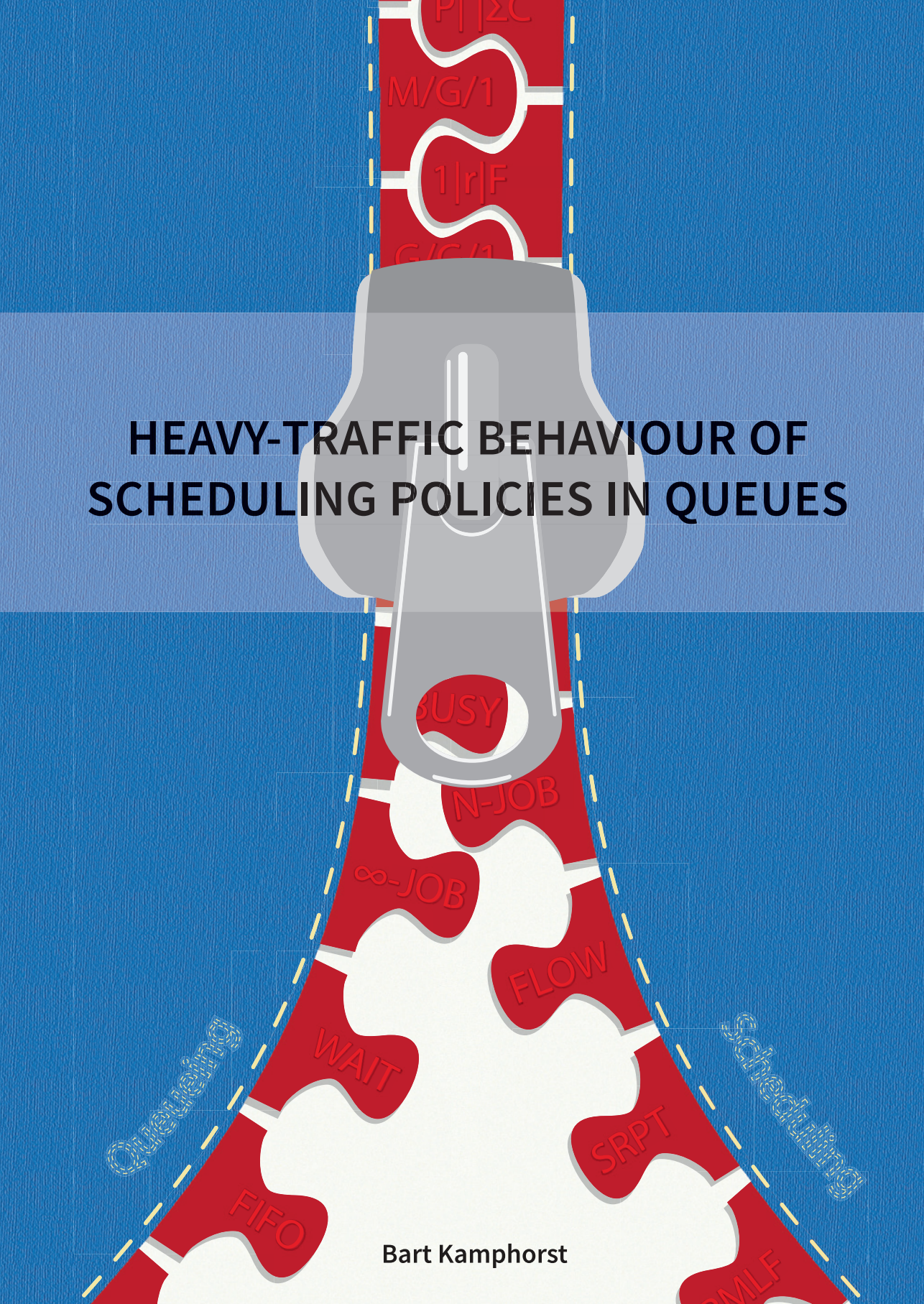
Heavy-traffic behaviour of scheduling policies in queues

Kamphorst, Bart

Accepted/In press: 31/05/2018

Document Version
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)



HEAVY-TRAFFIC BEHAVIOUR OF SCHEDULING POLICIES IN QUEUES

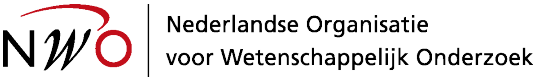
Bart Kamphorst

HEAVY-TRAFFIC BEHAVIOUR OF SCHEDULING POLICIES IN QUEUES

Bart Kamphorst

Dit werk maakt deel uit van het vrije competitie onderzoeksprogramma met projectnummer 613-001-219, welk gefinancierd is door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

Dit werk is uitgevoerd aan het Centrum Wiskunde & Informatica (CWI) in Amsterdam. Het CWI is het nationaal onderzoeksinstituut voor wiskunde en informatica en is onderdeel van de institutenorganisatie van NWO, NWO-I.



Heavy-traffic behaviour of scheduling policies in queues

Mathematical Subject Classification 2010: 60K25, 68M20, 90B22, 90B36

Gedrukte exemplaren van dit proefschrift zijn opgenomen in de collecties van de bibliotheken van het CWI en van de Technische Universiteit Eindhoven.
ISBN: 978-90-386-4515-5

Kaftontwerp door Remco Wetzels, www.remcowetzels.nl
Gedrukt door Ipskamp Printing, Enschede

HEAVY-TRAFFIC BEHAVIOUR OF SCHEDULING POLICIES IN QUEUES

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof. dr. ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op donderdag 31 mei 2018 om 16:00 uur

door

Bart Kamphorst

geboren te Westvoorne

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof. dr. ir. B. Koren
1e promotor:	prof. dr. A.P. Zwart
2e promotor:	prof. dr. N. Bansal
leden:	prof. dr. ir. S.C. Borst
	prof. dr. ir. O.J. Boxma
	prof. dr. Z. Palmowski (Wroclaw University of Science and Technology)
	prof. dr. L. Stougie (Vrije Universiteit Amsterdam)
	dr. T. Vredeveld (Maastricht University)

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Dankwoord – Acknowledgements

Bekende volkswijsheden stellen dat gedeelde smart halve smart is, terwijl geluk zich juist vermenigvuldigt wanneer men dit deelt. Dergelijke stellingen zijn volledig misplaatst in een wiskundig proefschrift, ware het niet dat dit dankwoord enige artistieke vrijheid geniet. Vandaar dat ik, met de bovenstaande filosofieën in gedachten, van deze gelegenheid gebruik wil maken om velen te bedanken voor hun aandeel in mijn promotietraject.

Het doen van onderzoek kent periodes van stagnatie en periodes van vooruitgang. De transitie naar vooruitgang werd doorgaans ingezet door een bespreking met een deelverzameling van mijn promotoren Bert Zwart en Nikhil Bansal. Bert, jouw ervaring, kennis en intuïtie stellen je als geen ander in staat om inzicht te krijgen in een complex wiskundig vraagstuk en richting te geven aan de bijbehorende analyse. Daarnaast heb je meerdere keren laten zien hoe waardevol een wetenschappelijk netwerk is, al is het maar om dat ene resultaat in een Siberisch tijdschrift te vinden. Het was een eer om te leren van een excellente wetenschapper zoals jijzelf. Nikhil, your vision on the interaction between queueing theory and scheduling theory is ahead of its time. This vision manifested itself into the problem that you posed at the very beginning of my PhD track and let to the most innovative chapter in this dissertation. I am grateful for the opportunity to have worked with a highly-esteemed researcher as yourself.

The trustworthiness and quality of a dissertation should always be validated by a panel of independent experts. It is therefore that I am indebted to all members of my PhD committee. In particular, I would like to express my gratitude to Sem Borst, Onno Boxma, Zbigniew Palmowski, Leen Stougie and Tjark Vredeveld for their efforts in assessing this monograph, providing me with detailed and valuable feedback, and discussing the contents of this dissertation during my defence.

Toen ik begon aan mijn promotietraject dacht ik toe te treden tot een onderzoeksgroep van dertig onderzoekers. Het internationale en verbindende karakter van de wetenschap heeft mij echter in contact gebracht met vele onderzoekers in het buiten- en binnenland. Een lijst met hun namen zou talrijke pagina's in beslag nemen, waardoor ik mij genoodzaakt zie om mij te beperken tot groepen van collega's en enkele specifieke dankbetuigingen.

Of my international colleagues I specifically thank Adam Wierman and his department for hosting my two weeks' visit to Caltech. Even though the scientific results were not quite what we hoped for, I enjoyed every minute of our collaboration. I also treasure the memory of fascinating, yet surprisingly tasty, oriental dishes that the junior researchers in the department made me taste. More generally, I extend my appreciation to the many colleagues that contributed to my educative and enjoyable recollections of international conferences and workshops.

Bij de bovengenoemde conferenties en workshops was het aantal deelnemers dat werkzaam is in Nederland buitenproportioneel groot. Het is dan ook niet vreemd dat ik promovendi heb ontmoet van bijna alle Nederlandse universiteiten. In het bijzonder koester ik de positieve bijdragen van TU/e, VU, UM en UvA collega's aan mijn herinneringen van INFORMS, LNMB en andere bijeenkomsten.

De meeste herinneringen van mijn carrière als promovendus deel ik uiteraard met CWI'ers, en dan hoofdzakelijk de CWI'ers die werkzaam zijn of waren in de Stochastics groep. Gedurende mijn gehele promotietraject voelde ik mij vrij om met jullie te discussiëren over onderzoek, maar ook zeker om te praten over minder wetenschappelijke materie. Daarbij hebben de talrijke lunches, theekransjes en tafeltennissessies op een zeer prettige manier voor afwisseling gezorgd, waarvoor ik jullie allen dank. Specifiek bedank ik mijn kamergenoot Ewan voor zijn prettige gezelschap, zijn internationale betrokkenheid en zijn bereidheid om mijn incidentele uitingen van frustratie of euforie over zich heen te laten komen. Ook bedank ik Thije, voor zijn vriendschap, zijn droge humor en zijn bereidheid om jarenlang met mij te carpoolen.

Mijn laatste woorden richt ik tot degenen die het dichtst bij mij staan. De mensen die onophoudelijk en elk op hun eigen manier hebben geholpen om mijn smart te blijven halveren en mijn geluk te blijven verdubbelen. Ik beschouw mijzelf als een bevoorrecht man tussen de warme persoonlijkheden van Esther, Heleen, Lisa, Maarten, Mathijs, Niels en Tanita. Ik beschouw mijzelf als een rijk man door de onvoorwaardelijke liefde en steun van mijn ouders Gerard en Anita, zusjes Richelle en Iris, schoonouders Jos en José en schoonzusje Kate. Tot slot beschouw ik mijzelf als de gelukkige man van mijn vrouw Robin, van wiens onuitputtelijke liefde, begrip, zorgzaamheid en enthousiasme ik heb genoten gedurende elke fase van het promotietraject en elke pagina van dit proefschrift. Pas nu, op dit moment waarin ik jou en mijn andere naasten bedank, is dit proefschrift compleet.

Bart Kamphorst
Oegstgeest, april 2018

Table of contents

CHAPTER 1

INTRODUCTION

1.1 Motivation

Schedules give structure to the dynamic, chaotic world that we live in. With our daily schedules, people know at what time to wake up, when to bring the kids to school, at what time the next meeting commences, and when to be at the doctor's office. More often than not, unfortunately, something causes you to deviate from that schedule. You tend to miss that important meeting because you find yourself in a traffic jam. Or perhaps your doctor's appointment starts late because an earlier patient required more than the scheduled amount of time. Of course you knew that this could happen, but what are the odds?

All in all, we find ourselves in a constant conflict between planning on the one hand, and uncertainty on the other. It is therefore not quite incidental that both aspects are represented in mathematical areas. First, there is the area of scheduling theory. Scheduling theory concerns itself with the planning of a set of tasks under various restrictions. Second, there is the area of queueing theory. This area focusses on congestion phenomena; situations, subject to uncertainty, in which customers arrive and require some kind of service. Both queueing and scheduling theory have developed individually and there are still many challenges at their interface. It is at this interface that one finds the contributions of this dissertation.

As it is conceivable that the reader is not an expert in both queueing and scheduling theory, the following two sections aim to provide a first step into introducing them. Specifically, Sections 1.2 and 1.3 present some history, notation, techniques and key results in queueing and scheduling theory, respectively. We occasionally include material that is not quite essential for later chapters, but that gives a more complete picture of the areas of discussion. Additionally, as we aim at a broader audience, the technical level in these sections is kept at a minimum. This is compensated for by the more technical nature of all other chapters, each of which is self-contained.

Once the basis is established, Section 1.4 provides a discussion of the overlap between the two areas, and discusses some results therein. Section 1.5 subsequently concludes the chapter with an overview of this dissertation's contributions.

1.2 Basics of queueing theory

Many researchers consider Agner Krarup Erlang as the founding father of the area of queueing theory. Erlang was a Danish mathematician and engineer who worked as head of the laboratory at the Copenhagen Telephone Company. There, he wrote an article in which he considered the following problem: if customers initiated a telephone connection at random moments during some time interval, and every operator required a fixed amount of time to put the customer through, then what is the statistical beha-

viour of the incoming calls and of customer waiting times? The mathematical answers to these questions were published in Erlang [52], where he ignored the fact that all lines to the company might be occupied. If indeed all lines are occupied, then a customer is unable to connect to the company; he or she is *blocked*. In 1917, Erlang appended his prior work by considering a model that incorporated blocking events, presenting his famous "Erlang-B" formula on the probability that a customer is blocked upon dialling [53]. These key results were later collected and translated [34].

Although the foundations of queueing theory lie in telecommunication systems, there are many more applications that share some or all characteristic components of the above telecommunication system and therefore benefit from the analyses of similar models. Specifically, one may think of queueing phenomena at the check-in counters at airports or at counters in supermarkets, but also of more general congestion phenomena such as data packets waiting for transmission in communication networks [79] or emergency patients in hospitals [35]. In all of these examples, there are *customers* that have some *service requirement*, which can be fulfilled by one or more *servers*. These notions are directly related to the following key components that characterise a classical queueing model:

a The *arrival process* describes the dynamics that underlie the arrival of customers. It is often more convenient to consider the amount of time A_i that expires between two consecutive (the i-th and $(i+1)$-th) arrival instances, where it is generally assumed that one customer arrives per arrival instance. We refer to the random variable A_i as the i-th inter-arrival time, and to $F_{A_i}(x) := \mathbb{P}(A_i \leq x)$ as the *inter-arrival distribution*.

b The *service-requirement distribution* $F_{B_i}(x) := \mathbb{P}(B_i \leq x)$ quantifies the probability that the i-th customer requires at most x units of service, where the exact service requirement is denoted by the positive random variable B_i. Once the server has served the i-th customer for B_i units of time, the customer is finished and leaves the system.

c The *number c of servers* in the system. Standard models assume that the servers are parallel and homogeneous, meaning that every server is able to help any customer and that the time required to help a customer is independent of the server. These assumptions may be relaxed in more complex models.

d The *scheduling policy* or *service discipline* is a rule or algorithm that, at any point in time, indicates which customer or customers are served by the server.

A typical characteristic of queueing models is that the arrivals never stop; either new customers keep emerging, customers return after some time, or both. Equivalently, one can say that there is an *infinite stream of arrivals*.

Property	Description
blind	policy does not base its decisions on the processing requirement of a job
pre-emptive	policy may pause the service of a customer momentarily, only to resume service later without losing any progress
work-conserving	policy assigns full capacity of the server to customers at any time

Table 1.1: Three properties of scheduling policies that are repeatedly mentioned throughout this dissertation.

Policy	Description
FB	Foreground-Background simultaneously serves all customers that have received the least amount of service thus far at equal, possibly reduced, rate
FIFO	First In First Out serves customers in order of arrival
LIFO	(Pre-emptive) Last In First Out serves the customer that arrived most recently, thereby pre-empting service of all other customers
PS	Processor Sharing simultaneously serves all customers in the system at equal, possibly reduced, rate
SRPT	Shortest Remaining Processing Time serves the customer that requires the least amount of service until completion; breaking ties arbitrarily

Table 1.2: A brief overview of classical scheduling policies used in this dissertation. SRPT is work-conserving but not blind. All other policies are work-conserving and blind. FB, LIFO and SRPT may pre-empt service, and both FB and PS can serve any number x of customers simultaneously at rate $1/x$.

Initial research generally assumed that customers were served in order of arrival; a policy that is typically referred to as the First In First Out (FIFO) or First Come First Served policy. FIFO is a natural policy in many applications, but is also attractive due to its analytical simplicity. The FIFO policy is *blind* and *work-conserving*, but never *pre-empts* service; see Table 1.1. Other well-studied policies are listed in Table 1.2. Without specification of the scheduling policy, it is generally assumed that customers are served in accordance with the FIFO policy.

By specifying the components that characterise a classical queueing model, we are able to define many queueing models and pursue numerous research questions. Prior

to further discussion, however, we present an established classification scheme and simultaneously touch upon several standard models. The classification scheme then facilitates further discussions due to its compact notation.

1.2.1 Kendall's classification

The coming paragraphs introduce Kendall's a/b/c classification scheme [80]. Proper use of this notation allows us to compactly denote some of the most studied classical queueing models. In particular, the 'a' field corresponds to the inter-arrival distribution, the 'b' field relates to the service-requirement distribution, and the 'c' field describes the number c of servers.

The following abbreviations are often used as input for the 'a' and 'b' fields: D, M, E_k, G and GI. Here, D denotes a **D**egenerate distribution, i.e. the associated random variable is deterministic; M is short for **M**arkovian or **M**emoryless, i.e. the associated random variable is exponentially distributed; and E_k stands for the **E**rlang distribution with k phases. If the associated random variable has a **G**eneral, unspecified distribution, then this is denoted by G. Here, we note that it is possible that the associated random variables may not be independent of each other; for example, if G is the input for field 'a', then it is possible that the random variables $A_1, A_2, \ldots$ are mutually dependent. If this is not allowed, then we say that the distribution is **G**eneral but **I**ndependent, and abbreviate this by GI.

Let us have a quick look at some examples. An M/M/1 queueing model is short for a queueing model where both the inter-arrival and the service-requirement distributions are exponentially distributed, and there is one server to serve the customers. If instead there are c servers and the service-requirement of every customer is fixed, then this is denoted by M/D/c. Note that this is the first model studied by Erlang.

Kendall's classification covered the mostly used queueing models at the time and has served for many years without alterations. However, it also suffers from serious limitations. For example, we are not able to capture Erlang's second model in Kendall's classification as it does not allow for a limited number of customers at a time. Also, if customers are served in an order different from FIFO, then this cannot be captured in the classification scheme. It is for those reasons that the notation was augmented to a/b/c/d/e/f, where 'd' represents the scheduling policy, and 'e' and 'f' respectively denote [92, 131]

e the *buffer size*, which is an integer that indicates how many customers can reside in the queue, including the customers in service, at any time. If unspecified, this number is assumed to be infinite.

f the *calling source*, which is an integer that equals the number of potential customers from which the actual customers originate. If this number is finite, then

no outsiders can enter the system (it is *closed* to the outside world). When unspecified, this number is assumed to be infinite (*open* system).

Again, we consider two examples to illustrate the extended Kendall's classification. First, the M/M/c/FIFO/c/∞ queueing model assumes that both the inter-arrival and the service-requirement distributions are exponential, that there are c servers, that customers originate from an infinite pool and are served in order of arrival, and finally that no more than c customers can reside in the system. This latter assumption implies that no more customers may enter the system once all servers are occupied. We note that this is the model studied in Erlang [53]. Second, the M/M/c/FIFO/∞/k queueing model only has a customer pool of size k, and has no buffer capacity[1]. This model is a special case of the machine repair model, where there are k machines and one repairman [70]. Only the k existing machines (customer pool) can break, and once they do they require service from the repairman (the server).

Authors often omit one or more fields of the extended Kendall's classification. Its meaning should then be clear from either the notation or from the context. In this dissertation, we consistently omit the 'e' and 'f' fields as they would always be replaced by their default, infinite value. We also omit the d field if the scheduling policy is FIFO. With these remarks in mind, we are ready to move on to a deeper understanding of queueing models.

1.2.2 Waiting time, system stability and related areas

When analysing a system, it is of interest to measure the performance of this system. A natural performance metric for assessing a queueing system is the *waiting time* of a customer. This metric has been studied extensively for many models and is closely associated to many key results in queueing theory. It is intuitively defined in a system that obeys the FIFO policy, where the waiting time W_i of the i-th customer equals the amount of time that the customer is in the system but has not yet received any service. The customer clearly benefits from a short waiting time; however, for general queueing models, it is possible that $\liminf_{i\to\infty} W_i$ is unbounded. To exclude this undesired behaviour, one may be interested in the *system stability*.

We call a queueing system *stable* if a proper steady-state distribution exists for the sequence of waiting times; i.e. that there exists a random variable W such that $\lim_{i\to\infty} \mathbb{P}(W_i \leq x) = \mathbb{P}(W \leq x)$ for all $x \in \mathbb{R}$. It turns out that there is a remarkably elegant and intuitive condition that ensures system stability for basic models. Specifically, one of the most important results in queueing theory is that a GI/GI/c queue is stable if $\rho := \mathbb{E}[B_1]/(c\mathbb{E}[A_1]) < 1$. The *traffic intensity* ρ equals the long-term fraction of time

[1] Note that, in this model, this is equivalent to a buffer capacity of size k.

that a work-conserving server is serving customers. Indeed, it is conceivable that this fraction should be less than unity for the system to be stable. Among the exceptions to this intuition is the D/D/c model, which remains stable even for $\rho = 1$.

System stability is a fundamental subject of investigation in several mathematical areas. The relatively young area of queueing theory could therefore benefit from known stability results in its early days. For example, the evolution of the number of customers in a M/M/1 queue is described by a continuous-time Markov process. Ergodicity of the embedded Markov chain then implies system stability; a result that Kendall [80] later applied to the non-Markovian GI/M/c model in a non-trivial way. Another parallel can be found between the evolution of the waiting time and a random walk with absorbing or reflecting barriers. Khintchine[2] [81], Lindley [96] and Kiefer and Wolfowitz [82] successfully exploited such parallels to study stability of the M/GI/1, GI/GI/1 and GI/GI/c models, respectively. Here, we should note that the stability result for the M/GI/1 model was obtained independently by Khintchine [81] and Pollaczek [110, 111], where the latter pioneered in applying functional-analytic methods. A survey of more recent stability analysis methods, such as the Lyapunov function and fluid approximation method, can be found in the survey by Foss and Konstantopoulos [57].

The above examples clearly illustrate the intrinsic relation of queueing theory to other areas in applied probability and suggest that many tools and techniques from other areas may also be applicable in queueing theory. This is indeed true, and these parallels will be exploited on several occasions in this dissertation.

1.2.3 Steady-state and time-dependent analysis

By our notion of system stability, the sequence $(W_i)_{i\in\mathbb{N}}$ of waiting times in a stable queueing system converges weakly to a random variable W. We refer to this random variable as the *steady-state* waiting time. The literature that we described in the previous section was not only dedicated to the existence of W, but also made an effort to uncover its probabilistic behaviour. This effort emerged into several important results, of which we discuss only one: the Pollaczek-Khintchine formula, derived independently by both researchers [81, 110, 111].

The Pollaczek-Khintchine formula is an expression for the steady-state waiting time in the M/GI/1 queueing model. In its classical form, it relates the Laplace-Stieltjes transform (LST) of W to the LST of B_1 and the rate of customer arrivals. Since the LST of a random variable uniquely defines the random variable, this result contains all information about W. Another form of the Pollaczek-Khintchine formula represents

[2] The phonetically more accurate, but seldom used, translation of his name would be Hintchin [130]. For reasons of convenience, we comply with the more frequently used spelling Khintchine.

the waiting-time distribution as the geometric sum of independent random variables B_i^*, where the distribution of the B_i^* is related to that of the B_i.

A direct consequence of the LST representation is that all moments of the steady-state waiting time can be deduced; in particular, it yields

$$\mathbb{E}[W] = \frac{\rho}{1-\rho} \cdot \frac{\mathbb{E}[B_1^2]}{2\mathbb{E}[B_1]} \tag{1.1}$$

for all $\rho < 1$. At this point, we emphasise the dependence on the traffic intensity ρ. Clearly, the expected steady-state waiting time diverges to infinity as ρ increases to unity. This supports the claim that ρ must be smaller than unity for system stability. Perhaps more interesting is the fact that the $(1-\rho)^{-1}$ scaling shows up in many analyses of queueing models, despite its deceivingly simple dependence on only the mean of the inter-arrival and service-time distributions.

We have not yet said anything about the time required for the system to achieve steady state, or about the behaviour of W_i for specific i. Our account on this part will be very limited, as it is concerned with the *time-dependent* or *transient* analysis of queueing models; topics that are outside the scope of this dissertation. It is generally much harder to obtain explicit expressions for $\mathbb{P}(W_i \leq x)$ than for $\mathbb{P}(W \leq x)$. This is clearly illustrated by Kleinrock [85] in his expression (2.163), which presents an explicit expression for the number of customers in the system at time t in the M/M/1 model. The expression involves an infinite sum of modified Bessel functions of the first kind, and is described as "most disheartening" by its author. The time-dependent expressions for the virtual waiting time in the M/GI/1 model are no more attractive [22, 132].

In fact, even the analysis of queueing models in steady-state rapidly complicates as we move away from the standard M/GI/1 setting. It is therefore that many researchers restrict their analyses to certain regimes that allow for further analysis, but still yield meaningful insights.

1.2.4 Large deviations and heavy traffic

The literature on queueing theory recognises several well-studied regimes, of which we consider two: the *large-deviation* and the *heavy-traffic* regime. First, the large-deviation regime concerns itself with the behaviour of $\mathbb{P}(W > x)$ as x grows large. Second, the heavy-traffic regime investigates the probabilistic behaviour of W as the traffic intensity ρ tends to one. The results from both regimes contribute to the literature by both quantifying the behaviour of a performance metric of interest, and obtaining insight into the circumstances under which rare events occur. Quite different techniques are used among these two regimes, which we shortly discuss.

Among others, the probability $\mathbb{P}(W > x)$ is of interest in telecommunication systems. It assesses the probability that a customer has to wait longer than x units of time,

which is related to customer satisfaction and is often part of the company's targets. Alternatively, the waiting time can be shown to relate to risk analyses. It then translates to the probability that an insurance company with capital x goes bankrupt.

A technique employed in this dissertation to derive a large-deviation result, is a *sample-path* analysis. In such analyses, one tries to recognise a likely way to obtain the event of interest, and then shows that the event of interest is unlikely to occur in any other way as x grows large (i.e. $x \to \infty$). For example, if we expect that a long waiting time is caused by having a single high-demanding customer in the system, then we try to show that it is unlikely that a customer has a long waiting time if there are no high-demanding customers in the system. The reader interested in general large deviations theory is referred to the books by Ganesh et al. [61] and Dembo and Zeitouni [44].

On the other hand, one might be interested in the behaviour of W as the traffic intensity ρ increases to unity. This corresponds to a change in the system input distributions: either the rate of arriving customers increases or the customers become more demanding. One may think of an increasing number of internet users, or the transfer of larger data files. In either case, the server is pushed to the limits of its capacity[3] and we have seen in relation (1.1) that this may have a considerable impact on the system performance. Heavy-traffic results aim to quantify this impact [135].

The heavy-traffic regime received ample attention after Kingman [83, 84] published two fundamental papers on the waiting time in GI/GI/1 queues. Specifically, he found that the scaled waiting time $(1-\rho)W$ in such a model converges to an exponential random variable as $\rho \uparrow 1$, provided that both A_1 and B_1 have finite variance. Kingman derived his results by means of analytical methods, whereas similar results in this area have been obtained by *diffusion approximations*. This technique approximates the discrete process of arriving and leaving customers by a continuous process, thereby greatly simplifying the analytical structure [63]. A general development of the theory and techniques associated with the heavy-traffic regime can be found in the books by Chen and Yao [37] and Whitt [136].

On a final note, we would like to point out that care needs to be taken in the steady-state analysis of queueing models in heavy traffic. Heavy traffic is achieved by changing the input distributions; however, alteration of these distributions will also disrupt the (steady) state of the system. One way to overcome this problem is by considering a family of queueing models where every next model has slightly different, but fixed, input distributions. Every system is then assumed to be in steady state, and one investigates how the metric of interest changes over the various models. This technique is exploited repeatedly throughout this dissertation.

[3] Here, we assume that the system remains stable.

The above models and techniques by no means form a complete picture of queueing theory. Nonetheless, we hope that the unfamiliar reader has gotten a flavour of the basic concepts and most fundamental results in the area. The interested reader is referred to the introductory lecture notes of Adan and Resing [6] or the more advanced books by Asmussen [8], Cohen [39], Harchol-Balter [71], Kleinrock [85, 86] and Wolff [141].

1.3 Basics of scheduling theory

Scheduling theory emerged from inventory management and manufacturing theory and gained momentum in the 1950s [40, 123]. At the time, the problems that we would now refer to as scheduling problems used to be solved by means of linear programming. Scheduling theory progressed as interest grew for techniques that allowed for more flexibility, and researchers longed for a more fundamental understanding of scheduling problems. As one might expect, the first papers that could be classified as scheduling literature focussed on the latter issue. Among these papers are the influential works of Jackson [74] and Johnson [75], who both considered a manufacturing problem.

The problem considered by Jackson can be illustrated as follows. Suppose that a carpenter sells several types of wooden furniture upon request, and a number of customers submit an order at about the same time. Since the carpenter crafts and coats all furniture himself, he decides to take no new orders until he has finished the current orders. Now, what is the best system to complete the orders?

The answer depends on the interpretation of "best". Jackson [74] assumes that every customer also mentioned when they would like to pick up the order. He additionally assumes that the carpenter would ideally deliver all furniture on time, but would otherwise aim to reduce the maximum tardiness over all orders. The optimal solution to this problem is a very simple rule, stating that the carpenter should always work on the order that needs to be finished first. This result may not sound surprising; however, if the carpenter had decided to take some more orders along the way, then there exists no computationally efficient manner to solve this problem. We will get back to this notion in Section 1.3.4.

Johnson [75], on the other hand, presumes that the carpenter hires an assistant for the second part of his manufacturing process. This introduces some dependency in the system, since the assistant can only start working once the carpenter finished crafting some furniture. Johnson also assumes that their common goal is to work as efficient as possible, that is, to finish all furniture as soon as possible. He then presents a surprisingly elegant and easy-to-use rule to optimise this. An interesting feature of his seminal paper is that it was written prior to the aforementioned article, even though the considered model is relatively complex.

In both of the above examples, we tried to allocate time to the tasks at hand in

order to minimise a certain objective. These characteristics are in fact common to all scheduling problems. More specifically, Pinedo [109] defines scheduling as "(...) a decision-making process that (...) deals with the allocation of resources to tasks over given time periods and its goal is to optimise one or more objectives." As one might guess from this definition, scheduling theory has found its way into numerous applications. Among others, its insights have been applied in computer operating systems [26, 86], web servers [72, 126], aircraft landings [18] and healthcare [67].

We will generally say that tasks are assigned to or scheduled on *machines*, rather than that resources are allocated to tasks. This perspective facilitates the similarity between queueing and scheduling theory. The due date of the j-th task is denoted by d_j, which is assumed to be infinite if unspecified. Also, we should note that scheduling problems traditionally consider *a finite set of tasks* that needs to be scheduled.

In parallel with Section 1.2, we continue by presenting some standard classification and seize the opportunity to give a taste of the endless model variations.

1.3.1 Three-field classification

Conway et al. [40] made a first attempt to structure scheduling problems. They proposed a four-field classification that, among others, had one field dedicated to the arrival process of tasks. By removing this field and improving the use of other fields, Graham et al.[4] [66] lay basis to the three-field classification $\alpha \mid \beta \mid \gamma$ that is now standard. In this notation, the field indicated by

- α denotes the *machine environment*. Common inputs for parallel machine environments are: 1 (single machine), P_m (m identical machines), Q_m (m machines that have different speeds) and R_m (m machines whose speed depends on the task assigned). Alternatively, jobs might have to run on several machines. This is usually denoted by F_m ("flow shop"; each job is scheduled on all m machines in series), J_m ("job shop"; each job is scheduled on all m machines in an individual, predetermined order) or O_m ("open shop"; each job is scheduled on an individual, predetermined set of machines in arbitrary order).

- β indicates *processing restrictions* and *relaxations*. There may be several inputs in this field simultaneously, presumably among r_j (the j-th task can not be scheduled prior to its *release date* r_j), *pmtn* (tasks may be *pre-empted* even though they have not yet finished) and *prec* (*precedence constraints* among tasks). There are many more restrictions and relaxations possible, of which this dissertation focusses primarily on scheduling problems with the *online* relaxation. This paradigm is described in Section 1.3.2.

[4] Graham et al.'s paper was later revised by Lawler et al. [91].

γ symbolises the *optimality criteria*. It is a function of the scheduling output that we intend to minimize. Usually, the scheduling output of interest is related to a task's *completion time* C_j, defined as the time when it has received all required resources; *flow time* F_j, defined as the difference between its completion time and release date; or *tardiness* L_j, defined as the difference between its completion time and due date. Common optimality criteria are then: $C_{\max}$ ("makespan", maximum over all C_j), $\sum_j C_j$, $F_{\max}$, $\sum_j F_j$, $L_{\max}$ and $\sum_j L_j$. Possibly, the sums are *weighted* by individual task weights w_j; this is typically denoted by $\sum_j w_j C_j$.

From the above descriptions, one may recognise the carpenter model of Jackson [74] as a $1 \mid\mid L_{\max}$ problem, the extension where the carpenter accepted more orders as a $1 \mid r_j \mid L_{\max}$ problem, and the carpenter model by Johnson [75] as a $F_2 \mid\mid C_{\max}$ problem.

It is generally assumed that every machine can process at most one task at a time, and every task can be processed at only one machine at a time. This makes sense from a technical perspective. However, by repeatedly allocating a set of tasks to a machine during very short intervals, one approaches a model where machines seemingly work on several tasks at the same time. From this perspective, the machine works on every task at reduced speed (cf. Table 1.2). It should be noted that, in reality, there may be overhead incurred by switching jobs on a machine. This approach may then be infeasible or expensive.

1.3.2 Off-line, on-line and scenario scheduling

In the following paragraphs, we discuss the off-line, on-line and scenario scheduling paradigms. Especially the on-line paradigm is of importance in this dissertation, as it is closely connected to queueing models.

Traditional scheduling problems assume that all information is available a priori. The scheduler is aware of the number of tasks, their release dates, and their requirements and is therefore able to make a schedule *off-line*. This may be a realistic approximation of certain real-life problems; for example, all of this information may be estimated very accurately when it comes to the day-to-day planning in a factory of mass production. A survey on this account was written by Potts and Strusevich [112], and standard textbooks cover this paradigm extensively [40, 109].

Alternatively, a scheduler may be unable to design a schedule if some of the information is missing prior to the process commences. The scheduler may then have to resort to *on-line* scheduling. In this paradigm, any of the task's properties may be (partially) unknown. If only the release dates are unknown, indicated by *online-time*, then the scheduler is oblivious of the task's existence until the time of its release. If, additionally, the processing requirements of a task are not revealed until it is actually finished, then the scheduler is said to be *non-clairvoyant* and this is denoted by *online-time-nclv*. In

this case, it is often assumed that the distribution of task requirements is known. We invite the reader to realise the resemblance between this latter setting and a queueing model. A well-written survey of on-line scheduling can be found in Pruhs et al. [116] and Sgall [128], whereas Borodin and El-Yaniv [27] and Pinedo [109] dedicated several chapters to the development of the subject.

Feuerstein et al. [55] and Kasperski et al. [77] recently considered a paradigm that lives between the off-line and on-line setting. Their papers describe *scenario* scheduling, where the scheduler designs a schedule before the actual problem instance is sampled from a known, finite set of problem instances. It is yet to be seen how this interesting paradigm evolves.

The above descriptions suggest that the off-line scheduler has a big advantage over the on-line scheduler. The off-line scheduler has all information needed to make a good schedule, whereas the on-line scheduler unknowingly makes decisions that may later turn out to be sub-optimal. The study of *competitive analysis* [65] is dedicated to the investigation of this phenomenon.

1.3.3 Competitive analysis

Consider an *on-line* scheduling problem. Let $\mathcal{S}(\mathcal{I})$ denote a schedule for a given problem instance $\mathcal{I}$ and let $f(\mathcal{S}(\mathcal{I}))$ denote the corresponding value of the function that we wish to minimize. Assume that $\mathcal{S}^*(\mathcal{I})$ is a schedule that minimizes this value; i.e. $\mathcal{S}^*(\mathcal{I})$ is an optimal schedule. A *deterministic algorithm* ALG is then said to be c-competitive, $c \geq 1$, if the algorithm designs a schedule $\mathcal{S}(\mathcal{I}) = \text{ALG}(\mathcal{I})$ satisfying $f(\text{ALG}(\mathcal{I})) \leq c f(\mathcal{S}^*(\mathcal{I}))$ for any problem instance $\mathcal{I}$. Here, c is allowed to depend on the problem parameters, such as the number of tasks in the instance.

There are several remarks to be made about this definition. First, we note that an optimal schedule $\mathcal{S}^*(\mathcal{I})$ is oblivious to the on-line nature of the problem. One may think of $\mathcal{S}^*(\mathcal{I})$ as an optimal schedule that was designed by an off-line scheduler that first observed the on-line instance materialize. Clearly, this off-line scheduler can never do worse than the on-line scheduler. Second, we observe that the competitive ratio is a worst-case classification, meaning that c is large if the algorithm performs very well on all but a few problem instances, and poorly on these few. It is this aspect upon which *randomised algorithms* provide an advantage.

Randomised algorithms flip internal coins during their execution and base their decisions on the corresponding outcomes. Consequently, the algorithm may design different schedules during repetitive executions on the same problem instance, and, as a result, the performance on a fixed problem instance is a random variable. A randomised algorithm rALG is now said to be c-competitive if it satisfies $\mathbb{E}[f(\text{rALG}(\mathcal{I}))] \leq c f(\mathcal{S}^*(\mathcal{I}))$ for all problem instances $\mathcal{I}$, where $\mathbb{E}[f(\text{rALG}(\mathcal{I}))]$ is the expected function value with

respect to the internal random choices of the algorithm. That is, the competitive ratio of a randomised algorithm quantifies its worst-case *expected* performance. One therefore finds that the competitive ratio of the best randomised algorithm is never worse, and generally better, than that of the best deterministic algorithm.

The competitive ratio can be thought of as follows: it is the worst-case (expected) relative performance of an on-line algorithm that plays against an *oblivious adversary*. This adversary is able to decide upon all task's characteristics, necessarily including the task's release date and its requirements. He may do so with full knowledge of the algorithm; that is, he knows the code of the algorithm but *not the outcome of randomised choices*. Armed with this knowledge, he designs a problem instance that maximizes the ratio of the expected performance of the algorithm to that of the optimal schedule in hindsight. The advantage of randomization in an algorithm then lies in the fact that the adversary is not quite sure how the algorithm responds to his problem instance.

To illustrate the difference between the competitive ratio of deterministic and stochastic algorithms, we consider the scheduling problem $1 \mid \textit{online-time-nclv}, r_j, \textit{pmtn} \mid \sum_j F_j$. Motwani et al. [102] showed that the competitive ratio of every deterministic algorithm is $\Omega(n^{1/3})$. Here, $\Omega(n^{1/3})$ indicates that any achievable competitive ratio grows at least as fast in the number of tasks n as some function of the form $c_1 n^{1/3}$. In their proof, Motwani et al. exploit their knowledge of the algorithm, whichever deterministic algorithm is given, to construct a problem instance where it performs badly. If, instead, one considers randomised algorithms, then Motwani et al. [102] found that all algorithms have a competitive ratio of at least $\Omega(\log n)$; a significantly lower ratio. Kalyanasundaram and Pruhs [76] subsequently presented a randomised algorithm that actually achieves this ratio; see also Chapter 2.

The above example makes it clear that an on-line scheduler can be significantly disadvantaged compared to an off-line scheduler. A similar observation holds for numerous scheduling problems, so that one may wonder whether on-line schedulers can actually achieve optimal off-line performance. The answer to this question is positive, as the (deterministic!) SRPT algorithm optimally solves the $1 \mid \textit{online-time}, r_j, \textit{pmtn} \mid \sum_j C_j$ scheduling problem. A more elaborative introduction to competitive on-line scheduling is presented in Pruhs [115].

Yao's minimax principle

A key instrument in Motwani et al.'s "$\Omega(\log n)$"-result is Yao's celebrated minimax principle [142]. This principle relates the competitive ratio to a game between two players. The first player may choose from a set of deterministic algorithms $\mathscr{A}$. The second player, the adversary, may then choose a problem instance from a set $\mathscr{I}$. The competitive ratio of the algorithm chosen by player one is then lower bounded by the relative perform-

ance of this algorithm on the problem instance chosen by player two (the "cost" of the game). Player one and two respectively aim to minimise and maximise this cost.

If player one reveals his strategy first, then player two may exploit this knowledge in his own strategy. Player one makes this harder by playing a random ("mixed") strategy, which corresponds to a randomised algorithm. Player two is aware of this random strategy but does not know the outcome. He will thus pick the instance that maximizes the cost for the randomised algorithm. Here, it can be shown that player two does not benefit from a mixed strategy. Alternatively, if player two defines a probability distribution F over $\mathscr{I}$ rather than fixing an instance, then player one can minimise the cost by selecting an appropriate deterministic algorithm. The key insight is now that if both players play their best strategy, then it does not matter which player goes first (Von Neumann [104]).

Yao used Von Neumann's result to show that the competitive ratio of a randomised algorithm is lower bounded by the best expected relative performance over all deterministic algorithms on the randomised problem instance. That is, if the expected performance of *all* deterministic algorithms is at least C, then it is impossible for the randomised algorithm to perform better than C on *all* instances. The power in this approach lies in the freedom for the researcher to decide upon F.

1.3.4 Algorithmic complexity and approximation algorithms

We finish our introduction to scheduling theory with a short account on *algorithmic complexity*; a concept that has had significant impact on the field.

Algorithmic complexity is related to the observation that all off-line scheduling problems are combinatorial problems. That is, given the inputs for the model, there is only a finite number of meaningful schedules and we could write down all of them. By specifying this list for any scheduling problem, examining all schedules and selecting one that is optimal, we could solve any off-line problem to optimality. Unfortunately, the number of meaningful schedules may be so large that even with all computing power on earth it could easily take hundreds if not thousands of years to list them all. It is for this reason that researchers philosophised about a way to quantify the efficiency of scheduling algorithms, which led to the study of algorithmic complexity [62].

The main idea in algorithmic complexity, as described by Edmonds [49], is that an algorithm is efficient if the number of computations needed to execute the algorithm is polynomial in the number of binaries needed to encode the problem input. Among these are the algorithms presented by Jackson [74] and Johnson [75] that correspond to the carpenter problems at the beginning of this section. Such algorithms are called *polynomial-time algorithms*.

For many problems, however, nobody has yet succeeded in designing a polynomial-

time algorithm, nor to show that such algorithm can not exist (the P equals NP problem). Among these is the problem of the carpenter that took more orders [93]. It is for this set of hard problems that researchers quest for polynomial-time algorithms that are in some sense close to optimal; the study of *approximation algorithms*.

The study of approximation algorithms applies to polynomial-time algorithms on off-line problems, and aims to quantify the worst-case sub-optimality that is caused by the restriction on their running time. Note that this is quite different from the competitive ratio, which relates to on-line problems and (possibly non-polynomial-time) algorithms that are restricted in their knowledge of the future. Also, one may show that the best achievable approximation ratio of randomised algorithms is no better than that of deterministic algorithms, which is in sharp contrast to several results in competitive analysis. The reader eager for more information on approximation algorithms is referred to the books of Vazirani [134] and Williamson and Shmoys [140].

This concludes our introduction to scheduling theory. As before, the models and techniques described in this section by no means cover all facets of scheduling theory but merely serve to illustrate the richness of its literature and possibilities. The next section facilitates the reader in organising his thoughts on the connection between queueing and scheduling theory, and discusses some results at their interface.

1.4 Fusing queueing and scheduling

Now that a basic understanding of the areas of queueing and scheduling theory has been established, we briefly discuss the major similarities and differences among them. We then examine several results at their interface and set the stage for all later chapters. The final paragraphs contain key results and references to more elaborate surveys and textbooks, rather than an independent overview of all results at the interface. Every next chapter contains a more detailed literature review specified to that chapter's contributions.

1.4.1 Overlap between both areas

Both queueing and scheduling theory are concerned with models where jobs (customers, tasks) require service (resources) from one or more servers (machines). This similarity remains intact when the models become more complex. For example, both queueing networks and job shop models consider customers that visit multiple – possibly heterogeneous – servers in some order. The role of the scheduler in scheduling theory is in both cases equivalent to the role of the scheduling policy in queueing systems. The key differences, however, lie in the job arrival characteristics, the information available to the scheduler, and the performance metrics.

First, the process according to which jobs arrive is quite different. The arrival process in queueing systems is typically *stochastic*. This implies that jobs have release dates, which are unknown to the scheduler, and that there is an infinite arrival stream of jobs. As a consequence of this latter characteristic, one needs to consider system stability. Scheduling problems instead assume a fixed number of jobs which may either be available from the start or released individually over time. When released over time, the release dates may be known or unknown to the scheduler. An algorithm from scheduling literature is therefore only a feasible scheduling policy for a queueing model if it makes on-line decisions.

Second, the stochastic nature of queueing processes is usually reflected in the performance metrics. Standard performance metrics for queueing models are the moment and distribution-tail characteristics of random variables associated to the performance of the system, such as queue length, waiting time and flow time. The probabilistic properties of the inter-arrival and service-requirement distributions are often reflected in these metrics. Also, these characteristics are meaningful even with, or primarily due to, the infinite stream of arriving jobs.

The competitive ratio, on the other hand, assumes a finite number of jobs and quantifies the worst-case behaviour of an (on-line) algorithm compared to the off-line optimal schedule. It therefore provides no immediate insights into the behaviour of an algorithm in a queueing model, where the number of jobs is infinite. Additionally, the competitive ratio does not take the characteristics of the queueing model into consideration. For example, the FIFO policy is both n-competitive for $1 \mid r_j \mid \sum_j C_j$ (worst possible) and optimal for minimising the expected waiting time in a stable D/D/1 queueing model. More generally, the competitive ratio is not designed to give any indication about the most likely or long-time average behaviour of an algorithm in a stochastic environment. This implies that an algorithm with a high competitive ratio may actually perform quite well with respect to a different performance metric. We will elaborate on this notion shortly, when we consider the large-deviations behaviour of the waiting time in M/GI/1.

1.4.2 Results at the interface of queueing and scheduling

As only on-line algorithms can be applied in queueing models and stochastic analysis of such algorithms tends to become intractable, there are relatively few well-understood scheduling policies. Nevertheless, current literature displays a rich and fascinating variety in the behaviour among the most classical scheduling policies (cf. Table 1.2). In the following paragraphs, we denote the steady-state waiting time and sojourn time (a.k.a. flow time, response time) in a queueing model with scheduling policy P by W_{P} and T_{P}, respectively. The reason for occasionally considering the sojourn time rather

than the waiting time is that the definition of the waiting time becomes less intuitive or meaningless for scheduling policies that may serve customers at reduced rate (e.g. FB, PS); however, since $T_{\mathrm{P}} \stackrel{d}{=} W_{\mathrm{P}} + B_1$ and W_{P} asymptotically stochastically dominates B_1, the waiting time results easily translate to sojourn time results.

The best-understood scheduling policy is FIFO, and even this policy is still not completely understood in the GI/GI/1 model. Our understanding improves if either the inter-arrival or the service-requirement distribution is exponential. In particular, it is known that the waiting time in the GI/M/1 model is exponentially distributed with some parameter that is obtained as the solution to an integral equation [8, Theorem X.5.1]. Furthermore, the waiting time in the M/GI/1 model has distribution $\mathbb{P}(W_{\mathrm{FIFO}} > x) = \sum_{n=0}^{\infty}(1-\rho)\rho^n \mathbb{P}(B_1^* + \ldots + B_n^* > x)$. The B_i^* in this expression are independently distributed as a *residual service requirement*; $\mathbb{P}(B_i^* > x) = \mathbb{E}[B_1]^{-1} \int_x^{\infty} \mathbb{P}(B_1 > y)\,\mathrm{d}y$.

Although the representation of the M/GI/1 waiting time is explicit, it does not quite indicate how well FIFO performs relative to other scheduling policies. In particular, this representation does not reveal how FIFO compares to frequently investigated scheduling policies like LIFO, PS and SRPT. A comparison would obviously benefit from waiting-time representations in models with these scheduling policies. Known representations, unfortunately, typically discourage one from attempting such comparisons. One is then forced to limit the comparison to asymptotic regimes, where it is easier to analyse the behaviour of scheduling policies. In particular, we focus on their behaviour in the large-deviations and heavy-traffic regimes.

The behaviour in either regime depends on the service-requirement distribution; specifically, queueing models exhibit fundamentally different behaviour for *light-tailed* and *heavy-tailed distributions* [50, 59]. We say that the service-requirement distribution is light-tailed if $\mathbb{E}[e^{sB}]$ is finite for some $s > 0$. This class includes all distributions with finite support and all phase-type distributions. The service-requirement distribution is heavy-tailed if it is not light-tailed; however, we will focus here on the subset of heavy-tailed distributions that satisfy $\lim_{x\to\infty}(1 - F_B(\mu x))/(1 - F_B(x)) = \mu^{-\alpha}$ for some index $\alpha > 2$ and all fixed $\mu \geq 1$. This is the class of *regularly-varying distributions*.

Large deviations

Let $f(x) \sim g(x)$ denote $\lim_{x\to\infty} f(x)/g(x) = 1$ and recall that the large-deviations regime corresponds to the asymptotic behaviour of tail probabilities for fixed traffic intensities ρ.

