbehorende toelatingsrijen regelmatig zijn met dichtheden d en $1 - d$, respectievelijk, waarbij $0 \leq d \leq 1$. Uit het bovenstaande volgt dat er een rationale dichtheid d is die voldoet. De bijbehorende optimale routeringsrij is dan periodiek met een totale gemiddelde wachttijd gelijk aan de eerder berekende ondergrens.

In hoofdstuk 5 beschouwen we dezelfde systemen als in hoofdstuk 4 en we nemen nu altijd aan dat de verkeersintensiteit kleiner dan 1 is. We tonen aan dat het probleem om een optimale routeringsstrategie te vinden tot een Markov beslissingsprobleem herleid kan worden met als functie het minimaliseren van de gemiddelde kosten. We tonen aan dat er altijd een optimale deterministische stationaire strategie is voor deze problemen. Hieruit leiden we af dat er een optimale periodieke routeringsrij bestaat in het geval dat alle bedieningsduren rationaal zijn.

In hoofdstuk 6 vergelijken we de prestatie van (periodieke) toelatings- en routeringsrijen met dezelfde dichtheid of dichtheden. We proberen voor een algemene klasse van parallelle wachtrijsystemen aan te tonen dat een bepaalde rij een lagere totale gemiddelde wachttijd heeft dan een andere rij, waarbij we puur combinatorische eigenschappen van de rijen gebruiken. Hiertoe gebruiken we enkele (partiële) ordeningen op zulke rijen. Van belang is dat de totale gemiddelde wachttijd een zogenaamde multimodulaire functie van de routeringsrij is. We breiden de definitie van onbalans uit tot periodieke rijen van niet-negatieve gehele getallen en tonen aan dat dit een verschuivingsinvariante multimodulaire functie geeft.

CURRICULUM VITAE

Ik ben geboren op 8 september 1976 te Alphen aan den Rijn. Aan het Christelijk Lyceum te Alphen aan den Rijn behaalde ik in 1994 mijn VWO diploma. Vervolgens ben ik aan de Rijksuniversiteit Leiden begonnen met een dubbele propedeuse wiskunde en sterrenkunde. Nadat ik in 1995 beide propedeuses behaald had, ben ik verder gegaan met de wiskundestudie. In januari 1999 studeerde ik cum laude af nadat ik mijn scriptie getiteld “Assigning jobs to servers with deterministic service times” onder begeleiding van prof. dr. R. Tijdeman geschreven had. Van september 1996 tot en met december 1998 ben ik ook studentassistent aan de Rijksuniversiteit Leiden geweest. In deze periode heb ik diverse werkgroepen zoals Lineaire Algebra, Algebra en Discrete Wiskunde begeleid. Van 1 februari 1999 tot 1 februari 2003 was ik werkzaam als assistent in opleiding aan de Universiteit Leiden. Mijn onderzoek werd begeleid door prof. dr. A. Hordijk en prof. dr. R. Tijdeman. De resultaten hiervan staan in dit proefschrift beschreven. Naast het onderzoek heb ik als assistent in opleiding diverse werkgroepen begeleid. Bovendien heb ik van september 1999 tot juni 2001 de cursussen van het Landelijk Netwerk voor Mathematische Besliskunde gevolgd.

In mijn vrije tijd houd ik van het spelen van (bord)spellen en dan met name het schaakspel. Vanaf 1987 ben ik lid van de Alphense schaakclub. Met het eerste team van deze schaakclub speel ik op dit moment in de derde klasse van de landelijke schaakcompetitie.