The Cramér-Lundberg theorem [50, Theorem 1.2.2] states that, for almost all light-tailed service-time distributions, there exists $c_2 > 0$ and $\gamma_{\mathrm{FIFO}}(\rho) = \gamma_{\mathrm{FIFO}} > 0$ such that $\mathbb{P}(W > x) \sim c_2 e^{-\gamma_{\mathrm{FIFO}} x}$. Sequentially, Boxma and Zwart [32] proved that no scheduling policy can improve the waiting time tail by more than a multiplicative constant. This is

in sharp contrast with the tail of T_{PS}, for which Mandjes and Zwart [99] showed that, for a certain subclass of light-tailed service-requirement distributions, $\log \mathbb{P}(T_{\text{PS}} > x) \sim \gamma_{\text{BP}}$, where $\gamma_{\text{BP}} < \gamma_{\text{FIFO}}$ is the worst (lowest) decay rate that any work-conserving scheduling policy can have. The same result holds for the tails of T_{LIFO} and T_{SRPT} [106]. One may thus conclude that FIFO outperforms LIFO, PS and SRPT in the large-deviations regime for most light-tailed service-requirement distributions. This illustrates the fact that scheduling policies with poor competitive ratios (FIFO is n-competitive), may actually perform well in queueing models.

If we instead consider the large-deviations regime with regularly-varying distributions, we observe the opposite relations. In that case, Borovkov [28] proved that

$$\mathbb{P}(W_{\text{FIFO}} > x) \sim c_3 (1-\rho)^{-1} x \mathbb{P}(B_1 > x) \tag{1.2}$$

for all fixed ρ and some $c_3 > 0$. Thus, if the service-requirement distribution is regularly varying with index $-\alpha$, then W_{FIFO} is regularly varying with index $1-\alpha$. This is the heaviest tail index possible, meaning that FIFO performs very poorly. The PS and SRPT scheduling policies instead achieve a sojourn-time distribution tail satisfying $\mathbb{P}(T_{\text{PS}} > x) \sim \mathbb{P}(T_{\text{SRPT}} > x) \sim (1-\rho)^{-\alpha} \mathbb{P}(B > x)$, which has the lightest tail index possible [106, 145]. The tail of T_{LIFO} has the same tail index, but a slightly larger multiplicative factor of $(1-\rho)^{1-\alpha}$.

We have observed that scheduling policies that perform well in the large-deviations regime for some distributions, perform poorly for other distributions. Indeed, if no information of the service-requirement distribution is given or learned, then Wierman and Zwart [139] proved that it is impossible for a scheduling policy to perform near-optimal for both light-tailed and regularly-varying service-time distributions. Nair et al. [103] subsequently showed that a hybrid between LIFO and PS, namely the LPS policy, remains robust over both classes in the sense that the decay rate and the tail index are better than γ_{BP} and $1-\alpha$, respectively. Their implementation of the LPS policy only requires knowledge of the traffic intensity ρ.

Two worthwhile surveys on the large-deviations performance of scheduling policies were presented by Borst et al. [30] and Boxma and Zwart [32].

Heavy traffic

We now consider the heavy-traffic regime; specifically, we limit our focus to the dependence of the expected sojourn time on the traffic intensity ρ as it tends to one. We alter our previous definition, by now denoting $f(\rho) \sim g(\rho)$ if $\lim_{\rho \uparrow 1} f(\rho)/g(\rho) = 1$.

As opposed to the large-deviations regime, it is possible for a scheduling policy to minimise the expected sojourn time over all GI/GI/1 models. Specifically, Schrage [125] showed that the SRPT policy is 1-competitive and therefore, as it works on-line,

minimizes $\mathbb{E}[T]$ over all scheduling policies. However, it took several more decades before the dependence of $\mathbb{E}[T_{\text{SRPT}}]$ on ρ was studied.

Pollaczek [110, 111] and Khintchine [81] showed that the expected sojourn time $\mathbb{E}[T_{\text{FIFO}}]$ in the M/GI/1 model is identical to $\rho\mathbb{E}[B_1^2]/(2(1-\rho)\mathbb{E}[B_1]) + \mathbb{E}[B_1]$ (cf. relation (1.1)). Kleinrock [86] found an even more appealing formula for the sojourn time under PS, namely $\mathbb{E}[T_{\text{PS}}] = (1-\rho)^{-1}\mathbb{E}[B_1]$. However, it wasn't until 2005 that Bansal [14] was able to quantify the superiority of SRPT in the M/M/1 model. For this model, Bansal showed that

$$\mathbb{E}[T_{\text{SRPT}}] \sim \frac{\mathbb{E}[B_1]}{1-\rho}\frac{1}{\log(e/(1-\rho))}; \tag{1.3}$$

that is, the expected sojourn time in M/M/1/SRPT is improved over FIFO, PS and in fact all blind scheduling policies [40] by a logarithmic factor.

Bansal was able to derive his result by examining the conditional sojourn time $\mathbb{E}[T_{\text{SRPT}}(x)]$ for a job of size exactly x and integrating this expression over all possible job sizes, while making several approximations along the way. Similar approaches were applied successfully to quantify the asymptotic behaviour of the expected sojourn time in general M/GI/1 models; specifically, Bansal and Gamarnik [15] considered the M/GI/1 model for the Foreground-Background and pre-emptive Shortest Job First policies, and Lin et al. [95] studied the M/GI/1/SRPT model.

Heavy-traffic results that concern tail probabilities in non-FIFO models are scarce. If the scheduling policy is LIFO, then Abate and Whitt [5] show that the scaled probability $(1-\rho)^{-1}\mathbb{P}((1-\rho)^2 W_{\text{LIFO}} > y)$ converges to a non-degenerate function of y. Zhang and Zwart [143] show that the same scaling is appropriate if the scheduling policy is LPS. On the other hand, we recall that Kingman [83, 84] showed that $\mathbb{P}((1-\rho)W_{\text{FIFO}} > x) \sim e^{-\mathbb{E}[B_1^2]/(2\mathbb{E}[B])x}$ as $\rho \uparrow 1$ in GI/GI/1 models. This result may be surprising at first sight, as we just established that the large-deviations behaviour of $\mathbb{P}(W_{\text{FIFO}} > x)$ is regularly varying if the service-requirement distribution is regularly varying. This brings us to our final observations.

Transition between asymptotic regimes

If the service-requirement distribution is regularly varying, then the tail distribution of the waiting time behaves in fundamentally different ways depending on which regime is considered. This observation is reflected in relation (1.2), which decreases as function of x and (even though ρ is fixed) increases as a function of ρ. This is explained by the fact that the event $\{W_{\text{FIFO}} > x\}$ can be caused in different ways, with different associated probabilities.

For the large-deviations regime, where ρ is fixed and x is sufficiently large, the event of a long waiting time is mainly determined by the presence of a single customer with exceptional processing requirements, prior to one's own arrival. This behaviour is

caused by the heavy tail of the service-time distribution, and is known as the principle of a single big jump [101, 144]. Relation (1.2) reflects this intuition, since the expected number of jobs between two idle periods is of the order $(1-\rho)^{-1}$, and $x\mathbb{P}(B_1 > x)$ relates to the probability that a customer requires at least x more units of service before completion.

Alternatively, consider the heavy-traffic regime where x is fixed and ρ is sufficiently large. Then the probability of having to wait for x units of time is most likely caused by an accumulation of many waiting customers that all require a moderate amount of processing. With this many customers, the work that is still in the system can be shown to behave as the all-time supremum of a Brownian motion, which is exponentially distributed.

One may now wonder how robust the large-deviations and heavy-traffic limits are, i.e. how large does x need to be before the heavy tail of B is reflected in $\mathbb{P}(W_{\text{FIFO}} > x)$; or how large does ρ need to be in order for the heavy tail to vanish into a general diffusion process? More specifically, one may ask whether there exist functions $x_1(\rho)$ and $x_2(\rho)$ such that the large-deviations limit remains intact for all $x \geq x_1(\rho)$, and the heavy-traffic still remains true for all $x \leq x_2(\rho)$ as $\rho \uparrow 1$. Olvera-Cravioto et al. [107] gave a positive answer to this question and in fact showed that the transition from one regime to the other is sharp; i.e. $x_1(\rho)$ coincides with $x_2(\rho)$. This is contrasting with M/GI/1 models for a class of service-requirement distributions with slightly lighter tails, where there is an intermediate regime [108]. To the best of our knowledge, no results of this kind are known for other service disciplines.

1.5 Contributions of this dissertation

The previous section presented a rather diverse collection of results on scheduling policies in queueing models. Specifically, it was concerned with large-deviations and heavy-traffic behaviour of such policies, and concluded with a note on the robustness of asymptotic results if both x and ρ tend to their limiting values. The contributions in this dissertation are equally diverse, as will become apparent in this final section.

In Chapter 2, we aim to understand how well blind scheduling policies can perform compared to the optimal non-blind policy. In particular, we consider a GI/GI/1 model that operates under the Randomised Multilevel Feedback (RMLF) algorithm; a blind, pre-emptive and work-conserving algorithm that efficiently balances the number of pre-emptions and the prioritising of the jobs that have received the least service. We show that the expected sojourn time under the RMLF algorithm is at most a factor $c_4 \log(1/(1-\rho))$ larger than $\mathbb{E}[T_{\text{SRPT}}]$. Here, c_4 is a constant that may depend on the $(2+\varepsilon)$-th moment of the service-requirement distribution. Also, we show that this result is sharp in the sense that it cannot be improved by more than a multiplicative constant

for the M/M/1 model.

The importance of this result is two-fold. First, as opposed to the SRPT algorithm, the RMLF algorithm does not require knowledge of the processing requirements. Processing requirement information may be inaccurate or unavailable in real-life applications, thereby affecting the performance or even the applicability of SRPT. Second, the proof is based on a novel combination of techniques from queueing and scheduling theory. The proof independently analyses busy periods with less or more than N_0 jobs, exploits the $\log(n)$-competitiveness of RMLF to analyse the first type and employs applied probability techniques to inspect the second type. As the obtained bounds are quite loose, it is conceivable that a similar approach may yield results in other queueing models.

Chapter 3 intends to substantiate the latter suggestion in multi-server queueing models. We follow a straightforward, but naive approach to derive the queueing-theoretic equivalent of the competitive ratio for multi-server SRPT. Since the regenerative properties of the GI/GI/1 model – which were essential for exploiting the competitive ratio – do not easily translate to the GI/GI/c model, this approach concludes with a negative result. The chapter ends with a discussion on approaches that are potentially more successful.

In Chapter 4, our focus shifts from relative performance of algorithms to absolute performance. We derive the heavy-traffic behaviour of the expected sojourn time in a broad class of M/GI/1/FB models, and provide the reader with the intuition behind the technical derivations. Additionally, we show that $T_{\text{FB}}/\mathbb{E}[T_{\text{FB}}]$ converges to zero in probability and subsequently obtain the non-trivial heavy-traffic behaviour of the tail probability $\mathbb{P}((1-\rho)^2 T_{\text{FB}} > y)$ for fixed $y > 0$. Both the analyses and the results in this chapter depend on assumptions that are commonly encountered in extreme value theory. Also, the proof of the latter result exploits a fine connection between $(1-\rho)^2 T_{\text{FB}}$ and the supremum of a Lévy process that is stopped at an exponential time.

Lévy processes also form the basis of Chapter 5. The terminology in Chapter 5 transcends the more specified terminology of the earlier chapters, but its contributions translate to our understanding of M/GI/1 models in three ways. First, the described all-time supremum M_∞ is equivalent to the waiting time W_{FIFO}. As such, the uniform asymptotic presented in Theorem 5.3.1 is a "local" analogue of the large-deviations robustness result in Olvera-Cravioto et al. [107]. Second, and most importantly, is the large-deviations result concerning the busy period duration τ, Theorem 5.3.5. Since τ is equal in distribution to T_{LIFO}, this result may be interpreted as a large-deviations result for the LIFO scheduling algorithm. Finally, we provide a "local" analogue of Kingman's heavy-traffic approximation in Lemma 5.3.6. Among the techniques in this chapter are a sample-path analysis, and a derivation involving q-scale functions [90].

CHAPTER 2

ACHIEVABLE PERFORMANCE OF BLIND POLICIES IN HEAVY TRAFFIC

For a GI/GI/1 queueing model, we show that the expected sojourn time under the (blind) Randomised Multilevel Feedback algorithm is no worse than that under the Shortest Remaining Processing Time algorithm times a logarithmic function of the traffic intensity. Moreover, it is verified that this bound is tight in heavy traffic, up to a constant multiplicative factor. We obtain this result by means of a novel combination of techniques from competitive analysis and applied probability.

Based on Bansal et al. [S1].

2.1 Introduction

One of the most relevant and widely studied measures of quality of service in a GI/GI/1 queue is the expected sojourn time, also known as response time or flow time, defined as the expected time spent by a job from its arrival in the system until its completion [14–16, 19, 32, 95, 103, 106, 137–139]. We consider the most basic setting of a single machine with pre-emption, i.e. jobs can be interrupted arbitrarily and resumed later without any penalty. Schrage [125] showed that the Shortest Remaining Processing Time (SRPT) policy, that at any time works on the job with the least remaining processing time, is the optimal policy for every problem instance (or equivalently for every sample path) for minimising the expected sojourn time. However, SRPT can only be executed appropriately if all exact job sizes are known upon arrival. This information may not be available in many settings; specifically, jobs sizes may only be known approximately, or may not be known at all [98]. In such settings, one may have to be content with more generally applicable policies.

In this chapter we are interested in policies that do not require the knowledge of job sizes in their scheduling decisions. We refer to such policies as *blind* policies. More formally, in a blind policy the scheduler is only aware of the existence of a job and how much processing it has received thus far. The size of the job becomes known to the scheduler only when it terminates and leaves the system. Observe that the class of blind policies contains several well-studied policies, such as First In First Out [8], Foreground-Background [105] and Processor Sharing [86].

It is natural to ask how much this inability to use the knowledge of job sizes can hurt performance. In particular, how much can the expected sojourn time differ between SRPT and an optimal blind policy for a given GI/GI/1 queue? As an illustration, let us consider the M/M/1 queueing model. In this setting, all blind policies are identical due to the memoryless nature of the job-size distribution. More precisely, Conway et al. [40] state that any blind policy has an expected sojourn time equal to $\mathbb{E}[B]/(1-\rho)$, where $\mathbb{E}[B]$ is the average job size and ρ is the traffic intensity of the system. On the other hand, if job sizes are known upon arrival, then Bansal [14] derives that the expected sojourn time $\overline{T}_{\text{SRPT}}$ under M/M/1/SRPT is

$$\overline{T}_{\text{SRPT}} = (1+o(1))\frac{1}{\log\left(\frac{e}{1-\rho}\right)}\frac{\mathbb{E}[B]}{1-\rho}, \tag{2.1}$$

where $o(1)$ vanishes as ρ approaches one. That is, the SRPT policy outperforms all blind policies in M/M/1 models by a factor $\log(e/(1-\rho))$ in heavy traffic.

The performance of SRPT as a function of the traffic intensity can be dramatically different for heavy-tailed distributions. Bansal and Gamarnik [15] and Lin et al. [95] show that the growth factor of the expected sojourn time in heavy traffic can be much smaller than $1/(1-\rho)$ even in M/GI/1 models. For example, if the job sizes follow a

Pareto(β) distribution with $\beta \in (1,2)$, then the growth factor of the expected sojourn time $\overline{T}_{\text{SRPT}}$ is $\mathbb{E}[B]\log(1/(1-\rho))$, up to constant factors depending on β. On the other hand, Kleinrock [86] states that Processor Sharing has an expected sojourn time of $\mathbb{E}[B]/(1-\rho)$ in any M/GI/1 model. As this example illustrates, it is conceivable that for a general distribution, the gap between blind policies and SRPT can be much larger than in the M/M/1 case.

Another subfield of computer science where the performance improvement of SRPT over blind policies has been studied is *competitive analysis* [27, 56, 116], which generally regards worst case analyses of algorithms. The study of competitive analysis of blind scheduling policies was initiated by Motwani et al. [102], who showed that no blind deterministic algorithm[1] can have a better competitive ratio than $\Omega(m^{1/3})$ for the problem of minimising the expected sojourn time, where m is the number of jobs in an instance. Motwani et al. also showed that no blind randomised algorithm can have a competitive ratio better than $\Omega(\log(m))$.

In a breakthrough, Kalyanasundaram and Pruhs [76] gave an elegant and non-trivial randomised algorithm that they called Randomised Multilevel Feedback (RMLF) and proved that it has a competitive ratio of at most $O(\log(m)\log(\log(m)))$. Later, Becchetti and Leonardi [19] showed that RMLF is in fact an $O(\log(m))$-competitive randomised algorithm and hence the best possible (up to constant factors). Becchetti and Leonardi derive their result under the assumption that job sizes are bounded from below by a strictly positive constant, an assumption which is removed in this chapter. The resulting "extended" version of RMLF is denoted by eRMLF and introduced in Section 2.3.2. Additional background on multilevel algorithms can be found in Kleinrock [86], and an analysis of the expected sojourn time under such algorithms is performed in Aalto and Ayesta [4].

The insights from applied probability and competitive analysis concerning the relation between blind policies and SRPT can be combined when m is taken as the number of jobs in a regeneration cycle, which has an expected value of the order $1/(1-\rho)$. We make this precise in our main theorem and its proof. In Section 2.4, the main theorem shows that, for a GI/GI/1 queue, the gap between SRPT and the best blind policy $\mathcal{A}$ for that system is at most $\log(1/(1-\rho))$ up to a constant factor. More specifically, we show that this growth factor is a guaranteed upper bound on the gap between SRPT and the eRMLF algorithm. That is, we show that

$$\mathbb{E}[\overline{T}_{\mathcal{A}}] \leq \mathbb{E}[\overline{T}_{\text{eRMLF}}] = O\left(\log\left(\frac{1}{1-\rho}\right)\right) \cdot \overline{T}_{\text{SRPT}} \tag{2.2}$$

as ρ grows to one. Note that the eRMLF algorithm makes random decisions, and as such the outcome of $\overline{T}_{\text{eRMLF}}$ is a stochastic random variable for any given instance. The same

[1] Note that SRPT is deterministic, but not blind.

may hold for the optimal blind policy $\mathcal{A}$. Also, we emphasise that the implementation of the RMLF algorithm does not depend on the distributions of inter-arrival times and job sizes and is therefore applicable to every GI/GI/1 queue. The optimal blind policy $\mathcal{A}$ may not have this appealing property.

The second main contribution of this chapter is the proof of (2.2) itself. It involves a novel combination of techniques from competitive analysis and applied probability. Using a renewal argument, we consider the expected sojourn time $\mathbb{E}[\overline{T}_{\mathrm{RMLF}}]$ of jobs in a general busy period, and subsequently distinguish two types of busy periods (small and large) by the number of jobs. For small busy periods, we apply a worst-case performance bound of RMLF from the study of competitive analysis. For large busy periods, we derive the heavy-traffic behaviour of moments of two functionals: the busy period duration and the number of jobs in a busy period. In particular, we show that the κ-th moment of both of these functionals behaves like $O((1-\rho)^{1-2\kappa})$ for $\kappa \geq 1$. These new results are presented in Section 2.5.4 and may facilitate future instances where competitive analysis and regenerative process theory are combined to obtain information about algorithms under uncertainty. To prove these bounds, we rely on properties of ladder-height distributions derived in Asmussen [8] and Lotov [97].

This chapter is organised as follows. A detailed model description and notation are introduced in Section 2.2. Section 2.3 clarifies the concept of a competitive ratio and describes the RMLF algorithm. Additionally, Section 2.3.2 relaxes the constraints on RMLF while preserving the competitive ratio. The main result, Theorem 2.4.1, is presented in Section 2.4, whereas its proof is given in Section 2.5. Propositions required for the main theorem are proven in Section 2.6. Finally, Section 2.7 concludes the chapter.

2.2 Preliminaries

This section introduces a general framework for sequences of GI/GI/1 queueing models, so that we may analyse their limiting behaviour in further sections. In particular, the model allows for a heavy-traffic analysis of the expected sojourn time and various other functionals.

Sequence of queues

Consider a sequence of GI/GI/1 queueing systems, indexed by $n \geq 1$, where jobs arrive sequentially with independent and identically distributed (i.i.d.) sizes $B_i^{(n)}$, $i \in \{1,2,\ldots\}$, chosen from a distribution $F_B^{(n)}$. The jobs are then processed by a single server with unit speed. The times between two consecutive job arrivals are given by the i.i.d. inter-arrival times $A_i^{(n)}$, $i \in \{1,2,\ldots\}$, chosen from a distribution $F_A^{(n)}$. All job sizes and inter-arrival

times are assumed to be positive, i.e. the support of $F_A^{(n)}$ and $F_B^{(n)}$ is contained in $(0,\infty)$. For notational convenience, we define $A^{(n)} := A_1^{(n)}$ and $B^{(n)} := B_1^{(n)}$.

In order for every queueing system to be stable, we require $\mathbb{E}[A^{(n)}] > \mathbb{E}[B^{(n)}]$ for all $n \geq 1$. The traffic intensity of the n-th system is denoted by $\rho^{(n)} := \mathbb{E}[B^{(n)}]/\mathbb{E}[A^{(n)}] \in (0,1)$ and is interpreted as the fraction of time that the server is busy. As is customary in the literature on heavy-traffic analysis, we assume $\lim_{n\to\infty} \rho^{(n)} = 1$. The expected change in backlog between two consecutive arrivals is represented by $\mu^{(n)} := \mathbb{E}[A^{(n)}] - \mathbb{E}[B^{(n)}] = \mathbb{E}[A^{(n)}](1-\rho^{(n)})$.

Furthermore, we require that the inter-arrival times have finite variance for all n and additionally that $\limsup_{n\to\infty} \mathbb{E}[(A^{(n)})^2] < \infty$. Since a queue can only form when a job arrives to a non-empty system, we pose the final requirement that there exist two constants $\delta > 0$ and $\gamma > 0$, both independent of n, such that $\mathbb{P}(B^{(n)} - A^{(n)} \geq \delta) \geq \gamma$ is satisfied for all $n \geq 1$. The independence of n is a technical restriction that we exploit in Lemma 2.6.1.

Example model. In order to interpret some of our obtained results, one may compare them to a M/GI/1 queue that is sent into heavy traffic in a natural manner. Specifically, assume that both the A_i's and B_i's have unit mean and that the inter-arrival times in the r-th system are given by $A_i^{(r)} = A_i/r$, $i \in \{1,2,\ldots\}$, $r \in (0,1)$. This model experiences a traffic intensity of $\rho = \mathbb{E}[B]/\mathbb{E}[A^{(r)}] = r$ and is exposed to heavy traffic as r tends to one due to decreasing inter-arrival times. The model fits in the framework described above by letting $A_i^{(n)} = A_i/(1-1/n)$, $B_i^{(n)} = B_i$ and $\rho^{(n)} = 1-1/n$, and is referred to as the *Example Model*. All further references to the Example Model are recognised by superscripts r for all related variables and functionals.

Queueing functionals

The sojourn time of a job is the amount of time it spends in the system, i.e. the difference between its service completion time and its arrival time. Given a scheduling policy π, we denote the expected sojourn time of a generic job by $\mathbb{E}[\overline{T}_\pi^{(n)}]$ or just $\overline{T}_\pi^{(n)}$ if the scheduling policy is deterministic. The steady-state cumulative amount of work in the system is represented by $V^{(n)}$, whose distribution has an atom at zero that corresponds to the times when the server is idle. The steady-state duration of such an idle period is denoted by $I^{(n)}$.

Idle periods are ended by the arrival of a new job, which initiates a busy period. A busy period finishes at the earliest subsequent time for which the system is empty again. The steady-state duration of a busy period is represented by $P^{(n)}$, whereas the total number of arrivals between two subsequent idle periods is denoted by $N^{(n)}$. Finally, the steady-state cumulative amount of work in the system *at an arrival instance* is represented by $W^{(n)}$.

Scheduling policies

A scheduling policy π is an algorithm or a rule which specifies which job receives service at any time in the system. For the GI/GI/1 queue under consideration, such a policy prescribes the behaviour of a single server under the relaxation that jobs can be *pre-empted*; that is, jobs can be interrupted at any point during their execution and can be resumed later from this point without any penalty. Of the large class of scheduling policies that apply to this system, we consider only those policies π that satisfy the following two criteria (quoted from Wierman and Zwart [139], after Stolyar and Ramanan [129]):

1. π is *non-anticipative*: a scheduling decision at time t does not depend on information about jobs that arrive beyond time t.

2. π is *non-learning*: the scheduling decisions cannot depend on information about previous busy periods. That is, a scheduling decision on a sample path cannot change when the history before the current busy period is changed.

Of special interest are those scheduling policies π that additionally obey the following characteristic:

3. π is *blind*: the scheduling decisions do not depend on the sizes of the jobs. That is, the scheduling decisions on a sample path up to time t cannot change when the sizes of jobs that have not finished at that time are altered (in such a way that the jobs remain unfinished).

Policies that satisfy all above criteria are very common: First In First Out (FIFO), Foreground-Background (FB) and Processor Sharing are all blind policies within the specified subclass of scheduling policies. On the other hand, policies like Shortest Job First or Shortest Remaining Processing Time (SRPT) are *non-blind* elements of the specified subclass as they require knowledge of the job sizes when making a scheduling decision.

We let $\mathcal{A}^{(n)}$ denote a blind policy that minimizes the expected sojourn time over the space of all blind policies for the n-th GI/GI/1 queue. In general, $\mathcal{A}^{(n)}$ could depend on the distributions $F_A^{(n)}$ and $F_B^{(n)}$ that specify the GI/GI/1 queue. The implementation of the RMLF and eRMLF algorithms, which are respectively formalised in Sections 2.3.1 and 2.3.2, does not depend on $F_A^{(n)}$ and $F_B^{(n)}$ and is therefore independent of the system index n.

Finally, we call a scheduling policy π *work-conserving* if it always has the server working at unit speed whenever work is present in the system. One can easily verify that all above policies, including $\mathcal{A}^{(n)}$, are work-conserving.

Asymptotic relations

We use the standard notation that for two functions $f(n)$ and $g(n)$, $f(n) = O(g(n))$ and $f(n) = o(g(n))$ if $\limsup_{n\to\infty} f(n)/g(n) < \infty$ and $\limsup_{n\to\infty} f(n)/g(n) = 0$, respectively. Similarly, $f(n) = \Omega(g(n))$ means $\liminf_{n\to\infty} f(n)/g(n) > 0$ and $f(n) = \Theta(g(n))$ is equivalent to $0 < \liminf_{n\to\infty} f(n)/g(n) \le \limsup_{n\to\infty} f(n)/g(n) < \infty$.

This chapter's final notational conventions are the floor-function $\lfloor x \rfloor$, defined as $\sup\{m \in \mathbb{N} : m \le x\}$, and the indicator function $\mathbb{1}(\textit{[logical expression]})$ that assumes value 1 if the logical expression is true, and value 0 otherwise.

2.3 Competitive analysis of scheduling policies

In this section, we describe some relevant definitions and results from the area of competitive analysis, which deals with the worst case analysis of algorithms. We restrict our presentation here to the competitive analysis of scheduling algorithms with respect to expected sojourn time. Subsequently, we introduce the original RMLF algorithm and its extension eRMLF.

A scheduling problem instance $\mathcal{I}$ consists of a collection of jobs specified by their sizes and their arrival times. We say that an instance has size m, i.e. $|\mathcal{I}| = m$, if it consists of m jobs. For an instance $\mathcal{I}$, we denote the optimal expected sojourn time possible for this instance by $\overline{T}_{\mathrm{OPT}}(\mathcal{I})$, which for our purposes is the same as $\overline{T}_{\mathrm{SRPT}}(\mathcal{I})$.

For a deterministic algorithm π, we let $\overline{T}_{\pi}(\mathcal{I})$ denote the expected sojourn time when the instance $\mathcal{I}$ is executed according to the algorithm π. We say that the algorithm π has competitive ratio $c(m)$ if

$$\sup_{\mathcal{I}:|\mathcal{I}|\le m} \frac{\overline{T}_{\pi}(\mathcal{I})}{\overline{T}_{\mathrm{OPT}}(\mathcal{I})} \le c(m).$$

Thus, the competitive ratio of an algorithm (possibly a function of m) is the worst case ratio over all input instances of length at most m of the sojourn time achieved by π and the optimal sojourn time on that instance. Observe that the definition of the competitive ratio is rather strict, in that even if an algorithm is close to optimal on all but one input instance, its competitive ratio will be lower bounded by its performance on the bad input instance.

For this reason it is useful to consider randomised algorithms. A randomised algorithm $\widetilde{\pi}$ can toss coins internally and base its decisions on the outcome of these internal random variables. Such an algorithm can thus be interpreted as a random variable on a space of deterministic algorithms π_i [27]. It then follows that the expected sojourn time of instance $\mathcal{I}$ under a randomised algorithm $\widetilde{\pi}$ equals $\mathbb{E}[\overline{T}_{\widetilde{\pi}}(\mathcal{I})] = \mathbb{E}_i[\overline{T}_{\pi_i}(\mathcal{I})]$, where the expectation is over the internal random choices of the algorithm. We say that

$\tilde{\pi}$ has competitive ratio $c(m)$ if

$$\sup_{\mathcal{I}:|\mathcal{I}|\leq m} \frac{\mathbb{E}[\overline{T}_{\tilde{\pi}}(\mathcal{I})]}{\overline{T}_{\mathrm{OPT}}(\mathcal{I})} \leq c(m). \tag{2.3}$$

Observe that the expectation is only over the random choices made by the algorithm, and the competitive ratio is still determined by the worst possible instance. However, the competitive ratio of a blind randomised algorithm can be substantially lower, e.g. in models where no single blind deterministic algorithm is good for all instances, but a suitable combination of algorithms is close to optimal for all instances.

2.3.1 Randomised Multilevel Feedback algorithm

This section introduces Kalyanasundaram and Pruhs's Randomised Multilevel Feedback (RMLF) algorithm [76]. As the name suggests, it is a randomised version of the Multilevel Feedback (MLF) algorithm proposed by Corbató et al. [41]. Both algorithms are blind and can therefore only learn the size of a job upon completion.

The general idea of both MLF and RMLF is to prioritise potential short jobs (e.g. jobs that have not received much service) and reduce the priority of a job as it receives more service. This prioritisation is embodied by assigning every job J_j to a virtual high priority queue Q_i, and move it to a lower priority queue Q_{i+1} once it has received $U_{i,j}$ units of service. The performance of the algorithm may suffer from a poor choice of the so-called targets $U_{i,j}$; in particular, if the job sizes are slightly above their targets, then jobs are moved to lower priority queues just prior to completion. The improvement of RMLF over MLF is due to randomization of the targets, thereby reducing the possibility of such events over general instances.

We now provide a mathematical representation of the RMLF algorithm. Assume first that there is a universal lower bound on the job sizes in every instance $\mathcal{I}$, say with value 2. For every instance of size m, the j-th job J_j is released at time r_j and has size B_j. The process $w_j(t)$ denotes the amount of time that RMLF has run J_j before time t. For some symbolic constant θ, fixed at $\theta := 4/3$, we define the independent exponentially distributed variables β_j with $\mathbb{P}(\beta_j \leq x) = 1 - \exp[-\theta x \ln j]$. Finally, the targets are defined as $U_{i,j} = 2^i \max\{1, 2-\beta_j\}$ for all $i \in \{1,2,\ldots\}$, $j \in \{1,\ldots,m\}$. RMLF is then formalised in Figure 2.1, similar to Kalyanasundaram and Pruhs [76] and Becchetti and Leonardi [19].

Kalyanasundaram and Pruhs [76] proved that the RMLF algorithm has a competitive ratio of $O(\log(m)\log(\log(m)))$. This result was later strengthened by Becchetti and Leonardi [19] to a competitive ratio of $O(\log(m))$:

Theorem 2.3.1 (Becchetti and Leonardi [19])**.** *The* RMLF *algorithm is* $\log(m)$*-competitive. That is,*

$$\mathbb{E}[\overline{T}_{\mathrm{RMLF}}(\mathcal{I})] \leq C_1 \log(m) \overline{T}_{\mathrm{SRPT}}(\mathcal{I}) \tag{2.4}$$

Algorithm RMLF**:** At all times the collection of released, but uncompleted, jobs is partitioned into queues, $Q_0, Q_1, \ldots$ We say that Q_i is lower than Q_j for $i < j$. For each job $J_j \in Q_i, U_{i,j} \in [2^i, 2^{i+1}]$ when it entered Q_i. RMLF maintains the invariant that it is always running the earliest released job in the lowest non-empty queue.
When a job J_h is released at time r_h, RMLF takes the following actions:

- Job J_h is enqueued on Q_0.
- The target $U_{0,h}$ is set to $\max\{1, 2-\beta_h\}$.
- If, just prior to r_h, it was the case that Q_0 was empty, and that RMLF was running a job J_j, RMLF then takes the following actions:
  - Job J_j is pre-empted. Note that J_j remains at the front of its queue.
  - RMLF begins running J_h.

If at some time t, a job $J_j \in Q_{i-1}$ is being run when $w_j(t)$ becomes equal to $U_{i-1,j}$, then RMLF takes the following actions:

- Job J_j is dequeued from Q_{i-1}.
- Job J_j is enqueued on Q_i.
- The target $U_{i,j}$ is set to $2U_{i-1,j} = 2^i \max\{1, 2-\beta_j\}$.

Whenever a job is completed, it is removed from its queue.

Figure 2.1: Formal statement of RMLF algorithm.

for all instances $\mathcal{I}$ of size at most m and a universal constant $C_1 > 0$.

The competitive ratio lower bound of $\Omega(\log(m))$ as shown by Motwani et al. [102] implies that, up to multiplicative factors, this is the best bound possible for randomised algorithms in the current model. Note that this competitive ratio is significantly lower than the best possible ratio for blind deterministic algorithms: $\Omega(m^{1/3})$.

In the next section we propose a variant on RMLF that makes the assumption of a universal lower bound on job sizes obsolete.

2.3.2 Extending the RMLF algorithm

In a general GI/GI/1 queue there may not be a strictly positive lower bound on the job sizes. The RMLF algorithm is not directly applicable in that case. This problem is solved in an extension of the RMLF algorithm, which we will refer to as the eRMLF algorithm. The eRMLF algorithm defines queues $\widetilde{Q}_1, \widetilde{Q}_2, \ldots$ that are identical to the queues $Q_1, Q_2, \ldots$ of the RMLF algorithm, but splits the first queue Q_0 into many queues $\widetilde{Q}_0, \widetilde{Q}_{-1}, \ldots$ Additionally, it considers a "new job" queue $\widetilde{Q}^*$. The concept of the eRMLF algorithm is described below; the formal statement is presented in Appendix 2.A.

Let a problem instance $\widetilde{\mathcal{I}}$ for eRMLF be given. A target $\widetilde{U}_{*,j} = 2^{z_j^*} \max\{1, 2 - \widetilde{\beta}_j\}$ is assigned to every job $\widetilde{J}_j$ upon arrival, where $\widetilde{\beta}_j$ is an exponentially distributed random variable and $z_j^* \in \mathbb{Z}$ depends on the current state of the system. When the target has been assigned to the new job, it receives service in $\widetilde{Q}^*$ until either the job is completed, the obtained service equals the target, or a new job arrives. Once either of the latter two events happens, the job in $\widetilde{Q}^*$ is assigned to a queue $\widetilde{Q}_z, z \in \mathbb{Z}$.

If there are no jobs in queue $\widetilde{Q}^*$, the eRMLF algorithm serves the queues $\widetilde{Q}_z$ in a similar fashion as the RMLF algorithm. Moreover, at any time the problem instance $\widetilde{\mathcal{I}}$ can be converted to a problem instance $\mathcal{I}$ for RMLF by a scaling argument, and under this scaling the sojourn times of all jobs are identical for both algorithms. From this perspective, it is only natural that eRMLF inherits the competitive ratio of RMLF:

Theorem 2.3.2. *The* eRMLF *algorithm is* $\log(m)$*-competitive. That is,*

$$\mathbb{E}[\overline{T}_{\text{eRMLF}}(\mathcal{I})] \leq C_1 \log(m) \overline{T}_{\text{SRPT}}(\mathcal{I}) \tag{2.5}$$

for all instances $\mathcal{I}$ *of size at most* m *for a universal constant* $C_1 > 0$*. This constant is identical to the constant* C_1 *in Theorem 2.3.1.*

The proof of Theorem 2.3.2 is given in Appendix 2.A.

2.4 Main result and discussion

We are now ready to present the main result, Theorem 2.4.1. The main result states that the expected sojourn time under SRPT is at most a factor $\log(1/(1-\rho^{(n)}))$ better than that under eRMLF in heavy traffic:

Theorem 2.4.1. *For a* GI/GI/1 *queue, the* eRMLF *algorithm satisfies the relation*

$$\mathbb{E}[\overline{T}^{(n)}_{\text{eRMLF}}] = O\left(\log\left(\frac{1}{1-\rho^{(n)}}\right)\right) \cdot \overline{T}^{(n)}_{\text{SRPT}} \tag{2.6}$$

as $n \to \infty$*, provided that* $\sup_{n\in\{1,2,\ldots\}} \mathbb{E}[(B^{(n)})^\alpha] < \infty$ *for some* $\alpha > 2$.

The proof of the theorem is postponed until the next section. It relies on techniques from both competitive analysis and applied probability.

As a consequence of Theorem 2.4.1, the optimal blind policy $\mathcal{A}^{(n)}$ also satisfies the above performance bound. We emphasise the fact that the implementation of eRMLF does not depend on the inter-arrival and job-size distributions, whereas this may not be true for the optimal blind policy $\mathcal{A}^{(n)}$. This property may pose a considerable advantage over a system-dependent optimal blind policy with similar expected performance, for example when the input distributions are only approximately known. Also, we note that Theorem 2.4.1 remains true if eRMLF is replaced by RMLF, provided that the support of

the job-size distribution $F_B^{(n)}$ is uniformly bounded away from zero (i.e. $B_i^{(n)} \geq B_{\min}$ for some $B_{\min} > 0$ independent of i and n). We conclude this section with some remarks:

Remark 1. Recall that the expected sojourn time under any blind policy in an M/M/1 queue is $\mathbb{E}[B^{(n)}]/(1-\rho^{(n)})$, whereas the expected sojourn time under SRPT is [16]

$$\overline{T}^{(n)}_{\text{SRPT}} = (1+o(1))\frac{1}{\log\left(e/(1-\rho^{(n)})\right)}\frac{\mathbb{E}[B^{(n)}]}{1-\rho^{(n)}}. \tag{2.7}$$

In this case, our result is tight up to a multiplicative factor.

Remark 2. There may be sequences of GI/GI/1 queues for which $\mathbb{E}[\overline{T}^{(n)}_{\text{eRMLF}}]$ has a worse heavy-traffic scaling than $\mathbb{E}[\overline{T}^{(n)}_{\mathcal{A}^{(n)}}]$. For example, it is known that the FB policy minimizes the expected sojourn time over all blind policies in a M/GI/1 queue if $F_B^{(n)}$ has a decreasing failure rate [122]. Moreover, if $F_B^{(n)}(x) = 1 - x^{-\beta}, x \geq 1, \beta \in (1,\infty)/\{2\}$, then $\overline{T}^{(n)}_{\text{FB}} = \Theta(\overline{T}^{(n)}_{\text{SRPT}})$ displays the best possible scaling in heavy traffic [95, 105]. The heavy-traffic behaviour of $\mathbb{E}[\overline{T}^{(n)}_{\text{eRMLF}}]$ is unknown for any GI/GI/1 queue and could scale worse than $\overline{T}^{(n)}_{\text{FB}}$ (although no worse than $\log(1/(1-\rho)) \cdot \overline{T}^{(n)}_{\text{FB}}$ by Theorem 2.4.1).

On the other hand, the optimal blind policy $\mathcal{A}^{(n)}$ may not be robust under different input distributions $F_A^{(n)}$ and $F_B^{(n)}$. Continuing the FB example, we see that it is optimal if $F_B^{(n)}$ is the Pareto distribution, yet $\overline{T}^{(n)}_{\text{FB}} = \Theta((1-\rho)^{-2}) = \Theta((1-\rho)^{-1}) \cdot \overline{T}^{(n)}_{\text{SRPT}}$ if $F_B^{(n)}$ is deterministic [95, 105].

2.5 Proof of the main theorem

The current section presents the proof of Theorem 2.4.1.

2.5.1 Proof strategy

The competitive ratio of the eRMLF algorithm provides an upper bound on the suboptimality of eRMLF. Specifically, it guarantees an upper bound of $O(\log(m))$ on the ratio of the expected sojourn time under eRMLF over the expected sojourn time under SRPT, for instances of length at most m. Unfortunately, a general GI/GI/1 queue corresponds to an infinite-length problem instance and hence the competitive ratio result can not be applied directly.

The key idea of the proof is that a GI/GI/1 queue is a regenerative process, and as such one would like to analyse individual busy periods rather than the infinite problem instance. This approach is justified by the fact that for a single server, any two work-conserving scheduling policies π_1 and π_2 generate the same busy periods, i.e. $V^{(n)}_{\pi_1}(t) \equiv V^{(n)}_{\pi_2}(t)$. This means that the server is simultaneously active under both policies, and hence that every busy period under π_1 can be compared to the same busy period under π_2.

Still, regarding every busy period as an individual problem instance does not bound the problem instance length. One way to circumvent the unbounded problem instances is by discriminating between "small" busy periods with at most $N_0^{(n)}$ jobs, and "large" busy periods. Busy periods with at most $N_0^{(n)}$ jobs can be analysed with the competitive ratio, yielding a bound of $O(\log(N_0^{(n)}))$. This leaves us with the analysis of large busy periods.

Since the GI/GI/1 queue induces a distribution over problem instances, the probability of experiencing busy periods with at least $N_0^{(n)}$ jobs can be made arbitrarily small by choosing the threshold $N_0^{(n)}$ properly. The combined sojourn time of all the jobs in such a large busy period is dominated by the product of the number of jobs $N^{(n)}$ in the busy period and the duration $P^{(n)}$ of the busy period. Therefore, the contribution of large busy periods to the overall expected sojourn time is at most $\mathbb{E}[N^{(n)}P^{(n)}\mathbb{1}(N^{(n)} > N_0^{(n)})]/\mathbb{E}[N^{(n)}]$. We will show that, for an appropriate choice of $N_0^{(n)}$, the contribution of the large busy periods is $o(\log(N_0^{(n)}))$.

The second part of this section formalizes the above strategy. In the analysis of the expectation $\mathbb{E}[N^{(n)}P^{(n)}\mathbb{1}(N^{(n)} > N_0^{(n)})]$ we greatly rely on Hölder's inequality for decoupling the given expectation into individual moments of $P^{(n)}$ and $N^{(n)}$. The behaviour of these moments is then the subject of Propositions 2.5.1 and 2.5.2, both of which are proven in Section 2.6.

2.5.2 Small and large busy periods

We begin by specifying the threshold that distinguishes small and large busy periods based on the number of jobs. Fix $s \in \left(\frac{\alpha}{\alpha-1}, 2\right)$ and $\zeta > \frac{4+2s}{2-s}$. The threshold $N_0^{(n)}$ is now defined as $N_0^{(n)} := (1-\rho^{(n)})^{-\zeta}$.

Let $T^{(n)}_{\text{eRMLF},i}, T^{(n)}_{\text{SRPT},i}, i \in \{1,\dots,N^{(n)}\}$, be the sojourn time of job i under algorithm eRMLF and SRPT, respectively. Using the fact that a GI/GI/1 queue is a regenerative process, we only need to consider a general busy period when analysing the expected sojourn time [8, Theorem VI.1.2, Proposition X.1.3]:

$$\mathbb{E}[\overline{T}^{(n)}_{\text{eRMLF}}] = \frac{1}{\mathbb{E}[N^{(n)}]}\mathbb{E}\left[\sum_{i=1}^{N^{(n)}} T^{(n)}_{\text{eRMLF},i}\right]. \tag{2.8}$$

Discriminating between small and large busy periods then yields

$$\begin{aligned}\mathbb{E}[\overline{T}^{(n)}_{\text{eRMLF}}] = \frac{1}{\mathbb{E}[N^{(n)}]}\mathbb{E}\left[\sum_{i=1}^{N^{(n)}} T^{(n)}_{\text{eRMLF},i}\mathbb{1}(N^{(n)} \le N_0^{(n)})\right] \\ + \frac{1}{\mathbb{E}[N^{(n)}]}\mathbb{E}\left[\sum_{i=1}^{N^{(n)}} T^{(n)}_{\text{eRMLF},i}\mathbb{1}(N^{(n)} > N_0^{(n)})\right].\end{aligned} \tag{2.9}$$

As described in the strategy, we will bound the first term by means of the competitive ratio of eRMLF and show that the second term vanishes asymptotically as $n \to \infty$. These analyses are the subjects of the following two subsections.

2.5.3 Small busy periods: competitive ratio

The first term in (2.9) considers busy periods with at most $N_0^{(n)}$ jobs. Theorem 2.3.2 ensures that, for any problem instance $\mathcal{I}$ with $N^{(n)} \le N_0^{(n)}$ jobs, the expected sojourn time $\mathbb{E}[\overline{T}_{\mathrm{eRMLF}}(\mathcal{I})]$ is bounded by $C_1 \log(N_0^{(n)}) \overline{T}_{\mathrm{SRPT}}(\mathcal{I})$. In particular,

$$\begin{aligned}\frac{1}{\mathbb{E}[N^{(n)}]}\mathbb{E}\left[\sum_{i=1}^{N^{(n)}} T^{(n)}_{\mathrm{eRMLF},i}\mathbb{1}(N^{(n)} \le N_0^{(n)})\right] &\le \frac{C_1}{\mathbb{E}[N^{(n)}]}\log(N_0^{(n)})\mathbb{E}\left[\sum_{i=1}^{N^{(n)}} T^{(n)}_{\mathrm{SRPT},i}\mathbb{1}(N^{(n)} \le N_0^{(n)})\right] \\ &\le \frac{C_1}{\mathbb{E}[N^{(n)}]}\log(N_0^{(n)})\mathbb{E}\left[\sum_{i=1}^{N^{(n)}} T^{(n)}_{\mathrm{SRPT},i}\right] \\ &= C_1 \log(N_0^{(n)}) \cdot \overline{T}^{(n)}_{\mathrm{SRPT}}.\end{aligned}$$

The proof is complete once we show that the second term in (2.9) is dominated by $\log(N_0^{(n)})\overline{T}^{(n)}_{\mathrm{SRPT}}$ as $n \to \infty$.

2.5.4 Large busy periods: Hölders inequality

For any work-conserving scheduling policy, the sojourn time of an individual job is bounded by the duration $P^{(n)}$ of the busy period. Therefore, the second term in (2.9) is bounded by

$$\frac{1}{\mathbb{E}[N^{(n)}]}\mathbb{E}\left[\sum_{i=1}^{N^{(n)}} T^{(n)}_{\mathrm{eRMLF},i}\mathbb{1}(N^{(n)} > N_0^{(n)})\right] \le \frac{1}{\mathbb{E}[N^{(n)}]}\mathbb{E}\left[N^{(n)}P^{(n)}\mathbb{1}(N^{(n)} > N_0^{(n)})\right]. \tag{2.10}$$

The functionals $N^{(n)}$ and $P^{(n)}$ are dependent, which makes an exact analysis of the expectation troublesome. This complication is avoided by applying Hölder's inequality, which allows us to approximate the dependent expectation by the product of two expectations. In particular, for $\widetilde{s} = \frac{s}{s-1} \in (2, \alpha)$ we have $1/s + 1/\widetilde{s} = 1$ and hence

$$\mathbb{E}\left[\sum_{i=1}^{N^{(n)}} T^{(n)}_{\mathrm{eRMLF},i}\mathbb{1}(N^{(n)} > N_0^{(n)})\right] \le \mathbb{E}[(P^{(n)})^{\frac{s}{s-1}}]^{\frac{s-1}{s}}\mathbb{E}[(N^{(n)})^s\mathbb{1}(N^{(n)} > N_0^{(n)})]^{\frac{1}{s}}. \tag{2.11}$$

Applying Hölder's inequality once more with parameters $2/s$ and $2/(2-s)$, we get

$$\mathbb{E}\left[\sum_{i=1}^{N^{(n)}} T^{(n)}_{\mathrm{eRMLF},i}\mathbb{1}(N^{(n)} > N_0^{(n)})\right] \le \mathbb{E}[(P^{(n)})^{\frac{s}{s-1}}]^{\frac{s-1}{s}}\mathbb{E}[(N^{(n)})^2]^{\frac{1}{2}}\mathbb{P}(N^{(n)} > N_0^{(n)})^{\frac{2-s}{2s}}. \tag{2.12}$$

Finally, the tail probability of $N^{(n)}$ is bounded by Markov's inequality. We therefore obtain the following upper bound for the second term in (2.9):

$$\frac{1}{\mathbb{E}[N^{(n)}]}\mathbb{E}\left[\sum_{i=1}^{N^{(n)}} T^{(n)}_{\mathrm{eRMLF},i}\mathbb{1}(N^{(n)} > N_0^{(n)})\right] \leq \mathbb{E}[(P^{(n)})^{\frac{s}{s-1}}]^{\frac{s-1}{s}}\mathbb{E}[(N^{(n)})^2]^{\frac{1}{2}}\frac{\mathbb{E}[N^{(n)}]^{\frac{2-s}{2s}-1}}{(N_0^{(n)})^{\frac{2-s}{2s}}}. \tag{2.13}$$

The analysis of the expected sojourn time for large busy periods is now reduced to the analysis of moments of $N^{(n)}$ and $P^{(n)}$. The following two propositions quantify the behaviour of these moments.

Proposition 2.5.1. *Assume* $\sup_{n\in\{1,2,\ldots\}}\mathbb{E}[(B^{(n)})^\alpha] < \infty$ *for some* $\alpha \geq 2$. *Then*

$$\mathbb{E}[(P^{(n)})^\kappa] = O\left((1-\rho^{(n)})^{1-2\kappa}\right) \tag{2.14}$$

for all $\kappa \in [1,\alpha]$. *Moreover,* $\mathbb{E}[P^{(n)}] = \Theta\left((1-\rho^{(n)})^{-1}\right)$.

Proposition 2.5.2. *Assume* $\sup_{n\in\{1,2,\ldots\}}\mathbb{E}[(B^{(n)})^\alpha] < \infty$ *for some* $\alpha \geq 2$. *Then*

$$\mathbb{E}[(N^{(n)})^\kappa] = O\left((1-\rho^{(n)})^{1-2\kappa}\right) \tag{2.15}$$

for all $\kappa \in [1,\alpha]$. *Moreover,* $\mathbb{E}[N^{(n)}] = \Theta\left((1-\rho^{(n)})^{-1}\right)$.

Both propositions are proven in Section 2.6.

Remark 3. When applied to the Example Model, Proposition 2.5.1 states that $\mathbb{E}[(P^{(r)})^\kappa]$ is uniformly bounded from above by $C_2(1-r)^{1-2\kappa}$, for some constant $C_2 > 0$. Alternatively, the integer moments of the busy period duration in an M/GI/1 queue can be calculated explicitly from its Laplace-Stieltjes transform, yielding $\mathbb{E}[P^{(r)}] = \mathbb{E}[B](1-r)^{-1}$ and $\mathbb{E}[(P^{(r)})^2] = \mathbb{E}[B^2](1-r)^{-3}$. One may therefore conclude that the asymptotic behaviour of the bound in Proposition 2.5.1 is in fact sharp for the first two moments of the busy period duration $P^{(r)}$ in the Example Model.

From Propositions 2.5.1 and 2.5.2 it follows that, for some constant $C_3 > 0$,

$$\begin{aligned}\frac{1}{\mathbb{E}[N^{(n)}]}\mathbb{E}\left[P^{(n)}N^{(n)}\mathbb{1}(N^{(n)} > N_0^{(n)})\right] &\\ \leq C_3(1-\rho^{(n)})^{\left(1-2\frac{s}{s-1}\right)\frac{s-1}{s}}(1-\rho^{(n)})^{-\frac{3}{2}}(1-\rho^{(n)})^{1-\frac{2-s}{2s}}(1-\rho^{(n)})^{\frac{2-s}{2s}\zeta} &\\ = C_3(1-\rho^{(n)})^{-1-\frac{1}{s}}(1-\rho^{(n)})^{-\frac{3}{2}}(1-\rho^{(n)})^{\frac{3}{2}-\frac{1}{s}}(1-\rho^{(n)})^{\frac{2-s}{2s}\zeta} &\\ = C_3(1-\rho^{(n)})^{\frac{2-s}{2s}\zeta-\frac{2+s}{s}}. &\end{aligned} \tag{2.16}$$

Now, by choice of ζ, this expression tends to zero as $n \to \infty$. In particular, the contribution of large busy periods to the overall expected sojourn time is negligible compared to the contribution of small busy periods. This completes the proof of Theorem 2.4.1.

2.6 Moments of busy period functionals

This section proves several results on the moments of functionals. First, we introduce some new notation in Section 2.6.1. We then state two lemmas in Section 2.6.2 in order to prove Propositions 2.5.1 and 2.5.2. Subsequently, the propositions are proven in Sections 2.6.3 and 2.6.4. We emphasise that all of the functionals considered are independent of the scheduling policy, provided that it is work-conserving. This makes the results of this section applicable to a wide range of queueing models.

2.6.1 Counting and netput processes

For any non-negative random variable Y, we define a random variable Y_e that is distributed as the excess of Y; i.e. $\mathbb{P}(Y_e \leq x) := \int_0^x \mathbb{P}(Y > x)\,\mathrm{d}x/\mathbb{E}[Y]$. Next, we define two counting processes in the GI/GI/1 queue under consideration. The first process $N^{(n)}(t) := \inf\{m \in \{1,2,\ldots\} : A_1^{(n)} + \ldots + A_m^{(n)} \geq t\}$, $t \geq 0$, counts the number of arrivals in t time units, starting from a reference arrival that is also the first count. The second process $\widetilde{N}^{(n)}(t)$, $t \geq 0$, is similar and only differs by initialising the count at an arbitrary point in time. Specifically,

$$\widetilde{N}^{(n)}(t) := \begin{cases} 0 & \text{if } t < A_e^{(n)}, \\ \inf\{m \in \{1,2,\ldots\} : A_e^{(n)} + A_1^{(n)} + \ldots + A_m^{(n)} \geq t\} & \text{otherwise.} \end{cases}$$

We subsequently introduce the two netput processes $X^{(n)}(t) := \sum_{i=1}^{N^{(n)}(t)} B_i^{(n)} - t$ and $\widetilde{X}^{(n)}(t) := \sum_{i=1}^{\widetilde{N}^{(n)}(t)} B_i^{(n)} - t$, that quantify the net amount of work that could have been processed by the server in the t time units after an arrival, or respectively after an arbitrary point in time. Note that $X(t)$ becomes negative right after the first time that the queue is emptied. One may verify that $\mathbb{P}(\widetilde{X}^{(n)}(t) > x) \leq \mathbb{P}(X^{(n)}(t) > x)$ for all $t \geq 0$, which will be denoted by $\widetilde{X}^{(n)}(t) \leq_{st} X^{(n)}(t)$.

Similarly, we define two discrete processes that quantify the netput process *at an arrival instance*. The process $S_m^{(n)}$, $m \geq 0$, is defined as $S_0^{(n)} := 0$, $S_m^{(n)} := \sum_{i=1}^m [B_i^{(n)} - A_i^{(n)}]$ and quantifies all the work that has entered the system between the arrival of the reference job and the m-th next arrival, minus the work that it could have addressed during this time. The process $\widetilde{S}_m^{(n)}$, $m \geq 0$, instead starts observing at an arbitrary point in time and is defined as $\widetilde{S}_0^{(n)} := -A_e^{(n)}$, $\widetilde{S}_m^{(n)} := -A_e^{(n)} + \sum_{i=1}^m [B_i^{(n)} - A_i^{(n)}]$. Again, we obtain the relation $\widetilde{S}_m^{(n)} \leq_{st} S_m^{(n)}$.

One may additionally verify that $\sup_{t\geq 0} X^{(n)}(t) = \sup_{m\in\{1,2,\ldots\}} S_m^{(n)}$ and hence, by Asmussen [8, Corollary III.6.5], we have $\sup_{t\geq 0} X^{(n)}(t) = \sup_{m\in\{0,1,\ldots\}} S_m^{(n)} \stackrel{d}{=} W^{(n)}$ where $\cdot \stackrel{d}{=} \cdot$ denotes equality in distribution.

In the remainder of this chapter, all sums $\sum_{i=1}^0$ are understood to be zero.

2.6.2 Preliminary lemmas

We present and prove two lemmas that facilitate the proof of Propositions 2.5.1 and 2.5.2. Lemma 2.6.1 concerns the first moment of $N^{(n)}$ and $I^{(n)}$, whereas Lemma 2.6.2 considers general moments of $W^{(n)}$.

Lemma 2.6.1. *The relations*

$$(1-\rho^{(n)})\mathbb{E}[N^{(n)}] = \Theta(1) \text{ and } \mathbb{E}[I^{(n)}] = \Theta(1) \tag{2.17}$$

both hold as $n \to \infty$.

Proof of Lemma 2.6.1. Since we have $\mu^{(n)} := \mathbb{E}[A^{(n)}](1-\rho^{(n)}) = \Theta(1-\rho^{(n)})$, it suffices to prove the relation $\mu^{(n)}\mathbb{E}[N^{(n)}] = \Theta(1)$. Proposition X.3.1 in Asmussen [8], stating $\mathbb{E}[I^{(n)}] = \mu^{(n)}\mathbb{E}[N^{(n)}]$, then implies that this is equivalent to the relation $\mathbb{E}[I^{(n)}] = \Theta(1)$.

Both the upper and the lower bound follow from Lotov [97], who considers the ladder height of a random walk. Specifically, Lotov obtains upper bounds for the moments of the ladder epochs and the moments of overshoot over an arbitrary non-negative level if the expectation of jumps is positive and close to zero. As such, his results apply to the random walk $-S_m^{(n)}$ with ladder epochs $N^{(n)}$.

The upper bound is implied by Theorem 2 in Lotov [97]. This theorem states that $\mu^{(n)}\mathbb{E}[N^{(n)}] \le C_4$ for some constant $C_4 > 0$ and all n, provided that the supremum $\sup_{n\in\{1,2,\ldots\}} \mathbb{E}[(\max\{A^{(n)} - B^{(n)}, 0\})^2]$ is bounded. Accordance with this condition follows directly from $\sup_{n\in\{1,2,\ldots\}} \mathbb{E}[(A^{(n)})^2] < \infty$.

The lower bound is implied by inequality (2) in Lotov [97]. In our model, we assumed that there exist constants $\delta > 0, \gamma > 0$ such that $\mathbb{P}(B^{(n)} - A^{(n)} \ge \delta) \ge \gamma$ for all n. Lotov's inequality then states

$$\mu^{(n)}\mathbb{E}[N^{(n)}] \ge \int_0^\infty x \, \mathrm{d}\mathbb{P}(B^{(n)} - A^{(n)} \le x) \ge \delta\gamma$$

for all n. This completes the proof. □

Lemma 2.6.2. *Let* $p > 0$ *and define* $q := \max\{2, p+1\}$. *Assume* $\sup_{n\in\{1,2,\ldots\}} \mathbb{E}[(B^{(n)})^q] < \infty$. *Then*

$$\limsup_{n\to\infty} (1-\rho^{(n)})^p \mathbb{E}[(W^{(n)})^p] < \infty. \tag{2.18}$$

Remark 4. Consider the Example Model, and assume that jobs are served according to the FIFO discipline. Then $W^{(r)} = \overline{W}^{(r)}_{\mathrm{FIFO}}$ represents the waiting time of a generic job, and hence for some constant $C_5 > 0$ the expected sojourn time $\overline{T}^{(r)}_{\mathrm{FIFO}} = \overline{W}^{(r)}_{\mathrm{FIFO}} + \mathbb{E}[B] \le C_5(1-r)^{-1} + \mathbb{E}[B]$ scales no worse than $1/(1-r)$. Lemma 2.6.2 provides bounds on more general moments of the waiting time $\overline{W}^{(r)}_{\mathrm{FIFO}}$ provided that a sufficiently high moment of the job-size distribution exists.

Proof of Lemma 2.6.2. Since $\sup_{n\in\{1,2,\ldots\}} \mathbb{E}[A^{(n)}] < \infty$, relation (2.18) is equivalent to

$$\limsup_{n\to\infty} (\mu^{(n)})^p \mathbb{E}[(W^{(n)})^p] < \infty, \tag{2.19}$$

which is proven below.

Assume $p \geq 1$ and let $E^{(n)}, E_i^{(n)}$, $i \in \{1,2,\ldots\}$, be independent exponentially distributed random variables with mean

$$\mathbb{E}[E^{(n)}] = \frac{\mathbb{E}[A^{(n)}] + \mathbb{E}[B^{(n)}]}{2} < \mathbb{E}[A^{(n)}].$$

By Asmussen [8, Corollary III.6.5] and subadditivity of suprema, $W^{(n)}$ is upper bounded as

$$\begin{aligned} W^{(n)} &\stackrel{d}{=} \sup_{m\in\{0,1,2,\ldots\}} \sum_{i=1}^{m} \left[B_i^{(n)} - A_i^{(n)} \right] \\ &\leq \sup_{m\in\{0,1,2,\ldots\}} \sum_{i=1}^{m} \left[B_i^{(n)} - E_i^{(n)} \right] + \sup_{m\in\{0,1,2,\ldots\}} \sum_{i=1}^{m} \left[E_i^{(n)} - A_i^{(n)} \right] =: W_1^{(n)} + W_2^{(n)}, \end{aligned}$$

where $W_1^{(n)}$ can be interpreted as the total work in an M/GI/1 queue as observed by an arrival and $W_2^{(n)}$ as the total work in an GI/M/1 queue as observed by an arrival. As a consequence, $\mathbb{P}(W^{(n)} > x) \leq \mathbb{P}(W_1^{(n)} + W_2^{(n)} > x) \leq \mathbb{P}(W_1^{(n)} > x/2) + \mathbb{P}(W_2^{(n)} > x/2)$ and thus

$$\begin{aligned} \mathbb{E}[(W^{(n)})^p] &= p \int_0^\infty x^{p-1} \mathbb{P}(W^{(n)} > x)\, \mathrm{d}x \\ &\leq p \int_0^\infty x^{p-1} \mathbb{P}(W_1^{(n)} > x/2)\, \mathrm{d}x + p \int_0^\infty x^{p-1} \mathbb{P}(W_2^{(n)} > x/2)\, \mathrm{d}x \\ &= 2^p \left(\mathbb{E}[(W_1^{(n)})^p] + \mathbb{E}[(W_2^{(n)})^p] \right). \end{aligned}$$

First, we consider $W_1^{(n)}$. Define the geometrically distributed random variable $K_1^{(n)}$ with support $\{0,1,\ldots\}$ and fail parameter

$$\xi_1^{(n)} := \frac{\mathbb{E}[B^{(n)}]}{\mathbb{E}[E^{(n)}]} = \frac{2\mathbb{E}[B^{(n)}]}{\mathbb{E}[A^{(n)}] + \mathbb{E}[B^{(n)}]}.$$

For notational convenience, we drop the superscript (n) of $\xi_1^{(n)}$ for the remainder of this section. Theorem VIII.5.7 in Asmussen [8] presents a random sum representation of the functional $W_1^{(n)}$ in terms of $K_1^{(n)}$ and $B_{e,i}^{(n)}$:

$$W_1^{(n)} \stackrel{d}{=} \sum_{i=1}^{K_1^{(n)}} B_{e,i}^{(n)}.$$

Since $f(x) = x^p$ is a convex function for all $p \geq 1$, Lemma 5 in Remerova et al. [118] implies

$$\mathbb{E}[(W_1^{(n)})^p] \leq \mathbb{E}[(K_1^{(n)})^p] \mathbb{E}[(B_e^{(n)})^p]. \tag{2.20}$$

The conditions of Lemma 2.6.2 ensure that the p-th moment of $B_e^{(n)}$ is finite as $n \to \infty$:

$$\begin{aligned}\mathbb{E}[(B_e^{(n)})^p] &= \int_0^\infty x^p \,\mathrm{d}\mathbb{P}(B_e^{(n)} \le x) = \frac{1}{\mathbb{E}[B^{(n)}]} \int_0^\infty x^p \mathbb{P}(B^{(n)} > x)\,\mathrm{d}x \\ &= \frac{1}{(p+1)\mathbb{E}[B^{(n)}]} \int_0^\infty x^{p+1} \,\mathrm{d}\mathbb{P}(B^{(n)} \le x) = \frac{1}{(p+1)\mathbb{E}[B^{(n)}]} \mathbb{E}[(B^{(n)})^{p+1}]. \quad (2.21)\end{aligned}$$

Therefore, we need to show that $(\mu^{(n)})^p \mathbb{E}[(K_1^{(n)})^p]$ is uniformly bounded as $n \to \infty$. Let $k = \lfloor p \rfloor$. Then

$$\begin{aligned}\mathbb{E}[(K_1^{(n)})^p] &= \frac{1-\xi_1}{(1-\xi_1)^p} \sum_{m=0}^{\infty} ((1-\xi_1)m)^p \xi_1^m \\ &\le \frac{1-\xi_1}{(1-\xi_1)^p} \sum_{m=0}^{\left\lfloor \frac{1}{1-\xi_1} \right\rfloor} ((1-\xi_1)m)^k \xi_1^m + \frac{1-\xi_1}{(1-\xi_1)^p} \sum_{m=\left\lfloor \frac{1}{1-\xi_1} \right\rfloor + 1}^{\infty} ((1-\xi_1)m)^{k+1} \xi_1^m \\ &\le \frac{(1-\xi_1)^{k+1}}{(1-\xi_1)^p} \sum_{m=0}^{\infty} m^k \xi_1^m + \frac{(1-\xi_1)^{k+2}}{(1-\xi_1)^p} \sum_{m=0}^{\infty} m^{k+1} \xi_1^m. \quad (2.22)\end{aligned}$$

On the one hand, for any $\ell \in \{1, 2, \ldots\}$, we have

$$(1-\xi_1)^{\ell+1} \xi_1^\ell \frac{\mathrm{d}^\ell}{\mathrm{d}\xi_1^\ell} \sum_{m=0}^{\infty} \xi_1^m = (1-\xi_1)^{\ell+1} \xi_1^\ell \frac{\mathrm{d}^\ell}{\mathrm{d}\xi_1^\ell} (1-\xi_1)^{-1} = \ell! \xi_1^\ell. \quad (2.23)$$

On the other hand we have

$$\begin{aligned}(1-\xi_1)^{\ell+1} \xi_1^\ell \frac{\mathrm{d}^\ell}{\mathrm{d}\xi_1^\ell} \sum_{m=0}^{\infty} \xi_1^m &= (1-\xi_1)^{\ell+1} \sum_{m=\ell}^{\infty} m(m-1)\cdots(m-\ell+1)\xi_1^m \\ &= (1-\xi_1)^{\ell+1} \sum_{m=0}^{\infty} m^\ell \xi_1^m - (1-\xi_1)^{\ell+1} \sum_{m=0}^{\ell-1} m^\ell \xi_1^m \\ &\quad + (1-\xi_1)^{\ell+1} \sum_{m=\ell}^{\infty} o(m^\ell) \xi_1^m. \quad (2.24)\end{aligned}$$

Combining equalities (2.23) and (2.24), we find that

$$(1-\xi_1)^{\ell+1} \sum_{m=0}^{\infty} m^\ell \xi_1^m = \ell! \xi_1^\ell + (1-\xi_1)^{\ell+1} \sum_{m=0}^{\ell-1} m^\ell \xi_1^m + (1-\xi_1)^{\ell+1} \sum_{m=\ell}^{\infty} o(m^\ell) \xi_1^m.$$

Now, for any $\nu > 0$ there exists a $M_\nu \in \{1, 2, \ldots\}$ independent of the system index n such that for all $m \ge M_\nu$ the $o(m^\ell)$ term is dominated by νm^ℓ. Fix such $\nu \in (0,1)$ and M_ν. Then

$$\begin{aligned}(1-\xi_1)^{\ell+1} \sum_{m=0}^{\infty} m^\ell \xi_1^m &\le \ell! + \ell^{\ell+1} + (1-\xi_1)^{\ell+1} \nu \sum_{m=M_\nu}^{\infty} m^\ell \xi_1^m + (1-\xi_1)^{\ell+1} \sum_{m=0}^{M_\nu} o(m^\ell) \xi_1^m \\ &\le C_6 + (1-\xi_1)^{\ell+1} \nu \sum_{m=0}^{\infty} m^\ell \xi_1^m\end{aligned}$$

for some constant $C_6 > 0$, and hence

$$(1-\xi_1)^{\ell+1} \sum_{m=0}^{\infty} m^{\ell} \xi_1^m \leq \frac{C_6}{1-\nu}. \tag{2.25}$$

Since $(\mu^{(n)}/(1-\xi_1))^p = (\mathbb{E}[A^{(n)}] + \mathbb{E}[B^{(n)}])^p$, we may conclude from relations (2.22) and (2.25) that $(\mu^{(n)})^p \mathbb{E}[(K_1^{(n)})^p]$ is uniformly bounded from above as $n \to \infty$ and so is $(\mu^{(n)})^p \mathbb{E}[(W_1^{(n)})^p]$ by (2.20).

Second, we consider the functional $W_2^{(n)}$. Recall that $W_2^{(n)}$ denotes the steady-state workload in a GI/M/1 queue upon arrival. Theorem VIII.5.8 and page 296 in Asmussen [8] together state that

$$W_2^{(n)} \stackrel{d}{=} \sum_{i=1}^{K_2^{(n)}} E_i^{(n)}, \tag{2.26}$$

where $K_2^{(n)}$ is a geometrically distributed random variable with support $\{0,1,\ldots\}$ and unknown fail parameter $\xi_2^{(n)}$. Remerova et al. [118] again ensure that

$$\mathbb{E}[(W_2^{(n)})^p] \leq \mathbb{E}[(K_2^{(n)})^p] \mathbb{E}[(E^{(n)})^p], \tag{2.27}$$

where the latter expectation is finite uniformly in n as a property of exponential distributions. The p-th moment of $K_2^{(n)}$ is bounded by (2.22) and (2.25), so that

$$(\mu^{(n)})^p \mathbb{E}[(K_2^{(n)})^p] \leq C_7 \left(\frac{\mu^{(n)}}{1-\xi_2^{(n)}} \right)^p \tag{2.28}$$

for some constant $C_7 > 0$.

The proof is complete once we show $\frac{\mu^{(n)}}{1-\xi_2^{(n)}} = O(1)$. One may deduce from (2.26) that $\mathbb{P}(W_2^{(n)} = 0) = \mathbb{P}(K_2^{(n)} = 0) = 1 - \xi_2^{(n)}$. Additionally, by Theorem VIII.2.3 in Asmussen [8], we have $\mathbb{P}(W_2^{(n)} = 0) = 1/\mathbb{E}[N_2^{(n)}]$ and hence $\mathbb{E}[N_2^{(n)}] = 1/(1-\xi_2^{(n)})$. Here, $N_2^{(n)}$ is the steady-state number of jobs in a busy period of the GI/M/1 queue. Applying Lemma 2.6.1 with inter-arrival times $A_i^{(n)}$, job sizes $E_i^{(n)}$, and expected change in backlog between two consecutive arrivals $\frac{1}{2}\mu^{(n)}$ yields $\frac{1}{2}\mu^{(n)}\mathbb{E}[N_2^{(n)}] = \Theta(1)$, and therefore $\mu^{(n)}/(1-\xi_2^{(n)}) = O(1)$.

Finally, for $0 < p < 1$ the lemma follows directly from the case $p = 1$ after observing that $(\mu^{(n)})^p \mathbb{E}[(W^{(n)})^p] \leq \left(\mu^{(n)} \mathbb{E}[W^{(n)}]\right)^p$ by Jensen's inequality. □

Lemmas 2.6.1 and 2.6.2 provide the asymptotic behaviour of functionals that are closely related to $P^{(n)}$ and $N^{(n)}$. The remainder of this section utilizes these results in order to prove Propositions 2.5.1 and 2.5.2.

2.6.3 Busy period duration $P^{(n)}$

This section is devoted to the proof of Proposition 2.5.1. We wish to show that

$$\mathbb{E}[(P^{(n)})^{\kappa}] = O\left((1-\rho^{(n)})^{1-2\kappa}\right) \tag{2.14, revisited}$$

for all $\kappa \in [1,\alpha]$, provided that $\sup_{n\in\{1,2,\ldots\}} \mathbb{E}[(B^{(n)})^{\alpha}] < \infty$ for some $\alpha \geq 2$. Moreover, we claim that $\mathbb{E}[P^{(n)}] = \Theta\big((1-\rho^{(n)})^{-1}\big)$.

First, consider $\kappa = 1$. Due to Little's law for a busy server, we have

$$1-\rho^{(n)} = \frac{\mathbb{E}[I^{(n)}]}{\mathbb{E}[I^{(n)}]+\mathbb{E}[P^{(n)}]}, \tag{2.29}$$

so that

$$\mathbb{E}[P^{(n)}] = \frac{\rho^{(n)}\mathbb{E}[I^{(n)}]}{1-\rho^{(n)}}. \tag{2.30}$$

The result now follows from Lemma 2.6.1.

Second, consider $\kappa > 1$. Similar to (2.21), one obtains $\mathbb{E}[(P_e^{(n)})^{\kappa-1}] = \frac{\mathbb{E}[(P^{(n)})^{\kappa}]}{\kappa\mathbb{E}[P^{(n)}]}$ and hence it suffices to show that $\mathbb{E}[(P_e^{(n)})^{\kappa-1}] = O((1-\rho^{(n)})^{2(1-\kappa)})$. We have the following convenient representation for $P_e^{(n)}$ [8, Theorem X.3.4]:

$$P_e^{(n)} \stackrel{d}{=} \inf\{\tau \geq 0 : \widetilde{X}^{(n)}(\tau) \leq -V^{(n)} \mid V^{(n)} > 0\} \stackrel{d}{=} \inf\{\tau \geq 0 : B_e^{(n)} + W^{(n)} + \widetilde{X}^{(n)}(\tau) \leq 0\}.$$

This representation allows us to bound $\mathbb{P}(P_e^{(n)} > t)$ as

$$\begin{aligned}
\mathbb{P}(P_e^{(n)} > t) &= \mathbb{P}(\inf\{\tau \geq 0 : B_e^{(n)} + W^{(n)} + \widetilde{X}^{(n)}(\tau) \leq 0\} > t) \\
&= \mathbb{P}(B_e^{(n)} + W^{(n)} + \widetilde{X}^{(n)}(\tau) > 0, \forall \tau \leq t) \\
&\leq \mathbb{P}(B_e^{(n)} + W^{(n)} + \widetilde{X}^{(n)}(t) + (1-\rho^{(n)})t/2 > (1-\rho^{(n)})t/2) \\
&\leq \mathbb{P}(B_e^{(n)} > (1-\rho^{(n)})t/6) + \mathbb{P}(W^{(n)} > (1-\rho^{(n)})t/6) \\
&\quad + \mathbb{P}\left(\sup_{\tau\geq 0}[\widetilde{X}^{(n)}(\tau) + (1-\rho^{(n)})\tau/2] > (1-\rho^{(n)})t/6\right).
\end{aligned}$$

In Section 2.6.1 we derived the relations $\widetilde{X}^{(n)}(\tau) \leq_{st} X^{(n)}(\tau)$ and $W^{(n)} \stackrel{d}{=} \sup_{\tau\geq 0} X^{(n)}(\tau)$, which now imply

$$\mathbb{P}(P_e^{(n)} > t) \leq \mathbb{P}(B_e^{(n)} > (1-\rho^{(n)})t/6) + 2\mathbb{P}\left(\sup_{\tau\geq 0}[X^{(n)}(\tau) + (1-\rho^{(n)})\tau/2] > (1-\rho^{(n)})t/6\right). \tag{2.31}$$

Consequently,

$$\begin{aligned}
\mathbb{E}[(P_e^{(n)})^{\kappa-1}] &= (\kappa-1)\int_0^{\infty} t^{\kappa-2}\mathbb{P}(P_e^{(n)} > t)\,\mathrm{d}t \\
&\leq (\kappa-1)\int_0^{\infty} t^{\kappa-2}\mathbb{P}(B_e^{(n)} > (1-\rho^{(n)})t/6)\,\mathrm{d}t \\
&\quad + 2(\kappa-1)\int_0^{\infty} t^{\kappa-2}\mathbb{P}\left(\sup_{\tau\geq 0}[X^{(n)}(\tau) + (1-\rho^{(n)})\tau/2] > (1-\rho^{(n)})t/6\right)\mathrm{d}t.
\end{aligned}$$

To deal with the first term, note that

$$\begin{aligned}
(\kappa-1)\int_0^{\infty} t^{\kappa-2}\mathbb{P}(B_e^{(n)} > (1-\rho^{(n)})t/6)\,\mathrm{d}t &= (\kappa-1)\int_0^{\infty} t^{\kappa-2}\mathbb{P}(6B_e^{(n)}/(1-\rho^{(n)}) > t)\,\mathrm{d}t \\
&= \mathbb{E}\left[\left(\frac{6B_e^{(n)}}{1-\rho^{(n)}}\right)^{\kappa-1}\right] = O\big((1-\rho^{(n)})^{1-\kappa}\big)\mathbb{E}[(B_e^{(n)})^{\kappa-1}]
\end{aligned}$$

since $\mathbb{E}[(B^{(n)})^\kappa] < \infty$ implies $\mathbb{E}[(B_e^{(n)})^{\kappa-1}] < \infty$ (cf. relation (2.21)). Consequently, the first term is of order $o\big((1-\rho^{(n)})^{2(1-\kappa)}\big)$.

For the second term, observe that

$$\begin{aligned}\sup_{\tau\geq 0}[X^{(n)}(\tau)+(1-\rho^{(n)})\tau/2] &= \sup_{\tau\geq 0}\left[\sum_{i=1}^{N^{(n)}(\tau)} B_i^{(n)} - \tau + (1-\rho^{(n)})\tau/2\right]\\ &= \sup_{\tau\geq 0}\left[\sum_{i=1}^{N^{(n)}(\tau)} B_i^{(n)} - \frac{1+\rho^{(n)}}{2}\tau\right] = \sup_{\eta\in\{0,1,2,\ldots\}}\left[\sum_{i=1}^{\eta}\left\{B_i^{(n)} - \frac{1+\rho^{(n)}}{2}A_i^{(n)}\right\}\right]\\ &=: \widetilde{W}^{(n)},\end{aligned}$$

where $\widetilde{W}^{(n)}$ is equal in distribution to the steady-state cumulative amount of work at an arrival moment in a GI/GI/1 queue with inter-arrival times $\frac{1+\rho^{(n)}}{2}A_i^{(n)}$ and job sizes $B_i^{(n)}$. Furthermore, the expected change in backlog between two consecutive arrivals in this system is given by $\widetilde{\mu}^{(n)} := \frac{1+\rho^{(n)}}{2}\mathbb{E}[A^{(n)}] - \mathbb{E}[B^{(n)}] = \frac{1-\rho^{(n)}}{2}\mathbb{E}[A^{(n)}]$. We therefore obtain

$$\begin{aligned}&(\kappa-1)\int_0^\infty t^{\kappa-2}\mathbb{P}\left(\sup_{\tau\geq 0}[X^{(n)}(\tau)+(1-\rho^{(n)})\tau/2] > (1-\rho^{(n)})t/6\right)\mathrm{d}t\\ &\quad= (\kappa-1)\int_0^\infty t^{\kappa-2}\mathbb{P}\left(\frac{6}{1-\rho^{(n)}}\widetilde{W}^{(n)} > t\right)\mathrm{d}t = \frac{6^{\kappa-1}}{(1-\rho^{(n)})^{\kappa-1}(\widetilde{\mu}^{(n)})^{\kappa-1}}\mathbb{E}[(\widetilde{\mu}^{(n)}\widetilde{W}^{(n)})^{\kappa-1}]\\ &\quad\leq \frac{C_8}{(1-\rho^{(n)})^{2(\kappa-1)}}\mathbb{E}[(\widetilde{\mu}^{(n)}\widetilde{W}^{(n)})^{\kappa-1}]\end{aligned}$$

for some constant $C_8 > 0$. Finally, $\mathbb{E}[(\widetilde{\mu}^{(n)}\widetilde{W}^{(n)})^{\kappa-1}]$ is bounded due to Lemma 2.6.2, which completes the proof of the proposition.

2.6.4 Arrivals in a busy period $N^{(n)}$

This section contains the proof of Proposition 2.5.2. The proposition states that

$$\mathbb{E}[(N^{(n)})^\kappa] = O\big((1-\rho^{(n)})^{1-2\kappa}\big) \qquad \text{(2.15, revisited)}$$

for all $\kappa \in [1,\alpha]$, provided that $\sup_{n\in\{1,2,\ldots\}}\mathbb{E}[(B^{(n)})^\alpha] < \infty$ for some $\alpha \geq 2$. Moreover, we claim that $\mathbb{E}[N^{(n)}] = \Theta\big((1-\rho^{(n)})^{-1}\big)$.

The structure of the proof is identical to the proof of Proposition 2.5.1; in particular, if $\kappa = 1$ then the result follows directly from Lemma 2.6.1. Therefore, we consider $\mathbb{E}[(N^{(n)})^\kappa]$ for $\kappa > 1$ and exploit the relations $\mathbb{E}[(N_e^{(n)})^{\kappa-1}] = \mathbb{E}[(N^{(n)})^\kappa]/(\kappa\mathbb{E}[N^{(n)}])$ and

$$\begin{aligned}N_e^{(n)} &\overset{d}{=} \inf\{\eta\in\{0,1,2,\ldots\} : \widetilde{S}_\eta^{(n)} \leq -V^{(n)} \mid V^{(n)} > 0\}\\ &\overset{d}{=} \inf\{\eta\in\{0,1,2,\ldots\} : B_e^{(n)} + W^{(n)} + \widetilde{S}_\eta^{(n)} \leq 0\}.\end{aligned}$$

As before, the relations $\widetilde{S}_\eta^{(n)} \leq_{st} S_\eta^{(n)}$ and $W^{(n)} \stackrel{d}{=} \sup_{\eta\in\{0,1,\ldots\}} S_\eta^{(n)}$ are exploited in order to obtain an equivalent of (2.31):

$$\mathbb{P}(N_e^{(n)} > m) \leq \mathbb{P}(B_e^{(n)} > (1-\rho^{(n)})m/6) + 2\mathbb{P}\left(\sup_{\eta\in\{0,1,2,\ldots\}} [S_\eta^{(n)} + (1-\rho^{(n)})\eta/2] > (1-\rho^{(n)})m/6\right).$$

For any $\eta \in \{0,1,2,\ldots\}$, let $\tau_\eta^{(n)} := A_1^{(n)} + \ldots + A_\eta^{(n)}$. Then $S_\eta^{(n)} = X^{(n)}(\tau_\eta^{(n)})$, so in particular

$$\mathbb{P}(N_e^{(n)} > m) \leq \mathbb{P}(B_e^{(n)} > (1-\rho^{(n)})m/6) + 2\mathbb{P}\left(\sup_{\tau\geq 0}[X^{(n)}(\tau) + (1-\rho^{(n)})\tau/2] > (1-\rho^{(n)})m/6\right).$$

The remainder of the proof is identical to that of Proposition 2.5.1.

2.7 Conclusion

In this chapter, we proved a result about the expected performance of (an extension of) the Randomised Multilevel Feedback (RMLF) algorithm in a GI/GI/1 queue. Specifically, the gap in expected sojourn time between the RMLF algorithm and the Shortest Remaining Processing Time algorithm behaves like $O(\log(1/(1-\rho^{(n)})))$ and this bound is tight for the M/M/1 queue. An appealing property of the RMLF algorithm is that its implementation does not depend on the input distributions $F_A^{(n)}$ and $F_B^{(n)}$; however, if $F_A^{(n)}$ and $F_B^{(n)}$ are known then there can be blind algorithms with a better performance than RMLF (e.g. Foreground-Background if $F_B^{(n)}$ has decreasing failure rate). The result was established by using techniques from both competitive analysis and applied probability. As the structure of the proof is quite general, it would be interesting to explore other possibilities in the intersection of these areas.

2.A The eRMLF algorithm

The eRMLF algorithm is presented after the introduction of some notation. Define the virtual queues $\widetilde{Q}_z$, $z \in \mathbb{Z}$, and a "new job" queue $\widetilde{Q}^*$. Let the targets $\widetilde{U}_{z,j}$ be given by $\widetilde{U}_{z,j} = 2^z \max\{1, 2-\widetilde{\beta}_j\}$, where the $\widetilde{\beta}_j$'s are independent random variables with exponential cumulative distribution function $\mathbb{P}(\widetilde{\beta}_j \leq x) = 1 - \exp[-\theta x \ln j]$. Similar to the RMLF algorithm, θ is a symbolic constant fixed at $\theta = 4/3$. All symbols $\widetilde{J}_j, \widetilde{r}_j, \widetilde{B}_j$ and $\widetilde{w}_j(t)$ are defined analogously to the symbols without accent in the RMLF algorithm. All release times $\widetilde{r}_j$ must be distinct (e.g. all inter-arrival times are strictly positive), and jobs may have any size $\widetilde{B}_j \geq 0$. Note that the original RMLF algorithm requires the job sizes to be uniformly bounded from below, but does not restrict the inter-arrival times to be non-zero.

Every job $\widetilde{J}_h$ is assigned an initial target $\widetilde{U}_{*,h}$ upon arrival, after which it is immediately served in $\widetilde{Q}^*$ by a dedicated server. It departs from $\widetilde{Q}^*$ on three occasions:

- The amount of service received equals the size $\widetilde{B}_h$ of the job. In this case, $\widetilde{J}_h$ is completed and leaves the system.
- A new job enters the system. In this case, $\widetilde{J}_h$ is moved to a queue $\widetilde{Q}_z$, $z \in \mathbb{Z}$, that it naturally belongs to based on the amount of service $\widetilde{w}_h(t)$ it has obtained thus far; that is, it is moved to the unique queue $\widetilde{Q}_{z_h^*}$ that satisfies $\widetilde{U}_{z_h^*-1,h} \le \widetilde{w}_h(t) < \widetilde{U}_{z_h^*,h}$.
- The amount of service received equals the initial target $\widetilde{U}_{*,h}$. In this case, similar to the previous case, $\widetilde{J}_h$ is moved to a queue $\widetilde{Q}_z$ that it naturally belongs to.

The choice of the initial target $\widetilde{U}_{*,h}$ depends on the system state:

- If the system is empty upon arrival, then the server is dedicated to $\widetilde{J}_h$ regardless of the queue that $\widetilde{J}_h$ is in. In this case, the target can be chosen arbitrarily; we set it to $\widetilde{U}_{*,h} = \widetilde{U}_{0,h}$.
- If the system is not empty upon arrival, then there must be a lowest-index non-empty queue $\widetilde{Q}_{z_h^*}$ (possibly after moving the job originally in $\widetilde{Q}^*$ to another queue). $\widetilde{J}_h$ may now experience a dedicated server until the moment when it would enter queue $\widetilde{Q}_{z_h^*}$ based on its obtained service and the (z_h^*-1)-th target $\widetilde{U}_{z_h^*-1,h}$. Therefore, $\widetilde{J}_h$ should be moved no later then after $\widetilde{U}_{*,h} = \widetilde{U}_{z_h^*-1,h}$ units of obtained service.

If $\widetilde{Q}^*$ is empty, then eRMLF always works on the earliest released job in the non-empty queue $\widetilde{Q}_z$ with the lowest index $z \in \mathbb{Z}$.

The eRMLF algorithm is formally presented in Figures 2.2 and 2.3. Observe that both RMLF and eRMLF preserve the ordering of the jobs; that is, if job $\widetilde{J}_j$ is released prior to job $\widetilde{J}_k$ then as long as both jobs are incomplete:

- job $\widetilde{J}_j$ will never be in a lower queue than job $\widetilde{J}_k$, and
- if both jobs are in the same queue, then job $\widetilde{J}_j$ has priority over job $\widetilde{J}_k$.

We are now ready to prove Theorem 2.3.2, stating that

$$\mathbb{E}[\overline{T}_{\text{eRMLF}}(\mathcal{I})] \le C_1 \log(m) \cdot \overline{T}_{\text{SRPT}}(\mathcal{I}) \qquad \text{(2.5, revisited)}$$

for all instances $\mathcal{I}$ of size at most m for a universal constant C_1. This constant is identical to the constant C_1 in Theorem 2.3.1.

Proof of Theorem 2.3.2. Consider any instance $\widetilde{\mathcal{I}}$ for eRMLF of size at most m. Since all jobs of size zero are immediately served in queue $\widetilde{Q}^*$ upon arrival, we assume without loss of generality that the instance does not contain any jobs of size zero. As a consequence, the minimum job size $\widetilde{B}_{\min} = \min_{j=1,\dots,|\widetilde{\mathcal{I}}|} \widetilde{B}_j$ is strictly positive. We now transform the instance $\widetilde{\mathcal{I}}$ for eRMLF to a corresponding instance $\mathcal{I}$ for RMLF.

Algorithm eRMLF**:** At all times the collection of released, but uncompleted, jobs is partitioned into queues, $\widetilde{Q}^*, \widetilde{Q}_z, z \in \mathbb{Z}$. We say that $\widetilde{Q}_i$ is lower than $\widetilde{Q}_j$ for $i < j$. $\widetilde{Q}^*$ is the lowest queue. For each job $\widetilde{J}_j \in \widetilde{Q}_i, \widetilde{U}_{i,j} \in [2^i, 2^{i+1}]$ when it entered $\widetilde{Q}_i$. eRMLF maintains the invariant that it is always running the earliest released job in the lowest non-empty queue. When a job $\widetilde{J}_h$ is released at time $\widetilde{r}_h$, eRMLF takes the following actions:

- If, just prior to $\widetilde{r}_h$, all queues were empty, then
  - Job $\widetilde{J}_h$ is enqueued on $\widetilde{Q}^*$.
  - The initial target $\widetilde{U}_{*,h}$ is set to $\widetilde{U}_{0,h} = \max\{1, 2 - \widetilde{\beta}_h\}$.
- If, just prior to $\widetilde{r}_h$, there are unfinished jobs in the system but $\widetilde{Q}^*$ is empty, then
  - Job $\widetilde{J}_h$ is enqueued on $\widetilde{Q}^*$.
  - The initial target $\widetilde{U}_{*,h}$ is set to $\widetilde{U}_{z^*_h - 1, h} = 2^{z^*_h - 1} \max\{1, 2 - \widetilde{\beta}_h\}$, where the queue index $z^*_h = \min\{z \in \mathbb{Z} : \widetilde{Q}_z \text{ non-empty at time } t\}$ corresponds to the lowest non-empty queue.
- If, just prior to $\widetilde{r}_h$, $\widetilde{Q}^*$ is non-empty, then $\widetilde{Q}^* = \{\widetilde{J}_{h-1}\}$ at that time. Now,
  - The target $\widetilde{U}_{z^*_h, h-1} = 2^{z^*_h} \max\{1, 2 - \widetilde{\beta}_{h-1}\}$ with $z^*_h := \min\{z \in \mathbb{Z} : \widetilde{w}_{h-1}(\widetilde{r}_h) \le \widetilde{U}_{z,h-1}\}$ is the lowest target not yet reached by job $\widetilde{J}_{h-1}$.
  - Job $\widetilde{J}_{h-1}$ is dequeued from $\widetilde{Q}^*$.
  - Job $\widetilde{J}_{h-1}$ is enqueued on $\widetilde{Q}_{z^*_h}$.
  - Job $\widetilde{J}_h$ is enqueued on $\widetilde{Q}^*$.
  - The initial target $\widetilde{U}_{*,h}$ is set to $\widetilde{U}_{z^*_h - 1, h} = 2^{z^*_h - 1} \max\{1, 2 - \widetilde{\beta}_h\}$.
- If, just prior to $\widetilde{r}_h$, it was the case that eRMLF was running a job $\widetilde{J}_j$, then $\widetilde{J}_j$ is preempted.
- eRMLF begins running $\widetilde{J}_h$.

If at some time t, a job $\widetilde{J}_j \in \widetilde{Q}_{z-1}$ is being run when $\widetilde{w}_j(t)$ becomes equal to $\widetilde{U}_{z-1,j}$, then eRMLF takes the following actions:

- Job $\widetilde{J}_j$ is dequeued from $\widetilde{Q}_{z-1}$.
- Job $\widetilde{J}_j$ is enqueued on $\widetilde{Q}_z$.
- The target $\widetilde{U}_{z,j}$ is set to $2\widetilde{U}_{z-1,j} = 2^z \max\{1, 2 - \widetilde{\beta}_j\}$.

<algorithm continues on next page>

Figure 2.2: Formal statement of eRMLF algorithm, part I. Continued in Figure 2.3.

<algorithm continued from previous page>

If at some time t, a job $\widetilde{J}_j \in \widetilde{Q}^*$ is being run when $\widetilde{w}_j(t)$ becomes equal to $\widetilde{U}_{*,j}$, then eRMLF takes the following actions:

- Job $\widetilde{J}_j$ is dequeued from $\widetilde{Q}^*$.
- Job $\widetilde{J}_j$ is enqueued on $\widetilde{Q}_{z_h^*}$, where $z_h^* = \log_2\left(\widetilde{w}_j(t)/\max\{1,2-\widetilde{\beta}_j\}\right)+1$.
- The target $\widetilde{U}_{z_h^*,j}$ is set to $2\widetilde{U}_{*,j} = 2^{z_h^*}\max\{1,2-\widetilde{\beta}_j\}$.

Whenever a job is completed, it is removed from its queue.

Figure 2.3: Formal statement of eRMLF algorithm, part II. Continuation from Figure 2.2.

Define the scaling parameter $g := \lfloor \log_2(\widetilde{B}_{\min})\rfloor - 1 \in \mathbb{Z}$, satisfying $2^{-g}\widetilde{B}_{\min} \geq 2$. The instance $\mathcal{I}$ consists of $|\widetilde{\mathcal{I}}|$ jobs that are scaled versions of the original $|\widetilde{\mathcal{I}}|$ jobs; specifically, job J_j has size $B_j := 2^{-g}\widetilde{B}_j$ and release date $r_j := 2^{-g}\widetilde{r}_j$. Then, the smallest job is of size at least 2 and the RMLF algorithm may be applied to the instance $\mathcal{I}$.

Since the jobs are released in the same order as in the original instance, we note that the random variables β_j assigned by RMLF have the same distribution as the $\widetilde{\beta}_j$ assigned by eRMLF. We therefore couple these random variables in a trivial way: $\beta_j \equiv \widetilde{\beta}_j$ for all $j = 1,\ldots,|\widetilde{\mathcal{I}}|$. It immediately follows that the targets $\widetilde{U}_{z,j}$ as assigned to $\widetilde{\mathcal{I}}$ by eRMLF and the targets $U_{i,j}$ as assigned to $\mathcal{I}$ by RMLF satisfy $\widetilde{U}_{i,j} = U_{i,j}$ for all $i \in \{0,1,2,\ldots\}$. Additionally, the initial RMLF target $U_{0,j}$ satisfies $U_{0,j} = \widetilde{U}_{0,j} = 2^{g-g}\max\{1,2-\widetilde{\beta}_j\} = 2^{-g}\widetilde{U}_{g,j}$.

We will show that the above construction implies an equivalence between RMLF and eRMLF. For all $z \in \mathbb{Z}$ and $t \geq 0$ define the sets

$$\widetilde{Q}_z(t) := \left\{\widetilde{J}_j : \widetilde{U}_{z-1,j} \leq \widetilde{w}_j(t) < \widetilde{U}_{z,j}\right\} \tag{2.32}$$

that contain all jobs in the system at time t that are in queue $\widetilde{Q}_z$, *or the most recently released job if it has received a similar amount of service.* The equivalence is first observed between the initial RMLF queue $Q_0(t)$ and the augmented eRMLF queue $\widehat{Q}(t)$, defined as

$$Q_0(t) := \left\{J_j : w_j(t) < U_{0,j}\right\} \tag{2.33}$$

and

$$\widehat{Q}(t) := \bigcup_{z=-\infty}^{g} \widetilde{Q}_z(t) = \left\{\widetilde{J}_j : \widetilde{w}_j(t) < \widetilde{U}_{g,j}\right\}, \tag{2.34}$$

respectively. One may observe that since Q_0 is the highest priority queue, it experiences a dedicated, work-conserving server that works at unit speed on jobs J_j with sizes $U_{0,j}$. Therefore, the event that the RMLF server works on Q_0 is equivalent to the event $\{Q_0(t) > 0\}$.

We assume without loss of generality that job J_1 arrives at time $r_1 = 0$. The arrival of this job initiates a first busy period for $Q_0(t)$ of N_1 jobs, where N_1 is such that the cumulative targets $U_{0,j}$ of the first N_1 jobs can be served before the (N_1+1)-th job is released. It is defined as $N_1 = \inf\{k \geq 1 : \sum_{j=1}^{k} U_{0,j} - r_{j+1} \leq 0\}$, where $r_{|\widetilde{\mathcal{I}}|+1}$ is understood as plus infinity. The duration of the busy period is given by $P_1 = \sum_{j=1}^{N_1} U_{0,j} = \sum_{j=1}^{N_1} 2^{-g} \widetilde{U}_{g,j}$. The server may then work on jobs in higher queues (perceived as idle time by Q_0) until time r_{N_1+1}, when a new busy period is initiated. For $t \in [0, r_{N_1+1})$ we have now obtained

$$Q_0(t) > 0 \Leftrightarrow t \leq P_1 \Leftrightarrow 2^g t \leq \sum_{j=1}^{N_1} \widetilde{U}_{g,j}. \tag{2.35}$$

By a similar analysis of the augmented queue $\widehat{Q}$ we find that for all $t \in [0, 2^g r_{N_1+1}) = [0, \widetilde{r}_{N_1+1})$ the relation $\widehat{Q}(t) > 0 \Leftrightarrow Q_0(2^{-g} t) > 0$ holds, and for all $t \geq 0$ by a straightforward generalisation of the above procedure.

After observing that both algorithms preserve the ordering of jobs, we may similarly show that the eRMLF server processes job $\widetilde{J}_j$ in queue $\widetilde{Q}_{g+i}$ at time t if and only if the RMLF server processes job J_j in queue Q_i at time $2^g t$ for all $i \in \{1, 2, \ldots\}$ and $t \geq 0$.

From the above results, one may deduce that the expected sojourn time $\mathbb{E}[\overline{T}_{\text{eRMLF}}(\widetilde{\mathcal{I}})]$ of instance $\widetilde{\mathcal{I}}$ under eRMLF equals 2^g times the expected sojourn time $\mathbb{E}[\overline{T}_{\text{RMLF}}(\mathcal{I})]$ of instance $\mathcal{I}$ under RMLF. The competitive ratio of RMLF as stated in Theorem 2.3.1 hence guarantees that, for all instances $\widetilde{\mathcal{I}}$ of size at most m,

$$\mathbb{E}[\overline{T}_{\text{eRMLF}}(\widetilde{\mathcal{I}})] = 2^g \mathbb{E}[\overline{T}_{\text{RMLF}}(\mathcal{I})] \leq C_1 \log(m) 2^g \cdot \overline{T}_{\text{SRPT}}(\mathcal{I}). \tag{2.36}$$

The competitive ratio of eRMLF is concluded by verifying

$$2^g \cdot \overline{T}_{\text{SRPT}}(\mathcal{I}) = \overline{T}_{\text{SRPT}}(\widetilde{\mathcal{I}}), \tag{2.37}$$

which is a direct consequence of our scaling. In particular, the constant C_1 in the upper bound is the same for RMLF and eRMLF. □

CHAPTER 3

BARRIERS TO ANALYSING THE PERFORMANCE OF MULTI-SERVER POLICIES

In the previous chapter we derived a theorem that compares the expected sojourn time under blind scheduling policies in GI/GI/1 queueing systems to the minimum achievable expected sojourn time (in hindsight). Some of the bounds that facilitated the proof were quite loose (e.g. bounding the sojourn time by the duration of the busy period) or exploited a large body of existing literature (e.g. the competitive ratio). One may thus wonder whether similar approaches may facilitate novel results for more general queueing systems. The current chapter results from the pursuit of one such approach in the setting of multi-server queueing systems.

More specifically, we observe that the approach in Chapter 2 crucially depends on the regenerative structure of the GI/GI/1 queue and the fact that the regeneration points are independent of the employed scheduling policy. When aiming to mirror that approach, it is thus essential to find such an invariant regenerative structure in the multi-server queueing system. The current chapter quests to find an intuitive regenerative structure that allows for further analysis of the multi-server SRPT algorithm, but instead concludes that SRPT may be both slower and faster than an expected sojourn time minimising assignment.

3.1 Introduction

Consider a job scheduling problem with a given number of identical servers. If there is only one server, then all work-conserving schedules yield the same makespan; i.e. the time at which all jobs have been fully processed is independent of the schedule, provided that the schedule never idles the server if there are jobs available for processing[1]. In Chapter 2 we exploited this property to analyse the expected time that a job spends in a GI/GI/1 queueing system (i.e. sojourn time, flow time) if the server employs either of two scheduling policies. In particular, we showed that the expected sojourn time under the optimal blind scheduling policy is at most a factor $C\log(1/(1-\rho))$ larger than the expected sojourn time in the optimal off-line schedule, where $\rho < 1$ is the long-run fraction of capacity needed to complete the jobs and $C > 0$ is a constant independent of ρ. Some of the analyses in the previous chapter seem applicable to more general queueing systems; however, as the analyses depend heavily on the above invariance property it could prove beneficial to have access to some – perhaps weaker – version of the invariance property in the multi-server setting.

To illustrate the desired properties, we discuss the role of the invariance property in Chapter 2. There, the invariance property allowed us to exploit the competitive ratio of an algorithm. The competitive ratio $c^{\mathrm{P}}(n)$ of an algorithm P quantifies the performance of the algorithm relative to the best performance that could have been achieved in hindsight (OPT), and bounds the worst possible ratio over all problem instances with at most n jobs. It is for this reason that results from competitive analysis are generally not applicable in a queueing setting, where the number of jobs is infinite. We overcame this problem by noting that the work in a GI/GI/1 queueing system is a *regenerative process*, implying that the infinitely many jobs can be partitioned into finite sets of jobs that can each be regarded as a separate scheduling sub-problem, oblivious to and independent of all other jobs.

A regenerative process is characterised by an *embedded renewal process*, which indicates at which times the regenerative process regenerates [8]. If one may choose the embedded renewal process under scheduling policy P identical to the embedded renewal process under the OPT, then the sub-problems coincide and, consequently, the competitive ratio can relate the performance of P to that of OPT for every sub-problem. These results may then be used to compare the relative performance in the infinite jobs problem. Finding a common embedded renewal process, however, is generally non-trivial.

Chapter 2 considered a typical embedded renewal process for an GI/GI/1 queue, characterised by the times at which a job arrives in an empty system. In this case, every

[1] This is readily seen if one considers the total amount of available, unprocessed work in the system and notes that this amount is reduced at fixed rate.

period between two consecutive renewal points consists of a period during which the server is busy and then idle. Since the server idles at the same time under all work-conserving scheduling policies, the given embedded renewal process is independent of the scheduling policy. Sequentially, we analysed busy periods with less than N_0 jobs by means of the competitive ratio, and presented a probabilistic analysis to show that busy periods with more than N_0 jobs are negligible.

Unfortunately, the described renewal process does generally not apply to multi-server systems. A work-conserving policy P_1 may idle servers at different times than another work-conserving policy P_2 due to different usage of its total capacity. This implies that the busy periods are no longer independent of the scheduling policy. One may hope, however, that there exists a relationship between embedded renewal processes for particular policies P_1 and P_2, thereby constructing sets of jobs upon which P_1 and P_2 can be compared.

A relationship that would potentially allow for further analysis, is that a well-chosen embedded renewal process under P_2 is a refinement of an embedded renewal processes under P_1. An example of this relation translates to a policy P_2 that always idles if P_1 idles, where the embedded renewal processes again correspond to moments at which a job arrives in an empty system. In this case, P_2 may experience several regenerative cycles during a single regeneration cycle of P_1, so that both policies may be compared between two renewal times for P_1. Note that this relation is equivalent to showing that the makespan $C_{\max}^{\mathrm{P}_1}$ under P_1 is never larger than the makespan $C_{\max}^{\mathrm{P}_2}$ under P_2.

The presented relation holds true in the GI/GI/1 queue and is an intuitive candidate for relating the embedded renewal processes in the multi-server queues. Recalling that we ultimately wish to exploit competitive ratios, we limit ourselves to examining the makespan under an appropriate policy P and OPT. As Leonardi and Raz [94] have shown that the Shortest Remaining Processing Time (SRPT) algorithm has the best possible competitive ratio[2] [102], it seems natural to examine the makespan of P = SRPT. SRPT tends to work intensively on small jobs, but may consequently end up with some idle machines and some machines working on stalled large jobs. It is therefore conceivable that the makespan under SRPT may suffer from poor use of capacity, whereas the makespan under OPT could potentially be consistently low due to its efficient use of capacity.

Finding instances where SRPT has strictly larger makespan than OPT is not hard. In order to show that SRPT never has strictly smaller makespan than OPT, one is required to show that this is true for any version of OPT; i.e. if multiple assignments minimise the expected sojourn time, then SRPT should not have smaller makespan than any of them. This chapter, however, presents an instance where this is not true, implying that our

[2] Up to a multiplicative factor independent of n.

candidate for the embedded renewal process can not facilitate the intended sojourn time comparison between SRPT and OPT.

The rest of this section is organised as follows. A formal model description is presented in Section 3.2. Section 3.3 supports the above discussion by presenting two problem instances that illustrate opposing relations between the makespan of SRPT and OPT. Finally, Section 3.4 discusses the implications of this chapter and discusses alternative approaches to the problem at hand.

3.2 Preliminaries

We consider a scheduling problem where the scheduler has access to m identical servers. A problem instance is completely characterised by the number n of jobs in the instance and the vectors $\mathbf{r} = (r_1, \ldots, r_n)$ and $\mathbf{p} = (p_1, \ldots, p_n)$, where r_j and p_j represent the release date and the processing requirements of job j, respectively. The scheduler becomes aware of a job's existence and processing requirements only after the job has been released, and is allowed to pre-empt and migrate jobs without any penalty. For any schedule π, the completion time of the j-th job is indicated by C_j^π, whereas the makespan is denoted by $C_{\max}^\pi := \max_{j=1,\ldots,n} C_j^\pi$. We refer to $\sum_{j=1}^n C_j^\pi$ as the total completion time. The sojourn time of job j is defined as $T_j^\pi = C_j^\pi - r_j$.

Let OPT denote a policy that always minimizes the total completion time, i.e. a solution to $P_m \mid r_j, pmtn \mid \sum_j C_j$, and note that this policy also minimizes the expected sojourn time. Finding such policy for $m \geq 2$ is an NP-hard problem [47] and therefore one commonly resorts to scheduling policies that are known to perform well in some sense.

A common benchmark for the performance of a scheduling policy is its *competitive ratio*. In the current setting, the competitive ratio $c^\pi(m,n)$ of an algorithm π is defined as the supremum of the ratio of the total completion time under π over that under OPT, where the supremum is taken over all instances with m servers and at most n jobs. That is, it is a function that satisfies $\sum_{j=1}^k C_j^\pi \leq c^\pi(m,n) \sum_{j=1}^k C_j^{\mathrm{OPT}}$ for every problem instance with $k \leq n$ jobs. Note that the competitive ratio only quantifies the worst possible instance and may consequently be overly pessimistic for generic instances. We denote $c^\pi(m,n) = O(f(m,n))$ if $c^\pi(m,n) \leq C f(m,n)$ for some constant $C > 0$ independent of m and n.

We consider the multi-server variant of the SRPT policy, which at any time serves the at most m released, unfinished jobs that require the least remaining amount of processing. SRPT is known to minimise the expected sojourn time if there is only a single server [125]. Contrastingly, if there are at least two servers, then Leonardi and Raz [94] state that SRPT has a competitive ratio of at most $O(\log(n/m))$. They also show

that the dependence on n and m can not be improved, which makes SRPT an appealing policy.

A potential downside of the SRPT policy is that it migrates jobs, i.e. jobs may be processed by multiple servers during their stay. The incurred overhead may be a major disadvantage and as such one may be interested in policies that circumvent this. Awerbuch et al. [12] introduce a nameless scheduling policy that refrains from migrating jobs while still achieving a competitive ratio of at most $O(\log(n))$. Two key administrative elements of their policy are that there is a central pool containing all jobs that have not yet received any processing, and that all jobs are categorised according to their remaining processing time. The Smallest Group policy by Chekuri et al. [36] potentially improves upon Awerbuch et al.'s policy by categorising jobs according to their initial processing time. Their analysis is simpler and allows for fine-tuning of the categories, resulting in a competitive ratio of $O(\log(n/m))$. Finally, Avrahami and Azar's Immediate Dispatching policy [10] is $O(\log(n))$-competitive while refraining from both migrating jobs and administrating a central pool. The interested reader is referred to Pruhs et al. [116] for a more detailed discussion on the topic.

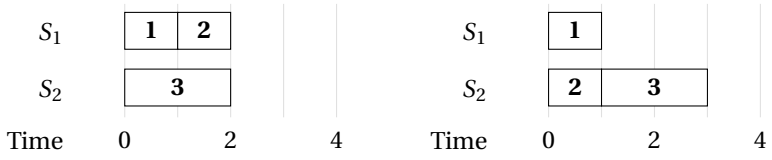
As a cost for not migrating jobs, however, the choices made by any of the above three policies depend on choices that it made earlier. This dependence on the past is a potential obstacle if one wishes to regard the queueing process as a regenerative processes, where regenerative cycles are required to be mutually independent. As such, the SRPT policy seems to be best suited for further analysis.

The next section considers the makespan of both SRPT and OPT for some key instances. We will abuse the notation introduced in this section by replacing the schedule superscript π by the scheduling policy superscript SRPT or OPT. We note that internal choices of a scheduling policy may affect the schedule that it produces, and may therewith affect characteristics such as the completion time of a job. Conveniently, neither the makespan nor the total completion time of the instances considered in the next section are affected by the internal choices of SRPT and OPT.

3.3 Results

We argued before that SRPT and OPT have identical makespan for any instance if $m = 1$. We will show that this contrasts with the $m = 2$ setting, where neither $C_{\max}^{\text{OPT}} \geq C_{\max}^{\text{SRPT}}$ nor $C_{\max}^{\text{OPT}} \leq C_{\max}^{\text{SRPT}}$ is generally true.

The first case can be discarded with an elementary example. Consider $n = 3$ jobs with release dates $\mathbf{r} = (0,0,0)$ and processing requirements $\mathbf{p} = (1,1,2)$. Then OPT may yield the schedule shown in Figure 3.1(a), which has total completion time $\sum_{j=1}^{n} C_j^{\text{OPT}} = 5$ and makespan $C_{\max}^{\text{OPT}} = 2$. On the other hand, SRPT yields the schedule in Figure 3.1(b),



(a) Schedule by OPT. (b) Schedule by SRPT.

Figure 3.1: Three-job problem instance for which $C_{\max}^{\text{OPT}} < C_{\max}^{\text{SRPT}}$. A job can be processed by either of servers S_1 and S_2 at any time after its release.

which has total completion time[3] $\sum_{j=1}^{n} C_j^{\text{SRPT}} = 5$ and $C_{\max}^{\text{SRPT}} = 3$.

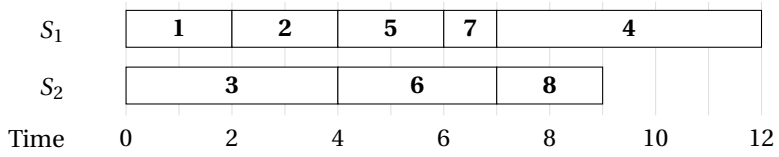
The above example shows that $C_{\max}^{\text{OPT}} \geq C_{\max}^{\text{SRPT}}$ is generally not true. The reason for this is that SRPT postponed working on job 3 and consequentially worked at half capacity from $t = 1$ onward. In fact, there are numerous instances where SRPT finds itself working on postponed jobs that occupy only a fraction of its servers. OPT, on the other hand, is designed to minimise the total completion time. This makes it imaginable that OPT never has larger makespan than SRPT. The following example, however, shows that this is not true.

Consider $n = 8$ jobs with release dates $\mathbf{r} = (0,0,0,0,4,4,6,6)$ and processing requirements $\mathbf{p} = (2,2,4,5,2,3,1,2)$. Now, a schedule that minimizes the total completion time is shown in Figure 3.2(a) and has total completion time $\sum_{j=1}^{n} C_j^{\text{OPT}} = 51$ and makespan $C_{\max}^{\text{OPT}} = 12$. One may verify that any schedule that achieves this total completion time also has the same makespan. SRPT, on the other hand, yields a schedule equivalent to the one in Figure 3.2(b) and has total completion time $\sum_{j=1}^{n} C_j^{\text{SRPT}} = 52$ and makespan $C_{\max}^{\text{SRPT}} = 11$. As such, the example shows that the makespan under SRPT may be strictly smaller than the makespan of a schedule that minimizes the total completion time. This confirms our claim that neither $C_{\max}^{\text{OPT}} \geq C_{\max}^{\text{SRPT}}$ nor $C_{\max}^{\text{OPT}} \leq C_{\max}^{\text{SRPT}}$ is generally true if there are at least two servers.

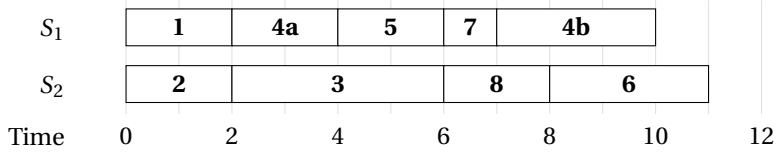
3.4 Discussion

To facilitate extension of the results in Chapter 2 to multi-server queueing systems, we investigated the existence of a specific common embedded renewal process underlying the workload process of a multi-server queueing system. Ideally, we would show that the makespan induced by SRPT is never smaller than the makespan induced by any scheduling policy OPT that minimizes the expected sojourn time, or vice versa. Our result is negative, in the sense that we presented problem instances for which opposing

[3] The identical total completion times in this example are not surprising; indeed, Conway et al. [40, p.77] show that OPT and SRPT always yield the same total completion time in case of uniform release dates.



(a) Schedule by OPT.



(b) Schedule by SRPT.

Figure 3.2: Eight-job problem instance for which $C_{\max}^{\mathrm{OPT}} > C_{\max}^{\mathrm{SRPT}}$. SRPT preempts the processing of job 4 by server S_1 at time 4, and resumes its service on the same server at time 7 without any penalty.

relations hold. This implies that our candidate for the embedded renewal process can not facilitate the intended sojourn time comparison between SRPT and OPT.

Alternatively, one may consider one of the following methods to analyse the behaviour of scheduling policies in multi-server queueing systems:

1. Find a more suitable regenerative process representation of the amount of work in a multi-server queueing system. There might very well be more suitable ways that transform the queueing systems into regenerative processes and that lend themselves for further analysis. However, it may be far from trivial how the corresponding regenerative cycles can be analysed; for example, to deduce the stochastic properties of the number of jobs in every cycle.

2. Consider a policy different from SRPT. We listed several other policies that may have a more favourable relationship with OPT. However, these policies also exhibit the property that their internal choices depend on past choices. This property may complicate the analysis, but does not necessarily rule out the regenerative process approach pursued in the previous chapter.

3. Consider a policy different from OPT. The competitive ratio guarantees that a scheduling policy P_1 performs at most a factor $c_{\mathrm{P}_1}(m,n)$ worse than the off-line optimal schedule. In particular, it performs at most that factor worse than *any* scheduling policy P_2. If one is able to find a scheduling policy P_2 that idles at the same time as P_1, then one could possibly derive statements of the performance of P_1 by instead analysing the performance of P_2.

4. Modify the OPT policy in a negligible manner. The examples in this chapter suggested that it is far more likely for OPT to have smaller makespan than SRPT than the other way around. As such, consider the (unknown) policy P_2 that minimizes the sum of completion times (in hindsight, similar to OPT) *under the constraint that the makespan under* P_2 *is at most the makespan that* SRPT *would incur.* Then any embedded renewal process that underlies the regenerative workload process under SRPT is also a valid choice for the embedded renewal process that underlies the workload process under P_2. The challenge in this approach is to show that the contribution of instances where P_2 is different from OPT to the overall expected sojourn time is asymptotically negligible.

In summary, it seems non-trivial to extend the intuitive approach of Chapter 2 to multi-server queueing systems; however, there are many alternative, unexplored approaches that may prove to be more successful.

CHAPTER 4

HEAVY-TRAFFIC ANALYSIS OF SOJOURN TIME UNDER THE FOREGROUND-BACKGROUND POLICY

This chapter considers the steady-state distribution of the sojourn time of a job entering an M/GI/1 queue with the foreground-background scheduling policy in heavy traffic. The growth rate of its mean, as well as the limiting distribution, are derived under broad conditions. Assumptions commonly used in extreme value theory play a key role in both the analysis and the results.

Based on Kamphorst and Zwart [S2].

4.1 Introduction

One of the main insights from queueing theory is that the M/GI/1 queue length and sojourn time grow at the order of $1/(1-\rho)$ as the traffic intensity of the system ρ approaches 100 percent utilization. This insight dates back to Kingman [83] and Prokhorov [114] and, appropriately reformulated, remains valid for queueing networks and multiple server queues [33, 60, 136]. However, the growth factor can be very different when the scheduling policy is no longer First In First Out (FIFO). This observation specifically applies to the Foreground-Background (FB) algorithm, which we investigate in this chapter.

Bansal [14] was the first to point out that the expected sojourn time (a.k.a. response time, flow time) of a user is of $o(1/(1-\rho))$ in the M/M/1 queue when the scheduling policy is Shortest Remaining Processing Time (SRPT). In particular, he showed that the growth factor of the expected sojourn time under SRPT is $\log(1/(1-\rho))$ smaller than the growth factor under FIFO. However, since SRPT requires information on service times in advance, the question was raised if the same growth rate in heavy traffic can be reached with a blind scheduling policy.

Bansal et al. [S1] answered this question negatively for the general GI/GI/1 queueing model. Specifically, the authors showed that for every blind scheduling policy there exists a service-time distribution under which the growth rate in heavy traffic of the expected sojourn time is at least a factor $\log(1/(1-\rho))$ larger than the growth rate of SRPT. Bansal et al. also constructed a scheduling policy that achieves this growth rate, but this policy is rather complicated as it involves randomization.

One might wonder whether the SRPT growth rate can be achieved by a deterministic blind algorithm for specific service-time distributions. To the best of the author's knowledge, no comprehensive answer to this question has been issued for the GI/GI/1 queue. However, researchers have derived the growth rate of the expected sojourn time in specific queueing models, thereby giving more insight into their behaviour and allowing for a comparison with SRPT. On this account, there have been several contributions: for certain M/GI/1 models, there are expected sojourn-time results for the FB [15, 105, 137], Pre-emptive Shortest Job First [15], and SRPT [14, 95] scheduling policies. All of these results utilize an explicit expression, focusing on a narrow class of job size distributions. Furthermore, these results only concern the mean sojourn time, and it is of interest to obtain information about the distribution of the sojourn time as well.

Motivated by these developments, we consider the sojourn-time distribution in the M/GI/1 queue with the FB scheduling policy. Like in previous works, we exploit explicit expression for this distribution, but will do so for a comprehensive class of job-size distributions, aiming to provide as much insight as possible in how the job-size

distribution affects the behavior in heavy traffic. The FB policy operates as follows: priority is given to the customer with the least-attained service, and when multiple customers satisfy this property, they are served at an equal rate. The only heavy-traffic results for FB we are aware of are of "big-O" type and are known in case of deterministic, exponential, Pareto and specific finite-support service times [15, 105]. For deterministic service times, it is straightforward to see that all customers under FB depart in one batch at the end of every busy period, and as a result the growth rate in heavy traffic in this case, $O((1-\rho)^{-2})$, is very poor. The behaviour of FB is much better for service-time distributions with a decreasing failure rate, as FB then optimizes the expected sojourn time among all blind policies [121]. For more background on the FB policy we refer to the survey by Nuyens and Wierman [105].

The main results of this chapter are of three types:

1. We characterise the exact growth rate (up to a constant independent of ρ) of the sojourn time in heavy traffic under very general assumptions on the service-time distribution. As in Bansal and Gamarnik [15] and Lin et al. [95], we find a dichotomy: when the service-time distribution has finite variance, the expected sojourn time $\mathbb{E}[T^{\rho}_{\mathrm{FB}}] = \Theta\left(\frac{\overline{F}(G^{\leftarrow}(\rho))}{(1-\rho)^2}\right)$. Here $\overline{F}(x) = 1 - F(x)$ is the tail of the service-time distribution and $G^{\leftarrow}$ is the right-inverse of the distribution function of a residual service time; a detailed overview of notation can be found in Section 4.2. In the infinite variance case, we find that $\mathbb{E}[T^{\rho}_{\mathrm{FB}}] = \Theta\left(\log \frac{1}{1-\rho}\right)$. This result is formally stated in Theorem 4.3.1. The precise conditions for these results to hold involve Matuszewska indices, a concept that will be reviewed in Section 4.2. The behaviour of $\overline{F}(G^{\leftarrow}(\rho))$ is quite rich, as will be illustrated by several examples.

2. Contrary to the results in Bansal and Gamarnik [15] and Lin et al. [95], we have been able to obtain a more precise estimate of the growth rate of $\mathbb{E}[T^{\rho}_{\mathrm{FB}}]$. It turns out that extreme value theory plays an essential role in our analysis, and the limiting constant factor in front of the growth rate $\frac{\overline{F}(G^{\leftarrow}(\rho))}{(1-\rho)^2}$ crucially depends on in which domain of attraction the service-time distribution is. This result is summarised in Theorem 4.3.2 and appended in Theorem 4.3.4. When the service-time distribution tail is regularly varying, it is shown that the growth rate of the sojourn time under FB is equal to that of SRPT up to a multiplicative constant. A comparison of the sojourn times under FB and SRPT is given in Corollary 4.3.5.

3. When analysing the distribution, we first show that $T^{\rho}_{\mathrm{FB}}/\mathbb{E}[T^{\rho}_{\mathrm{FB}}]$ converges to zero in probability as $\rho \uparrow 1$. To still get a heavy-traffic approximation for $\mathbb{P}(T^{\rho}_{\mathrm{FB}} > y)$, we state a sample path representation for the sojourn-time distribution for a job that requires a known amount of service. We then use fluctuation theory for spectrally negative Lévy processes to rewrite this representation into an expression that is

amenable to analysis; in particular, we obtain a representation for the Laplace transform of the *residual* sojourn-time distribution from which a heavy-traffic limit theorem follows. Finally, this Laplace transform provides an estimate for the tail distribution of T_{FB}.

More specifically, our results show that $\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}} > y)/\overline{F}(G^{\leftarrow}(\rho))$ converges to a non-trivial function $g^*(y)$, for which we give an integral expression in terms of error functions. Along the way, we derive a heavy-traffic limit for the total workload in an M/GI/1 queue with truncated service times that also seems to be of independent interest (see Proposition 4.7.1). As in the analysis for the expected sojourn time, ideas from extreme value theory play an important role in the analysis, and the limit function g^* depends on which domain of attraction the service-time distribution falls into. A precise description of this result can be found in Theorem 4.3.7.

The function $\overline{F}(G^{\leftarrow}(\rho))$ that shows up in many of our results corresponds to the probability that a customer requires at least $G^{\leftarrow}(\rho)$ units of service. Our analyses indicate that customers who require at least $G^{\leftarrow}(\rho)$ units of service determine the generic sojourn time characteristics, whereas the contribution of smaller customers is negligible. Although not mentioned explicitly, a similar phenomenon (with a different function G) can be observed in the analysis of the mean sojourn time under SRPT by Lin et al. [95].

Even though our analysis relies on an explicit representation of the sojourn-time distribution, we hope that the insights given by our results (apart from how to separate small and large jobs, also the determination of the right scaling, which we think will not be affected by the inter-arrival time distribution), will help to design proofs that do not require explicit expressions. Hopefully, such proofs can also deal with non-Poisson arrival streams and process limit theorems. An example of such a proof for the queue-length process for SRPT with light-tailed job sizes can be found in Puha et al. [117]. A similar comment applies to the extension of our results from FB to a broader class of scheduling disciplines, like the class of SMART scheduling policies considered in Wierman et al. [137] and Nuyens et al. [106]. Developing a more probabilistic proof of our result potentially would also clarify the precise role of extreme value theory, which we feel is not entirely clear from the analysis in this chapter. Finally, we want to point out that the methodology in this chapter does seem to be applicable to the class of size-based scheduling disciplines which is introduced and analyzed in Scully et al. [127].

The rest of the chapter is organised as follows. Section 4.2 formally introduces the model that is considered. Section 4.3 presents all our main results on the asymptotic behaviour of the expectation and the tail of the sojourn-time distribution under FB. The results concerning the expectation are then proven in Sections 4.4 and 4.5, whereas

the results on the tail distribution are supported in Sections 4.6 and 4.7.

4.2 Preliminaries

Consider a sequence of M/GI/1 queues, indexed by n, where the i-th job requires B_i units of service for all n. For convenience, we say that a job that requires x units of service is a *job of size* x. All B_i are independent and identically distributed (i.i.d.) random variables with cumulative distribution function (c.d.f.) $F(x) = \mathbb{P}(B_i \leq x)$ and finite mean $\mathbb{E}[B_1]$. We assume that $F(0) = 0$, and denote $x_R := \sup\{x \geq 0 : F(x) < 1\} \leq \infty$. Jobs in the n-th queue arrive with rate $\lambda^{(n)}$, where $\lambda^{(n)} < 1/\mathbb{E}[B_1]$ to ensure that the n-th system experiences traffic intensity $\rho^{(n)} := \lambda^{(n)}\mathbb{E}[B_1] < 1$. For notational convenience, we let B denote a random variable with c.d.f. F.

Let $\overline{F}(x) := 1 - F(x)$ and $F^{\leftarrow}(y) := \inf\{x \geq 0 : F(x) \geq y\}$ denote the complementary c.d.f. (c.c.d.f.) and the right-inverse of F respectively. The random variable B^* is defined by its c.d.f. $G(x) := \mathbb{P}(B^* \leq x) = \int_0^x \overline{F}(t)/\mathbb{E}[B]\,\mathrm{d}t$ and has k-th moment $\mathbb{E}[(B^*)^{k-1}] = \mathbb{E}[B^k]/(k\mathbb{E}[B])$. Since $G^{\leftarrow}(y)$ is continuous and strictly increasing, its (right-)inverse $G^{\leftarrow}(y)$ satisfies $G^{\leftarrow}(G(x)) = x$. Also, we recognise $h^*(x) := \frac{\overline{F}(x)}{\mathbb{E}[B]\overline{G}(x)}$ as the failure rate of B^*. One may deduce that $h^*(x)$ equals the reciprocal of the expected residual time; $h^*(x) = 1/\mathbb{E}[B - x \mid B > x]$.

Foreground-Background scheduling policy

Jobs are served according to the Foreground-Background (FB) policy, meaning that at any moment in time, the server equally shares its capacity over all available jobs that have received the least amount of service thus far. First, we are interested in characteristics of the sojourn time $T_{\mathrm{FB}}^{(n)}$, defined as the duration of time that a generic job spends in the system. In order to analyse this, we consider an expression for the expected sojourn time of a generic job *of size* x, $\mathbb{E}[T_{\mathrm{FB}}^{(n)}(x)]$, for which Schrage [124, relation (18)] states that

$$\mathbb{E}[T_{\mathrm{FB}}^{(n)}(x)] = \frac{x}{1-\rho_x^{(n)}} + \frac{\mathbb{E}[W^{(n)}(x)]}{1-\rho_x^{(n)}} = \frac{x}{1-\rho_x^{(n)}} + \frac{\lambda^{(n)} m_2(x)}{2(1-\rho_x^{(n)})^2}, \tag{4.1}$$

where $\rho_x^{(n)} := \lambda^{(n)}\mathbb{E}[B \wedge x] = \rho\mathbb{P}(B^* \leq x)$ and $m_2(x) := \mathbb{E}[(B \wedge x)^2] = 2\int_0^x t\overline{F}(t)\,\mathrm{d}t$ are functions of the first and second moments of $B \wedge x := \min\{B, x\}$, and $W^{(n)}(x)$ is the steady-state waiting time in a M/GI/1/FIFO queue with arrival rate $\lambda^{(n)}$ and jobs of size $B_i \wedge x$. As a consequence of (4.1), the expected sojourn time $\mathbb{E}[T_{\mathrm{FB}}^{(n)}]$ of a generic job is given by

$$\mathbb{E}[T_{\mathrm{FB}}^{(n)}] = \int_0^\infty \frac{x}{1-\rho_x^{(n)}}\,\mathrm{d}F(x) + \int_0^\infty \frac{\lambda^{(n)} m_2(x)}{2(1-\rho_x^{(n)})^2}\,\mathrm{d}F(x). \tag{4.2}$$

The intuition behind relation (4.1) is that a job J_1 of size x experiences a system where all job sizes are truncated. Indeed, if another job J_2 of size $x+y, y>0$, has received at least x service, then FB will never dedicate its resources to job J_2 while job J_1 is incomplete. The expected sojourn time of job J_1 can now be salvaged from its own service requirement x, the truncated work already in the system upon arrival $W^{(n)}(x)$, and the rate $1-\rho_x^{(n)}$ at which it is expected to be served.

Second, we focus attention on the tail behaviour of $T_{\mathrm{FB}}^{(n)}$. Write $X \stackrel{d}{=} Y$ if the relation $\mathbb{P}(X \le x) = \mathbb{P}(Y \le x)$ is satisfied for all $x \in \mathbb{R}$ and let $\mathscr{L}_x(y)$ denote the time required by the server to empty the system if all job sizes are truncated to $B_i \wedge x$ and the current amount of work is y. The analysis of the tail behaviour is then facilitated by relation (4.28) in Kleinrock [86], stating

$$T_{\mathrm{FB}}^{(n)}(x) \stackrel{d}{=} \mathscr{L}_x(W_x^{(n)} + x). \tag{4.3}$$

For both the expectation and tail behaviour of $T_{\mathrm{FB}}^{(n)}$, we take specific interest in systems that experience *heavy traffic*, that is, systems where $\rho^{(n)} \uparrow 1$ as $n \to \infty$. In the current setting, this is equivalent to sequences $\lambda^{(n)}$ that converge to $1/\mathbb{E}[B]$. Most results in this chapter make no assumptions on sequence $\lambda^{(n)}$, in which case we drop the superscript n for notational convenience and just state $\rho \uparrow 1$.

The remainder of this section introduces some notation related to Matuszewska indices and extreme value theory.

Matuszewska indices

We now introduce the notion of the upper and lower Matuszewska index.

Definition 4.2.1. Suppose that $f(\cdot)$ is positive.

- The *upper Matuszewska index* $\alpha(f)$ is the infimum of those α for which there exists a constant $C = C(\alpha)$ such that for each $\mu^* > 1$,

$$\lim_{x\to\infty} f(\mu x)/f(x) \le C\mu^{\alpha} \tag{4.4}$$

  uniformly in $\mu \in [1, \mu^*]$ as $x \to \infty$.

- The *lower Matuszewska index* $\beta(f)$ is the supremum of those β for which there exists a constant $D = D(\beta) > 0$ such that for each $\mu^* > 1$,

$$\lim_{x\to\infty} f(\mu x)/f(x) \ge D\mu^{\beta} \tag{4.5}$$

  uniformly in $\mu \in [1, \mu^*]$ as $x \to \infty$.

One may note from the above definitions that $\beta(f) = -\alpha(1/f)$ holds for any positive f. Intuitively, a function f with upper and lower Matuszewska indices $\alpha(f)$ and $\beta(f)$ is bounded between functions $Dx^{\beta(f)}$ and $Cx^{\alpha(f)}$ for appropriate constants $C, D > 0$. More accurately, however, C and D could be unbounded or vanishing functions of x. Of special interest is the class of functions that satisfy $\beta(f) = \alpha(f)$.

Definition 4.2.2. A measurable function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is *regularly varying (at infinity) with index* $\alpha \in \mathbb{R}$ (denoted by $f \in \mathrm{RV}_\alpha$) if for all $\mu > 0$

$$\lim_{x\to\infty} f(\mu x)/f(x) = \mu^\alpha. \tag{4.6}$$

If (4.6) holds with $\alpha = 0$, then f is called *slowly varying*. If (4.6) holds with $\alpha = -\infty$, then f is called *rapidly varying*.

The following result elegantly characterises functions of regular variation.

Theorem 4.2.3 (Bingham et al. [24], Theorem 1.4.1)**.** *A measurable function $f(x)$ is regularly varying with index $\alpha \in \mathbb{R}$ if and only if there exists a slowly varying function $l(x)$ such that $f(x) = l(x)x^\alpha$.*

Extreme value theory

The next paragraphs introduce some notions and results from extreme value theory. The field of extreme value theory generally aims to assess the probability of an extreme event; however, for our purposes we restrict attention to the limiting distribution of $\max\{X_1, \ldots, X_m\}$. A key result on this functional is the Fisher-Tippett theorem:

Theorem 4.2.4 (Resnick [119], Proposition 0.3)**.** *Let $(X_m)_{m\in\mathbb{N}}$ be a sequence of i.i.d. random variables and define $M_m := \max\{X_1, \ldots, X_m\}$. If there exist norming sequences $c_m > 0$, $d_m \in \mathbb{R}$ and some non-degenerate H such that*

$$\mathbb{P}(c_m^{-1}(M_m - d_m) \leq x) = F^m(c_m x + d_m) \to H(x) \tag{4.7}$$

weakly as $m \to \infty$, then H belongs to the type of one of the following three c.d.f.'s:

$$\textit{Fréchet:} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0, \\ \exp\{-x^{-\alpha}\}, & x > 0, \end{cases} \qquad \alpha > 0,$$

$$\textit{Weibull:} \quad \Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^\alpha\}, & x \leq 0, \\ 1, & x > 0, \end{cases} \qquad \alpha > 0, \textit{ and}$$

$$\textit{Gumbel:} \quad \Lambda(x) = \exp\{-e^{-x}\}, \qquad x \in \mathbb{R}.$$

The three distributions above are referred to as the extreme value distributions.

A c.d.f. F is said to be in the *maximum domain of attraction of* H if there exist norming sequences c_m and d_m such that (4.7) holds. In this case, we write $F \in \mathrm{MDA}(H)$. A large body of literature has identified conditions on F such that $F \in \mathrm{MDA}(H)$ and excellent collections of such and related results can be found in Embrechts et al. [50] and Resnick [119].

The following theorems show a particularly elegant characterisation of the classes $\mathrm{MDA}(\Phi_\alpha)$ and $\mathrm{MDA}(\Psi_\alpha)$ as classes of regularly-varying distributions.

Theorem 4.2.5 (Embrechts et al. [50], Theorem 3.3.7)**.** *The c.d.f. F belongs to the maximum domain of attraction of $\Phi_\alpha, \alpha > 0$ if and only if $x_R = \infty$ and $\overline{F}$ is regularly varying with index $-\alpha$. If $F \in \mathrm{MDA}(\Phi_\alpha)$, then the norming constants can be chosen as $c_n = F^{\leftarrow}(1 - n^{-1})$ and $d_n = 0$.*

Theorem 4.2.6 (Embrechts et al. [50], Theorem 3.3.12)**.** *The c.d.f. F belongs to the maximum domain of attraction of $\Psi_\alpha, \alpha > 0$ if and only if $x_R < \infty$ and $\overline{F}(x_R - (\cdot)^{-1})$ is regularly varying with index $-\alpha$. If $F \in \mathrm{MDA}(\Psi_\alpha)$, then the norming constants can be chosen as $c_n = x_R - F^{\leftarrow}(1 - n^{-1})$ and $d_n = x_R$.*

The class $\mathrm{MDA}(\Lambda)$ is not quite as closely related to regularly-varying distributions, and can be characterised as follows:

Theorem 4.2.7 (Embrechts et al. [50], Theorem 3.3.26)**.** *The c.d.f. F with right endpoint $x_R \leq \infty$ belongs to the maximum domain of attraction of Λ if and only if there exists some $z < x_R$ such that F has representation*

$$\overline{F}(x) = c(x) \exp\left\{ -\int_z^x \frac{g(t)}{f(t)}\, \mathrm{d}t \right\}, \quad z < x < x_R, \tag{4.8}$$

where c and g are measurable functions satisfying $c(x) \to c > 0, g(x) \to 1$ as $x \uparrow x_R$, and $f(\cdot)$ is a positive, absolutely continuous function (with respect to the Lebesgue measure) with density $f'(x)$ having $\lim_{x \uparrow x_R} f'(x) = 0$.

If $F \in \mathrm{MDA}(\Lambda)$, then the norming constants can be chosen as $c_m = f(d_m)$ and $d_m = F^{\leftarrow}(1 - m^{-1})$. A possible choice for the function $f(\cdot)$ is $f(\cdot) = 1/h^*(\cdot)$.

The function $f(\cdot)$ in the above definition is unique up to asymptotic equivalence. We refer to f as the *auxiliary function* of $\overline{F}$. Also, we note the following property of $f(\cdot)$:

Lemma 4.2.8 (Resnick [119], Lemma 1.2)**.** *Suppose that $f(\cdot)$ is an absolutely continuous auxiliary function with $f'(x) \to 0$ as $x \uparrow x_R$.*

(i) If $x_R = \infty$, then $\lim_{x \to \infty} \frac{f(x)}{x} = 0$.

(ii) If $x_R < \infty$, then $\lim_{x \uparrow x_R} \frac{f(x)}{x_R - x} = 0$.

Although MDA(Λ) does not coincide with a class of regularly-varying distributions, the following lemma shows that it is related to the class of rapidly varying distributions.

Corollary 4.2.9 (Embrechts et al. [50], Corollary 3.3.32)**.** *Assume that* $F \in \text{MDA}(\Lambda)$*. If* $x_R = \infty$*, then* $\overline{F} \in \text{RV}_{-\infty}$*. If* $x_R < \infty$*, then* $\overline{F}(x_R - (\cdot)^{-1}) \in \text{RV}_{-\infty}$*.*

This section's final lemma presents a useful property for c.d.f.'s in MDA(Λ):

Lemma 4.2.10. *Suppose that the c.d.f.* F *is in* MDA(Λ) *and let* $G(x) = \int_0^x \overline{F}(t)/\mathbb{E}[B]\,\mathrm{d}t$*. Then* $G \in \text{MDA}(\Lambda)$ *and any auxiliary function for* F *is also an auxiliary function for* G*.*

Proof. According to Theorem 3.3.27 in Embrechts et al. [50], $G \in \text{MDA}(\Lambda)$ with auxiliary function $f(\cdot)$ if and only if $\lim_{x \uparrow x_R} \overline{G}(x + t f(x))/\overline{G}(x) = e^{-t}$ for all $t \in \mathbb{R}$. It is straightforward to check that the above relation holds for any auxiliary function $f(\cdot)$ of F by using l'Hôpital's rule and $\lim_{x \uparrow x_R} f'(x) = 0$. □

Asymptotic relations

Let $f(\cdot)$ and $g(\cdot)$ denote two positive functions and X and Y two random variables. We write $f \sim g$ if $\lim_{z \uparrow z^*} f(z)/g(z) = 1$, where the appropriate limit $z \uparrow z^*$ should be clear from the context; it usually equals $x \uparrow x_R$ or $\rho \uparrow 1$. Similarly, we adopt the conventions $f = o(g)$ if $\limsup_{z \uparrow z^*} f(z)/g(z) = 0$, $f = O(g)$ if $\limsup_{z \uparrow z^*} f(z)/g(z) < \infty$ and $f = \Theta(g)$ if $0 < \liminf_{z \uparrow z^*} f(z)/g(z) \leq \limsup_{z \uparrow z^*} f(z)/g(z) < \infty$. We write $X \leq_{st} Y$ if the relation $P(X > x) \leq \mathbb{P}(Y > x)$ is satisfied for all $x \in \mathbb{R}$.

Finally, the complementary error function is defined as $\text{Erfc}(x) := 2\pi^{-1/2} \int_x^\infty e^{-u^2}\,\mathrm{d}u$.

4.3 Main results and discussion

This section presents and discusses our main results. Theorems 4.3.1 and 4.3.2 consider the asymptotic behaviour of the expected sojourn time $\mathbb{E}[T_{\text{FB}}]$ for various classes of service-time distributions. Theorem 4.3.4 connects the asymptotic behaviour of $\overline{F}(G^{\leftarrow}(\rho))$ to the literature on extreme value theory. As a consequence, the expressions obtained in Theorem 4.3.2 can be specified for many distributions in MDA(Λ). Theorem 4.3.6 shifts focus to the distribution of T_{FB} and states that the scaled sojourn time $T_{\text{FB}}/\mathbb{E}[T_{\text{FB}}]$ tends to zero in probability. Instead, Theorem 4.3.7 shows that a certain fraction of jobs experiences a sojourn time of order $(1-\rho)^{-2}$. This result is achieved through the Laplace transform of the remaining sojourn time T^*_{FB}, for which we give an integral representation. The proofs of the theorems are postponed to later sections.

Recall that $\overline{F}(G^{\leftarrow}(\rho)) = \mathbb{E}[B](1-\rho)h^*(G^{\leftarrow}(\rho))$. Our first theorem presents the growth rate of $\mathbb{E}[T_{\text{FB}}]$.

Theorem 4.3.1. *Assume that either $x_R = \infty$ and $-\infty < \beta(\overline{F}) \le \alpha(\overline{F}) < -2$, or that $x_R < \infty$ and $-\infty < \beta(\overline{F}(x_R - (\cdot)^{-1})) \le \alpha(\overline{F}(x_R - (\cdot)^{-1})) < 0$. Then the relations*

$$\mathbb{E}[T_{\mathrm{FB}}] = \Theta\left(\frac{\overline{F}(G^{\leftarrow}(\rho))}{(1-\rho)^2}\right) = \Theta\left(\frac{h^*(G^{\leftarrow}(\rho))}{1-\rho}\right) \tag{4.9}$$

hold as $\rho \uparrow 1$, where $\lim_{\rho\uparrow 1} h^*(G^{\leftarrow}(\rho)) = 0$ if $x_R = \infty$ and $\lim_{\rho\uparrow 1} h^*(G^{\leftarrow}(\rho)) = \infty$ if $x_R < \infty$. Alternatively, assume $x_R = \infty$ and $\beta(\overline{F}(x)) > -2$. Then the relation

$$\mathbb{E}[T_{\mathrm{FB}}] = \Theta\left(\log\frac{1}{1-\rho}\right) \tag{4.10}$$

holds as $\rho \uparrow 1$.

Theorem 4.3.1 shows that the behaviour of $\mathbb{E}[T_{\mathrm{FB}}]$ is fundamentally different for $\alpha(\overline{F}) < -2$ and $\beta(\overline{F}(x)) > -2$. In the first case, the variance of B_1 is bounded and therefore the expected remaining busy period duration is of order $\Theta((1-\rho)^{-2})$. Our analysis roughly shows that all jobs of size $G^{\leftarrow}(\rho)$ and larger will remain in the system until the end of the busy period, and hence experience a sojourn time of order $\Theta((1-\rho)^{-2})$. The threshold $G^{\leftarrow}(\rho)$ itself originates as the solution of $1-\rho_x = 1-\rho^2$, which indicates that – as the traffic intensity increases to unity – jobs of size at least $G^{\leftarrow}(\rho)$ experience a truncated system that is almost as heavily congested as the non-truncated system. The theorem indicates that these jobs determine the asymptotic growth of the overall expected sojourn time.

The above argumentation does not apply in case $\beta(\overline{F}(x)) > -2$, since then the expected remaining busy period duration is infinite. It turns out that in this case the expected sojourn time of a large job of size x is of the same order as the time that the job is in service, which has expectation $x/(1-\rho_x)$. The result follows after integrating over the service-time distribution.

Additionally, it can be shown that the statements in Theorem 4.3.1 also hold if $F \in \mathrm{MDA}(\Lambda)$, which is a special case of either $\alpha(\overline{F}) = \beta(\overline{F}) = -\infty$ or $\alpha(\overline{F}(x_R - (\cdot)^{-1})) = \beta(\overline{F}(x_R - (\cdot)^{-1}) = -\infty$ (cf. Corollary 4.2.9). In this case, and equivalently in case $\overline{F}(\cdot)$ or $\overline{F}(x_R - (\cdot)^{-1})$ is regularly varying, one can show that $(1-\rho)^2\mathbb{E}[T_{\mathrm{FB}}]/\overline{F}(G^{\leftarrow}(\rho))$ converges. Theorem 4.3.2 specifies Theorem 4.3.1 for the aforementioned cases, as well as for distributions with an atom in their endpoint.

Theorem 4.3.2. *The following relations hold as $\rho \uparrow 1$:*

(i) If $F \in \mathrm{MDA}(\Phi_\alpha), \alpha \in (1,2)$, then $\mathbb{E}[T_{\mathrm{FB}}] \sim \frac{\alpha}{2-\alpha}\mathbb{E}[B]\log\frac{1}{1-\rho}$.

(ii) If $F \in \mathrm{MDA}(H)$, then $\mathbb{E}[T_{\mathrm{FB}}] \sim \frac{r(H)\mathbb{E}[B^*]\overline{F}(G^{\leftarrow}(\rho))}{(1-\rho)^2} = \frac{r(H)\mathbb{E}[B^2]h^*(G^{\leftarrow}(\rho))}{2(1-\rho)}$ where

$$r(H) = \begin{cases} \frac{\pi/(\alpha-1)}{\sin(\pi/(\alpha-1))}\frac{\alpha}{\alpha-1} & \text{if } H = \Phi_\alpha, \alpha > 2, \\ 1 & \text{if } H = \Lambda, \text{ and} \\ \frac{\pi/(\alpha+1)}{\sin(\pi/(\alpha+1))}\frac{\alpha}{\alpha+1} & \text{if } H = \Psi_\alpha, \alpha > 0. \end{cases} \tag{4.11}$$

Additionally, if $H = \Phi_\alpha, \alpha > 2$, *then* $\lim_{\rho\uparrow 1} h^*(G^{\leftarrow}(\rho)) = 0$, *whereas if either* $H = \Lambda$ *and* $x_R < \infty$ *or if* $H = \Psi_\alpha, \alpha > 0$, *then* $\lim_{\rho\uparrow 1} h^*(G^{\leftarrow}(\rho)) = \infty$.

(iii) If F *has an atom in* $x_R < \infty$, *say* $\lim_{\delta\downarrow 0} \overline{F}(x_R - \delta) = p > 0$, *then* $\mathbb{E}[T_{\mathrm{FB}}] \sim \frac{p\mathbb{E}[B^*]}{(1-\rho)^2}$.

The expressions in Theorems 4.3.1 and 4.3.2 give insight into the asymptotic behaviour of $\mathbb{E}[T_{\mathrm{FB}}]$. The following corollary shows that the asymptotic expressions above may be specified further if the service times are Pareto distributed. This extends the result by Bansal and Gamarnik [15], who derived the growth factor of $\mathbb{E}[T_{\mathrm{FB}}]$ but not the exact asymptotics.

Corollary 4.3.3. *Assume* $\overline{F}(x) = (x/x_L)^{-\alpha}, x \geq x_L$. *Then the relations*

$$\mathbb{E}[T_{\mathrm{FB}}] \sim \begin{cases} \frac{\alpha}{2-\alpha}\mathbb{E}[B]\log\frac{1}{1-\rho} & \text{if } \alpha \in (1,2), \\ \frac{\pi/(\alpha-1)}{2\sin(\pi/(\alpha-1))}\frac{\mathbb{E}[B^2]\alpha^{\frac{\alpha}{\alpha-1}}}{x_L(1-\rho)^{\frac{\alpha-2}{\alpha-1}}} & \text{if } \alpha \in (2,\infty), \end{cases} \tag{4.12}$$

hold as $\rho \uparrow 1$.

Proof. One may derive that $\overline{G}(x) = \frac{1}{\alpha}\left(\frac{x}{x_L}\right)^{1-\alpha}$ for $x \geq x_L$. Consequentially, one deduces that $h^*(x) = \frac{\alpha-1}{x}$ for $x \geq x_L$ and $G^{\leftarrow}(\rho) = x_L(\alpha(1-\rho))^{\frac{-1}{\alpha-1}}$ for $\rho \geq 1 - 1/\alpha$. The result then follows from Theorem 4.3.2. □

Corollary 4.3.3 exemplifies that the asymptotic growth of $\mathbb{E}[T_{\mathrm{FB}}]$ can be specified in some cases. However, it is often non-trivial to analyse the behaviour of $\overline{F}(G^{\leftarrow}(\rho))$ or equivalently $h^*(G^{\leftarrow}(\rho))$. Theorem 4.3.4 aims to overcome this problem if $F \in \mathrm{MDA}(\Lambda)$ by presenting a relation between $h^*(G^{\leftarrow}(\rho))$ and norming constants c_n of F, which can often be found in the large body of literature on extreme value theory.

Theorem 4.3.4. *Assume* $F \in \mathrm{MDA}(\Lambda)$ *and* $x_R = \infty$, *and let* c_m *and* d_m *be such that* $F^m(c_m x + d_m) \to \Lambda(x)$ *weakly as* $m \to \infty$. *Define* $\lambda^{(n)} = (1 - n^{-1})/\mathbb{E}[B]$ *so that* $\rho^{(n)} = 1 - n^{-1}$.

(i) If there exists $\alpha > 0$ *and a slowly varying function* $l(x)$ *such that* $-\log\overline{F}(x) \sim l(x)x^\alpha$ *as* $x \to \infty$, *then* $h^*(x) \sim \alpha l(x)x^{\alpha-1}$ *if and only if*

$$\inf_{\lambda\downarrow 1} \liminf_{x\to\infty} \inf_{t\in[1,\lambda]} \{\log h^*(tx) - \log h^*(x)\} \geq 0. \tag{4.13}$$

If (4.13) *holds, then* $\mathbb{E}[T_{\mathrm{FB}}^{(n)}] \sim \frac{\mathbb{E}[B^2]}{2(1-\rho^{(n)})c_n}$ *as* $n \to \infty$.

(ii) If there exists a function $l(x) : [0,\infty) \to \mathbb{R}$, $\liminf_{x\to\infty} l(x) > 1$ *such that for all* $\lambda > 0$

$$\lim_{x\to\infty} \frac{-\log\overline{F}(\lambda x) + \log\overline{F}(x)}{l(x)} = \log(\lambda) \tag{4.14}$$

and $L = \lim_{x\to\infty} \frac{\log(x)}{l(x)}$ *exists in* $[0,\infty]$, *then* $\lim_{n\to\infty} \frac{2(1-\rho^{(n)})c_n}{\mathbb{E}[B^2]}\mathbb{E}[T_{\mathrm{FB}}^{(n)}] = e^{-L}$.

The same results hold if $x_R < \infty$, provided that the $\overline{F}(\cdot)$ and $h^*(\cdot)$ in (i) and (ii) are replaced by $\overline{F}(x_R - (\cdot)^{-1})$ and $(\cdot)^{-2}h^*(x_R - (\cdot)^{-1})$, respectively.

Remark 1. Condition (4.13) in part (i) of Theorem 4.3.4 is a *Tauberian condition*, and originates from Theorem 1.7.5 in Bingham et al. [24]. A Tauberian theorem makes assumptions on a transformed function (here h^*), and uses these assumptions to deduce the asymptotic behaviour of that transform. The interested reader is referred to Section 1.7 in Bingham et al. [24] or Section XIII.5 in Feller [54].

Theorem 4.2.7 implies that $c_n \sim 1/h^*(G^{\leftarrow}(1-n^{-1}))$ for many distributions in MDA(Λ). As c_n may be chosen as $1/h^*(F^{\leftarrow}(1-n^{-1}))$, Theorem 4.3.4 implicitly states conditions under which $\lim_{n\to\infty} h^*(G^{\leftarrow}(1-n^{-1}))/h^*(F^{\leftarrow}(1-n^{-1})) = \lim_{y\uparrow 1}(1-y)^{-2}\overline{F}(G^{\leftarrow}(y))\overline{G}(F^{\leftarrow}(y))$ exists, and exploits this limit to write $\mathbb{E}[T_{\mathrm{FB}}^{(n)}]$ as function of c_n rather than of the generally unknown $h^*(G^{\leftarrow}(1-n^{-1}))$. To illustrate the implications of Theorem 4.3.4, the exact asymptotic behaviour of several well-known distributions is presented in Table 4.1.

We take a brief moment to compare the asymptotic expected sojourn time under FB to that under SRPT in M/GI/1 models. Clearly, FB can perform no better than SRPT due to SRPT's optimality [125]. The ratio of their respective expected sojourn time is shown to be unbounded if the service times are exponentially distributed or if the service-time distribution has finite support [14, 86, 95, 105], but bounded if the service times are Pareto distributed [15, 95]. To the best of the authors' knowledge, no results of this nature are known if service times are Weibull distributed.

The following corollary specifies the asymptotic advantage of SRPT over FB if the service times are Pareto distributed, and presents the first such results for Weibull distributed service times. Its statements follow directly from Corollaries 1 and 2 in Lin et al. [95] and the results earlier in this section. Further results may be obtained by analysing their function $G^{-1}(\rho)$ for other service-time distributions.

Corollary 4.3.5. *The following relations hold as $\rho \uparrow 1$:*

(i) If $\overline{F}(x) = (x/x_L)^{-\alpha}, x \geq x_L > 0$ and $\alpha \in (1,2)$, then $\mathbb{E}[T_{\mathrm{FB}}]/\mathbb{E}[T_{\mathrm{SRPT}}] \sim \alpha^2$.

(ii) If $\overline{F}(x) = (x/x_L)^{-\alpha}, x \geq x_L > 0$ and $\alpha > 2$, then $\mathbb{E}[T_{\mathrm{FB}}]/\mathbb{E}[T_{\mathrm{SRPT}}] \sim \alpha^{\frac{\alpha}{\alpha-1}}$.

(iii) If $\overline{F}(x) = e^{-\mu x^{\beta}}, x \geq 0$ and $\mu, \beta > 0$, then $\mathbb{E}[T_{\mathrm{FB}}]/\mathbb{E}[T_{\mathrm{SRPT}}] \sim \beta \log\left(\frac{1}{1-\rho}\right)$.

On the other hand, we may also compare FB to the classic FIFO policy [40]. Since $\mathbb{E}[T_{\mathrm{FIFO}}] = \mathbb{E}[B] + \rho\mathbb{E}[B^*]/(1-\rho)$, Theorems 4.3.1 and 4.3.2 indicate that FB performs better than FIFO if the service-time distribution has a heavy tail, but also that FB performs worse than FIFO if the service-time distribution has finite support. If the service-time distribution has infinite support but no heavy tail, then Table 4.1 shows that their relationship depends on the tail of the service-time distribution. This is

Distribution	c.c.d.f. $\overline{F}$ or p.d.f. F'		L	$\mathbb{E}[T_{\mathrm{FB}}] \sim$
Exponential-like	$\overline{F}(x) \sim Ke^{-\mu x}$	$K, \mu > 0$	–	$\frac{\mathbb{E}[B^2]\mu}{2(1-\rho)}$
Weibull-like	$\overline{F}(x) \sim Kx^{\alpha} e^{-\mu x^{\beta}}$	$K, \mu, \beta > 0, \alpha \in \mathbb{R}$	–	$\frac{\beta \mu^{1/\beta} \mathbb{E}[B^2]}{2(1-\rho)\log\left(\frac{1}{1-\rho}\right)^{1/\beta-1}}$
Gamma	$F'(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\alpha, \beta > 0$	–	$\frac{\mathbb{E}[B^2]\beta}{2(1-\rho)}$
Normal	$F'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$		–	$\frac{\mathbb{E}[B^2]\log\left(\frac{1}{1-\rho}\right)^{1/2}}{\sqrt{2}(1-\rho)}$
Lognormal	$F'(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-(\log(x)-\mu)^2/(2\sigma^2)}$	$\sigma > 0, \mu \in \mathbb{R}$	σ^2	$\frac{\mathbb{E}[B^2]\log\left(\frac{1}{1-\rho}\right)^{1/2}}{\sigma\sqrt{2}(1-\rho)\exp\left[\mu+\sigma\left(\sigma+\sqrt{2\log\left(\frac{1}{1-\rho}\right)}-\frac{\log(4\pi)+\log\log(1/(1-\rho))}{2\sqrt{2\log(1/(1-\rho))}}\right)\right]}$
Finite exponential	$\overline{F}(x) = Ke^{-\frac{\mu}{x_R - x}}$	$K, \mu > 0, x < x_R$	–	$\frac{\mathbb{E}[B^2]\log\left(\frac{1}{1-\rho}\right)^2}{2\mu(1-\rho)}$
Benktander-I	$\overline{F}(x) = \left(1 + 2\frac{\beta}{\alpha}\log(x)\right) \times e^{-(\beta\log(x)^2 + (\alpha+1)\log(x))}$	$\alpha, \beta > 0, x > 1$	$\frac{1}{2\beta}$	$\frac{\mathbb{E}[B^2]\sqrt{\beta\log\left(\frac{1}{1-\rho}\right)}}{(1-\rho)\exp\left[-\frac{\alpha+1/2}{\beta}+\sqrt{\frac{\log\left(\frac{1}{1-\rho}\right)}{\beta}}\right]}$
Benktander-II	$\overline{F}(x) = x^{-(1-\beta)} e^{-\frac{\alpha}{\beta}(x^{\beta}-1)}$	$\alpha > 0, 0 < \beta < 1, x > 1$	–	$\frac{\alpha^{1/\beta}\mathbb{E}[B^2]}{2\beta^{1/\beta-1}(1-\rho)\log\left(\frac{1}{1-\rho}\right)^{1/\beta-1}}$

Table 4.1: Asymptotic expressions for the expected sojourn time for several well-known distributions in MDA(Λ), characterised by either their tail distribution or their probability density function (p.d.f.). These expressions follow from Table 3.4.4 in Embrechts et al. [50] through Theorem 4.3.4, where it is assumed that relation (4.13) holds.

exemplified by Weibull distributed service-times, $\overline{F}(x) = e^{-\mu x^\beta}, \beta > 0$, in which case $\mathbb{E}[T_{\mathrm{FB}}]/\mathbb{E}[T_{\mathrm{FIFO}}] \sim \beta\Gamma(1+1/\beta)(\mu^{-1}\log 1/(1-\rho))^{1-1/\beta}$. In fact, Table 4.1 seems to suggests that FB outperforms FIFO if $-\log\overline{F}(x)/x \to 0$ as $x \to \infty$ and vice versa if $-\log\overline{F}(x)/x \to 0$ as $x \to \infty$. However, investigating this observation is beyond the scope of this chapter.

Now that the asymptotic behaviour of the expected sojourn time under FB has been quantified, it is natural to investigate more complex characteristics. One such characteristic is the behaviour of the tail of the sojourn-time distribution, where one usually starts by analysing the distribution of the sojourn time normalised by its mean, $T_{\mathrm{FB}}/\mathbb{E}[T_{\mathrm{FB}}]$. The following theorem indicates that this random variable converges to zero in probability, meaning that almost every job experiences a sojourn time that is significantly shorter than the expected sojourn time as $\rho \uparrow 1$:

Theorem 4.3.6. *If either*

(i) $x_R = \infty$ *and either* $\beta(\overline{F}) > -2$ *or* $-\infty < \beta(\overline{F}) \le \alpha(\overline{F}) < -2$, *or*

(ii) $x_R < \infty$ *and* $-\infty < \beta(\overline{F}(x_R - (\cdot)^{-1})) \le \alpha(\overline{F}(x_R - (\cdot)^{-1})) < 0$, *or*

(iii) $F \in \mathrm{MDA}(\Lambda)$,

then $\frac{T_{\mathrm{FB}}}{\mathbb{E}[T_{\mathrm{FB}}]} \xrightarrow{p} 0$ *as* $\rho \uparrow 1$.

Theorem 4.3.6 indicates that a decreasing fraction of jobs experiences a sojourn time of at least duration $\mathbb{E}[T_{\mathrm{FB}}]$. Our final main result aims to specify both the size of this fraction, and the growth factor of the associated jobs' sojourn time.

The intuition behind Theorem 4.3.1 suggests that T_{FB} scales as $(1-\rho)^{-2}$, but only for jobs of size at least $G^{\leftarrow}(\rho)$. This makes it conceivable that the scaled probability $\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}} > y)/\overline{F}(G^{\leftarrow}(\rho))$ may be of $\Theta(1)$ as $\rho \uparrow 1$. Theorem 4.3.7 confirms this hypothesis, and additionally shows that the residual sojourn time T^*_{FB} with density $\mathbb{P}(T_{\mathrm{FB}} > x)/\mathbb{E}[T_{\mathrm{FB}}]$ scales as $(1-\rho)^{-2}$.

Theorem 4.3.7. *Assume* $F \in \mathrm{MDA}(H)$, *where* H *is an extreme value distributions with finite* $(2+\varepsilon)$*-th moment for some* $\varepsilon > 0$. *Let* $r(H)$ *be as in relation* (4.11). *Then* $(1-\rho)^2 T^*_{\mathrm{FB}}$ *converges to a non-degenerate random variable with monotone density* g^* *as* $\rho \uparrow 1$, *and*

$$\lim_{\rho\uparrow 1} \frac{\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}} > y)}{r(H)\mathbb{E}[B^*]\overline{F}(G^{\leftarrow}(\rho))} = g^*(y) \tag{4.15}$$

almost everywhere. Here,

$$g^*(t) = \int_0^1 8r(H)^{-1}\nu\left(\frac{1-\nu}{\nu}\right)^{p(H)} g(t,\nu)\,\mathrm{d}\nu, \tag{4.16}$$

$$g(t,\nu) = \frac{e^{-\frac{t}{4\mathbb{E}[B^*]\nu^2}}}{4\mathbb{E}[B^*]\nu^2}\left(\frac{\sqrt{t}}{\nu\sqrt{\pi\mathbb{E}[B^*]}} - \frac{t}{2\mathbb{E}[B^*]\nu^2} e^{\frac{t}{4\mathbb{E}[B^*]\nu^2}} \operatorname{Erfc}\left(\frac{1}{2\nu}\sqrt{\frac{t}{\mathbb{E}[B^*]}}\right)\right), \tag{4.17}$$

and $p(H) = \frac{\alpha}{\alpha-1}$ *if* $H = \Phi_\alpha$, $\alpha > 2$; $p(H) = 1$ *if* $H = \Lambda$ *and* $p(H) = \frac{\alpha}{\alpha+1}$ *if* $H = \Psi_\alpha$, $\alpha > 0$.

All theorems presented in this section are now proven in order. First, Theorems 4.3.1 and 4.3.2 are proven in Section 4.4. Then, Theorem 4.3.4 is justified in Section 4.5. Finally, Sections 4.6 and 4.7 respectively validate Theorems 4.3.6 and 4.3.7.

4.4 Asymptotic behaviour of the expected sojourn time

In this section, we prove Theorems 4.3.1 and 4.3.2 in order. The intuition behind the theorems is that jobs of size x can only be completed once the server has finished processing of all jobs of size at most x. Additionally, jobs of size x experience a system with job sizes $B_i \wedge x$ since no job will receive more than x units of processing as long as there are size x jobs in the system. One thus expects all jobs of size x to stay in the system for the duration of a remaining busy period in the truncated system, which is expected to last for $\Theta(\mathbb{E}[(B \wedge x)^2]/(1-\rho_x)^2)$ time.

Now, if $\mathbb{E}[B^2] < \infty$ and x_ρ^ν is such that $(1-\rho)/(1-\rho_{x_\rho^\nu}) = \nu \in (1-\rho, 1)$, then one can see from (4.1) that

$$(1-\rho)^2 \mathbb{E}[T_{\mathrm{FB}}(x_\rho^\nu)] = \nu(1-\rho)x_\rho^\nu + \nu^2 \frac{\lambda m_2(x_\rho^\nu)}{2}. \tag{4.18}$$

It turns out that the asymptotic behaviour of $(1-\rho)^2\mathbb{E}[T_{\mathrm{FB}}]$ is now determined by the fraction of jobs for which ν takes values away from zero.

If instead $\mathbb{E}[B^2] = \infty$, it will be shown that the growth rate of the second term in (4.1) is bounded by the growth rate of $x\overline{G}(x)$. It then turns out that the sojourn time is of the same order as the time that a job receives service, which is of order $\Theta(x/(1-\rho_x))$.

Both theorems follow after integrating $\mathbb{E}[T_{\mathrm{FB}}(x)]$ over all possible values of x, as shown in (4.2). By integrating by parts, we find that the first integral in (4.2) can be rewritten as

$$\begin{aligned}\int_0^\infty \frac{x}{1-\rho_x}\,\mathrm{d}F(x) &= \int_0^\infty \frac{\overline{F}(x)}{1-\rho_x}\,\mathrm{d}x + \lambda\int_0^\infty \frac{x\overline{F}(x)^2}{(1-\rho_x)^2}\,\mathrm{d}x \\ &= \frac{1}{\lambda}\log\frac{1}{1-\rho} + \lambda\int_0^\infty \frac{x\overline{F}(x)^2}{(1-\rho_x)^2}\,\mathrm{d}x.\end{aligned}$$

Similarly, the second integral can be rewritten as

$$\int_0^\infty \frac{\lambda m_2(x)}{2(1-\rho_x)^2}\,\mathrm{d}F(x) = \lambda\int_0^\infty \frac{x\overline{F}(x)^2}{(1-\rho_x)^2}\,\mathrm{d}x + \lambda^2\int_0^\infty \frac{m_2(x)\overline{F}(x)^2}{(1-\rho_x)^3}\,\mathrm{d}x,$$

and therefore

$$\begin{aligned}\mathbb{E}[T_{\mathrm{FB}}] &= \frac{1}{\lambda}\log\frac{1}{1-\rho} + 2\lambda\int_0^\infty \frac{x\overline{F}(x)^2}{(1-\rho_x)^2}\,\mathrm{d}x + \lambda^2\int_0^\infty \frac{m_2(x)\overline{F}(x)^2}{(1-\rho_x)^3}\,\mathrm{d}x \\ &= \frac{\mathbb{E}[B]}{\rho}\log\frac{1}{1-\rho} + 2\rho\int_0^\infty \frac{x\overline{F}(x)}{(1-\rho_x)^2}\,\mathrm{d}G(x) + \frac{\rho^2}{\mathbb{E}[B]}\int_0^\infty \frac{m_2(x)\overline{F}(x)}{(1-\rho_x)^3}\,\mathrm{d}G(x). \end{aligned} \tag{4.19}$$

We will now derive Theorems 4.3.1 and 4.3.2 from this relation.

4.4.1 General Matuszewska indices

This section proves Theorem 4.3.1. Relation (4.19) will be analysed separately for the cases $-\infty < \beta(\overline{F}) \le \alpha(\overline{F}) < -2$ and $-2 < \beta(\overline{F}) \le \alpha(\overline{F}) < 1$, which will be referred to as the finite and the infinite variance case, respectively. The finite variance case also considers $-\infty < \beta(\overline{F}(x_R - (\cdot)^{-1})$. Note that we always have $\beta(\overline{F}(x_R - (\cdot)^{-1})) \le \alpha(\overline{F}(x_R - (\cdot)^{-1})) \le 0$ since $\overline{F}(x_R - (\cdot)^{-1})$ is non-increasing. Prior to further analysis, however, we introduce several results that will facilitate the analysis.

Lemma 4.4.1. *Let $f_1(\cdot), f_2(\cdot)$ be positive.*

(i) If $\alpha(f_1), \alpha(f_2) < \infty$, then $\alpha(f_1 \cdot f_2) \le \alpha(f_1) + \alpha(f_2)$ and, assuming that f_1 is non-decreasing, $\alpha(f_1 \circ f_2) \le \alpha(f_1) \cdot \alpha(f_2)$.

(ii) If $\beta(f_1), \beta(f_2) > -\infty$, then $\beta(f_1 \cdot f_2) \ge \beta(f_1) + \beta(f_2)$ and, assuming that f_1 is non-increasing, $\beta(f_1 \circ f_2) \ge \beta(f_1) \cdot \beta(f_2)$.

Lemma 4.4.2. *Let f be positive. If $\alpha(f) < 0$, then $\lim_{x\to\infty} f(x) = 0$.*

Lemma 4.4.3 (Bingham et al. [24], Theorem 2.6.1)**.** *Let f be positive and locally integrable on $[X, \infty)$. Let $g(x) := \int_X^x f(t)/t \, \mathrm{d}t$. If $\beta(f) > 0$, then $\liminf_{x\to\infty} f(x)/g(x) > 0$.*

Lemma 4.4.4 (Bingham et al. [24], Theorem 2.6.3)**.** *Let f be positive and measurable. Let $g(x) := \int_x^\infty f(t)/t \, \mathrm{d}t$.*

(i) If $\alpha(f) < 0$, then $g(x) < \infty$ for all large x.

(ii) If $\beta(f) > -\infty$, then $\limsup_{x\to\infty} f(x)/g(x) < \infty$.

Lemma 4.4.5. *If $x_R = \infty$, then $\alpha(\overline{G}) \le \alpha(\overline{F}) + 1$ and $\beta(\overline{G}) \ge \beta(\overline{F}) + 1$. If $x_R < \infty$, then $\alpha(\overline{G}(x_R - (\cdot)^{-1})) \le \alpha(\overline{F}(x_R - (\cdot)^{-1})) - 1$ and $\beta(\overline{G}(x_R - (\cdot)^{-1})) \ge \beta(\overline{F}(x_R - (\cdot)^{-1})) - 1$.*

Lemma 4.4.6. *If $x_R = \infty$ and $\beta(\overline{F}) > -\infty$, then $\alpha(G^{\leftarrow}(1 - (\cdot)^{-1})) \le -\frac{1}{\alpha(\overline{F})+1}$ and $\beta(G^{\leftarrow}(1 - (\cdot)^{-1})) \ge -\frac{1}{\beta(\overline{F})+1}$. Alternatively, if $x_R < \infty$ and $\beta(\overline{F}(x_R - (\cdot)^{-1})) > -\infty$, then $\alpha(G^{\leftarrow}(1 - (\cdot)^{-1})) \le -\frac{1}{\alpha(\overline{F}(x_R-(\cdot)^{-1}))-1}$ and $\beta(G^{\leftarrow}(1 - (\cdot)^{-1})) \ge -\frac{1}{\beta(\overline{F}(x_R-(\cdot)^{-1}))-1}$.*

Corollary 4.4.7. *If $x_R = \infty$ and $\beta(\overline{F}) > -\infty$, then $\alpha(\overline{F}(G^{\leftarrow}(1 - (\cdot)^{-1}))) \le \frac{-\alpha(\overline{F})}{\alpha(\overline{F})+1}$ and $\beta(\overline{F}(G^{\leftarrow}(1 - (\cdot)^{-1}))) \ge \frac{-\beta(\overline{F})}{\beta(\overline{F})+1}$. Alternatively, if $x_R < \infty$ and $\beta(\overline{F}(x_R - (\cdot)^{-1})) > -\infty$, then $\alpha(\overline{F}(G^{\leftarrow}(1 - (\cdot)^{-1}))) \le \frac{-\alpha(\overline{F}(x_R-(\cdot)^{-1}))}{\alpha(\overline{F}(x_R-(\cdot)^{-1}))-1}$ and $\beta(\overline{F}(G^{\leftarrow}(1 - (\cdot)^{-1}))) \ge \frac{-\beta(\overline{F}(x_R-(\cdot)^{-1}))}{\beta(\overline{F}(x_R-(\cdot)^{-1}))-1}$.*

Lemma 4.4.1 states some closure properties of Matuszewska indices. Lemma 4.4.2 gives a sufficient condition for f to vanish. Lemmas 4.4.3 and 4.4.4 state helpful results on the asymptotic behaviour of the ratio between a function and certain integrals over this function, depending on its Matuszewska indices. Lemmas 4.4.5 and 4.4.6

and Corollary 4.4.7 specify the earlier lemmas by giving bounds on the Matuszewska indices of $\overline{G}$, $G^{\leftarrow}$ and the composition of $\overline{F}$ and $G^{\leftarrow}$. The proofs of Lemmas 4.4.1, 4.4.2, 4.4.5 and 4.4.6, along with several additional results, are postponed to Appendix 4.A. Corollary 4.4.7 follows immediately from Lemmas 4.4.1 and 4.4.6.

Finite variance

In this section, we assume either $x_R = \infty$ and $-\infty < \beta(\overline{F}) \leq \alpha(\overline{F}) < -2$, or $x_R < \infty$ and $\beta(\overline{F}(x_R - (\cdot)^{-1})) > -\infty$. If $x_R = \infty$, then $\alpha((\cdot)^2\overline{F}(\cdot)) < 0$ and thus $\mathbb{E}[B^2] = 2\int_0^\infty t\overline{F}(t)\,\mathrm{d}t < \infty$ by Lemma 4.4.4(i); if $x_R < \infty$ then clearly $\mathbb{E}[B^2] < \infty$.

Noting that $G^{\leftarrow}$ is a continuous, strictly increasing function, it follows that the function $x_\rho^\nu := G^{\leftarrow}\left(1 - \frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)$ is well-defined for all $\nu \in (1-\rho, 1)$. For this choice of x_ρ^ν, we have $\frac{1-\rho}{1-\rho_{x_\rho^\nu}} = \nu$ and $\frac{\mathrm{d}G(x_\rho^\nu)}{\mathrm{d}\nu} = \frac{1-\rho}{\rho}\frac{1}{\nu^2}$, and therefore relation (4.19) becomes

$$\begin{aligned}
(1-\rho)^2\mathbb{E}[T_{\mathrm{FB}}] &= \frac{\mathbb{E}[B](1-\rho)^2}{\rho}\log\frac{1}{1-\rho} + 2\rho\int_0^\infty \left(\frac{1-\rho}{1-\rho_x}\right)^2 x\overline{F}(x)\,\mathrm{d}G(x) \\
&\quad + \frac{\rho^2}{\mathbb{E}[B]}\int_0^\infty \left(\frac{1-\rho}{1-\rho_x}\right)^3 \frac{m_2(x)\overline{F}(x)}{1-\rho}\,\mathrm{d}G(x) \\
&= \frac{\mathbb{E}[B](1-\rho)^2}{\rho}\log\frac{1}{1-\rho} \\
&\quad + 2(1-\rho)\int_{1-\rho}^1 G^{\leftarrow}\left(1 - \frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\overline{F}\left(G^{\leftarrow}\left(1 - \frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)\mathrm{d}\nu \\
&\quad + \frac{\rho}{\mathbb{E}[B]}\int_{1-\rho}^1 \nu\, m_2\left(G^{\leftarrow}\left(1 - \frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)\overline{F}\left(G^{\leftarrow}\left(1 - \frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)\mathrm{d}\nu.
\end{aligned}$$

Dividing both sides by $\overline{F}(G^{\leftarrow}(\rho))$ yields

$$\begin{aligned}
\frac{(1-\rho)^2\mathbb{E}[T_{\mathrm{FB}}]}{\overline{F}(G^{\leftarrow}(\rho))} &= \frac{\mathbb{E}[B](1-\rho)^2}{\rho\overline{F}(G^{\leftarrow}(\rho))}\log\frac{1}{1-\rho} \\
&\quad + \frac{2(1-\rho)}{\overline{F}(G^{\leftarrow}(\rho))}\int_{1-\rho}^1 G^{\leftarrow}\left(1 - \frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\overline{F}\left(G^{\leftarrow}\left(1 - \frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)\mathrm{d}\nu \\
&\quad + \frac{\rho}{\mathbb{E}[B]}\int_{1-\rho}^1 \nu\, m_2\left(G^{\leftarrow}\left(1 - \frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)\frac{\overline{F}\left(G^{\leftarrow}\left(1 - \frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)}{\overline{F}(G^{\leftarrow}(\rho))}\mathrm{d}\nu \\
&= \mathrm{I}(\rho) + \mathrm{II}(\rho) + \mathrm{III}(\rho). \qquad (4.20)
\end{aligned}$$

We will show that $\mathrm{I}(\rho) + \mathrm{II}(\rho) = o(1)$ and $\mathrm{III}(\rho) = \Theta(1)$.

Assume $x_R = \infty$. Then, by Lemma 4.4.1 and Corollary 4.4.7 we find that

$$\begin{aligned}
\alpha(\mathrm{I}(1-(\cdot)^{-1}) &\leq \alpha((\cdot)^{-2}) + \alpha(1/\overline{F}(G^{\leftarrow}(1-(\cdot)^{-1}))) + \alpha(\log(\cdot)) \\
&= -2 - \beta(\overline{F}(G^{\leftarrow}(1-(\cdot)^{-1}))) + 0 \leq -2 + \frac{\beta(\overline{F})}{\beta(\overline{F})+1} < 0, \qquad (4.21)
\end{aligned}$$

and consequently $\mathrm{I}(\rho) = o(1)$ as $\rho \uparrow 1$ by Lemma 4.4.2.

Next, fix $0 \le \varepsilon < 2 - \frac{\beta(\overline{F})}{\beta(\overline{F})+1}$. Substitution of $w = \frac{\rho}{1-\rho}\frac{\nu}{1-\nu}$ in $\mathrm{II}(\rho)$ yields

$$\begin{aligned}\mathrm{II}(\rho) &= \frac{2(1-\rho)}{\overline{F}(G^{\leftarrow}(\rho))}\int_1^\infty \frac{\rho}{1-\rho}\left(\frac{\rho}{1-\rho}+w\right)^{-2} G^{\leftarrow}(1-w^{-1})\overline{F}(G^{\leftarrow}(1-w^{-1}))\,\mathrm{d}w \\ &\le \frac{2(1-\rho)^{2-\varepsilon}}{\rho^{1-\varepsilon}\overline{F}(G^{\leftarrow}(\rho))}\int_1^\infty w^{-\varepsilon} G^{\leftarrow}(1-w^{-1})\overline{F}(G^{\leftarrow}(1-w^{-1}))\,\mathrm{d}w.\end{aligned}$$

Let $q(w)$ denote the integrand in the last line. A similar analysis to (4.21) indicates that the term in front of the integral vanishes as $\rho \uparrow 1$, so we only need to show that the integral is bounded. This is implied by Lemma 4.4.4(i) after noting that

$$\alpha(q) \le -\varepsilon + \alpha(G^{\leftarrow}(1-(\cdot)^{-1})) + \alpha(\overline{F}(G^{\leftarrow}(1-(\cdot)^{-1}))) \le -1-\varepsilon < 0,$$

where the inequalities follow from Lemmas 4.4.1 and 4.4.6 and Corollary 4.4.7.

Lastly, we wish to show that $\mathrm{III}(\rho) = \Theta(1)$. Observe that

$$\begin{aligned}\mathrm{III}(\rho) &\le \lambda\mathbb{E}[B^2]\int_{1-\rho}^{\frac{1}{1+\rho}} \nu\frac{\overline{F}\left(G^{\leftarrow}\left(1-\frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)}{\overline{F}(G^{\leftarrow}(\rho))}\,\mathrm{d}\nu + \lambda\mathbb{E}[B^2]\int_{\frac{1}{1+\rho}}^1 \frac{\overline{F}\left(G^{\leftarrow}\left(1-\frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)}{\overline{F}(G^{\leftarrow}(\rho))}\,\mathrm{d}\nu \\ &\le 2\rho\mathbb{E}[B^*]\int_1^{\frac{1}{1-\rho}} \frac{\rho w}{1-\rho}\left(\frac{\rho}{1-\rho}+w\right)^{-3}\frac{\overline{F}(G^{\leftarrow}(1-w^{-1}))}{\overline{F}(G^{\leftarrow}(\rho))}\,\mathrm{d}w + \mathbb{E}[B^*] \\ &\le \frac{2\mathbb{E}[B^*]}{\rho}\int_1^{\frac{1}{1-\rho}} \frac{w\overline{F}(G^{\leftarrow}(1-w^{-1}))}{\frac{1}{(1-\rho)^2}\overline{F}(G^{\leftarrow}(\rho))}\,\mathrm{d}w + \mathbb{E}[B^*] = \frac{2\mathbb{E}[B^*]}{\rho}\int_1^{\frac{1}{1-\rho}} \frac{f(w)/w}{f(1/(1-\rho))}\,\mathrm{d}w + \mathbb{E}[B^*],\end{aligned}$$

where $f(w) = w^2\overline{F}(G^{\leftarrow}(1-w^{-1}))$. Lemma 4.4.1 and Corollary 4.4.7 then state that $\beta(f) \ge 2 - \frac{\beta(\overline{F})}{\beta(\overline{F})+1} > 0$, and therefore Lemma 4.4.3 implies

$$\limsup_{\rho\uparrow 1}\int_1^{\frac{1}{1-\rho}} \frac{f(w)/w}{f(1/(1-\rho))}\,\mathrm{d}\nu = \left[\liminf_{y\to\infty}\frac{f(y)}{\int_1^y f(w)/w\,\mathrm{d}w}\right]^{-1} < \infty.$$

As such, $\limsup_{\rho\uparrow 1}\mathrm{III}(\rho) < \infty$.

In order to show $\liminf_{\rho\uparrow 1}\mathrm{III}(\rho) > 0$, fix $c \in (0,1)$ and let $\delta_\rho := (1-\rho)/(c\rho+1-\rho)$. One may then readily verify that $\mathrm{III}(\rho) \ge \lambda m_2(G^{\leftarrow}(1-c))\int_{\delta_\rho}^{\frac{1}{1+\rho}} \nu\,\mathrm{d}\nu \to \frac{m_2(G^{\leftarrow}(1-c))}{8\mathbb{E}[B]} > 0$.

The $x_R = \infty$ case is concluded once we prove $\lim_{\rho\uparrow 1} h^*(G^{\leftarrow}(\rho)) = 0$. To this end, write $h^*(G^{\leftarrow}(\rho))$ as $x\overline{F}(G^{\leftarrow}(1-x^{-1}))/\mathbb{E}[B]$, where $x = (1-\rho)^{-1}$. The claim then follows from Lemma 4.4.2 after noting that

$$\alpha(h^*(G^{\leftarrow}(1-(\cdot)^{-1}))) \le \alpha(\cdot) + \alpha(\overline{F}(G^{\leftarrow}(1-(\cdot)^{-1}))) \le 1 - \frac{\alpha(\overline{F})}{\alpha(\overline{F})+1} = \frac{1}{\alpha(\overline{F})+1} < 0,$$

where the inequalities follow from Lemma 4.4.1 and Corollary 4.4.7.

The $x_R < \infty$ case can be proven similarly. One then fixes $1 < \varepsilon < 2 - \frac{\beta(\overline{F}(x_R-(\cdot)^{-1}))}{\beta(\overline{F}(x_R-(\cdot)^{-1}))-1}$ and obtains

$$\alpha(\mathrm{I}(1-(\cdot)^{-1}) \le -2 + \frac{\beta(\overline{F}(x_R-(\cdot)^{-1}))}{\beta(\overline{F}(x_R-(\cdot)^{-1}))-1} < 0,$$
$$\alpha(q) \le -\varepsilon - \frac{\alpha(\overline{F}(x_R-(\cdot)^{-1}))+1}{\alpha(\overline{F}(x_R-(\cdot)^{-1}))-1} \le 1-\varepsilon < 0,$$

and

$$\beta(f) \ge 2 - \frac{\beta(\overline{F}(x_R-(\cdot)^{-1}))}{\beta(\overline{F}(x_R-(\cdot)^{-1}))-1} > 0.$$

The claim $h^*(G^{\leftarrow}(\rho)) \to \infty$ follows from Lemma 4.2.8.

Infinite variance

Assume $\beta(\overline{F}) > -2$ and recall that $m_2(x) = 2\mathbb{E}[B]\int_0^x t\,\mathrm{d}G(t) = 2\mathbb{E}[B]\left(\int_0^x \overline{G}(t)\,\mathrm{d}t - x\overline{G}(x)\right)$. By Lemmas 4.4.1 and 4.4.5, one sees that $\beta((\cdot)\overline{G}(\cdot)) > 0$ and therefore it follows from Lemma 4.4.3 that

$$\limsup_{x\to\infty} \frac{m_2(x)}{2\mathbb{E}[B]x\overline{G}(x)} = \limsup_{x\to\infty} \frac{\int_0^x \overline{G}(t)\,\mathrm{d}t}{x\overline{G}(x)} - 1 < \infty. \tag{4.22}$$

Also, since $\beta((\cdot)\overline{F}(\cdot)) > -\infty$, Lemma 4.4.4(ii) indicates that

$$\limsup_{x\to\infty} \frac{x\overline{F}(x)}{\overline{G}(x)} = \limsup_{x\to\infty} \frac{\mathbb{E}[B]x\overline{F}(x)}{\int_x^\infty \overline{F}(t)\,\mathrm{d}t} < \infty.$$

Consequently, it follows from relation (4.19) that, for some $C, D > 0$ and all ρ sufficiently close to one, we have

$$\begin{aligned}\mathbb{E}[T_{\mathrm{FB}}] &\le \frac{\mathbb{E}[B]}{\rho}\log\frac{1}{1-\rho} + 2\int_0^\infty \frac{x\overline{F}(x)}{(1-\rho G(x))^2}\,\mathrm{d}G(x) + \frac{1}{\mathbb{E}[B]}\int_0^\infty \frac{m_2(x)}{x\overline{G}(x)}\frac{x\overline{F}(x)}{(1-\rho G(x))^2}\,\mathrm{d}G(x)\\ &\le \frac{\mathbb{E}[B]}{\rho}\log\frac{1}{1-\rho} + C\int_0^\infty \frac{x\overline{F}(x)}{\overline{G}(x)}\frac{1}{1-\rho G(x)}\,\mathrm{d}G(x) \le D\log\frac{1}{1-\rho},\end{aligned}$$

and therefore $\mathbb{E}[T_{\mathrm{FB}}] = \Theta\left(\log\frac{1}{1-\rho}\right)$.

4.4.2 Special cases

This section proves Theorem 4.3.2. The maximum domains of attraction of each of the extreme value distributions are considered in order, followed by a distribution with an atom in its right endpoint. The Fréchet and Weibull cases follow readily from Theorem 4.3.1 and the Dominated Convergence Theorem. The same approach works for the Gumbel case, although Theorem 4.3.1 is not directly applicable. Finally, the atom case follows readily by analysing the sojourn time of maximum-sized jobs.

Fréchet(α) and Weibull(α)

Theorems 4.2.3 and 4.2.5 together state that $F \in \text{MDA}(\Phi_\alpha)$ if and only if $\overline{F}(x) = L(x)x^{-\alpha}$. Karamata's theorem [24, Theorem 1.5.11] then states that $\mathbb{E}[B]\overline{G}(x) \sim x\overline{F}(x)/(\alpha-1)$ is regularly varying with index $-(\alpha-1)$. Consequently, Theorem 1.5.12 in Bingham et al. [24] states that $G^{\leftarrow}(1-1/x)$ is regularly varying with index $1/(\alpha-1)$ and therefore Proposition 1.5.7 in Bingham et al. states that $\overline{F}(G^{\leftarrow}(1-1/x))$ is regularly varying with index $-\alpha/(\alpha-1)$.

First assume $\alpha > 2$. We saw in Section 4.4.1 that the asymptotic behaviour of $\mathbb{E}[T_{\text{FB}}]$ is identical to the asymptotic behaviour of term III(ρ) (cf. relation (4.20)). Now, the Uniform Convergence Theorem [24, Theorem 1.5.2] states that $\frac{\overline{F}(G^{\leftarrow}(1-1/x))}{\overline{F}(G^{\leftarrow}(1-1/y))} \to \left(\frac{y}{x}\right)^{\alpha/(\alpha-1)}$ uniformly for all $0 < c < x, y < \infty$. We hence substitute $w = \frac{\nu-(1-\rho)}{\rho}$ and exploit the Dominated Convergence Theorem to obtain

$$\begin{aligned}
&\lim_{\rho\uparrow 1} \text{III}(\rho) \\
&\quad = \lim_{\rho\uparrow 1} \frac{\rho^2}{\mathbb{E}[B]} \int_0^1 (\rho w + 1 - \rho)\, m_2\left(G^{\leftarrow}\left(1 - \frac{(1-\rho)(1-w)}{1-\rho+\rho w}\right)\right) \frac{\overline{F}\left(G^{\leftarrow}\left(1 - \frac{(1-\rho)(1-w)}{1-\rho+\rho w}\right)\right)}{\overline{F}(G^{\leftarrow}(\rho))}\, \mathrm{d}w \\
&\quad = \frac{\mathbb{E}[B^2]}{\mathbb{E}[B]} \int_0^1 w \left(\frac{1-w}{w}\right)^{\alpha/(\alpha-1)} \mathrm{d}w \\
&\quad = \mathbb{E}[B^*] \frac{\pi/(\alpha-1)}{\sin(\pi/(\alpha-1))} \frac{\alpha}{\alpha-1}.
\end{aligned}$$

Similarly, Theorems 4.2.3 and 4.2.6 together state that $F \in \text{MDA}(\Psi_\alpha)$, $\alpha > 0$, if and only if $x_R < \infty$ and $\overline{F}(x_R - x^{-1}) = L(x)x^{-\alpha}$. The corresponding result then follows after noting that $\mathbb{E}[B]\overline{G}(x_R - x^{-1}) \sim L(x)x^{-\alpha-1}/(\alpha+1)$ is regularly varying with index $-(\alpha+1)$ and $\frac{\overline{F}(G^{\leftarrow}(1-1/x))}{\overline{F}(G^{\leftarrow}(1-1/y))} \to \left(\frac{y}{x}\right)^{\alpha/(\alpha+1)}$ uniformly for all $0 < c < x, y < \infty$.

Finally, assume that $F \in \text{MDA}(\Phi_\alpha)$, $\alpha \in (1,2)$. Then, Karamata's Theorem implies $m_2(x) = 2\int_0^x y\overline{F}(y)\,\mathrm{d}y \sim 2x^2\overline{F}(x)/(2-\alpha)$ as $x \to \infty$. We analyse relation (4.19) and again exploit the Dominated Convergence Theorem to find

$$\begin{aligned}
\mathbb{E}[T_{\text{FB}}] &= \frac{\mathbb{E}[B]}{\rho} \log\frac{1}{1-\rho} + 2\rho \int_0^\infty \frac{x\overline{F}(x)}{\overline{G}(x)} \frac{1-G(x)}{(1-\rho G(x))^2}\, \mathrm{d}G(x) \\
&\quad + \frac{\rho^2}{\mathbb{E}[B]} \int_0^\infty \frac{m_2(x)\overline{F}(x)}{\overline{G}(x)^2} \frac{(1-G(x))^2}{(1-\rho G(x))^3}\, \mathrm{d}G(x) \\
&\sim \mathbb{E}[B] \log\frac{1}{1-\rho} + 2(\alpha-1)\mathbb{E}[B] \int_0^1 \frac{1-y}{(1-\rho y)^2}\, \mathrm{d}y + \frac{2}{\mathbb{E}[B]} \frac{(\alpha-1)^2\mathbb{E}[B]^2}{2-\alpha} \int_0^1 \frac{(1-y)^2}{(1-\rho y)^3}\, \mathrm{d}y \\
&\sim \mathbb{E}[B] \log\frac{1}{1-\rho} + 2(\alpha-1)\mathbb{E}[B] \log\frac{1}{1-\rho} + \frac{2(\alpha-1)^2\mathbb{E}[B]}{2-\alpha} \log\frac{1}{1-\rho} = \frac{\alpha}{2+\alpha}\mathbb{E}[B] \log\frac{1}{1-\rho}
\end{aligned}$$

as $\rho \uparrow 1$.

Gumbel

If $F \in \text{MDA}(\Lambda)$, then so is G by Lemma 4.2.10 and we may choose h^* as the auxiliary function of G. Propositions 0.9(a), 0.10 and 0.12 in Resnick [119] together state that

$$a_G(x) := \frac{1}{h^*(G^{\leftarrow}(1-1/x))} = \frac{\mathbb{E}[B]}{x\overline{F}(G^{\leftarrow}(1-1/x))}$$

is 0-varying[1], implying that $\overline{F}(G^{\leftarrow}(1-1/x))$ is (-1)-varying.

Following the analysis in Section 4.4.1, we obtain $\alpha(\text{I}) = -1 < 0$ as before. Consider term $\text{II}(\rho)$. By Markov's inequality, we have $\overline{G}(x) \leq \mathbb{E}[B^*]/x$. Substituting $x = G^{\leftarrow}(1-w^{-1})$ then yields $G^{\leftarrow}(1-w^{-1}) \leq \mathbb{E}[B^*]w$, and hence

$$\begin{aligned}\text{II}(\rho) &= \frac{2(1-\rho)}{\overline{F}(G^{\leftarrow}(\rho))}\int_1^{\infty}\frac{\rho}{1-\rho}\left(\frac{\rho}{1-\rho}+w\right)^{-2}G^{\leftarrow}(1-w^{-1})\overline{F}(G^{\leftarrow}(1-w^{-1}))\,\mathrm{d}w \\ &\leq \frac{2\mathbb{E}[B^*](1-\rho)^{3/2}}{\rho^{1/2}\overline{F}(G^{\leftarrow}(\rho))}\int_1^{\infty} w^{1/2}\overline{F}(G^{\leftarrow}(1-w^{-1}))\,\mathrm{d}w.\end{aligned}$$

The term in front of the integral and the integrand both have upper Matuszewska index $-1/2$, and therefore $\text{II}(\rho) \to 0$.

Lastly, consider term $\text{III}(\rho)$. The relation $\limsup_{\rho\uparrow 1}\text{III}(\rho) < \infty$ follows analogously to the analysis in Section 4.4.1. Then, along the lines of the Fréchet and Weibull cases before, one may apply the Uniform Convergence Theorem and the Dominated Convergence Theorem to derive the theorem statement.

Atom in right endpoint

First, we show that $\text{I}(\rho)+\text{II}(\rho) = o(1)$. Lemma 4.2.8 states that $\lim_{x\uparrow x_R} h^*(x) = \infty$, and therefore $\lim_{\rho\uparrow 1}\text{I}(\rho) = \lim_{\rho\uparrow 1}\frac{(1-\rho)\log\frac{1}{1-\rho}}{\rho h^*(G^{\leftarrow}(\rho))} = 0$. Also, $G^{\leftarrow}$ is bounded from above by x_R and consequently $\lim_{\rho\uparrow 1}\text{II}(\rho) \leq \lim_{\rho\uparrow 1}\frac{2(1-\rho)}{\overline{F}(G^{\leftarrow}(\rho))}\cdot x_R = \lim_{\rho\uparrow 1}\frac{2x_R}{\mathbb{E}[B]h^*(G^{\leftarrow}(\rho))} = 0$.

It remains to show that $\text{III}(\rho) \to \mathbb{E}[B^*]$ and $\overline{F}(G^{\leftarrow}(\rho)) \to p$ as $\rho\uparrow 1$. The following lemma facilitates the analysis of this term. The proof of the lemma is postponed until the end of this section.

Lemma 4.4.8. *Let $f : D \to \mathbb{R}$ be any function that maps $D \subseteq \mathbb{R}$ onto $\mathbb{R}$, and assume that $\lim_{y\uparrow x} f(y) = p$ for some x in the closure $\overline{D}$ of D. Then, there exist $z > 0$ and $q > 0$ such that*

$$f(x-y) \leq p + qy \tag{4.23}$$

for all $y \in (0, z]$ that satisfy $x - y \in D$.

[1] The propositions regard Π- and Γ-varying functions; we consider these classes in Section 4.5.

Let $q>0$ and $\delta^*>0$ be such that $\overline{F}(x_R-\delta)\le p+q\delta$ for all $\delta\in(0,\delta^*]$. It follows that $\mathbb{E}[B]\overline{G}(x)=\int_x^{x_R}\overline{F}(y)\,\mathrm{d}y\sim p(x_R-x)$ as $x\uparrow x_R$, and hence $x_R-G^{\leftarrow}(u)\sim\mathbb{E}[B](1-u)/p$ as $u\uparrow 1$. Fix $\varepsilon>0$ and let $u^*\in(0,1)$ be such that $x_R-G^{\leftarrow}(u)\le(1+\varepsilon)\mathbb{E}[B](1-u)/p$ for all $u\in(u^*,1)$. Now, for all $u>\rho_0:=\max\{u^*,1-p\delta^*/((1+\varepsilon)\mathbb{E}[B])\}$ we have

$$p\le\overline{F}(G^{\leftarrow}(u))\le p+\frac{q}{p}(1+\varepsilon)\mathbb{E}[B](1-u)=:p+p\widetilde{q}(1-u) \tag{4.24}$$

and hence, for $\widetilde{q}=q(1+\varepsilon)\mathbb{E}[B]/p^2$, the relations

$$\frac{1}{1+\widetilde{q}(1-\rho)}\le\frac{\overline{F}\left(G^{\leftarrow}\left(1-\frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)}{\overline{F}(G^{\leftarrow}(\rho))}\le 1+\widetilde{q}\,\frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\le 1+\widetilde{q}\,\frac{1-\rho}{\rho}\frac{1}{\nu}$$

hold for all $\nu>\frac{1-\rho}{1-\rho\cdot\rho_0},\rho>\rho_0$.

Consider term $\mathrm{III}(\rho)$. On the one hand, we find

$$\begin{aligned}\limsup_{\rho\uparrow1}\mathrm{III}(\rho)&\le\limsup_{\rho\uparrow1}\frac{\mathbb{E}[B^2]}{\mathbb{E}[B]}\int_{1-\rho}^{1}\nu\,\frac{\overline{F}\left(G^{\leftarrow}\left(1-\frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)}{\overline{F}(G^{\leftarrow}(\rho))}\,\mathrm{d}\nu\\&\le\limsup_{\rho\uparrow1}\frac{\mathbb{E}[B^2]}{\mathbb{E}[B]}\int_{1-\rho}^{\frac{1-\rho}{1-\rho\cdot\rho_0}}\frac{1}{p}\,\mathrm{d}\nu+\frac{\mathbb{E}[B^2]}{\mathbb{E}[B]}\int_{\frac{1-\rho}{1-\rho\cdot\rho_0}}^{1}\left\{\nu+\widetilde{q}\,\frac{1-\rho}{\rho}\right\}\mathrm{d}\nu\\&\le\limsup_{\rho\uparrow1}\frac{\mathbb{E}[B^2]}{p\mathbb{E}[B]}\frac{1-\rho}{1-\rho\cdot\rho_0}+\frac{\mathbb{E}[B^2]}{2\mathbb{E}[B]}+\frac{\mathbb{E}[B^2]}{\mathbb{E}[B]}\widetilde{q}\,\frac{1-\rho}{\rho}=\mathbb{E}[B^*].\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\liminf_{\rho\uparrow1}\mathrm{III}(\rho)&\ge\liminf_{\rho\uparrow1}\frac{\rho m_2(G^{\leftarrow}(\rho_0))}{\mathbb{E}[B]}\int_{\frac{1-\rho}{1-\rho\cdot\rho_0}}^{1}\frac{\nu}{1+\widetilde{q}(1-\rho)}\,\mathrm{d}\nu\\&=\liminf_{\rho\uparrow1}\frac{m_2(G^{\leftarrow}(\rho_0))}{2\mathbb{E}[B]}\frac{\rho}{1+\widetilde{q}(1-\rho)}\left(1-\left(\frac{1-\rho}{1-\rho\cdot\rho_0}\right)^2\right)=\frac{m_2(G^{\leftarrow}(\rho_0))}{\mathbb{E}[B^2]}\cdot\mathbb{E}[B^*].\end{aligned}$$

Since ρ_0 may be chosen arbitrarily close to unity, we find $\mathbb{E}[T_{\mathrm{FB}}]\sim\frac{\mathbb{E}[B^*]\overline{F}(G^{\leftarrow}(\rho))}{(1-\rho)^2}\sim\frac{p\mathbb{E}[B^*]}{(1-\rho)^2}$ as $\rho\uparrow1$ where the last equivalence follows from (4.24). The section is concluded with the proof of Lemma 4.4.8.

Proof of Lemma 4.4.8. Without loss of generality, we assume that $(x-1,x)\subset D$. For sake of finding a contradiction, assume that the lemma statement is not true, i.e. for all $z>0$ and all $q>0$ there exists $\xi\in(0,z]$ such that

$$f(x-\xi)>p+q\xi. \tag{4.25}$$

Define $z_1:=1,q_1:=1$ and let $\xi_1\in(0,1]$ be such that (4.25) holds with $q=q_1$ and $\xi=\xi_1$. By definition of the left limit, for any $\varepsilon>0$ there exists $\eta^*>0$ such that $f(x-\eta)\le p+\varepsilon$ for all $\eta\in(0,\eta^*]$. In particular, by choosing $\varepsilon=q_1\xi_1$ we obtain $\eta^*=:\eta_2^*<\xi_1\le z_1$ such that $f(x-\eta)\le p+q_1\xi_1$ for all $\eta\in(0,\eta_2^*]$.

Define $z_2 := \min\{\eta_2^*, 1/2\}$ and set $q_2 := 1/z_2$. Again, there exists $\xi_2 \in (0, z_2]$ such that (4.25) holds for $q = q_2$ and $\xi = \xi_2$. By repeating the above procedure we obtain three sequences $(q_n)_{n\in\mathbb{N}}$, $(z_n)_{n\in\mathbb{N}}$ and $(\xi_n)_{n\in\mathbb{N}}$ such that $q_n = 1/z_n$, $0 < z_{n+1} < \xi_n < z_n \le 1/n$ and

$$f(x - \xi_n) > p + q_n \xi_n \tag{4.26}$$

for all $n \in \mathbb{N}$. From these properties, one may additionally deduce that $\xi_n > 1/q_{n+1}$, $\xi_n \downarrow 0$ and $q_n \to \infty$.

We will obtain a contradiction by showing that $(q_n)_{n\in\mathbb{N}}$ must also converge. If $\limsup_{n\to\infty} q_n\xi_n > 0$, then by relation (4.26) we must have $\limsup_{n\to\infty} f(x-\xi_n) \ge \limsup_{n\to\infty} p + q_n\xi_n > p$. However, this contradicts the lemma assumptions and therefore $\limsup_{n\to\infty} q_n\xi_n$ must equal zero. As such, we find $0 \le \limsup_{n\to\infty} q_n/q_{n+1} \le \limsup_{n\to\infty} q_n\xi_n = 0$ so that the sequence $(q_n)_{n\in\mathbb{N}}$ converges by the ratio test. □

Note that Lemma 4.4.8 can be applied generally to yield lower and upper bounds for $f(y)$ around any point $x \in \overline{D}$ for which either $\lim_{y\uparrow x} f(y)$ or $\lim_{y\downarrow x} f(y)$ exists.

4.5 Asymptotic relation for $h^*(G^{\leftarrow}(\rho))$ in the Gumbel case

This section is dedicated to the proof of Theorem 4.3.4. Theorem 4.2.7 states that c_n may be chosen as $1/h^*(F^{\leftarrow}(1-n^{-1}))$, so that Theorem 4.3.4 follows from Theorem 4.3.2 and an analysis of the limit $\lim_{n\to\infty} h^*(G^{\leftarrow}(1-n^{-1}))/h^*(F^{\leftarrow}(1-n^{-1})) = \lim_{y\uparrow 1}(1-y)^{-2}\overline{F}(G^{\leftarrow}(y))\overline{G}(F^{\leftarrow}(y))$. The proof heavily relies upon the work by De Haan [69] and Resnick [119], who both consider Γ- and Π-varying functions:

Definition 4.5.1. A function $U : (x_L, x_R) \to \mathbb{R}, \lim_{x\uparrow x_R} U(x) = \infty$ is in the class of Γ-varying functions if it is non-decreasing, and there exists a function $f : (x_L, x_R) \to \mathbb{R}_{\ge 0}$ satisfying

$$\lim_{x\uparrow x_R} \frac{U(x + tf(x))}{U(x)} = e^t \tag{4.27}$$

for all $t \in \mathbb{R}$. The function $f(\cdot)$ is called an *auxiliary function* and is unique up to asymptotic equivalence.

Definition 4.5.2. A function $V : (x_L, \infty) \to \mathbb{R}_{\ge 0}$ is in the class of Π-varying functions if it is non-decreasing, and there exist functions $a(x) > 0, b(x) \in \mathbb{R}$, such that

$$\lim_{x\to\infty} \frac{V(tx) - b(x)}{a(x)} = \log t \tag{4.28}$$

for all $t \in \mathbb{R}$. The function $a(\cdot)$ is called an *auxiliary function* and is unique up to asymptotic equivalence.

It turns out that Γ- and Π-varying functions are closely related to MDA(Λ). In particular, if $F \in \text{MDA}(\Lambda)$ with auxiliary function $1/h^*$, then Proposition 1.9 in Resnick [119] states that $U_F := 1/\overline{F} \in \Gamma$ with auxiliary function $f_F := 1/h^*$. Proposition 0.9(a) then states that $V_F(\cdot) := U_F^{\leftarrow}(\cdot) = \left(1/\overline{F}\right)^{\leftarrow}(\cdot) = F^{\leftarrow}\left(1-(\cdot)^{-1}\right) \in \Pi$ with auxiliary function $a_F(\cdot) := f_F(U_F^{\leftarrow}(\cdot)) = 1/h^*(F^{\leftarrow}(1-(\cdot)^{-1}))$. Similarly, using Lemma 4.2.10, we find that $U_G := 1/\overline{G} \in \Gamma$ and $V_G(\cdot) := U_G^{\leftarrow}(\cdot) = G^{\leftarrow}\left(1-(\cdot)^{-1}\right) \in \Pi$ with auxiliary function $a_G(\cdot) := 1/h^*(G^{\leftarrow}(1-(\cdot)^{-1}))$.

Now, since Theorem 4.2.7 states that the norming constants c_n may be chosen as $1/h^*(F^{\leftarrow}(1-n^{-1}))$, we are done once we show that $\lim_{n\to\infty} c_n h^*(G^{\leftarrow}(1-n^{-1})) = \lim_{x\to\infty} \frac{a_F(x)}{a_G(x)}$ tends to the right quantity for all cases in the theorem.

Corollary 3.4 in De Haan [69] states that[2] $\lim_{x\uparrow x_R} \frac{a_F(x)}{a_G(x)} = \xi^{-1} \in [0,\infty]$ if and only if there exist a positive function $b(x)$ with $\lim_{x\uparrow x_R} b(x) = \xi$ and constants $b_2 > 0$ and $b_3 \in \mathbb{R}$ such that[3] $P(x) = b_3 + \int_0^x b(t)\,\mathrm{d}t$ and $V_F^{\leftarrow}(x) \sim b_2 V_G^{\leftarrow}(P(x))$ as $x \uparrow x_R$. As $V_\bullet^{\leftarrow}(x) = (U_\bullet^{\leftarrow})^{\leftarrow}(x) \sim U_\bullet(x)$ [119, p.44], this is equivalent to finding a function $P(x)$, of the given form, that satisfies

$$\lim_{x\uparrow x_R} \frac{\overline{G}(P(x))}{b_2\overline{F}(x)} = \lim_{x\uparrow x_R} \frac{U_F(x)}{b_2 U_G(P(x))} = \lim_{x\uparrow x_R} \frac{V_F^{\leftarrow}(x)}{b_2 V_G^{\leftarrow}(P(x))} = 1. \tag{4.29}$$

We use the following lemma, proven at the end of this section, to construct a suitable $P(x)$:

Lemma 4.5.3. *Let F be a c.d.f. Then, there exists a strictly increasing, continuous c.d.f. $F_\uparrow(x)$ satisfying both $\overline{F}_\uparrow(x) \sim \overline{F}(x)$ and $\overline{G}(F_\uparrow(x)) \sim \overline{G}(F(x))$ as $x \uparrow x_R$.*

As $G^{\leftarrow}(F_\uparrow(x))$ is strictly increasing, there exists a positive function $b(\cdot)$ such that $\int_0^x b(t)\,\mathrm{d}t = G^{\leftarrow}(F_\uparrow(x))$. Therefore, we see that (4.29) is satisfied with $b_2 = 1$ and $b_3 = 0$. The result follows once we show that

$$\lim_{x\to\infty} b(x) = \lim_{x\to\infty} \frac{P(x)}{x} = \lim_{x\to\infty} \frac{G^{\leftarrow}(F(x))}{x} = \xi \tag{4.30}$$

if $x_R = \infty$, and once we show that

$$\lim_{x\uparrow x_R} b(x) = \lim_{x\uparrow x_R} \frac{P(x_R)-P(x)}{x_R - x} = \lim_{x\uparrow x_R} \frac{x_R - G^{\leftarrow}(F(x))}{x_R - x} = \xi \tag{4.31}$$

if $x_R < \infty$.

The right-hand sides of both (4.30) and (4.31) depend on the function $G^{\leftarrow}(F(x))$. The advantage of this representation is apparent from the following key relation, which connects $G^{\leftarrow}(F(x))$ to $h^*(x)$:

$$\mathbb{E}[B]h^*(x) = \exp\left[\int_{G^{\leftarrow}(F(x))}^{x} h^*(t)\,\mathrm{d}t\right]. \tag{4.32}$$

[2] Here, we denote $0^{-1} = +\infty$.

[3] Their paper only considers the $x_R = \infty$ case; however, the proof also holds for finite x_R.

Relation (4.32) follows readily from $h^*(x) = -\frac{\mathrm{d}}{\mathrm{d}x}\log\overline{G}(x)$. In the upcoming analysis, we first focus on (4.30) and then consider (4.31).

4.5.1 Infinite support

First assume $x_R = \infty$. The following theorem relates the assumptions on $\overline{F}(x)$ to properties of $h^*(x)$:

Theorem 4.5.4 (Beirlant et al. [20], Theorem 2.1)**.**

(i) If there exists $\alpha > 0$ and a slowly varying function $l(x)$ such that $-\log\overline{F}(x) \sim l(x)x^{\alpha}$ as $x \to \infty$, then $h^*(x) \sim \alpha l(x)x^{\alpha-1}$ as $x \to \infty$ if and only if

$$\lim_{\lambda\downarrow 1}\liminf_{x\to\infty}\inf_{t\in[1,\lambda]}\{\log h^*(tx) - \log h^*(x)\} \geq 0. \tag{4.33}$$

(ii) If there exists a function $l(x) : [0,\infty) \to \mathbb{R}$, $\liminf_{x\to\infty} l(x) > 1$ such that for all $\lambda > 0$

$$\lim_{x\to\infty}\frac{-\log\overline{F}(\lambda x) + \log\overline{F}(x)}{l(x)} = \log(\lambda), \tag{4.34}$$

then $l(x)$ is slowly varying and $h^*(x) \sim (l(x)-1)/x$ as $x \to \infty$.

The cases in Theorem 4.3.4 correspond to the cases in Theorem 4.5.4. We will consider the implications of Theorem 4.5.4 to derive the results presented in Theorem 4.3.4.

(i) Assume $h^*(x) \sim \alpha l(x)x^{\alpha-1}$, $\alpha > 0$, and note that

$$\lim_{x\to\infty}\frac{-\log(\mathbb{E}[B]h^*(x))}{xh^*(x)} = \lim_{x\to\infty}\frac{-\log(\mathbb{E}[B]\alpha l(x)) - (\alpha-1)\log(x)}{\alpha l(x)x^{\alpha}} = 0.$$

We will prove the relation $\lim_{x\to\infty} G^{\leftarrow}(F(x))/x = 1$ by contradiction. Specifically, if $\limsup_{x\to\infty} G^{\leftarrow}(F(x))/x > 1$ then there exists $\varepsilon > 0$ and a sequence $(x_n)_{n\in\mathbb{N}}$, $x_n \to \infty$, such that $G^{\leftarrow}(F(x_n))/x_n \geq 1+\varepsilon$ for all $n \in \mathbb{N}$. The Uniform Convergence Theorem [24, Theorems 1.2.1 and 1.5.2] then implies

$$\begin{aligned}\frac{-\log(\mathbb{E}[B]h^*(x_n))}{x_n h^*(x_n)} &= \int_{x_n}^{G^{\leftarrow}(F(x_n))}\frac{h^*(t)}{x_n h^*(x_n)}\,\mathrm{d}t = \int_1^{G^{\leftarrow}(F(x_n))/x_n}\frac{h^*(\tau x_n)}{h^*(x_n)}\,\mathrm{d}\tau \\ &\geq \int_1^{1+\varepsilon}\frac{h^*(\tau x_n)}{h^*(x_n)}\,\mathrm{d}\tau \sim \int_1^{1+\varepsilon}\tau^{\alpha-1}\,\mathrm{d}\tau = \alpha^{-1}((1+\varepsilon)^{\alpha}-1)\end{aligned}$$

for every $n \in \mathbb{N}$. However, this contradicts with $\lim_{x\to\infty}\frac{\log(\mathbb{E}[B]h^*(x))}{xh^*(x)} = 0$ and it follows that $\liminf_{x\to\infty} G^{\leftarrow}(F(x))/x \leq 1$. One may similarly verify the inequality $\liminf_{x\to\infty} G^{\leftarrow}(F(x))/x \geq 1$, so that $\lim_{x\to\infty} G^{\leftarrow}(F(x))/x = 1$ as claimed.

(ii) Alternatively, assume $h^*(x) \sim \frac{l(x)-1}{x}$ and denote $L = \lim_{x\to\infty} \log(x)/l(x) \in [0,\infty]$. Then Lemma 4.2.8 states that $l(x) \to \infty$ and as such

$$\lim_{x\to\infty} \frac{-\log(\mathbb{E}[B]h^*(x))}{xh^*(x)} = \lim_{x\to\infty} \frac{-\log(\mathbb{E}[B]) - \log(l(x)-1) + \log(x)}{l(x)-1} = L. \tag{4.35}$$

Now, if $L = 0$ then the analysis in (i) yields $\lim_{x\to\infty} G^{\leftarrow}(F(x))/x = 1$. If $L \in (0,\infty)$ then (4.32) and (4.35) imply

$$\begin{aligned} L &= \lim_{x\to\infty} \frac{-\log(\mathbb{E}[B]h^*(x))}{xh^*(x)} = \lim_{x\to\infty} \int_x^{G^{\leftarrow}(F(x))} \frac{h^*(t)}{xh^*(x)}\,\mathrm{d}t \\ &= \lim_{x\to\infty} \frac{1}{\log(x)} \int_x^{G^{\leftarrow}(F(x))} \frac{l(t)-1}{\log t} \cdot \frac{\log(x)}{l(x)-1} \cdot \frac{\log(t)}{t}\,\mathrm{d}t \\ &= \lim_{x\to\infty} \frac{1}{\log(x)} \int_x^{G^{\leftarrow}(F(x))} \frac{\log(t)}{t}\,\mathrm{d}t = \lim_{x\to\infty} \frac{\log^2(G^{\leftarrow}(F(x))) - \log^2(x)}{2\log(x)}. \end{aligned}$$

Writing $G^{\leftarrow}(F(x)) = u(x)x$, $u(x)x \to \infty$, now yields

$$L = \lim_{x\to\infty} \log(u(x)) \left(1 + \frac{\log(u(x))}{2\log(x)}\right),$$

from which we conclude $u(x) \to e^L$ and consequently $\lim_{x\to\infty} G^{\leftarrow}(F(x))/x = e^L$.

Finally, if $L = \infty$ then $h^*(x) \downarrow 0$ and therefore $G^{\leftarrow}(F(x)) \geq x$ by (4.32). For sake of contradiction, assume $\liminf_{x\to\infty} G^{\leftarrow}(F(x))/x < \infty$. Then there exists $M_0 \geq 1$ such that for all $M \geq M_0$ there exists a sequence $(x_n)_{n\in\mathbb{N}}$, $x_n \to \infty$, such that $G^{\leftarrow}(F(x_n))/x_n \leq M$ for every $n \in \mathbb{N}$. A similar analysis as in (i) then shows that this contradicts relation (4.35), and therefore $\lim_{x\to\infty} G^{\leftarrow}(F(x))/x = \infty$.

4.5.2 Finite support

Now assume $x_R < \infty$. Theorem 4.2.7 states that $\overline{F}(x)$ can be represented as

$$\overline{F}(x) = c(x) \exp\left\{-\int_z^x g(t)h^*(t)\,\mathrm{d}t\right\}, \quad z < x < x_R,$$

where c and g are measurable functions satisfying $c(x) \to c > 0$, $g(t) \to 1$ as $x \uparrow x_R$, and the auxiliary function $f_F(\cdot) = 1/h^*(\cdot)$ is positive, absolutely continuous and has density $f_F'(x)$ satisfying $\lim_{x\uparrow x_R} f_F'(x) = 0$. It is easily verified that the c.d.f. $\overline{F}_\infty(x) := \overline{F}(x_R - x^{-1})$, $x \geq (x_R - z)^{-1}$, is also in MDA(Λ) with auxiliary function $f_\infty(x) := x^2/h^*(x_R - x^{-1})$. From this representation it is straightforward to obtain a finite-support equivalent of Theorem 4.5.4:

Corollary 4.5.5. *Assume $x_R < \infty$.*

(i) If there exists $\alpha > 0$ and a slowly varying function $l(x)$ such that $-\log\overline{F}(x_R - x^{-1}) \sim l(x)x^\alpha$ as $x \to \infty$, then $h^*(x_R - x^{-1}) \sim \alpha l(x)x^{\alpha+1}$ as $x \to \infty$ if and only if

$$\lim_{\lambda\downarrow 1} \liminf_{x\to\infty} \inf_{t\in[1,\lambda]} \{\log h^*(x_R - (tx)^{-1}) - \log h^*(x_R - x^{-1}) - 2\log(t)\} \geq 0. \tag{4.36}$$

(ii) If there exists a function $l(x) : [0,\infty) \to \mathbb{R}$, $\liminf_{x\to\infty} l(x) > 1$, *such that for all* $\lambda > 0$

$$\lim_{x\to\infty} \frac{-\log \overline{F}(x_R - (\lambda x)^{-1}) + \log \overline{F}(x_R - x^{-1})}{l(x)} = \log(\lambda), \tag{4.37}$$

then $l(x)$ *is slowly varying and* $h^*(x_R - x^{-1}) \sim (l(x)-1)x$ *as* $x \to \infty$.

Again, the cases in Theorem 4.3.4 correspond to the cases in Corollary 4.5.5. The proof for the finite support case is similar to the infinite support case, yet we state it for completeness. Since $h^*(x) \to \infty$ as $x \uparrow x_R$ in both cases, relation (4.32) implies that $\frac{x_R - G^{\leftarrow}(F(x))}{x_R - x} \geq 1$ for all x sufficiently close to x_R.

(i) Assume $h^*(x_R - x^{-1}) \sim \alpha\, l(x) x^{\alpha+1}$, $\alpha > 0$, and note that

$$\begin{aligned}\lim_{x\uparrow x_R} \frac{-\log(\mathbb{E}[B] h^*(x))}{(x_R - x) h^*(x)} &= \lim_{y\to\infty} \frac{-\log(\mathbb{E}[B] h^*(x_R - y^{-1}))}{h^*(x_R - y^{-1})/y} \\ &= \lim_{y\to\infty} \frac{-\log(\mathbb{E}[B]\alpha l(y)) - (\alpha+1)\log(y)}{\alpha l(y) y^{\alpha}} = 0.\end{aligned}$$

We will show that $\lim_{x\to\infty} \frac{x_R - G^{\leftarrow}(F(x))}{x_R - x} = 1$ by contradiction. By our previous remark, we only need to show $\limsup_{x\to\infty} \frac{x_R - G^{\leftarrow}(F(x))}{x_R - x} \leq 1$. If this is false, then there exists $\varepsilon \in (0,1)$ and a sequence $(x_n)_{n\in\mathbb{N}}$, $x_n \uparrow x_R$, such that $\frac{x_R - x_n}{x_R - G^{\leftarrow}(F(x_n))} \leq 1 - \varepsilon$ for all $n \in \mathbb{N}$. As before, the Uniform Convergence Theorem [24, Theorems 1.2.1 and 1.5.2] then implies

$$\begin{aligned}-\frac{\log(\mathbb{E}[B] h^*(x_n))}{(x_R - x_n) h^*(x_n)} &= \int_{x_n}^{G^{\leftarrow}(F(x_n))} \frac{h^*(t)}{(x_R - x_n) h^*(x_n)}\, \mathrm{d}t \\ &= \int_{\frac{x_R - x_n}{x_R - G^{\leftarrow}(F(x_n))}}^{1} \frac{h^*(x_R - (x_R - x_n)\tau^{-1})}{\tau^2 h^*(x_R - (x_R - x_n))}\, \mathrm{d}\tau \\ &\geq \int_{1-\varepsilon}^{1} \frac{h^*(x_R - (x_R - x_n)\tau^{-1})}{\tau^2 h^*(x_R - (x_R - x_n))}\, \mathrm{d}\tau \\ &\sim \int_{1-\varepsilon}^{1} \tau^{\alpha-1}\, \mathrm{d}\tau = \alpha^{-1}(1 - (1-\varepsilon)^{\alpha})\end{aligned}$$

for every $n \in \mathbb{N}$, which contradicts with $\lim_{x\uparrow x_R} \frac{\log(\mathbb{E}[B] h^*(x))}{(x_R - x) h^*(x)} = 0$.

(ii) Now, assume $h^*(x_R - x^{-1}) \sim (l(x)-1)x$ and let $L = \lim_{x\to\infty} \log(x)/l(x) \in [0,\infty]$. Lemma 4.2.8 implies $l(x) \to \infty$, so that

$$\lim_{x\uparrow x_R} \frac{-\log(\mathbb{E}[B] h^*(x))}{(x_R - x) h^*(x)} = \lim_{y\to\infty} \frac{-\log(\mathbb{E}[B](l(y)-1)) - \log(y)}{l(y) - 1} = -L. \tag{4.38}$$

If $L=0$, then $\lim_{x\to\infty}\frac{x_R-G^{\leftarrow}(F(x))}{x_R-x}=1=e^0$ by the analysis in (i). Alternatively, if $L\in(0,\infty)$ then (4.32) and (4.38) imply

$$\begin{aligned}
L&=\lim_{x\uparrow x_R}\frac{\log(\mathbb{E}[B]h^*(x))}{(x_R-x)h^*(x)}=\lim_{x\uparrow x_R}\int_{G^{\leftarrow}(F(x))}^{x}\frac{h^*(t)}{(x_R-x)h^*(x)}\,\mathrm{d}t\\
&=\lim_{x\uparrow x_R}\int_{\frac{1}{x_R-G^{\leftarrow}(F(x))}}^{\frac{1}{x_R-x}}\frac{h^*(x_R-\tau^{-1})}{(x_R-x)\tau^2h^*(x_R-(x_R-x))}\,\mathrm{d}\tau\\
&=\lim_{x\uparrow x_R}\frac{1}{\log((x_R-x)^{-1})}\int_{\frac{1}{x_R-G^{\leftarrow}(F(x))}}^{\frac{1}{x_R-x}}\frac{l(\tau)-1}{\log(\tau)}\cdot\frac{\log((x_R-x)^{-1})}{l((x_R-x)^{-1})-1}\cdot\frac{\log(\tau)}{\tau}\,\mathrm{d}\tau\\
&=\lim_{x\uparrow x_R}\frac{1}{\log(x_R-x)}\int_{\frac{1}{x_R-x}}^{\frac{1}{x_R-G^{\leftarrow}(F(x))}}\frac{\log(\tau)}{\tau}\,\mathrm{d}\tau\\
&=\lim_{x\uparrow x_R}\frac{\log^2(x_R-G^{\leftarrow}(F(x)))-\log^2(x_R-x)}{2\log(x_R-x)}.
\end{aligned}$$

Write $G^{\leftarrow}(F(x))=x_R-(x_R-x)u(x)$ where $(x_R-x)u(x)\to 0$ for all x sufficiently close to x_R. One then obtains

$$L=\lim_{x\uparrow x_R}\log(u(x))\left(1+\frac{\log(u(x))}{2\log(x_R-x)}\right),$$

implying $u(x)\to e^L$ and subsequently $\lim_{x\to\infty}\frac{x_R-G^{\leftarrow}(F(x))}{x_R-x}=e^L$.

Lastly, consider $L=\infty$ and assume $\limsup_{x\to\infty}\frac{x_R-G^{\leftarrow}(F(x))}{x_R-x}<\infty$ for sake of contradiction. Then there exists $M_0\geq 1$ such that for all $M\geq M_0$ there exists a sequence $(x_n)_{n\in\mathbb{N}}$, $x_n\uparrow x_R$, such that $\frac{x_R-G^{\leftarrow}(F(x_n))}{x_R-x_n}\leq M$ for every $n\in\mathbb{N}$. A similar analysis as in (i) then shows that this contradicts relation (4.38), and therefore $\lim_{x\to\infty}\frac{x_R-G^{\leftarrow}(F(x))}{x_R-x}=\infty$.

4.5.3 Proof of Lemma 4.5.3

For any positive, non-increasing $\phi:[0,1)\to(0,1)$ that vanishes as the argument tends to unity, we may define

$$F_\phi(x):=\begin{cases}F(x) & \text{if } x<s_1, \text{ and}\\ F(x)+\frac{x-s_n}{s_{n+1}-s_n}(F(s_{n+1})-F(x)) & \text{if } s_n\leq x<s_{n+1}, n\geq 1,\end{cases}\tag{4.39}$$

where $s_1:=0$ and $s_{n+1}:=\inf\left\{x\geq 0: F(x)\geq\frac{F(s_n)+\phi(F(s_n))}{1+\phi(F(s_n))}\right\}$ forms a strictly increasing sequence. Now, if $s_n\uparrow s^*<x_R$ then $F(s_n)\uparrow p$ for some $p\in(0,1)$ and therefore, for any $\varepsilon\in(0,1)$ and all n sufficiently large, we have $(1-\varepsilon)p\leq F(s_n)\leq p$. Consequently, s_{n+1} must satisfy $p\geq F(s_{n+1})\geq\frac{(1-\varepsilon)p+\phi(p)}{1+\phi(p)}$, which yields a contradiction if $\varepsilon<\phi(p)\frac{1-p}{p}$. We conclude that F_ϕ is a strictly increasing, continuous c.d.f. that satisfies $\overline{F}_\phi(x)\leq\overline{F}(x)$ for all x.

Define $n(x) := \sup\{n \in \mathbb{N} : s_{n-1} \le x\}$. Then

$$\begin{aligned}\frac{\overline{F}_\phi(x)}{\overline{F}(x)} &= 1 - \frac{x - s_{n(x)}}{s_{n(x)+1} - s_{n(x)}} \frac{F(s_{n(x)+1}) - F(x)}{\overline{F}(x)} \ge 1 - \frac{F(s_{n(x)+1}) - F(s_{n(x)})}{1 - F(s_{n(x)+1})} \\ &\ge 1 - \frac{\frac{F(s_{n(x)})+\phi(F(s_{n(x)}))}{1+\phi(F(s_{n(x)}))} - F(s_{n(x)})}{1 - \frac{F(s_{n(x)})+\phi(F(s_{n(x)}))}{1+\phi(F(s_{n(x)}))}} = 1 - \phi(F(s_{n(x)})) \to 1 \end{aligned} \tag{4.40}$$

as $x \uparrow x_R$, so that $\overline{F}_\uparrow(x) \sim \overline{F}(x)$ by our earlier remark.

Let $(s_n)_{n\in\mathbb{N}}$ and $(\widetilde{s}_n)_{n\in\mathbb{N}}$ be the sequences associated with F_ϕ and $F_{\widetilde{\phi}}$ and assume $\widetilde{\phi}(y) \le \phi(y)$ for all $y \in [0,1)$. We prove $\widetilde{s}_n \le s_n$ for all $n \in \mathbb{N}$ by induction. The inequality $\widetilde{s}_1 \le s_1$ is immediate from the definition. Now, assume that $\widetilde{s}_n \le s_n$ and observe that $(F(s)+q)/(1+q)$ is non-decreasing in s for every $q \ge 0$, and in q for every $s \in \mathbb{R}$. Thus, any x that satisfies $F(x) \ge (F(s_n) + \phi(F(s_n)))/(1+\phi(F(s_n)))$ evidently satisfies $F(x) \ge (F(\widetilde{s}_n) + \widetilde{\phi}(F(\widetilde{s}_n)))/(1+\widetilde{\phi}(F(\widetilde{s}_n)))$ and hence $\widetilde{s}_{n+1} \le s_{n+1}$.

As $F_\phi(x) \ge F(x)$ implies $G^{\leftarrow}(F_\phi(x)) \ge G^{\leftarrow}(F(x))$, the proof is complete once we show that there is a version of ϕ such that $\limsup_{x\uparrow x_R} \frac{G^{\leftarrow}(F(x))}{G^{\leftarrow}(F_\phi(x))} \ge 1$. To this end, we construct a suitable ϕ inductively.

Fix $\phi_1 := 1/2$. Then, for $n = 1,2,\ldots$, let $r_{n+1} := \inf\left\{x \ge 0 : F(x) \ge \frac{F(s_n)+\phi_n}{1+\phi_n}\right\}$, denote $\phi_{n+1} := \min\{\phi_n, \overline{F}(G^{\leftarrow}(r_{n+1}))^2/(4\mathbb{E}[B]^2)\}$ and define $\phi(y) := \phi_{n+1}$ for $y \in [F(s_n), F(s_{n+1}))$.

Since $\phi(F(s_n)) \le \phi_n$, it must be that $s_n \le r_n$ for all $n \in \mathbb{N}$. As a consequence, $\phi(F(s_n)) \le 2^{-2}\mathbb{E}[B]^{-2}\overline{F}(G^{\leftarrow}(s_{n+1}))^2$. Writing $\eta(x) := \phi(F(s_{n(x)}))$ for notational convenience, one may now use (4.40) to deduce

$$\begin{aligned} G^{\leftarrow}(F(x)) &= \inf\{z \in \mathbb{R} : G(z) \ge F(x)\} = \inf\{z \in \mathbb{R} : \overline{G}(z) \le \overline{F}(x)\} \\ &\ge \inf\left\{z \in \mathbb{R} : \overline{G}(z) \le \frac{\overline{F}_\phi(x)}{1-\eta(x)}\right\} \\ &= \inf\left\{z - \sqrt{\eta(x)} \in \mathbb{R} : \overline{G}(z) + \mathbb{E}[B]^{-1}\int_{z-\sqrt{\eta(x)}}^{z} \overline{F}(t)\,\mathrm{d}t \le \overline{F}_\phi(x) + \frac{\eta(x)}{1-\eta(x)}\overline{F}_\phi(x)\right\} \\ &\ge \inf\left\{z \in \mathbb{R} : \overline{G}(z) + \mathbb{E}[B]^{-1}\sqrt{\eta(x)\overline{F}(z)} \le \overline{F}_\phi(x) + \frac{\eta(x)}{1-\eta(x)}\right\} - \sqrt{\eta(x)} \\ &\ge G^{\leftarrow}(F_\phi(x)) - \sqrt{\eta(x)}, \end{aligned}$$

where the last inequality follows from the relation

$$\begin{aligned} \frac{\eta(x)}{1-\eta(x)} - \mathbb{E}[B]^{-1}\sqrt{\eta(x)\overline{F}(z)} &\le \frac{\phi(F(s_{n(x)}))}{1-\phi(F(s_{n(x)}))} - \mathbb{E}[B]^{-1}\sqrt{\phi(F(s_{n(x)}))\overline{F}(G^{\leftarrow}(F(s_{n(x)+1})))} \\ &\le \sqrt{\phi(F(s_{n(x)}))}\left[2\sqrt{\phi(F(s_{n(x)}))} - \mathbb{E}[B]^{-1}\overline{F}(G^{\leftarrow}(F(s_{n(x)+1})))\right] \le 0 \end{aligned}$$

for all $z \le G^{\leftarrow}(F_\phi(x)) \le G^{\leftarrow}(F(s_{n(x)+1}))$. We conclude that $\overline{G}(F_\uparrow(x)) \sim \overline{G}(F(x))$ as $x \uparrow x_R$.

4.6 Scaled sojourn time tends to zero in probability

The current section is dedicated to the proof of Theorem 4.3.6. The intuition behind the proof is that the sojourn times of all jobs of size at most $\widetilde{x}_\rho$ grow slower than $\mathbb{E}[T_{\mathrm{FB}}]$, where $\widetilde{x}_\rho$ is a function that depends on F. Alternatively, the fraction of jobs of size at least $\widetilde{x}_\rho$ tends to zero, since $\widetilde{x}_\rho \to x_R$ as $\rho \uparrow 1$. Section 4.7 discusses the sojourn time of these jobs in more detail.

For any $\varepsilon > 0$ we have

$$\mathbb{P}\left(\frac{T_{\mathrm{FB}}}{\mathbb{E}[T_{\mathrm{FB}}]} > \varepsilon\right) = \int_0^\infty \mathbb{P}(T_{\mathrm{FB}}(x) > \varepsilon\mathbb{E}[T_{\mathrm{FB}}])\,\mathrm{d}F(x) \le \mathbb{P}(T_{\mathrm{FB}}(\widetilde{x}_\rho) > \varepsilon\mathbb{E}[T_{\mathrm{FB}}]) + \overline{F}(\widetilde{x}_\rho), \quad (4.41)$$

where the final term vanishes as $\rho \uparrow 1$ by choice of $\widetilde{x}_\rho$. The proof is completed if the first probability at the right-hand side also vanishes as $\rho \uparrow 1$.

In preparation for the analysis of $\mathbb{P}(T_{\mathrm{FB}}(\widetilde{x}_\rho) > \varepsilon\mathbb{E}[T_{\mathrm{FB}}])$, reconsider the busy period representation $T_{\mathrm{FB}}(x) \overset{d}{=} \mathscr{L}_x(W_x + x)$. This relation states that the sojourn time of a job of size x is equal in distribution to a busy period with job sizes $B_i \wedge x$, initiated by the job of size x itself and the time W_x required to serve all jobs already in the system up to level x. Here, the random variable W_x is equal in distribution to the steady-state waiting time in an M/GI/1/FIFO queue with job sizes $B_i \wedge x$.

Let $N_x(t)$ denote a Poisson process with rate $\rho_x/\mathbb{E}[B \wedge x]$. Then, it follows from the busy period representation of T_{FB} that

$$\begin{aligned}
\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}}(x) > y) &= \mathbb{P}(\mathscr{L}_x(W_x + x) > (1-\rho)^{-2} y) \\
&= \mathbb{P}\left(\inf\left\{t \ge 0 : \sum_{i=1}^{N(t)} (B_i \wedge x) - t \le -(W_x + x)\right\} > (1-\rho)^{-2} y\right) \\
&= \mathbb{P}\left(\inf_{t \in [0,(1-\rho)^{-2} y]} \left\{\sum_{i=1}^{N(t)} (B_i \wedge x) - t\right\} \ge -(W_x + x)\right) \\
&= \mathbb{P}\left(\sup_{t \in [0,y]} \left\{\frac{t}{(1-\rho)^2} - \sum_{i=1}^{N((1-\rho)^{-2} t)} (B_i \wedge x)\right\} \le W_x + x\right). \quad (4.42)
\end{aligned}$$

Additionally, application of Chebychev's inequality to the above relation yields

$$\begin{aligned}
\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}}(x) > y) &\le \mathbb{P}\left(\frac{y}{(1-\rho)^2} - \sum_{i=1}^{N((1-\rho)^{-2} y)} (B_i \wedge x) \le W_x + x\right) \\
&\le \mathbb{P}\Bigg(\left|W_x + \sum_{i=1}^{N((1-\rho)^{-2} y)} (B_i \wedge x) - \frac{\rho_x}{1-\rho_x}\mathbb{E}[(B \wedge x)^*] - \frac{\rho_x}{(1-\rho)^2} y\right| \ge \\
&\qquad \frac{1-\rho_x}{(1-\rho)^2} y - x - \frac{\rho_x}{1-\rho_x}\mathbb{E}[(B \wedge x)^*]\Bigg) \\
&\le \frac{\mathbb{V}\mathrm{ar}[W_x] + \mathbb{V}\mathrm{ar}\left[\sum_{i=1}^{N((1-\rho)^{-2} y)} (B_i \wedge x)\right]}{\left(\frac{1-\rho_x}{(1-\rho)^2} y - x - \frac{\rho_x}{1-\rho_x}\mathbb{E}[(B \wedge x)^*]\right)^2}
\end{aligned}$$

$$= \frac{\frac{\rho_x^2}{(1-\rho_x)^2}\mathbb{E}[(B\wedge x)^*]^2 + \frac{\rho_x}{1-\rho_x}\mathbb{E}[((B\wedge x)^*)^2] + \frac{2\rho_x\mathbb{E}[(B\wedge x)^*]}{(1-\rho)^2}y}{\left(\frac{1-\rho_x}{(1-\rho)^2}y - x - \frac{\rho_x}{1-\rho_x}\mathbb{E}[(B\wedge x)^*]\right)^2}. \quad (4.43)$$

At this point, similar to the approach in Section 4.4, we distinguish between the finite and infinite variance cases.

4.6.1 Finite variance

This section considers all functions F that satisfy one of the conditions in the theorem statement and have finite variance. Specifically, this excludes the case $x_R = \infty$ for $\beta(\overline{F}) > -2$. Fix

$$\widetilde{p}(F) := \begin{cases} \frac{\beta(\overline{F})}{\beta(\overline{F})+1} & \text{if } F \notin \mathrm{MDA}(\Lambda) \text{ and } x_R = \infty, \\ \frac{\beta(\overline{F}(x_R-(\cdot)^{-1})}{\beta(\overline{F}(x_R-(\cdot)^{-1}))-1} & \text{if } F \notin \mathrm{MDA}(\Lambda) \text{ and } x_R < \infty, \text{ and} \\ 1 & \text{if } F \in \mathrm{MDA}(\Lambda), \end{cases} \quad (4.44)$$

and $\widetilde{\gamma} \in (\widetilde{p}(F)/2, 1)$, and define $v(\rho) := (1-\rho)^{\widetilde{\gamma}}$ and $\widetilde{x}_\rho := x_\rho^{v(\rho)} = G^{\leftarrow}\left(1 - \frac{1-\rho}{\rho}\frac{1-v(\rho)}{v(\rho)}\right)$. Indeed $\widetilde{x}_\rho \to x_R$, and we proceed with the analysis in (4.43). Noting that $\mathbb{E}[((B\wedge x)^*)^2] = \frac{\mathbb{E}[(B\wedge x)^3]}{3\mathbb{E}[B]} \le \frac{x\mathbb{E}[B^2]}{3\mathbb{E}[B]} = \frac{2}{3}\mathbb{E}[B^*]x$ and substituting $x = \widetilde{x}_\rho$, gives

$$\begin{aligned}\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}}(\widetilde{x}_\rho) > y) &\le \frac{\left(\frac{1-\rho}{1-\rho_{\widetilde{x}_\rho}}\right)^2\mathbb{E}[B^*]^2 + \frac{1-\rho}{1-\rho_{\widetilde{x}_\rho}}\frac{2}{3}\mathbb{E}[B^*](1-\rho)\widetilde{x}_\rho + 2\mathbb{E}[B^*]y}{\left(\frac{1-\rho_{\widetilde{x}_\rho}}{1-\rho}y - (1-\rho)\widetilde{x}_\rho - \frac{1-\rho}{1-\rho_{\widetilde{x}_\rho}}\rho_{\widetilde{x}_\rho}\mathbb{E}[B^*]\right)^2} \\ &= \frac{\mathbb{E}[B^*]^2 v(\rho)^2 + \frac{2}{3}\mathbb{E}[B^*]v(\rho)(1-\rho)x_\rho^{v(\rho)} + 2\mathbb{E}[B^*]y}{\left(v(\rho)^{-1}y - (1-\rho)x_\rho^{v(\rho)} - \rho_{x_\rho^{v(\rho)}}\mathbb{E}[B^*]v\right)^2}.\end{aligned}$$

We now return to the probability $\mathbb{P}(T_{\mathrm{FB}}(\widetilde{x}_\rho) > \varepsilon\mathbb{E}[T_{\mathrm{FB}}])$ in relation (4.41). By Theorems 4.3.1 and 4.3.2, there exists $C > 0$ such that the inequality $(1-\rho)^2\mathbb{E}[T_{\mathrm{FB}}] \ge C\overline{F}(G^{\leftarrow}(\rho))$ holds true for all ρ sufficiently close to one. Denoting $\widetilde{\varepsilon} := \varepsilon C$, this gives

$$\begin{aligned}\mathbb{P}(T_{\mathrm{FB}}(\widetilde{x}_\rho) > \varepsilon\mathbb{E}[T_{\mathrm{FB}}]) &\le \mathbb{P}((1-\rho)^2 T_{\mathrm{FB}}(\widetilde{x}_\rho) > \widetilde{\varepsilon}\overline{F}(G^{\leftarrow}(\rho))) \\ &\le \frac{\mathbb{E}[B^*]^2 v(\rho)^2 + \frac{2}{3}\mathbb{E}[B^*]v(\rho)(1-\rho)x_\rho^{v(\rho)} + 2\widetilde{\varepsilon}\mathbb{E}[B^*]\overline{F}(G^{\leftarrow}(\rho))}{\left(\widetilde{\varepsilon}v(\rho)^{-1}\overline{F}(G^{\leftarrow}(\rho)) - (1-\rho)x_\rho^{v(\rho)} - \rho_{x_\rho^{v(\rho)}}\mathbb{E}[B^*]v(\rho)\right)^2} \\ &= \frac{\mathbb{E}[B^*]^2\frac{v(\rho)^4}{\overline{F}(G^{\leftarrow}(\rho))^2} + \frac{2\mathbb{E}[B^*]}{3}\frac{v(\rho)^3(1-\rho)x_\rho^{v(\rho)}}{\overline{F}(G^{\leftarrow}(\rho))^2} + 2\widetilde{\varepsilon}\mathbb{E}[B^*]\frac{v(\rho)^2}{\overline{F}(G^{\leftarrow}(\rho))}}{\left(\widetilde{\varepsilon} - \frac{v(\rho)(1-\rho)x_\rho^{v(\rho)}}{\overline{F}(G^{\leftarrow}(\rho))} - \rho_{x_\rho^{v(\rho)}}\mathbb{E}[B^*]\frac{v(\rho)^2}{\overline{F}(G^{\leftarrow}(\rho))}\right)^2}.\end{aligned}$$

Subsequently, we observe for any $\nu \in (0,1)$ that

$$\lim_{\rho\uparrow 1}(1-\rho)x_\rho^\nu = \lim_{\rho\uparrow 1}(1-\rho)G^{\leftarrow}\left(1-\frac{1-\nu}{\nu}\frac{1-\rho}{\rho}\right) = \lim_{z\to x_R}\frac{\frac{\nu}{1-\nu}\overline{G}(z)\cdot z}{1+\frac{\nu}{1-\nu}\overline{G}(z)} \le \lim_{z\to x_R}\frac{\nu\cdot z\overline{G}(z)}{1-\nu}, \tag{4.45}$$

where $z\overline{G}(z)\to 0$ as $z\to x_R$ since $\mathbb{E}[B^2]<\infty$ (cf. Section 4.4.1). It therefore follows that $(1-\rho)x_\rho^{\nu(\rho)} = o(\nu(\rho))$ as $\rho\uparrow 1$, and consequently $\lim_{\rho\uparrow 1}\mathbb{P}(T_{\rm FB} > \varepsilon\mathbb{E}[T_{\rm FB}]) = 0$ provided that $\lim_{\rho\uparrow 1}\frac{\nu(\rho)^2}{\overline{F}(G^{\leftarrow}(\rho))} = 0$.

Write $x = (1-\rho)^{-1}$. By Lemma 4.4.2, it suffices to show $\alpha\left((\cdot)^{-2\widetilde{\gamma}}\overline{F}(G^{\leftarrow}(1-(\cdot)^{-1}))\right) < 0$. This relation follows from Lemma 4.4.1, Corollary 4.4.7 and our choice of $\widetilde{\gamma}$:

$$\alpha\left(\frac{(\cdot)^{-2\widetilde{\gamma}}}{\overline{F}(G^{\leftarrow}(1-(\cdot)^{-1}))}\right) \le -2\widetilde{\gamma} - \beta\left(\overline{F}(G^{\leftarrow}(1-(\cdot)^{-1}))\right) \le -2\widetilde{\gamma} + \widetilde{p}(F) < 0.$$

4.6.2 Infinite variance

This section regards all functions F that satisfy $x_R=\infty, \beta(\overline{F}) > -2$. In this case, $\widetilde{x}_\rho$ can be any function that satisfies both $\lim_{\rho\uparrow 1}\widetilde{x}_\rho = \infty$ and $\lim_{\rho\uparrow 1}\frac{\widetilde{x}_\rho}{\overline{G}(\widetilde{x}_\rho)\log\left(\frac{1}{1-\rho}\right)} = 0$.

Theorem 4.3.1 implies that there exists $C>0$ such that $\mathbb{E}[T_{\rm FB}] \ge C\log\left(\frac{1}{1-\rho}\right)$ for all ρ sufficiently close to one. Again, denote $\widetilde{\varepsilon} = \varepsilon C$. The analysis resumes with relation (4.43), where we substitute y by $\widetilde{\varepsilon}(1-\rho)^2\log\left(\frac{1}{1-\rho}\right)$ to obtain

$$\begin{aligned}\mathbb{P}(T_{\rm FB}(x) > \varepsilon\mathbb{E}[T_{\rm FB}]) &\le \mathbb{P}\left((1-\rho)^2 T_{\rm FB}(x) > \widetilde{\varepsilon}(1-\rho)^2\log\left(\frac{1}{1-\rho}\right)\right)\\ &\le \frac{\frac{1}{(1-\rho_x)^2}\mathbb{E}[(B\wedge x)^*]^2 + \frac{1}{1-\rho_x}\mathbb{E}[((B\wedge x)^*)^2] + 2\widetilde{\varepsilon}\mathbb{E}[(B\wedge x)^*]\log\left(\frac{1}{1-\rho}\right)}{\left(\widetilde{\varepsilon}(1-\rho_x)\log\left(\frac{1}{1-\rho}\right) - x - \frac{\rho_x}{1-\rho_x}\mathbb{E}[(B\wedge x)^*]\right)^2}.\end{aligned}$$

By relation (4.22), there exists a function $b(x)$ that is bounded for all x sufficiently large and satisfies $m_2(x) = \mathbb{E}[B]b(x)x\overline{G}(x)$. As such, $\mathbb{E}[((B\wedge x)^*)^2] = \frac{\mathbb{E}[(B\wedge x)^3]}{3\mathbb{E}[B]} \le \frac{xm_2(x)}{3\mathbb{E}[B]} = b(x)x^2\overline{G}(x)/3$ and similarly $\mathbb{E}[(B\wedge x)^*] = \frac{m_2(x)}{2\mathbb{E}[B]} = b(x)x\overline{G}(x)/2$. Substituting this into the above relation yields

$$\mathbb{P}(T_{\rm FB}(x) > \varepsilon\mathbb{E}[T_{\rm FB}]) \le \frac{\frac{b(x)^2}{4}\frac{x^2\overline{G}(x)^2}{(1-\rho_x)^2} + \frac{b(x)}{3}\frac{x^2\overline{G}(x)}{1-\rho_x} + \widetilde{\varepsilon}b(x)x\overline{G}(x)\log\left(\frac{1}{1-\rho}\right)}{\left(\widetilde{\varepsilon}(1-\rho_x)\log\left(\frac{1}{1-\rho}\right) - x - \frac{\rho_x b(x)}{2}\frac{x\overline{G}(x)}{1-\rho_x}\right)^2},$$

so that

$$\begin{aligned}&\mathbb{P}(T_{\rm FB}(x) > \varepsilon\mathbb{E}[T_{\rm FB}])\\ &\le \frac{\frac{b(x)^2}{4}\frac{\overline{G}(x)^2}{(1-\rho_x)^2}\frac{x^2}{(1-\rho_x)^2\log^2\left(\frac{1}{1-\rho}\right)} + \frac{b(x)}{3}\frac{\overline{G}(x)}{1-\rho_x}\frac{x^2}{(1-\rho_x)^2\log^2\left(\frac{1}{1-\rho}\right)} + \widetilde{\varepsilon}b(x)\frac{\overline{G}(x)}{1-\rho_x}\frac{x}{(1-\rho_x)\log\left(\frac{1}{1-\rho}\right)}}{\left(\widetilde{\varepsilon} - \frac{x}{(1-\rho_x)\log\left(\frac{1}{1-\rho}\right)} - \frac{\rho_x b(x)}{2}\frac{\overline{G}(x)}{1-\rho_x}\frac{x}{(1-\rho_x)\log\left(\frac{1}{1-\rho}\right)}\right)^2}.\end{aligned}$$

The result follows after noting that $1-\rho_x = 1-\rho G(x) \geq \overline{G}(x)$ and substituting $\widetilde{x}_\rho$ for x.

4.7 Asymptotic behaviour of the sojourn time tail

In this section, we prove Theorem 4.3.7 after presenting two facilitating propositions. The proofs of the propositions are postponed to Sections 4.7.1 and 4.7.2. Throughout this section, $\mathbf{e}(q)$ will denote an exponentially distributed random variable with rate $q>0$. We abuse notation by writing $\mathbf{e}(0) = +\infty$.

Reconsider the relation $T_{\mathrm{FB}}(x) \overset{d}{=} \mathscr{L}_x(W_x + x)$ to gain some intuition. A rough approximation of the duration of a busy period, given $W_x + x$ units of work at time $t=0$, is $(W_x+x)/(1-\rho_x)$. The scaled sojourn time $(1-\rho)^2 T_{\mathrm{FB}}(x)$ is then approximated by $\frac{1-\rho}{1-\rho_x}(1-\rho)(W_x+x)$. As in Section 4.4, define $x_\rho^\nu = G^{\leftarrow}\left(1-\frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right), \nu \in (1-\rho,1)$, so that $\frac{1-\rho}{1-\rho_x} = \nu$. Then for all $\nu \in (0,1)$, we have $(1-\rho)^2 T_{\mathrm{FB}}(x_\rho^\nu) \overset{d}{\approx} \nu(1-\rho)(W_{x_\rho^\nu} + x_\rho^\nu)$. We will show that $(1-\rho)x_\rho^\nu \to 0$ for all fixed $\nu \in (0,1)$. Instead, the following proposition shows that $(1-\rho)W_{x_\rho^\nu}$ behaves as an exponentially distributed random variable as $\rho \uparrow 1$:

Proposition 4.7.1. *Let $x_\rho^\nu = G^{\leftarrow}\left(1-\frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right), \nu \in (1-\rho,1)$, and let W_x^ρ denote the steady-state waiting time in an* M/GI/1/FIFO *queue with job sizes $B_i \wedge x$ and arrival rate $\rho_x/\mathbb{E}[B\wedge x]$. Then, for any fixed $\nu \in (0,1)$, $(1-\rho)W_{x_\rho^\nu} \xrightarrow{d} \mathrm{Exp}((\nu\mathbb{E}[B^*])^{-1})$ as $\rho \uparrow 1$.*

Kingman [83] proved that if $W^\rho = W_\infty^\rho$ denotes the steady-state waiting time in the non-truncated system, then $(1-\rho)W^\rho \xrightarrow{d} \mathrm{Exp}(\mathbb{E}[B^*]^{-1})$. Proposition 4.7.1 shows how jobs can be truncated such that the exponential behaviour is preserved, and quantifies how the truncation affects the parameter of the exponential distribution.

Substituting the result in Proposition 4.7.1 into our approximation above yields $(1-\rho)^2 T_{\mathrm{FB}}(x_\rho^\nu) \overset{d}{\approx} \mathrm{Exp}((\nu^2\mathbb{E}[B^*])^{-1})$ for every fixed $\nu \in (0,1)$. We will show that the fraction of jobs for which ν is in $(\varepsilon, 1-\varepsilon)$ scales as $\overline{F}(G^{\leftarrow}(\rho))$, and that the contribution of other jobs to the tail of $(1-\rho)^2 T_{\mathrm{FB}}$ is negligible. The result is presented in Proposition 4.7.2, where we focus on the probability $\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}} > \mathbf{e}(q))$ for its connection to the Laplace transform of T_{FB}^*.

Proposition 4.7.2. *Assume $F \in \mathrm{MDA}(H)$, where H is an extreme value distribution. Let $p(H) = \frac{\alpha}{\alpha-1}$ if $H = \Phi_\alpha$, $\alpha > 2$; $p(H) = 1$ if $H = \Lambda$ and $p(H) = \frac{\alpha}{\alpha+1}$ if $H = \Psi_\alpha$, $\alpha > 0$. Then*

$$\lim_{\rho\uparrow 1} \frac{\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}} > \mathbf{e}(q))}{\overline{F}(G^{\leftarrow}(\rho))} = \int_0^1 \frac{8\mathbb{E}[B^*]q\nu}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}\left(\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1\right)^2}\left(\frac{1-\nu}{\nu}\right)^{p(H)} \mathrm{d}\nu \tag{4.46}$$

for all $q \geq 0$. Here, the integral is finite for all $q \geq 0$.

We are now ready to prove Theorem 4.3.7. Using the relation $\mathbb{E}[e^{-qY}] = \mathbb{P}(\mathbf{e}(q) > Y)$, one sees that $\mathbb{P}((1-\rho)^2 T^{\rho}_{\mathrm{FB}} > \mathbf{e}(q)) = 1 - \mathbb{E}[e^{-q(1-\rho)^2 T^{\rho}_{\mathrm{FB}}}]$ and consequently

$$\begin{aligned}\frac{\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}} > \mathbf{e}(q))}{\overline{F}(G^{\leftarrow}(\rho))} &= \frac{(1-\rho)^2\mathbb{E}[T_{\mathrm{FB}}]}{\overline{F}(G^{\leftarrow}(\rho))} \cdot \frac{1 - \mathbb{E}\left[e^{-q(1-\rho)^2 T_{\mathrm{FB}}}\right]}{(1-\rho)^2\mathbb{E}[T_{\mathrm{FB}}]} \\ &= \frac{(1-\rho)^2\mathbb{E}[T_{\mathrm{FB}}]}{\overline{F}(G^{\leftarrow}(\rho))} \cdot q \cdot \mathbb{E}\left[e^{-q(1-\rho)^2 T^*_{\mathrm{FB}}}\right],\end{aligned}$$

where T^*_{FB} is the residual sojourn time and has density $\mathbb{P}(T_{\mathrm{FB}} > t)/\mathbb{E}[T_{\mathrm{FB}}]$. Consequently,

$$\begin{aligned}&\lim_{\rho\uparrow 1} \mathbb{E}\left[e^{-q(1-\rho)^2 T^*_{\mathrm{FB}}}\right] \\ &\quad= \lim_{\rho\uparrow 1} \frac{\overline{F}(G^{\leftarrow}(\rho))}{(1-\rho)^2\mathbb{E}[T_{\mathrm{FB}}]} \int_0^1 \frac{8\mathbb{E}[B^*]\nu}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}\left(\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1\right)^2} \left(\frac{1-\nu}{\nu}\right)^{p(H)} \mathrm{d}\nu \\ &\quad= r(H)^{-1} \int_0^1 \frac{8\nu}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}\left(\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1\right)^2} \left(\frac{1-\nu}{\nu}\right)^{p(H)} \mathrm{d}\nu \qquad (4.47)\end{aligned}$$

for all $q \geq 0$, where $r(H)$ was introduced in Theorem 4.3.2. It follows from Section 4.4.2 that $\lim_{q\downarrow 0}\lim_{\rho\uparrow 1}\mathbb{E}\left[e^{-q(1-\rho)^2 T^*_{\mathrm{FB}}}\right] = 1$. Additionally, the right-hand side is continuous in q, so that $(1-\rho)^2 T^*_{\mathrm{FB}}$ converges to some non-degenerate random variable by the Continuity Theorem [54, Section XIII.1, Theorem 2a].

The Laplace transform inversion formula (12) in Bateman [17, p.234] states that $f(t) = \frac{2\sqrt{t}}{\sqrt{\pi}} - 2te^t \operatorname{Erfc}(\sqrt{t})$ is the Laplace inverse of $s^{-1/2}(s^{1/2}+1)^{-2}$, i.e. $\int_0^\infty e^{-qt} f(t)\,\mathrm{d}t = \frac{1}{\sqrt{q}(\sqrt{q}+1)^2}$. Consequently, we have

$$\int_0^\infty e^{-qt} g(t,\nu)\,\mathrm{d}t = \frac{1}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}\left(\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1\right)^2} \qquad (4.48)$$

for $g(t,\nu) = \frac{e^{-\frac{t}{4\mathbb{E}[B^*]\nu^2}}}{4\mathbb{E}[B^*]\nu^2} f\left(\frac{t}{4\mathbb{E}[B^*]\nu^2}\right)$, and hence relation (4.47) may be rewritten as

$$\begin{aligned}\lim_{\rho\uparrow 1}\mathbb{E}\left[e^{-q(1-\rho)^2 T^*_{\mathrm{FB}}}\right] &= \int_0^\infty e^{-qt}\left[\int_0^1 8r(H)^{-1}\nu\left(\frac{1-\nu}{\nu}\right)^{p(H)} g(t,\nu)\,\mathrm{d}\nu\right]\mathrm{d}t \\ &=: \int_0^\infty e^{-qt} g^*(t)\,\mathrm{d}t.\end{aligned}$$

We conclude that the limiting random variable $\lim_{\rho\uparrow 1}(1-\rho)^2 T^*_{\mathrm{FB}}$ has density g^*. Furthermore, as

$$\begin{aligned}\lim_{\rho\uparrow 1}\mathbb{E}\left[e^{-q(1-\rho)^2 T^*_{\mathrm{FB}}}\right] &= \lim_{\rho\uparrow 1}\int_0^\infty e^{-q\tau}\frac{\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}} > \tau)}{(1-\rho)^2\mathbb{E}[T_{\mathrm{FB}}]}\,\mathrm{d}\tau \\ &= \lim_{\rho\uparrow 1}\int_0^\infty e^{-q\tau}\frac{\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}} > \tau)}{r(H)\mathbb{E}[B^*]\overline{F}(G^{\leftarrow}(\rho))}\,\mathrm{d}\tau,\end{aligned}$$

for all $q \geq 0$, we also see that $\lim_{\rho\uparrow 1} \frac{\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}} > y)}{r(H)\mathbb{E}[B^*]\overline{F}(G^{\leftarrow}(\rho))} = g^*(y)$ almost everywhere.

To see that g^* is monotone, it suffices to show that $f(t)$ is monotone. To this end, we exploit the continued fraction representation (13.2.20a) in Cuyt et al. [42] and find

$$\mathrm{Erfc}(x) = \frac{x}{\sqrt{\pi}} e^{-x^2} \cfrac{1}{x^2 + \cfrac{1/2}{1 + \cfrac{1}{x^2 + \cfrac{3/2}{1+\ldots}}}} \geq \frac{e^{-x^2}}{x\sqrt{\pi}} \left(1 - \frac{x^2+3/2}{2x^4+6x^2+3/2}\right). \tag{4.49}$$

As a consequence, one sees that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} f(t) &= \frac{1+2t}{\sqrt{\pi}\sqrt{t}} - 2(1+t)e^t \,\mathrm{Erfc}(\sqrt{t}) \\ &\leq \frac{1+2t-2(1+t)\left(1-\frac{t+3/2}{2t^2+6t+3/2}\right)}{\sqrt{\pi}\sqrt{t}} = \frac{-1+\frac{2t^2+5t+3}{2t^2+6t+3/2}}{\sqrt{\pi}\sqrt{t}}, \end{aligned}$$

which is negative for all $t \geq 0$. We conclude the section with the postponed proofs of Propositions 4.7.1 and 4.7.2.

4.7.1 Proof of Proposition 4.7.1

The Pollaczek-Khintchine formula states that $\mathbb{E}[e^{-s(1-\rho)W_x}] = \frac{1-\rho_x}{1-\rho_x \mathbb{E}[e^{-s(1-\rho)(B\wedge x)^*}]}$. In this representation, we expand the Laplace-Stieltjes transform $\mathbb{E}[e^{-s(1-\rho)(B\wedge x)^*}]$ around $\rho = 1$ to find

$$\mathbb{E}[e^{-s(1-\rho)W_x}] = \frac{1-\rho_x}{1-\rho_x\left(1-\mathbb{E}[(B\wedge x)^*](1-\rho)s + o(1-\rho)\right)}$$

and hence

$$\mathbb{E}[e^{-s(1-\rho)W_{x_\rho^\nu}}] = \frac{1}{1+\frac{1-\rho}{1-\rho_{x_\rho^\nu}}\rho_{x_\rho^\nu}\mathbb{E}[(B\wedge x_\rho^\nu)^*]s + o\left(\frac{1-\rho}{1-\rho_{x_\rho^\nu}}\right)} = \frac{1}{1+\nu\rho_{x_\rho^\nu}\mathbb{E}[(B\wedge x_\rho^\nu)^*]s + o(1)},$$

where $o(1)$ vanishes as $\rho \uparrow 1$. By definition of x_ρ^ν, $x_\rho^\nu \to \infty$ and $\rho_{x_\rho^\nu} \uparrow 1$ as $\rho \uparrow 1$ for any fixed $\nu \in (0,1)$. In particular, $\lim_{\rho\uparrow 1}\mathbb{E}[e^{-s(1-\rho)W_{x_\rho^\nu}}] = \frac{1}{1+\nu\mathbb{E}[B^*]s}$. The proof is completed by applying the Continuity Theorem [54, Section XIII.1, Theorem 2a].

4.7.2 Proof of Proposition 4.7.2

We require functions $\nu_l(\rho) \downarrow 0$ and $\nu_u(\rho) \uparrow 1$ that distinguish the jobs that significantly contribute to the tail of $(1-\rho)^2 T_{\mathrm{FB}}$, and those that do not. For the former function, fix $\gamma \in (p(H)/2, 1)$ and let $\nu_l(\rho) = (1-\rho)^\gamma$ as in Section 4.6.1. This is possible as $p(H) < 2$ for

all H to which the theorem applies. For the latter function, we refer to relation (4.45) to verify that there exists a function $v(\rho) \uparrow 1$ such that $(1-\rho)x_\rho^{v(\rho)} \to 0$. Let $v_u(\rho)$ be a function with this property, and write

$$\begin{aligned}
&\frac{\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}} > \mathbf{e}(q))}{\overline{F}(G^{\leftarrow}(\rho))} \\
&\qquad = \int_{v=0}^{v_l(\rho)} \mathbb{P}((1-\rho)^2 T_{\mathrm{FB}}(x_\rho^v) > \mathbf{e}(q)) \frac{\mathrm{d}F(x_\rho^v)}{\overline{F}(G^{\leftarrow}(\rho))} \\
&\qquad\quad + \int_{v=v_l(\rho)}^{v_u(\rho)} \mathbb{P}((1-\rho)^2 T_{\mathrm{FB}}(x_\rho^v) > \mathbf{e}(q)) \frac{\mathrm{d}F(x_\rho^v)}{\overline{F}(G^{\leftarrow}(\rho))} \\
&\qquad\quad + \int_{v=v_u(\rho)}^{1} \mathbb{P}((1-\rho)^2 T_{\mathrm{FB}}(x_\rho^v) > \mathbf{e}(q)) \frac{\mathrm{d}F(x_\rho^v)}{\overline{F}(G^{\leftarrow}(\rho))} \\
&\qquad =: \widehat{\mathrm{I}}(\rho) + \widehat{\mathrm{II}}(\rho) + \widehat{\mathrm{III}}(\rho). \qquad (4.50)
\end{aligned}$$

The next paragraphs study the behaviour of $\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}}(x) > \mathbf{e}(q))$, which will then facilitate the analysis of the above three regions. Specifically, we will derive the asymptotic behaviour of $\widehat{\mathrm{II}}(\rho)$ in terms of q, and show that $\widehat{\mathrm{I}}(\rho) + \widehat{\mathrm{III}}(\rho) = o(1)$ for any $q \geq 0$.

Define $X_x^\rho(t) := \frac{t}{1-\rho} - \sum_{i=1}^{N((1-\rho)^{-2}t)} (1-\rho)(B_i \wedge x)$. Then $X_x^\rho(t)$ is a spectrally negative Lévy process and we obtain

$$\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}}(x) > \mathbf{e}(q)) = \mathbb{P}\left(\sup_{t \in [0, \mathbf{e}(q)]} X_x^\rho(t) \leq (1-\rho) W_x + (1-\rho)x \right) \qquad (4.51)$$

from relation (4.42). The Laplace exponent of $X_x^\rho(t)$ is given by $\psi(s) := t^{-1} \log \mathbb{E}[e^{sX_x^\rho(t)}]$, and has right-inverse $\varphi(x, \rho, q) := \sup\{s \geq 0 : \psi(x, \rho, s) = q\}$. With these notions, relation (8.4) in Kyprianou [90] states that

$$\mathbb{P}((1-\rho)^2 T_{\mathrm{FB}}^\rho(x) > \mathbf{e}(q)) = \mathbb{P}\left(\mathbf{e}(\varphi(x, \rho, q)) \leq (1-\rho) W_x + (1-\rho)x \right). \qquad (4.52)$$

Since

$$\begin{aligned}
\psi(x, \rho, s) &= t^{-1} \log \mathbb{E}\left[e^{\frac{st}{1-\rho} - \sum_{i=1}^{N((1-\rho)^{-2}t)} (1-\rho) s (B_i \wedge x)} \right] \\
&= \frac{s}{1-\rho} + t^{-1} \log \mathbb{E}\left[e^{-\sum_{i=1}^{N((1-\rho)^{-2}t)} (1-\rho) s (B_i \wedge x)} \right]
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[e^{-\sum_{i=1}^{N((1-\rho)^{-2}t)} (1-\rho) s (B_i \wedge x)}] &= \sum_{n=0}^{\infty} \mathbb{E}[e^{-(1-\rho)s(B \wedge x)}]^n \frac{\left(\frac{\lambda t}{(1-\rho)^2}\right)^n}{n!} e^{-\frac{\lambda t}{(1-\rho)^2}} \\
&= e^{-\frac{\lambda t}{(1-\rho)^2}\left(1 - \mathbb{E}[e^{-(1-\rho)s(B \wedge x)}]\right)},
\end{aligned}$$

we obtain $\psi(x,\rho,s) = \frac{s}{1-\rho} - \frac{\lambda}{(1-\rho)^2}\left(1-\mathbb{E}[e^{-(1-\rho)s(B\wedge x)}]\right)$. A Taylor expansion around $\rho = 1$ now yields

$$\begin{aligned}
&\psi(x,\rho,s)\\
&\quad= \frac{s}{1-\rho} - \frac{\lambda}{(1-\rho)^2}\left(1-\left(1-(1-\rho)s\mathbb{E}[B\wedge x] + \frac{(1-\rho)^2 s^2}{2}\mathbb{E}[(B\wedge x)^2] + o((1-\rho)^2 s^2)\right)\right)\\
&\quad= \frac{s}{1-\rho} - \left(\frac{\rho_x s}{1-\rho} - \frac{\lambda\mathbb{E}[(B\wedge x)^2]}{2}s^2 + o(s^2)\right) = \frac{1-\rho_x}{1-\rho}s + \frac{\rho\mathbb{E}[(B\wedge x)^2]}{2\mathbb{E}[B]}s^2 + o(s^2),
\end{aligned}$$

so that $\lim_{\rho\uparrow 1}\psi(x_\rho^\nu,\rho,s) = \nu^{-1}s + \mathbb{E}[B^*]s^2$ for all $\nu > 0$, and consequently

$$\lim_{\rho\uparrow 1}\varphi(x_\rho^\nu,\rho,q) = \frac{\sqrt{\nu^{-2}+4\mathbb{E}[B^*]q} - \nu^{-1}}{2\mathbb{E}[B^*]} = \frac{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}-1}{2\mathbb{E}[B^*]\nu} =: \varphi(\nu,q). \tag{4.53}$$

Similarly, one deduces that $\lim_{\rho\uparrow 1}\nu_l(\rho)\psi(x_\rho^{\nu_l(\rho)},\rho,s) = s$ and

$$\lim_{\rho\uparrow 1}\nu_l(\rho)^{-1}\varphi(x_\rho^{\nu_l(\rho)},\rho,q) = q. \tag{4.54}$$

We now gathered sufficient tools to analyse the asymptotic behaviour of $\widehat{\mathrm{II}}(\rho)$.

Fix $\varepsilon \in (0,1/3)$. We have already proven the relations $(1-\rho)W^\rho_{x_\rho^\nu} \to \mathbf{e}((\nu\mathbb{E}[B^*])^{-1})$ and $(1-\rho)x_\rho^\nu \to 0$ as $\rho\uparrow 1$ for all $\nu\in(0,1)$. Since $\mathbf{e}(q_1) \le_{st} \mathbf{e}(q_2)$ whenever $q_1 \ge q_2$, relations (4.52) and (4.53) imply

$$\begin{aligned}
\mathbb{P}((1-\rho)^2 T^\rho_{\mathrm{FB}}(x_\rho^\nu) > \mathbf{e}(q)) &\le \mathbb{P}\left(\mathbf{e}((1+\varepsilon)\varphi(\nu,q)) \le \mathbf{e}((1-\varepsilon)(\nu\mathbb{E}[B^*])^{-1}) + \varepsilon\right)\\
&= e^{-(1+\varepsilon)\varepsilon\varphi(\nu,q)}\frac{(1+\varepsilon)\varphi(\nu,q)}{(1+\varepsilon)\varphi(\nu,q)+(1-\varepsilon)(\nu\mathbb{E}[B^*])^{-1}} + 1 - e^{-\varepsilon(1+\varepsilon)\varphi(\nu,q)}\\
&\le \frac{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}-1}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1-\frac{4\varepsilon}{1+\varepsilon}} + 1 - e^{-\varepsilon\cdot\frac{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}-1}{\mathbb{E}[B^*]\nu}}
\end{aligned}$$

for all $\rho \ge \rho_\varepsilon$, where $\rho_\varepsilon \in (0,1)$ is fixed sufficiently close to one. Consequently, for all $\rho \ge \rho_\varepsilon$,

$$\begin{aligned}
\widehat{\mathrm{II}}(\rho) &\le \int_{\nu_l(\rho)}^{\nu_u(\rho)} \frac{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}-1}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1-\frac{4\varepsilon}{1+\varepsilon}}\frac{\mathrm{d}F\left(G^{\leftarrow}\left(1-\frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)}{\overline{F}(G^{\leftarrow}(\rho))}\\
&\quad+\int_{\nu_l(\rho)}^{\nu_u(\rho)}\left(1-e^{-\varepsilon\cdot\frac{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}-1}{\mathbb{E}[B^*]\nu}}\right)\frac{\mathrm{d}F\left(G^{\leftarrow}\left(1-\frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)}{\overline{F}(G^{\leftarrow}(\rho))}\\
&\le -\left[\frac{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}-1}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1-\frac{4\varepsilon}{1+\varepsilon}}\frac{\overline{F}\left(G^{\leftarrow}\left(1-\frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)}{\overline{F}(G^{\leftarrow}(\rho))}\right]_{\nu=\nu_l(\rho)}^{\nu_u(\rho)}\\
&\quad+\int_{\nu_l(\rho)}^{\nu_u(\rho)}\frac{8\mathbb{E}[B^*]q\nu}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}\left(\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1-\frac{4\varepsilon}{1+\varepsilon}\right)^2}\frac{\overline{F}\left(G^{\leftarrow}\left(1-\frac{1-\rho}{\rho}\frac{1-\nu}{\nu}\right)\right)}{\overline{F}(G^{\leftarrow}(\rho))}\,\mathrm{d}\nu
\end{aligned}$$

$$- \left[\left(1 - e^{-\varepsilon \cdot \frac{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}-1}{\mathbb{E}[B^*]\nu}} \right) \frac{\overline{F}\left(G^{\leftarrow}\left(1 - \frac{1-\rho}{\rho}\frac{1-\nu}{\nu} \right) \right)}{\overline{F}(G^{\leftarrow}(\rho))} \right]_{\nu=\nu_l(\rho)}^{\nu_u(\rho)}$$
$$+ 4q \int_{\nu_l(\rho)}^{\nu_u(\rho)} \varepsilon \cdot e^{-\varepsilon \cdot \frac{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}-1}{\mathbb{E}[B^*]\nu}} \frac{\overline{F}\left(G^{\leftarrow}\left(1 - \frac{1-\rho}{\rho}\frac{1-\nu}{\nu} \right) \right)}{\overline{F}(G^{\leftarrow}(\rho))} \, \mathrm{d}\nu.$$

In Section 4.4.2, we deduced that $\overline{F}(G^{\leftarrow}(1-(\cdot)^{-1}))$ is regularly varying with index $-p(H)$. The Uniform Convergence Theorem hence implies

$$\begin{aligned}
\limsup_{\rho\uparrow 1} \widehat{\mathrm{II}}(\rho) \leq & - \left[\frac{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}-1}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1-\frac{4\varepsilon}{1+\varepsilon}} \left(\frac{1-\nu}{\nu} \right)^{p(H)} \right]_{\nu=0}^{1} \\
& + \int_0^1 \frac{8\mathbb{E}[B^*]q\nu}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}\left(\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1-\frac{4\varepsilon}{1+\varepsilon} \right)^2} \left(\frac{1-\nu}{\nu} \right)^{p(H)} \mathrm{d}\nu \\
& - \left[\left(1 - e^{-\varepsilon \cdot \frac{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}-1}{\mathbb{E}[B^*]\nu}} \right) \left(\frac{1-\nu}{\nu} \right)^{p(H)} \right]_{\nu=0}^{1} \\
& + 4q \int_0^1 \varepsilon \cdot e^{-\varepsilon \cdot \frac{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}-1}{\mathbb{E}[B^*]\nu}} \left(\frac{1-\nu}{\nu} \right)^{p(H)} \mathrm{d}\nu \\
= & \int_0^1 \frac{8\mathbb{E}[B^*]q\nu}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}\left(\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1-\frac{4\varepsilon}{1+\varepsilon} \right)^2} \left(\frac{1-\nu}{\nu} \right)^{p(H)} \mathrm{d}\nu \\
& + 4q \int_0^1 \varepsilon \cdot e^{-\varepsilon \cdot \frac{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}-1}{\mathbb{E}[B^*]\nu}} \left(\frac{1-\nu}{\nu} \right)^{p(H)} \mathrm{d}\nu.
\end{aligned}$$

Both these integrals are bounded for all $\varepsilon \in (0, 1/3)$ and all $q \geq 0$. Additionally, both integrands are increasing in ε for all ε sufficiently small. One may thus take the limit $\varepsilon \downarrow 0$ and apply the Dominated Convergence Theorem to find

$$\limsup_{\rho\uparrow 1} \widehat{\mathrm{II}}(\rho) \leq \int_0^1 \frac{8\mathbb{E}[B^*]q\nu}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}\left(\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1 \right)^2} \left(\frac{1-\nu}{\nu} \right)^{p(H)} \mathrm{d}\nu. \tag{4.55}$$

Similarly, one may show that

$$\liminf_{\rho\uparrow 1} \widehat{\mathrm{II}}(\rho) \geq \int_0^1 \frac{8\mathbb{E}[B^*]q\nu}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}\left(\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1+\frac{4\varepsilon}{1-\varepsilon} \right)^2} \left(\frac{1-\nu}{\nu} \right)^{p(H)} \mathrm{d}\nu,$$

and we conclude

$$\lim_{\rho\uparrow 1} \widehat{\mathrm{II}}(\rho) = \int_0^1 \frac{8\mathbb{E}[B^*]q\nu}{\sqrt{1+4\mathbb{E}[B^*]q\nu^2}\left(\sqrt{1+4\mathbb{E}[B^*]q\nu^2}+1 \right)^2} \left(\frac{1-\nu}{\nu} \right)^{p(H)} \mathrm{d}\nu. \tag{4.56}$$

Second, consider $\widehat{\mathrm{I}}(\rho)$. Define $M(\rho) := (1-\rho)^{-\widehat{\gamma}}$ for some $\widehat{\gamma} \in \left(p(H)/2, \gamma \right)$ and recall that $(1-\rho)x \to 0$ and $(1-\rho)W_x^\rho \xrightarrow{d} 0$ (and hence in probability) for all $x \leq x_\rho^{\nu_l(\rho)}$. Thus,

for all $x \le x_\rho^{\nu_l}$ and all ρ sufficiently large, we have

$$\begin{aligned}\widehat{\mathrm{I}}(\rho) &= \int_{v=0}^{\nu_l(\rho)} \mathbb{P}\left(\mathbf{e}(\varphi(x_\rho^v,\rho,q)) \le (1-\rho)W_{x_\rho^v} + (1-\rho)x_\rho^v\right) \frac{\mathrm{d}F(x_\rho^v)}{\overline{F}(G^{\leftarrow}(\rho))} \\ &\le \frac{\mathbb{P}\left(\mathbf{e}(\varphi(x_\rho^{\nu_l(\rho)},\rho,q)) \le 2M(\rho)\right)}{\overline{F}(G^{\leftarrow}(\rho))} + \frac{\mathbb{P}((1-\rho)W_x^\rho \ge M(\rho))}{\overline{F}(G^{\leftarrow}(\rho))} =: \widehat{\mathrm{Ia}}(\rho) + \widehat{\mathrm{Ib}}(\rho).\end{aligned}$$

Fix $\delta \in \left(0, p(H)-\gamma-\widehat{\gamma}\right)$. Potter's Theorem [24, Theorem 1.5.6] states that $\overline{F}(G^{\leftarrow}(\rho)) \ge C(1-\rho)^{p(H)+\delta}$ for some constant $C>0$ and all ρ sufficiently close to one. Also, one may readily deduce from relation (4.54) that $\mathbf{e}(\varphi(x_\rho^{\nu_l(\rho)},\rho,q)) \ge_{st} \mathbf{e}(2q\nu_l(\rho))$ for all $x \le x_\rho^{\nu_l(\rho)}$ and ρ sufficiently large. Consequently,

$$\begin{aligned}\limsup_{\rho\uparrow 1} \widehat{\mathrm{Ia}}(\rho) &\le \limsup_{\rho\uparrow 1} \frac{1-e^{-4q\nu_l(\rho)M(\rho)}}{\overline{F}(G^{\leftarrow}(\rho))} \le \lim_{\rho\uparrow 1} \frac{1-e^{-4q(1-\rho)^{\gamma-\widehat{\gamma}}}}{C(1-\rho)^{p(H)+\delta}} \\ &= \lim_{\rho\uparrow 1} \frac{4q(\gamma-\widehat{\gamma})(1-\rho)^{\gamma-\widehat{\gamma}-1}e^{-4q(1-\rho)^{\gamma-\widehat{\gamma}}}}{C\left(p(H)+\delta\right)(1-\rho)^{p(H)-1+\delta}} \\ &= \lim_{\rho\uparrow 1} \frac{4q(\gamma-\widehat{\gamma})}{C\left(p(H)+\delta\right)} \cdot \exp\left[-4q(1-\rho)^{\gamma-\widehat{\gamma}} + \left(\gamma-\widehat{\gamma}-p(H)-\delta\right)\log(1-\rho)\right] = 0.\end{aligned}$$

For term $\widehat{\mathrm{Ib}}(\rho)$, we apply Markov's inequality and Potter's Theorem to obtain

$$\begin{aligned}\limsup_{\rho\uparrow 1} \widehat{\mathrm{Ib}}(\rho) &\le \limsup_{\rho\uparrow 1} \frac{\frac{1-\rho}{1-\rho_x}\rho_x \mathbb{E}[(B\wedge x)^*]}{M(\rho)\overline{F}(G^{\leftarrow}(\rho))} \le \lim_{\rho\uparrow 1} C_1 \frac{\mathbb{E}[B^*]\nu_l(\rho)}{M(\rho)(1-\rho)^{p(H)+\delta}} \\ &= \lim_{\rho\uparrow 1} C_1 \mathbb{E}[B^*](1-\rho)^{\gamma+\widehat{\gamma}-p(H)-\delta} = 0.\end{aligned}$$

Finally, consider term $\widehat{\mathrm{III}}(\rho)$. For this term, the claim follows readily from the Uniform Convergence Theorem and the property $\nu_u(\rho) \uparrow 1$:

$$\begin{aligned}\limsup_{\rho\uparrow 1} \widehat{\mathrm{III}}(\rho) &\le \limsup_{\rho\uparrow 1} \frac{\overline{F}(x_\rho^{\nu_u})}{\overline{F}(G^{\leftarrow}(\rho))} = \limsup_{\rho\uparrow 1} \frac{\overline{F}\left(G^{\leftarrow}\left(1-\frac{1-\rho}{\rho}\frac{1-\nu_u(\rho)}{\nu_u(\rho)}\right)\right)}{\overline{F}(G^{\leftarrow}(\rho))} \\ &= \limsup_{\rho\uparrow 1} \left(\frac{1-\nu_u(\rho)}{\rho\nu_u(\rho)}\right)^{p(H)} = 0.\end{aligned}$$

This concludes the proof of Proposition 4.7.2. The chapter is concluded with some additional Matuszewska theory and the postponed proofs of the lemmas in Section 4.4.1.

4.A Additional Matuszewska theory

This appendix gathers some results on Matuszewska indices. Lemmas 4.4.1 and 4.4.2 are proven directly from Definition 4.2.1. Then, a generalised version of Potter's Theorem allows us to prove Lemmas 4.4.5 and 4.4.6.

Proof of Lemma 4.4.1. Let $\alpha_1 > \alpha(f_1)$ and $\alpha_2 > \alpha(f_2)$. Then, by definition of the upper Matuszewska index, there exist $C_1, C_2 > 0$ such that for all $\mu \in [1, \mu^*], \mu^* > 1$, and all x sufficiently large we have $f_1(\mu x) \le C_1 \mu^{\alpha_1} f_1(x)$ and $f_2(\mu x) \le C_2 \mu^{\alpha_2} f_2(x)$. Consequently, we have $\limsup_{x\to\infty} \frac{f_1(\mu x) f_2(\mu x)}{f_1(x) f_2(x)} \le C_1 C_2 \mu^{\alpha_1+\alpha_2}$ and thus $\alpha(f_1 \cdot f_2) \le \alpha(f_1) + \alpha(f_2)$.

Similarly, if f_1 is non-decreasing, we have

$$f_1(f_2(\mu x)) \le f_1(C_2 \mu^{\alpha_2} f_2(x)) \le C_1 C_2^{\alpha_1} \mu^{\alpha_1 \alpha_2} f_1(f_2(x))$$

and thus $\alpha(f_1 \circ f_2) \le \alpha(f_2) \cdot \alpha(f_2)$. The results on the lower Matuszewska indices are proven analogously. □

Proof of Lemma 4.4.2. As f is positive, it suffices to show that $\limsup_{x\to\infty} f(x) = 0$. For sake of contradiction, assume that this is false. Then there exists a constant $m > 0$ and a sequence $(x_n)_{n\in\mathbb{N}}, x_n \to \infty$, such that $f(x_n) \ge m$ for all $n \in \mathbb{N}$. Now, by definition of the upper Matuszewska index, there exists $C > 0$ such that for all $\mu \in [1, \mu^*], \mu^* > 1$, we have $f(x) \ge C\mu^{-\alpha(f)/2} f(\mu x)$ for all x sufficiently large. As a consequence, for some $N \in \mathbb{N}$ we have $f(x_N) \ge C(x_n/x_N)^{-\alpha(f)/2} f(x_n) \ge Cm(x_n/x_N)^{-\alpha(f)/2}$ for any fixed $n \ge N$. This is a contradiction for any x_n that satisfies $x_n > x_N(Cm/f(x_N))^{2/\alpha(f)}$. □

The following result is a generalised version of Potter's theorem and gives bounds on the ratio $f(y)/f(x)$:

Theorem 4.A.1 (Bingham et al. [24], Proposition 2.2.1)**.** *Let f be positive.*

(i) If $\alpha(f) < \infty$, then for every $\alpha > \alpha(f)$ there exist positive constants C and X such that $f(y)/f(x) \le C(y/x)^\alpha$ for all $y \ge x \ge X$.

(ii) If $\beta(f) > -\infty$, then for every $\beta < \beta(f)$ there exist positive constants D and X such that $f(y)/f(x) \ge D(y/x)^\beta$ for all $y \ge x \ge X'$.

Theorem 4.A.1 allows us to derive a relation between the Matuszewska indices of f to those of $f^{\leftarrow}$, which is presented as Lemma 4.A.2:

Lemma 4.A.2. *Let f be positive and locally integrable on $[X, \infty)$. If f is strictly increasing, unbounded above and $\alpha(f) < \infty$, then $\beta(f^{\leftarrow}) = 1/\alpha(f)$. If $\beta(f) > 0$, then $\alpha(f^{\leftarrow}) = 1/\beta(f)$.*

Proof. By definition of the upper Matuszewska index, for all $\alpha > \alpha(f)$ there exists a constant $C > 0$ such that for each $\mu^* > 1$, $f(\mu x)/f(x) \le C\mu^\alpha$ uniformly in $\mu \in [1, \mu^*]$ as $x \to \infty$. In particular, for all x sufficiently large we have $f((\mu/C)^{1/\alpha} x) \le \mu f(x)$. As f is strictly increasing and unbounded above, one can hence see that

$$\lim_{x\to\infty} \frac{f^{\leftarrow}(\mu x)}{f^{\leftarrow}(x)} = \lim_{y\to\infty} \frac{f^{\leftarrow}(\mu f(y))}{f^{\leftarrow}(f(y))} \ge \lim_{y\to\infty} \frac{f^{\leftarrow}(f((\mu/C)^{1/\alpha} y))}{y} \ge (C)^{-1/\alpha} \mu^{1/\alpha} \tag{4.57}$$

uniformly for $\mu \in [1, \mu^*]$. As a consequence, $\beta(f^{\leftarrow}) \ge 1/\alpha(f)$.

On the other hand, if $\beta(f^{\leftarrow}) > 1/\alpha(f), \alpha(f) > 0$, then Theorem 4.A.1(ii) claims that for some $\varepsilon > 0$ sufficiently small there exists a constant $C' > 0$ such that $f^{\leftarrow}(y)/f^{\leftarrow}(z) \geq C'(y/z)^{1/\alpha(f)+\varepsilon}$ for all $y \geq z$ sufficiently large. By substitution of $y = f(\mu x)$ and $z = f(x)$, we obtain

$$C'\left(\frac{f(\mu x)}{f(x)}\right)^{1/\alpha(f)+\varepsilon} \leq \frac{f^{\leftarrow}(f(\mu x))}{f^{\leftarrow}(f(x))} = \mu$$

and hence $\lim_{x\to\infty} f(\mu x)/f(x) \leq ((C')^{-1}\mu)^{\frac{\alpha(f)}{1+\varepsilon\alpha(f)}}$. This inequality, however, indicates that $\alpha(f)$ was not the infimum over all α satisfying (4.4), which is a contradiction.

The relation $\alpha(f^{\leftarrow}) = 1/\beta(f)$ is proven similarly. □

A more general version of this lemma has been stated in several other works [24, 95]; however, these works refer to an unpublished manuscript by De Haan and Resnick for the corresponding proof.

Our final results relate the Matuszewska indices of $\overline{F}$ to those of related functions. First, Lemma 4.4.5 relates the Matuszewska indices of $\overline{F}$ to those of $\overline{G}$. Its proof is similar to the proof of Lemma 6 in Lin et al. [95].

Proof of Lemma 4.4.5. Assume $x_R = \infty$. Then by definition of $\alpha(\overline{F})$, we have for all $\alpha > \alpha(\overline{F})$ that $\overline{F}(\mu t)/\overline{F}(t) \leq C(1+o(1))\mu^{\alpha}$ uniformly in $\mu \in [1, \mu^*]$ and hence

$$\begin{aligned}\mathbb{E}[B]\overline{G}(\mu x) &= \mu\int_x^{\infty} \overline{F}(\mu\tau)\,\mathrm{d}\tau \leq C(1+o(1))\mu^{\alpha+1}\int_x^{\infty}\overline{F}(\tau)\,\mathrm{d}\tau \\ &= C(1+o(1))\mu^{\alpha+1}\mathbb{E}[B]\overline{G}(x)\end{aligned}$$

as $x \to \infty$. On the other hand, if $x_R < \infty$ then

$$\begin{aligned}\mathbb{E}[B]\overline{G}(x_R - (\mu x)^{-1}) &= \int_{x_R-(\mu x)^{-1}}^{x_R}\overline{F}(t)\,\mathrm{d}t = \int_x^{\infty}\mu^{-1}\tau^{-2}\overline{F}(x_R-(\mu\tau)^{-1})\,\mathrm{d}\tau \\ &\leq C(1+o(1))\mu^{\alpha-1}\int_x^{\infty}\tau^{-2}\overline{F}(x_R-\tau^{-1})\,\mathrm{d}\tau \\ &= C(1+o(1))\mu^{\alpha-1}\mathbb{E}[B]\overline{G}(x_R-x^{-1})\end{aligned}$$

as $x \to \infty$. The claims on the lower Matuszewska index can be proven analogously. □

Second, Lemma 4.4.6 relates the Matuszewska indices of $\overline{F}$ to those of $G^{\leftarrow}$. It does so by combining Lemmas 4.4.1, 4.4.5 and 4.A.2.

Proof of Lemma 4.4.6. We only prove the relation between the lower Matuszewska indices, as the relation between the upper Matuszewska indices can be proven similarly.

First, assume $x_R = \infty$. Since $\beta(\overline{F}) > -\infty$, it follows from Lemma 4.4.5 that $\beta(\overline{G}) > -\infty$ and hence, by Lemma 4.4.1, that $\alpha(1/\overline{G}) = -\alpha(\overline{G}) \leq -\beta(\overline{G}) < \infty$. The result follows

readily from Lemma 4.A.2 through $\beta(G^{\leftarrow}(1-(\cdot)^{-1})) = \beta((1/\overline{G})^{\leftarrow}) = 1/\alpha(1/\overline{G}) = -1/\beta(\overline{G})$ and subsequent application of Lemma 4.4.5.

Similarly, if $x_R < \infty$ then $\alpha(1/\overline{G}(x_R-(\cdot)^{-1})) < \infty$ and

$$\begin{aligned}\frac{1}{x_R - G^{\leftarrow}(1-x^{-1})} &= \frac{1}{x_R - \inf\{z : G(z) > 1-x^{-1}\}} = \inf\left\{\frac{1}{x_R - z} : G(z) > 1-x^{-1}\right\} \\ &= \inf\{y : G(x_R - y^{-1}) > 1-x^{-1}\} = \inf\{y : 1/\overline{G}(x_R - y^{-1}) > x\} \\ &= \left(\frac{1}{\overline{G}\left(x_R - \frac{1}{\cdot}\right)}\right)^{\leftarrow}(x).\end{aligned}$$

The result then follows from $\beta\left(\frac{1}{x_R - G^{\leftarrow}(1-(\cdot)^{-1})}\right) = \beta\left(\left(\frac{1}{\overline{G}(x_R-(\cdot)^{-1})}(\cdot)\right)^{\leftarrow}\right) = 1/\alpha\left(\frac{1}{\overline{G}(x_R-(\cdot)^{-1})}\right) = -1/\beta(\overline{G}(x_R-(\cdot)^{-1}))$ and application of Lemma 4.4.5. □

CHAPTER 5

UNIFORM ASYMPTOTICS FOR COMPOUND POISSON PROCESSES WITH REGULARLY-VARYING JUMPS AND VANISHING DRIFT

This chapter addresses heavy-tailed large-deviation estimates for the tail distribution of functionals of a class of spectrally one-sided Lévy processes. Our contribution is to show that these estimates remain valid in a near-critical regime. This complements recent similar results that have been obtained for the all-time supremum of such processes. Specifically, we consider local asymptotics of the all-time supremum, the supremum of the process until exiting $[0,\infty)$, the maximum jump until that time, and the time it takes until exiting $[0,\infty)$. The proofs rely, among other things, on properties of scale functions.

The terminology in this chapter transcends the more specified terminology of the earlier chapters, but its contributions translate to our understanding of M/GI/1 models in three ways. First, the described all-time supremum is equivalent to the steady-state waiting time in a M/GI/1/FIFO model. Second, the time until exiting $[0,\infty)$ is identical to the steady-state sojourn time in a M/GI/1/LIFO model. Finally, we provide a local analogue of Kingman's heavy-traffic approximation.

Based on Kamphorst and Zwart [S3].

5.1 Introduction

The analysis of spectrally one-sided Lévy processes is a topic of fundamental interest in the stochastic processes literature [90] and arises in many applications, such as queueing [43] and insurance risk theory [9, 11, 51]. More generally, Lévy processes and various functionals have been studied extensively over the last decades through fluctuation theory, leading to many interesting and useful results. If the underlying Lévy measure is heavy-tailed, then exact expressions are harder to obtain and one often resorts to asymptotic estimates based on heavy-tailed large deviations. The goal of this chapter is to assess the robustness of several of these approximations in a regime where the underlying Lévy process has a small drift.

To make this more specific, consider the compound Poisson process with deterministic drift

$$X^\rho(t) := X_0 + \sum_{i=1}^{N^\rho(t)} B_i - t, \tag{5.1}$$

where $N^\rho(t)$, $t \geq 0$, is a Poisson process with a rate that depends on a drift parameter ρ. With a slight abuse of terminology, we call X^ρ a compound Poisson process throughout this chapter, and investigate the asymptotic behaviour of various functionals of X^ρ under the assumption that the i.i.d. nonnegative jump sizes B_i have a regularly-varying tail with index $\alpha > 2$. The initial condition X_0 is equal in distribution to B_i and independent of ρ; we present a more detailed model description in Section 5.2. The long-term drift $\mathbb{E}[X^\rho(1) - X_0]$ of the process is negative, and of order $1-\rho$. In the central limit regime, we let $\rho \uparrow 1$ so that the long-term drift tends to zero.

A functional that has received ample attention in the literature is the all-time supremum $M_\infty^\rho := \sup_{t\geq 0} X^\rho(t)$. For fixed ρ, as $x \to \infty$, the following estimate holds [51, 58, 87, 100]:

$$\mathbb{P}(M_\infty^\rho > x) \sim \frac{\rho}{\mathbb{E}[B_1](1-\rho)} \int_x^\infty \mathbb{P}(B_1 > t)\,\mathrm{d}t. \tag{5.2}$$

This approximation can be very inaccurate when ρ is not fixed. Specifically, if $\rho \uparrow 1$ and $x = y/(1-\rho)$ for fixed y, then $\mathbb{P}(M_\infty^\rho > x)$ will converge to $\exp[-2(\mathbb{E}[B_1]/\mathbb{E}[B_1^2])\,y]$ (this is a heavy-traffic limit, cf. [135, 136]). The contributions of the present chapter all relate to the validity of heavy-tail approximations like (5.2) when $\rho \uparrow 1$.

Motivated by the contrast between the two regimes, Olvera-Cravioto et al. [107] derive an explicit threshold,

$$\widetilde{x}_\rho := \mu(\alpha-2)\frac{1}{1-\rho}\log\frac{1}{1-\rho}, \tag{5.3}$$

for some $\mu > 0$ as specified in the next section, where the two regimes connect. In particular, they show that estimate (5.2) remains valid when $\rho \uparrow 1$ and $x \geq (1+\varepsilon)\widetilde{x}_\rho$. Similar results, including examinations when the heavy-traffic approximation remains

valid, can be found in the works of Blanchet and Lam [25], Denisov and Kugler [45] and Kugler and Wachtel [88].

The above-mentioned works all focus on global asymptotics of the all-time supremum functional M_∞^ρ, and one may wonder how robust the obtained insights are when other functionals of importance are considered. For example, another well-studied functional of Lévy processes is the first passage time of zero, τ^ρ, which among others may characterise a busy-period duration in queueing theory. A third functional of importance is $M_\tau^\rho := \sup_{t<\tau^\rho} X^\rho(t)$. A series of prior works [13, 101, 144] obtain useful asymptotic approximations for τ^ρ, while M_τ^ρ has been considered in Asmussen [7]. All these works focus on (a subclass of) subexponential jump sizes and fixed ρ. Our aim is to investigate how robust these asymptotic estimates are when also $\rho \uparrow 1$.

We feel that our main achievement is a description of the tail behaviour of $\mathbb{P}(\tau^\rho > x)$ as $x \to \infty$ while $\rho \uparrow 1$. For fixed ρ, Zwart [144] showed that

$$\mathbb{P}(\tau^\rho > x) \sim \frac{1}{1-\rho}\mathbb{P}(B_1 > (1-\rho)x) \tag{5.4}$$

as $x \to \infty$. In the current chapter, we show that this large-deviations approximation remains valid as $\rho \uparrow 1$ for all x above a certain threshold x_ρ^* which turns out to be much larger than threshold (5.3):

$$x_\rho^* := \frac{1}{(1-\rho)^2}\left(\log\frac{1}{1-\rho}\right)^{k^*}, \tag{5.5}$$

where $k^* > 2$. We actually show that the asymptotic behaviour of $\mathbb{P}(\tau^\rho > x)$ coincides with $\mathbb{P}(M_\tau^\rho > (1-\rho)x)$; intuitively, if the process hits zero after time x, then it is likely that the process obeyed the long-term drift after reaching level $(1-\rho)x$ early in time. Uniform heavy-tail approximations for M_τ^ρ, which are established in this chapter as by-product of independent interest, yield the given asymptotic. The gap between x_ρ^* and $\widetilde{x}_\rho/(1-\rho)$ is required for technical reasons; however, we show that our result does *not* hold for $k^* = 0$ in (5.5) (i.e. if x_ρ^* is proportional to $(1-\rho)^{-2}$).

Additional theorems that lead to our main result provide uniform heavy-tail approximations on the "local" tail probability $\mathbb{P}(M_\infty^\rho \in [x, x+T))$ of the all-time supremum functional M_∞^ρ, and uniform heavy-tail approximations on the tail distribution of the largest jump B_τ^ρ until time τ^ρ. The local asymptotics of M_∞^ρ provide a generalization of Corollary 2.1(b) in Olvera-Cravioto et al. [107] and are obtained in a similar fashion via a decomposition of the Pollaczek-Khintchine formula. Furthermore, we derive asymptotic expressions for the conditional expected time of reaching a high level a, given that level a is reached before time τ^ρ. The corresponding lemma relies heavily on fluctuation theory for Lévy processes; specifically, it relies on the theory of scale functions. A recent review article on and examples of scale functions can be found in Kuznetsov et al. [89] and Hubalek and Kyprianou [73], respectively.

The chapter is organised as follows. A precise description of the model and an introduction to the notation used can be found in Section 5.2. Section 5.3 presents and discusses our results; in particular, Theorems 5.3.1 and 5.3.5 display our main results. The four subsequent sections are each devoted to the proof of one theorem. Section 5.8 contains the extensive proof of a crucial lemma, and, finally, Section 5.9 provides the theoretical support for the discussion presented in Section 5.3.1. Finally, the deferred proof of a minor lemma is presented in Appendix 5.A.

5.2 Preliminaries

Let $\{B\} \cup \{B_i\}_{i=0}^{\infty}$ be a sequence of non-negative, independent and identically distributed (i.i.d.) regularly-varying random variables (cf. [24]) with mean $\mathbb{E}[B] > 0$ and finite variance σ_B^2. More specifically, their common cumulative distribution function (c.d.f.) $F_B : \mathbb{R} \to [0,1]$, $F_B(0) = 0$ is characterised by its tail

$$\overline{F}_B(x) := \mathbb{P}(B > x) = L(x)x^{-\alpha}, \tag{5.6}$$

where $\alpha > 2$, $\alpha \neq 3$, and $L(x)$ is a slowly varying function: $\lim_{x\to\infty} L(ax)/L(x) = 1$ for all $a > 0$. A key property of such distributions is that $\mathbb{E}[B^p] < \infty$ for $p < \alpha$ and $\mathbb{E}[B^p] = \infty$ for $p > \alpha$. The α-th moment can be either finite or infinite. For technical reasons, this chapter does not address the $\alpha = 3$ case. It should be noted that regularly-varying distributions are a subclass of subexponential distributions [64], and as such satisfy $\lim_{x\to\infty} \mathbb{P}(\max\{B_1, \ldots, B_n\} > x)/\mathbb{P}(B_1 > x) = n$.

Define the Poisson process $N^1(t), t \geq 0$, which is independent of the B_i and has rate $1/\mathbb{E}[B]$. Then $N^\rho(t) := N^1(\rho t), t \geq 0$, is a Poisson process with rate $\lambda^\rho := \rho/\mathbb{E}[B]$. We consider a family of Lévy processes $\{X^\rho(t)\}$, indexed by $\rho \in (0,1)$, where $X^\rho : [0,\infty) \to \mathbb{R}$ is characterised as

$$X^\rho(t) := X_0 + \sum_{i=1}^{N^\rho(t)} B_i - t. \tag{5.7}$$

We say that X^ρ is a compound Poisson process with initial value $X^\rho(0) = X_0 := B_0$ and long-term drift $\mathbb{E}[X^\rho(1) - X^\rho(0)] = -(1-\rho) < 0$. The process $X^\rho(t)$ experiences a deterministic decrease of $-t$ and has jumps of size B_i. For this reason we refer to F_B as the jump-size distribution.

The first passage time of level x is denoted by $\sigma^\rho(x) := \inf\{t \geq 0 : X^\rho(t) \geq x\}$, whereas the first hitting time of level zero is indicated by $\tau^\rho := \inf\{t \geq 0 : X^\rho(t) = 0\}$. Of primary interest in this chapter are the supremum M_τ^ρ until the first down-crossing of level zero, i.e. $M_\tau^\rho := \sup\{X^\rho(t) : 0 \leq t \leq \tau^\rho\}$, and the all-time supremum $M_\infty^\rho := \sup\{X^\rho(t) : t \geq 0\}$ of the Lévy process. We also derive a result on the largest jump B_τ^ρ before time τ^ρ: $B_\tau^\rho := \sup\{B_i : 0 \leq i \leq N^\rho(\tau^\rho)\}$.

Consider the sequence of i.i.d. random variables $\{B^*\} \cup \{B_i^*\}_{i=1}^{\infty}$ with c.d.f. F_{B^*}. F_{B^*} is the excess distribution of B and will be referred to as excess jump-size distribution. The excess jump-size distribution can be characterised by its probability density function (p.d.f.) $f_{B^*}(x) = \frac{1}{\mathbb{E}[B]}\mathbb{P}(B > x)$ and has finite mean $\mu := \mathbb{E}[B^2]/(2\mathbb{E}[B]) < \infty$. It is assumed that B^* and B_i^* are independent of N^ρ, B and B_i for all relevant indices.

Since B is regularly varying, Theorem 2.45 in Foss et al. [59] states that the tail distribution of B^*,

$$\overline{F}_{B^*}(x) = \frac{1}{\mathbb{E}[B]}\int_x^\infty \mathbb{P}(B > t)\,\mathrm{d}t \sim \frac{1}{(\alpha-1)\mathbb{E}[B]}L(x)x^{-\alpha+1}, \tag{5.8}$$

is also regularly varying, where $f(z) \sim g(z)$ if and only if $\lim_{z\uparrow z^*} f(z)/g(z) = 1$ for some limiting value $z^* \in \{1, \infty\}$. In this chapter, the limit of interest is either $\rho \uparrow 1, x \to \infty$ or $a \to \infty$. The proper limit should be clear from the context. Similarly, $f(z) \gtrsim (\lesssim) g(z)$ denotes the relation $\liminf_{z\uparrow z^*}(\limsup_{z\uparrow z^*}) f(z)/g(z) \geq (\leq) 1$. We adopt the common conventions $f(z) = O(g(z))$ if and only if $\limsup_{z\uparrow z^*} |f(z)/g(z)| < \infty$ and $f(z) = o(g(z))$ if and only if $\limsup_{z\uparrow z^*} f(z)/g(z) = 0$. If both $f(z) = O(g(z))$ and $g(z) = O(f(z))$, then this is denoted by $f(z) = \Theta(g(z))$.

Let $T \in (0, \infty)$ be any positive constant and define the interval $\Delta = [0, T)$. In the remainder of this chapter we will denote the "local" tail probability $\mathbb{P}(B^* \in [x, x+T))$ by $\mathbb{P}(B^* \in x + \Delta)$. Furthermore, we adopt the well-known conventions of the floor-function $\lfloor x \rfloor := \max\{n \in \mathbb{N} : n \leq x\}$ and the ceiling-function $\lceil x \rceil := \min\{n \in \mathbb{N} : n \geq x\}$, and denote $\mathbb{1}(\textit{[logical expression]})$ for the indicator function that assumes value 1 if the logical expression is true, and value 0 otherwise.

Most variables that have been introduced so far depend on the parameter ρ. Now that their dependence has been noted, we drop the superscripts ρ for the remainder of this chapter. Variables that are introduced in later sections and that depend on ρ will carry a sub- or superscript unless mentioned otherwise.

Finally, we note that many expressions in this chapter involve constants that do not provide additional insight, and that do not contribute to the global behaviour of the expressions. For this reason, many constants have been replaced by C: a constant whose value may change from line to line.

5.3 Main results and discussion

The purpose of this section is to present and discuss our main results. Theorem 5.3.1 presents an exact uniform asymptotic relation between the all-time supremum M_∞ and excess jump size B^*. Theorems 5.3.2 and 5.3.3 and Corollary 5.3.4 display exact uniform asymptotic relations between the supremum M_τ, the largest jump B_τ and the jump sizes B. Theorem 5.3.5 shows an exact uniform asymptotic for the relation between the first hitting time τ and M_τ. The tightness of the results is discussed in Section 5.9.

Our first theorem relates the local tail probability $\mathbb{P}(M_\infty \in x+\Delta)$ to the local tail probability $\mathbb{P}(B^* \in x+\Delta)$:

Theorem 5.3.1. *Suppose $\mathbb{P}(B > x) = L(x)x^{-\alpha}$ for some $\alpha > 2$, $\alpha \neq 3$ and $L(x)$ slowly varying. Let $\mu = \mathbb{E}[B^2]/(2\mathbb{E}[B])$ and define $x_\rho := k\mu(\alpha-1)\frac{1}{1-\rho}\log\frac{1}{1-\rho}$ for any $k > 1$. Then for any fixed interval $\Delta = [0,T)$ the relation*

$$\sup_{x\geq x_\rho}\left|\frac{\mathbb{P}(M_\infty \in x+\Delta)}{\frac{\rho}{1-\rho}\mathbb{P}(B^* \in x+\Delta)} - 1\right| \to 0 \tag{5.9}$$

holds as $\rho \uparrow 1$. Furthermore, (5.9) remains valid for $k = 1$ if $L(x)/(\log x)^\alpha \to \infty$.

Theorem 5.3.1 extends Corollary 2.3(b) of Olvera-Cravioto et al. [107], who considered the "global" tail probability $\Delta = [0,\infty)$. The similarity of the results is also reflected in the proof of the theorem, which greatly depends on the Pollaczek-Khintchine formula and the power law nature of the jump-size distribution. A key difference between the proofs is Olvera-Cravioto et al.'s application of the "global" big jump asymptotics as reported by Borovkov and Borovkov [29] versus our usage of the "local" analogues as derived by Denisov et al. [46].

The transition point $\widetilde{x}_\rho$ in Olvera-Cravioto et al. [107] (cf. expression (5.3)) differs from x_ρ by a factor $\frac{\alpha-1}{\alpha-2}$, which is an artefact of our analysis of the local tail probability (index α) as opposed to their analysis of the global tail probability (index $\alpha-1$). Similarly, their $k = 1$ case requires $L(x)$ to asymptotically dominate $(\log x)^{\alpha-1}$ instead of $(\log x)^\alpha$.

Our next result relates the tail behavior of M_τ to that of B:

Theorem 5.3.2. *Suppose that all conditions in Theorem 5.3.1 hold. Then*

$$\sup_{x\geq x_\rho}\left|\frac{\mathbb{P}(M_\tau > x)}{\frac{1}{1-\rho}\mathbb{P}(B > x)} - 1\right| \to 0 \tag{5.10}$$

holds as $\rho \uparrow 1$. Furthermore, (5.10) remains valid for $k = 1$ if $L(x)/(\log x)^\alpha \to \infty$.

Theorem 5.3.2 is related to a similar result for general random walks, derived for a larger class of subexponential distributions, see also Theorem 2.1 in Asmussen [7]. Again, the contribution in our setting is the validity of this asymptotic estimate in the near-critical regime. Also the intuition behind this result, that M_τ is comparable in size to the largest jump B_τ, remains valid:

Theorem 5.3.3. *Suppose $\mathbb{P}(B > x) = L(x)x^{-\alpha}$ for some $\alpha > 2$, $\alpha \neq 3$ and $L(x)$ slowly varying. Let $\widehat{x}_\rho$ satisfy $\mathbb{P}(B > \widehat{x}_\rho)/(1-\rho)^2 \to 0$ as $\rho \uparrow 1$. Then the relation*

$$\sup_{x\geq \widehat{x}_\rho}\left|\frac{\mathbb{P}(B_\tau > x)}{\frac{1}{1-\rho}\mathbb{P}(B > x)} - 1\right| \to 0 \tag{5.11}$$

holds as $\rho \uparrow 1$. In particular, the above statement holds for $\widehat{x}_\rho \geq 1/(1-\rho)$.

Corollary 5.3.4. *Suppose that all conditions in Theorem 5.3.1 hold. Then*

$$\sup_{x\geq x_\rho}\left|\frac{\mathbb{P}(M_\tau > x)}{\mathbb{P}(B_\tau > x)} - 1\right| \to 0 \tag{5.12}$$

holds as $\rho \uparrow 1$. *Furthermore,* (5.12) *remains valid for* $k = 1$ *if* $L(x)/(\log x)^\alpha \to \infty$.

Here, we note that $\mathbb{E}[\tau] = \mathbb{E}[B]/(1-\rho)$ and therefore one might guess $P(B_\tau > x) \approx \mathbb{P}(\max\{B_1,\dots,B_{1/(1-\rho)}\} > x)) \approx \frac{1}{1-\rho}\mathbb{P}(B > x)$ as a property of subexponential functions. Theorem 5.3.3 makes this relation explicit.

We are now ready to examine the asymptotic behaviour of the tail probability of the first hitting time of zero, $\mathbb{P}(\tau > x)$:

Theorem 5.3.5. *Suppose* $\mathbb{P}(B > x) = L(x)x^{-\alpha}$ *for some* $\alpha > 2$, $\alpha \neq 3$ *and* $L(x)$ *slowly varying. For any* $k^* > 2$ *define* $x_\rho^* := \frac{1}{(1-\rho)^2}\left(\log\frac{1}{1-\rho}\right)^{k^*}$. *Then both*

$$\sup_{x\geq x_\rho^*}\left|\mathbb{P}(\tau > x \mid M_\tau > (1-\rho)x) - 1\right| \to 0 \tag{5.13}$$

and

$$\sup_{x\geq x_\rho^*}\left|\mathbb{P}(M_\tau > (1-\rho)x \mid \tau > x) - 1\right| \to 0 \tag{5.14}$$

hold as $\rho \uparrow 1$. *In particular,* (5.13) *and* (5.14) *imply*

$$\sup_{x\geq x_\rho^*}\left|\frac{\mathbb{P}(\tau > x)}{\mathbb{P}(M_\tau > (1-\rho)x)} - 1\right| \to 0 \tag{5.15}$$

as $\rho \uparrow 1$.

For fixed ρ, related questions have been examined by Durrett [48] for random walks and Zwart [144] for queues. Their results lead to the insight that a large value of τ is caused by an 'early' big jump, after which the process drifts towards 0 at rate $1-\rho$ (see Figure 5.1). This suggests the approximation $\tau \approx M_\tau/(1-\rho)$, which was made rigorous by Zwart [144] using a sample-path analysis. The challenge in our setting is to show that the big jump occurs at a time that does not grow too large as $\rho \uparrow 1$. This is settled by the crucial technical Lemma 5.7.2 in Section 5.7, which essentially states that it takes $O(1/(1-\rho))$ time units until the largest jump. This lemma is proven by providing an estimate of the time until the big jump in terms of q-scale functions, which in turn need to be estimated in detail for various specific ranges of parameter values.

5.3.1 Tightness of bounds

It is natural to question the quality of our thresholds x_ρ and x_ρ^* in the results presented above. Corollary 5.3.7 and Lemma 5.3.8 show that our choices are close to optimal,

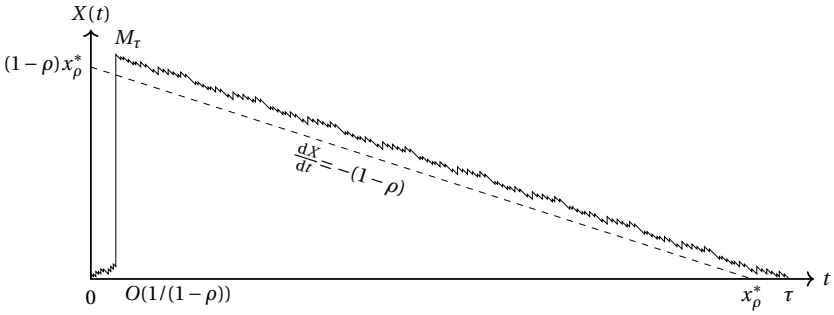


Figure 5.1: Illustration of a scenario where $X(t)$ stays positive for a long time due to a large jump early in the process. The largest jump of size $(1-\rho)x$ happens at time $O(1/(1-\rho))$. The long-term drift of $-(1-\rho)$ suggests that $\tau \approx (1+o(1))x$.

in the sense that our results are no longer valid if the logarithmic terms in x_ρ and x_ρ^* are dropped. Prior to these claims, however, we state a local analogue of Kingman's heavy-traffic approximation that is required in the proof of Corollary 5.3.7, but which is also of independent interest.

First, consider the function $x_\rho = k\mu(\alpha-1)\frac{1}{1-\rho}\log\frac{1}{1-\rho}$ as presented in Theorem 5.3.1, Theorem 5.3.2 and Corollary 5.3.4. As stated earlier, Theorem 5.3.1 is the local analogue of Corollary 2.3(b) in Olvera-Cravioto et al. [107] and the function x_ρ only differs by a constant from their function $\widetilde{x}_\rho$. Additionally, their Corollary 2.3(a) states that the tail probability $\mathbb{P}(M_\infty > x)$ asymptotically behaves as an exponential random variable for $x < (1-\varepsilon)\widetilde{x}_\rho$, $\varepsilon > 0$ sufficiently small. This result suggests that the local tail probability $\mathbb{P}(M_\infty \in x+\Delta)$ behaves as the density of an exponential random variable for x sufficiently small. The next lemma supports this suggestion by presenting a local analogue of Kingman's heavy-traffic approximation that appears to be new:

Lemma 5.3.6. *Suppose that the jump size p.d.f. $f_B(x)$ of B is completely monotone; i.e. $f_B(x)$ and all its derivatives exist and satisfy $(-1)^n\frac{\mathrm{d}^n}{\mathrm{d}x^n}f_B(x) \geq 0$ for all $x > 0$ and $n = 1,2,\ldots$ Fix $y > 0$. Then the all-time supremum p.d.f. of $f_{M_\infty}(x)$ on $(0,\infty)$ exists and satisfies*

$$\lim_{\rho\uparrow 1}\frac{1}{1-\rho}f_{M_\infty}\left(\frac{y}{1-\rho}\right) = \frac{1}{\mathbb{E}[B^*]}e^{-\frac{y}{\mathbb{E}[B^*]}}. \tag{5.16}$$

We hence expect $\mathbb{P}(M_\infty \in x+\Delta)$ to display exponential decay as $\rho \uparrow 1$ for x sufficiently smaller than x_ρ, similar to the results of Olvera-Cravioto et al. [107]. Analysing the local tail probability $\mathbb{P}(M_\infty \in x+\Delta)$ for general $x \leq x_\rho$ is beyond the scope of this chapter; however, the corollary below shows that $(1-\rho)x(\rho)$ must diverge to infinity in order for Theorem 5.3.1 to remain true:

Corollary 5.3.7. *Suppose* $\mathbb{P}(B > x) = L(x)x^{-\alpha}$ *for some* $\alpha > 2$ *and* $L(x)$ *slowly varying, and assume that the jump size p.d.f.* $f_B(x)$ *of* B *is completely monotone. Fix* $y > 0$*. Then for* $y_\rho = \frac{y}{1-\rho}$ *the limit*

$$\lim_{\rho \uparrow 1} \frac{\mathbb{P}(M_\infty \in y_\rho + \Delta)}{\frac{\rho}{1-\rho}\mathbb{P}(B^* \in y_\rho + \Delta)} \tag{5.17}$$

diverges to infinity.

The proof of Theorem 5.3.2 derives the estimates

$$\frac{1}{\lambda}\mathbb{P}(M_\infty \in [x, x+1)) \lesssim \mathbb{P}(M_\tau > x) \lesssim \frac{1}{\lambda}\mathbb{P}(M_\infty \in [x-1, x)) \tag{5.18}$$

as $x \to \infty$. As such, a similar necessary condition on any function $x(\rho)$ for which Theorem 5.3.1 holds is also necessary for Theorem 5.3.2. An analogue argument holds for Corollary 5.3.4.

We next discuss the function $x^*_\rho = \frac{1}{(1-\rho)^2}\left(\log\frac{1}{1-\rho}\right)^k$ which is of interest in Theorem 5.3.5. The proof of Theorem 5.3.5 greatly relies on Theorem 5.3.2 but considers $\mathbb{P}(M_\tau > (1-\rho)x)$ instead of $\mathbb{P}(M_\tau > x)$. We would therefore expect Theorem 5.3.5 to hold with $x(\rho) = x_\rho/(1-\rho)$. The current proof, however, requires the higher level x^*_ρ for technical reasons. In contrast, the following lemma gives a lower bound on $x(\rho)$ if it is to replace x^*_ρ. In particular, it states that $(1-\rho)^2 x(\rho)$ needs to diverge to infinity:

Lemma 5.3.8. *Suppose* $\mathbb{P}(B > x) = L(x)x^{-\alpha}$ *for some* $\alpha > 2$ *and* $L(x)$ *slowly varying. Fix* $y > 0$*. Then for* $y^*_\rho = \frac{y}{(1-\rho)^2}$ *the limit*

$$\lim_{\rho \uparrow 1} \frac{\mathbb{P}(\tau > y^*_\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B > (1-\rho)y^*_\rho)} \tag{5.19}$$

diverges to infinity.

5.4 Local asymptotics of the all-time supremum

This section contains the proof of Theorem 5.3.1. We consider the all-time supremum by its Pollaczek-Khintchine infinite-series representation. From this representation, we distinguish between few jumps and many jumps scenarios (small and large n), where the threshold is approximately $x/\mathbb{E}[B^*]$. It is shown that under the few jumps scenario, a large all-time supremum is most probably due to a large value of a single B^*_i. On the other hand, the many jumps scenario is shown to be negligible compared to the few jumps scenario.

Define $S^*_0 := 0$ and $S^*_n := \sum_{i=1}^n B^*_i$. By Theorem VIII.5.7 in Asmussen [8],

$$\mathbb{P}(M_\infty \in x + \Delta) = \sum_{n=0}^{\infty} (1-\rho)\rho^n \mathbb{P}(S^*_n \in x + \Delta) \tag{5.20}$$

for all $x>0$. An equivalent representation of (5.9) is therefore

$$\sup_{x\geq x_\rho}\left|\frac{\sum_{n=1}^{\infty}(1-\rho)\rho^n\left[\mathbb{P}(S_n^*\in x+\Delta)-n\mathbb{P}(B^*\in x+\Delta)\right]}{\frac{\rho}{1-\rho}\mathbb{P}(B^*\in x+\Delta)}\right|\to 0 \tag{5.21}$$

as $\rho\uparrow 1$. Fix δ such that $\max\{\frac{1}{2},\frac{1}{\alpha-1}\}<\delta<1$ and define $U_\delta(x):=\lfloor(x-x^\delta)/\mu\rfloor$. Then, the numerator in (5.21) can be decomposed as

$$\begin{aligned}
&\left|\sum_{n=1}^{\infty}(1-\rho)\rho^n\left[\mathbb{P}(S_n^*\in x+\Delta)-n\mathbb{P}(B^*\in x+\Delta)\right]\right| \\
&\qquad\leq\sum_{n=1}^{U_\delta(x)}(1-\rho)\rho^n\left|\mathbb{P}(S_n^*\in x+\Delta)-n\mathbb{P}(B^*\in x-(n-1)\mu+\Delta)\right| \\
&\qquad\quad+\sum_{n=1}^{U_\delta(x)}(1-\rho)\rho^n n\left|\mathbb{P}(B^*\in x-(n-1)\mu+\Delta)-\mathbb{P}(B^*\in x+\Delta)\right| \\
&\qquad\quad+\left|\sum_{n=U_\delta(x)+1}^{\infty}(1-\rho)\rho^n\left[\mathbb{P}(S_n^*\in x+\Delta)-n\mathbb{P}(B^*\in x+\Delta)\right]\right|.
\end{aligned} \tag{5.22}$$

Here, the first term corresponds to the few jumps scenario and the third term corresponds to the many jumps scenario. The second term corrects a shift in the argument of $\mathbb{P}(B^*\in\cdot)$, which is required for application of the following lemma:

Lemma 5.4.1. *Suppose ξ is a non-negative regularly-varying random variable whose c.d.f. has index $-\alpha_\xi<-2,\alpha_\xi\neq 3$; i.e. $\mathbb{P}(\xi>x)=L(x)x^{-\alpha_\xi}$. Let F_{ξ^*} be the excess distribution of ξ with index $-\alpha_\xi+1<-1$ and i.i.d. samples $\xi^*,\xi_1^*,\xi_2^*,\ldots$ For any $\max\left\{\frac{1}{\alpha_\xi-1},\frac{1}{2}\right\}<\Gamma<1$ denote $U_\Gamma(x)=\lfloor\frac{x-x^\Gamma}{\mathbb{E}[\xi^*]}\rfloor$. Then, there exists a non-increasing function $\phi(x)$ satisfying $\phi(x)\downarrow 0$ as $x\to\infty$ such that*

$$\sup_{1\leq n\leq U_\Gamma(x)}\left|\frac{\mathbb{P}(\xi_1^*+\ldots+\xi_n^*\in x+\Delta)}{n\mathbb{P}(\xi^*\in x-(n-1)\mathbb{E}[\xi^*]+\Delta)}-1\right|\leq\phi(x).$$

The proof is delayed until the end of this section and relies heavily on the machinery provided by Denisov et al. [46]. Lemma 5.4.1 is closely related to the subexponential property $\lim_{x\to\infty}\mathbb{P}(B_1^*+\ldots+B_n^*>x)/\mathbb{P}(B_1^*>x)=n$ and guarantees that, for some non-increasing $\phi(x)\downarrow 0$ as $x\to\infty$, expression (5.22) is dominated by

$$\begin{aligned}
&\phi(x)\sum_{n=1}^{U_\delta(x)}(1-\rho)\rho^n n\mathbb{P}(B^*\in x+\Delta) \\
&\quad+(1+\phi(x))\sum_{n=1}^{U_\delta(x)}(1-\rho)\rho^n n\left|\mathbb{P}(B^*\in x-(n-1)\mu+\Delta)-\mathbb{P}(B^*\in x+\Delta)\right| \\
&\quad+\sum_{n=U_\delta(x)+1}^{\infty}(1-\rho)\rho^n\left[1+n\mathbb{P}(B^*\in x+\Delta)\right] \\
&=:\phi(x)\mathrm{I}(x,\rho)+(1+\phi(x))\mathrm{II}(x,\rho)+\mathrm{III}(x,\rho).
\end{aligned}$$

Term $\mathrm{I}(x,\rho)$ is bounded by $\frac{\rho}{1-\rho}\mathbb{P}(B^*\in x+\Delta)$, so that $x_\rho\to\infty$ implies

$$\sup_{x\geq x_\rho}\frac{\phi(x)\mathrm{I}(x,\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B^*\in x+\Delta)}\leq\phi(x_\rho)\to 0 \tag{5.23}$$

as $\rho\uparrow 1$. We are done if terms II and III also vanish uniformly.

Term II is split into two parts. Fix γ such that $0<\gamma<\delta$ and define the function $V_\gamma(x):=\lfloor(1-\gamma)x/\mu\rfloor$. For x sufficiently large we have $V_\gamma(x)<U_\delta(x)$, so that II may be written as

$$\begin{aligned}\mathrm{II}(x,\rho)=&\sum_{n=1}^{V_\gamma(x)}(1-\rho)\rho^n n\Big|\mathbb{P}(B^*\in x-(n-1)\mu+\Delta)-\mathbb{P}(B^*\in x+\Delta)\Big|\\&+\sum_{n=V_\gamma(x)+1}^{U_\delta(x)}(1-\rho)\rho^n n\Big|\mathbb{P}(B^*\in x-(n-1)\mu+\Delta)-\mathbb{P}(B^*\in x+\Delta)\Big|\\=:&\,\mathrm{IIa}(x,\rho)+\mathrm{IIb}(x,\rho).\end{aligned}$$

For $1\leq n\leq V_\gamma(x)$, Newton's generalised binomial Theorem implies that

$$\begin{aligned}\frac{\mathbb{P}(B^*\in x-(n-1)\mu+\Delta)}{\mathbb{P}(B^*\in x+\Delta)}&\leq\frac{\mathbb{P}(B>x-(n-1)\mu)}{\mathbb{P}(B>x+T)}\sim\left(1-\frac{(n-1)\mu+T}{x+T}\right)^{-\alpha}\\&=1+\sum_{m=1}^{\infty}\frac{\alpha(\alpha+1)\cdots(\alpha+m-1)}{m!}\left(\frac{(n-1)\mu+T}{x+T}\right)^m\\&\leq 1+\alpha\left(\frac{(n-1)\mu+T}{x+T}\right)\left(1-\frac{(n-1)\mu+T}{x+T}\right)^{-\alpha-1}\\&\lesssim 1+\alpha\gamma^{-\alpha-1}\frac{(n-1)\mu+T}{x+T},\end{aligned}$$

as $x\to\infty$, and therefore

$$\frac{\mathbb{P}(B^*\in x-(n-1)\mu+\Delta)}{\mathbb{P}(B^*\in x+\Delta)}-1\lesssim C\frac{n-1}{x}$$

as $x\to\infty$. Substituting this into IIa gives

$$\begin{aligned}\mathrm{IIa}(x,\rho)&\lesssim C\mathbb{P}(B^*\in x+\Delta)\frac{1}{x}\sum_{n=1}^{V_\gamma(x)}(1-\rho)\rho^n n(n-1)\\&\leq C\mathbb{P}(B^*\in x+\Delta)\frac{2\rho^2}{(1-\rho)^2x}\left(1-\rho^{V_\gamma(x)}-(1-\rho)V_\gamma(x)\rho^{V_\gamma(x)}\right)\\&\leq\frac{C\rho}{(1-\rho)^2x}\mathbb{P}(B^*\in x+\Delta).\end{aligned}$$

We hence conclude

$$\sup_{x\geq x_\rho}\frac{\mathrm{IIa}(x,\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B^*\in x+\Delta)}\lesssim\frac{C}{\log\frac{1}{1-\rho}}\to 0 \tag{5.24}$$

as $\rho\uparrow 1$.

Next, consider term IIb. Since $\mathbb{P}(B^* \in y+\Delta)$ is decreasing in y, we find

$$\begin{aligned}\mathrm{IIb}(x,\rho) &\le C(1-\rho)\rho^{V_\gamma(x)+1}x \sum_{n=V_\gamma(x)+1}^{U_\delta(x)} \mathbb{P}(B^* \in x-(n-1)\mu+\Delta)\\ &\le C(1-\rho)\rho^{(1-\gamma)x/\mu}x\int_{x-\mu U_\delta(x)}^{x-\mu V_\gamma(x)} \mathbb{P}(B^* \in t+\Delta)\,\mathrm{d}t.\end{aligned}$$

Noting that $\mathbb{P}(B^* \in x+\Delta)$ is regularly varying with index $-\alpha < -2$, Theorem 1.5.11 in Bingham et al. [24] indicates that

$$\begin{aligned}\mathrm{IIb}(x,\rho) &\lesssim C(1-\rho)\rho^{(1-\gamma)x/\mu}x(x-\mu U_\delta(x))\mathbb{P}(B^* \in x-\mu U_\delta(x)+\Delta)\\ &\le C(1-\rho)\rho^{(1-\gamma)x/\mu}x^{1+\delta}\mathbb{P}(B^* \in x^\delta-\Delta).\end{aligned}$$

It remains to verify that IIb decreases sufficiently fast for $x \ge x_\rho$. To this end, we write

$$\begin{aligned}\sup_{x\ge x_\rho}\frac{\mathrm{IIb}(x,\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B^* \in x+\Delta)} &\lesssim C\sup_{x\ge x_\rho}(1-\rho)^2\rho^{(1-\gamma)\frac{x}{\mu}-1}x^{1+\delta}\frac{\mathbb{P}(B^* \in x^\delta+\Delta)}{\mathbb{P}(B^* \in x+\Delta)}\\ &\le C\sup_{x\ge x_\rho}(1-\rho)^2\rho^{(1-\gamma)\frac{x}{\mu}-1}x^{1+\delta}\frac{\mathbb{P}(B>x^\delta)}{\mathbb{P}(B>x+T)}\\ &\sim C\sup_{x\ge x_\rho}(1-\rho)^2 e^{\left((1-\gamma)\frac{x}{\mu}-1\right)\log\rho}x^{1+\delta+(1-\delta)\alpha}\\ &\le C\sup_{x\ge x_\rho}(1-\rho)^2 e^{-\left((1-\gamma)\frac{x}{\mu}-1\right)(1-\rho)}x^{1+\delta+(1-\delta)\alpha},\end{aligned}$$

where we exploited the inequality $\log\rho \le -(1-\rho)$. Additionally, for ρ sufficiently close to one, the supremum is achieved in $x = x_\rho$ and

$$\sup_{x\ge x_\rho}\frac{\mathrm{IIb}(x,\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B^* \in x+\Delta)} \lesssim C(1-\rho)^2 e^{-(1-\gamma)(1-\rho)\frac{x_\rho}{\mu}}x_\rho^{1+\delta+(1-\delta)\alpha}.$$

Substituting $x_\rho = k\mu(\alpha-1)\frac{1}{1-\rho}\log\frac{1}{1-\rho}$ now gives

$$\sup_{x\ge x_\rho}\frac{\mathrm{IIb}(x,\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B^* \in x+\Delta)} \lesssim C(1-\rho)^{k(1-\gamma)(\alpha-1)-(1-\delta)(\alpha-1)}\left(\log\frac{1}{1-\rho}\right)^{1+\delta+(1-\delta)\alpha} \to 0 \quad (5.25)$$

as $\rho \uparrow 1$, since $\gamma < \delta$. This verifies the convergence of term II to zero.

We continue with the analysis of term III. This term is rewritten into two specified terms:

$$\begin{aligned}\mathrm{III}(x,\rho) &= \rho^{U_\delta(x)+1} + \left[(U_\delta(x)+1)\,\rho^{U_\delta(x)+1} + \frac{\rho^{U_\delta(x)+2}}{1-\rho}\right]\mathbb{P}(B^* \in x+\Delta)\\ &\le \rho^{\frac{x-x^\delta}{\mu}} + \left[\left(\frac{x-x^\delta}{\mu}+1\right)+\frac{\rho}{1-\rho}\right]\mathbb{P}(B^* \in x+\Delta)\rho^{\frac{x-x^\delta}{\mu}}\\ &\le \rho^{\frac{x-x^\delta}{\mu}} + C\frac{(1-\rho)x+1}{1-\rho}\mathbb{P}(B^* \in x+\Delta)\rho^{\frac{x-x^\delta}{\mu}}\\ &=: \mathrm{IIIa}(x,\rho)+\mathrm{IIIb}(x,\rho).\end{aligned}$$

We consider terms IIIa and IIIb in order.

For term IIIa, we first assume that $k > 1$. Potter's Theorem (e.g. Theorem 1.5.6 in Bingham et al. [24]) suggests that $\mathbb{P}(B^* \in x+\Delta) \geq T\mathbb{P}(B \geq x+T) \geq TC(x+T)^{-\alpha-\nu}$ for any fixed $\nu > 0$ and x sufficiently large. In particular, for $0 < \nu < (k-1)(\alpha-1)$,

$$\sup_{x \geq x_\rho} \frac{\mathrm{IIIa}(x,\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B^* \in x+\Delta)} \leq \sup_{x \geq x_\rho} C(1-\rho)(x+T)^{\alpha+\nu} \rho^{\frac{x-x^\delta}{\mu}-1}.$$

Again, the supremum is achieved in $x = x_\rho$ for ρ sufficiently close to one and hence

$$\begin{aligned}\sup_{x \geq x_\rho} \frac{\mathrm{IIIa}(x,\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B^* \in x+\Delta)} &\leq C(1-\rho) e^{(\alpha+\nu)\log x_\rho + \left(\frac{x_\rho - x_\rho^\delta}{\mu}-1\right)\log\rho + (\alpha+\nu)\log\left(1+\frac{T}{x_\rho}\right)} \\ &\leq Ce^{(\alpha+\nu)\log x_\rho - \left(\frac{x_\rho - x_\rho^\delta}{\mu}-1\right)(1-\rho) - \log\frac{1}{1-\rho} + (\alpha+\nu)\log\left(1+\frac{T}{x_\rho}\right)}.\end{aligned}$$

Substitution of $x_\rho = k\mu(\alpha-1)\frac{1}{1-\rho}\log\frac{1}{1-\rho}$ now yields

$$\sup_{x \geq x_\rho} \frac{\mathrm{IIIa}(x,\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B^* \in x+\Delta)} \leq C(1-\rho) e^{(\alpha+\nu-1)\log\frac{1}{1-\rho} - k(\alpha-1)\log\frac{1}{1-\rho} + o\left(\log\frac{1}{1-\rho}\right)}, \tag{5.26}$$

which tends to zero as $\rho \uparrow 1$.

Second, assume $k = 1$ and $L(x)/(\log x)^\alpha \to \infty$. Then there exists a non-increasing function $\phi(x) \downarrow 0$ such that $L(x) \geq (\log^\alpha x)/\phi(x)$. Similar to the preceding analysis we find

$$\begin{aligned}\sup_{x \geq x_\rho} \frac{\mathrm{IIIa}(x,\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B^* \in x+\Delta)} &\leq \sup_{x \geq x_\rho} \frac{1}{L(x+T)}(1-\rho)(x+T)^\alpha \rho^{\frac{x-x^\delta}{\mu}-1} \\ &\lesssim \phi(x_\rho) \sup_{x \geq x_\rho} \frac{1}{\log^\alpha x} e^{\alpha\log x + \left(\frac{x-x^\delta}{\mu}-1\right)\log\rho - \log\frac{1}{1-\rho}} \\ &\leq \phi(x_\rho)\frac{1}{\log^\alpha x_\rho} e^{(\alpha-1)\log\frac{1}{1-\rho} + \alpha\log\log\frac{1}{1-\rho} - (\alpha-1)\left(1-x_\rho^{\delta-1}-x_\rho^{-1}\right)\log\frac{1}{1-\rho}} \\ &= C\phi(x_\rho)\frac{1}{\log^\alpha \frac{1}{1-\rho}} e^{\alpha\log\log\frac{1}{1-\rho} + (\alpha-1)\left(x_\rho^{\delta-1}+x_\rho^{-1}\right)\log\frac{1}{1-\rho}} \\ &= C\phi(x_\rho) e^{(\alpha-1)\left(x_\rho^{\delta-1}+x_\rho^{-1}\right)\log\frac{1}{1-\rho}} \to 0\end{aligned}$$

as $\rho \uparrow 1$ since $(\log x)/x^{1-\delta} \to 0$ for any $\delta < 1$.

Finally, for term IIIb one can see that

$$\sup_{x \geq x_\rho} \frac{\mathrm{IIIb}(x,\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B^* \in x+\Delta)} = C \sup_{x \geq x_\rho} ((1-\rho)x+1)\rho^{\frac{x-x^\delta}{\mu}-1}.$$

As before, the supremum is attained in $x = x_\rho$ for ρ sufficiently close to one. Thus,

$$\sup_{x\geq x_\rho} \frac{\mathrm{IIIb}(x,\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B^* \in x+\Delta)} = C((1-\rho)x_\rho+1)e^{\left(\frac{x_\rho - x_\rho^\delta}{\mu}-1\right)\log\rho}$$
$$\leq C\log\frac{1}{1-\rho}e^{-k(\alpha-1)(1+o(1))\log\frac{1}{1-\rho}} \to 0 \tag{5.27}$$

as $\rho \uparrow 1$. From relations (5.23) – (5.27), we may conclude that (5.21) and equivalently (5.9) converges to zero. This completes the proof of Theorem 5.3.1.

This section is concluded by the proof of Lemma 5.4.1.

5.4.1 Proof of Lemma 5.4.1

First consider the case $-\alpha_\xi < -3$. Then $\sigma^2_{\xi^*} = \mathbb{V}\mathrm{ar}(\xi^*) = \frac{\mathbb{E}[\xi^3]}{3\mathbb{E}[\xi]}$ is finite, and therefore $\overline{\xi}^*_i = \frac{\xi^*_i - \mathbb{E}[\xi^*]}{\sigma_{\xi^*}}$ and $\overline{S}^*_n = \frac{\xi^*_1+\ldots+\xi^*_n - n\mathbb{E}[\xi^*]}{\sigma_{\xi^*}}$ are well-defined for all $i \geq 1, n \geq 1$. Since

$$\frac{\mathbb{P}(\xi^*_1+\ldots+\xi^*_n \in x+\Delta)}{n\mathbb{P}(\xi^* \in x-(n-1)\mathbb{E}[\xi^*]+\Delta)} = \frac{\mathbb{P}\left(\overline{S}^*_n \in \frac{x-n\mathbb{E}[\xi^*]+\Delta}{\sigma_{\xi^*}}\right)}{n\mathbb{P}\left(\overline{\xi}^*_1 \in \frac{x-n\mathbb{E}[\xi^*]+\Delta}{\sigma_{\xi^*}}\right)}, \tag{5.28}$$

the result follows from Theorem 8.1 in Denisov et al. [46] once we show that the fraction $(x-n\mathbb{E}[\xi^*])/\sqrt{(\alpha_\xi-3)n\log n}$ diverges to infinity uniformly for $1 \leq n \leq U_\Gamma(x)$ as $x \to \infty$. As $\Gamma > \frac{1}{2}$, one may see that

$$\frac{x-n\mathbb{E}[\xi^*]}{\sqrt{(\alpha_\xi-3)n\log n}} \geq \frac{x-U_\Gamma(x)\mathbb{E}[\xi^*]}{\sqrt{(\alpha_\xi-3)U_\Gamma(x)\log U_\Gamma(x)}} \sim \sqrt{\frac{\mathbb{E}[\xi^*]}{\alpha_\xi-3}}\frac{x^{\Gamma-\frac{1}{2}}}{\log x}$$

indeed tends to infinity as $x \to \infty$.

Now assume $-3 < -\alpha_\xi < -2$. Let $\widetilde{\xi}^*_i = \xi^*_i - \mathbb{E}[\xi^*]$ and $\widetilde{S}^*_n = \xi^*_1+\ldots+\xi^*_n - n\mathbb{E}[\xi^*]$ for all $i \geq 1, n \geq 1$. Then

$$\frac{\mathbb{P}(\xi^*_1+\ldots+\xi^*_n \in x+\Delta)}{n\mathbb{P}(\xi^* \in x-(n-1)\mathbb{E}[\xi^*]+\Delta)} = \frac{\mathbb{P}\left(\widetilde{S}^*_n \in x-n\mathbb{E}[\xi^*]+\Delta\right)}{n\mathbb{P}\left(\widetilde{\xi}^*_1 \in x-n\mathbb{E}[\xi^*]+\Delta\right)}. \tag{5.29}$$

Fix Γ^* such that $\frac{1}{\alpha_\xi-1} < \Gamma^* < \Gamma$. Theorem 9.1 in Denisov et al. [46] implies that $\mathbb{P}(\widetilde{S}^*_n \in x+\Delta) \sim n\mathbb{P}(\widetilde{\xi}^*_1 \in x+\Delta)$ uniformly for $x \geq n^{\Gamma^*}$. The proof is concluded by showing that $(x-n\mathbb{E}[\xi^*])/n^{\Gamma^*} \to \infty$ uniformly for $1 \leq n \leq U_\Gamma(x)$, which follows from

$$\frac{x-n\mathbb{E}[\xi^*]}{n^{\Gamma^*}} \geq \frac{x-U_\Gamma(x)\mathbb{E}[\xi^*]}{U_\Gamma(x)^{\Gamma^*}} \sim \mathbb{E}[\xi^*]^{\Gamma^*}x^{\Gamma-\Gamma^*}.$$

5.5 Asymptotics of the supremum M_τ

This section is dedicated to the proof of Theorem 5.3.2. The proof quickly follows from Theorem 5.3.1 and the following lemma:

Lemma 5.5.1. *The inequalities*

$$\frac{1}{\lambda}\frac{\mathbb{P}(M_\infty \in [x,x+1))}{\mathbb{P}(M_\infty < x+1)} \leq \mathbb{P}(M_\tau > x) \leq \frac{1}{\lambda}\frac{\mathbb{P}(M_\infty \in [x-1,x))}{\mathbb{P}(M_\infty < x-1)} \tag{5.30}$$

are valid for all $x > 1$.

Lemma 5.5.1 is proven in Appendix 5.A by means of scale functions; a concept that is introduced in Section 5.8. Lemma 5.5.1 and Theorem 5.3.1 together state that

$$\mathbb{P}(M_\tau > x) \lesssim \frac{1}{\lambda}\frac{1}{\mathbb{P}(M_\infty < x-1)}\frac{\rho}{1-\rho}\mathbb{P}(B^* \in [x-1,x))$$

for $x \geq x_\rho$ as $\rho \uparrow 1$. Applying the simple bound $\mathbb{P}(B^* \in [x-1,x)) \leq \frac{1}{\mathbb{E}[B]}\mathbb{P}(B > x-1)$ yields

$$\mathbb{P}(M_\tau > x) \lesssim \frac{1}{\mathbb{P}(M_\infty < x-1)}\frac{1}{1-\rho}\mathbb{P}(B > x-1),$$

which, since B is long-tailed, is asymptotically equivalent to

$$\mathbb{P}(M_\tau > x) \lesssim \frac{1}{\mathbb{P}(M_\infty < x-1)}\frac{1}{1-\rho}\mathbb{P}(B > x). \tag{5.31}$$

It follows that $\mathbb{P}(M_\tau > x)/((1-\rho)^{-1}\mathbb{P}(B > x)) \lesssim 1$ for all $x \geq x_\rho$ as $\rho \uparrow 1$.

The asymptotic lower bound is proven similarly, thereby completing the proof of Theorem 5.3.2.

5.6 Asymptotics of the supremum jump size

This section contributes the proof of Theorem 5.3.3. The following equality is an interpretation of expression (3.4) in Boxma [31]:

$$\mathbb{P}(B_\tau > x) = \mathbb{P}(B > x) + \int_0^x \left[1 - e^{-\lambda\mathbb{P}(B_\tau > x)t}\right] \mathrm{d}\mathbb{P}(B \leq t). \tag{5.32}$$

From this equality it follows that

$$\begin{aligned}
\frac{\mathbb{P}(B > x)}{\mathbb{P}(B_\tau > x)} &= 1 - \int_0^x \left[\frac{1 - e^{-\lambda\mathbb{P}(B_\tau > x)t}}{\mathbb{P}(B_\tau > x)}\right] \mathrm{d}\mathbb{P}(B \leq t) \\
&= 1 - \lambda\int_0^x t\,\mathrm{d}\mathbb{P}(B \leq t) + \lambda\int_0^x \left[1 - \frac{1 - e^{-\lambda\mathbb{P}(B_\tau > x)t}}{\lambda\mathbb{P}(B_\tau > x)t}\right] t\,\mathrm{d}\mathbb{P}(B \leq t) \\
&= 1 - \rho + \lambda\int_x^\infty t\,\mathrm{d}\mathbb{P}(B \leq t) + \lambda\int_0^x \left[1 - \frac{1 - e^{-\lambda\mathbb{P}(B_\tau > x)t}}{\lambda\mathbb{P}(B_\tau > x)t}\right] t\,\mathrm{d}\mathbb{P}(B \leq t),
\end{aligned}$$

so that

$$\begin{aligned}
\frac{1}{1-\rho}\frac{\mathbb{P}(B > x)}{\mathbb{P}(B_\tau > x)} - 1 = \frac{\lambda}{1-\rho}\int_x^\infty t\,\mathrm{d}\mathbb{P}(B \leq t) \\
+ \frac{\lambda}{1-\rho}\int_0^x \left[1 - \frac{1 - e^{-\lambda\mathbb{P}(B_\tau > x)t}}{\lambda\mathbb{P}(B_\tau > x)t}\right] t\,\mathrm{d}\mathbb{P}(B \leq t).
\end{aligned} \tag{5.33}$$

Here, we note that the right-hand side of the latter expression is non-negative because $(1-e^{-y})/y \le 1$.

The first integral in (5.33) can be upper bounded as

$$\begin{aligned}\sup_{x\ge\widehat{x}_\rho}\frac{\lambda}{1-\rho}\int_x^\infty t\,\mathrm{d}\mathbb{P}(B\le t) &= \sup_{x\ge\widehat{x}_\rho}\frac{\lambda}{1-\rho}\mathbb{E}\left[B\mathbb{1}(B>x)\right]\\ &= \sup_{x\ge\widehat{x}_\rho}\left(\frac{\lambda}{1-\rho}\mathbb{E}\left[B-x\mid B>x\right]\mathbb{P}(B>x)+\frac{\lambda x\mathbb{P}(B>x)}{1-\rho}\right)\\ &\lesssim C\sup_{x\ge\widehat{x}_\rho}\frac{\lambda x\mathbb{P}(B>x)}{1-\rho}, \end{aligned}\tag{5.34}$$

since $\mathbb{E}\left[B-x\mid B>x\right]\sim\frac{x}{\alpha-1}$, as shown in Embrechts et al. [50, p.162]. Clearly, this upper bound tends to zero for all $x\ge\widehat{x}_\rho$ provided that $\sup_{x\ge\widehat{x}_\rho}\frac{x\mathbb{P}(B>x)}{1-\rho}\to 0$ as $\rho\uparrow 1$.

Sequentially, we consider the second integral in (5.33). The bound $e^y\ge 1+y+y^2/2$ for $y\ge 0$ implies

$$\begin{aligned}&\sup_{x\ge\widehat{x}_\rho}\frac{\lambda}{1-\rho}\int_0^x\left[1-\frac{1-e^{-\lambda\mathbb{P}(B_\tau>x)t}}{\lambda\mathbb{P}(B_\tau>x)t}\right]t\,\mathrm{d}\mathbb{P}(B\le t)\\ &\qquad\le\sup_{x\ge\widehat{x}_\rho}\frac{\lambda}{1-\rho}\int_0^x\left[1-\frac{1-\frac{1}{1+\lambda\mathbb{P}(B_\tau>x)t+\lambda^2\mathbb{P}(B_\tau>x)^2t^2/2}}{\lambda\mathbb{P}(B_\tau>x)t}\right]t\,\mathrm{d}\mathbb{P}(B\le t)\\ &\qquad=\sup_{x\ge\widehat{x}_\rho}\frac{\lambda}{2}\frac{1}{1-\rho}\int_0^x\frac{\lambda\mathbb{P}(B_\tau>x)t+\lambda^2\mathbb{P}(B_\tau>x)^2t^2}{1+\lambda\mathbb{P}(B_\tau>x)t+\lambda^2\mathbb{P}(B_\tau>x)^2t^2/2}t\,\mathrm{d}\mathbb{P}(B\le t)\\ &\qquad\le\sup_{x\ge\widehat{x}_\rho}\frac{\lambda^2}{2}\frac{\mathbb{P}(B_\tau>x)}{1-\rho}\int_0^x\left[t^2+\lambda\mathbb{P}(B_\tau>x)t^3\right]\mathrm{d}\mathbb{P}(B\le t)\\ &\qquad\le\sup_{x\ge\widehat{x}_\rho}\frac{\lambda^2}{2}\frac{\mathbb{P}(B_\tau>x)}{1-\rho}\left[1+\lambda\mathbb{P}(B_\tau>x)x\right]\mathbb{E}[B^2].\end{aligned}$$

From equation (5.33) we know that $\mathbb{P}(B_\tau>x)\le\frac{\mathbb{P}(B>x)}{1-\rho}$, and therefore

$$\begin{aligned}&\sup_{x\ge\widehat{x}_\rho}\frac{\lambda}{1-\rho}\int_0^x\left[1-\frac{1-e^{-\lambda\mathbb{P}(B_\tau>x)t}}{\lambda\mathbb{P}(B_\tau>x)t}\right]t\,\mathrm{d}\mathbb{P}(B\le t)\\ &\qquad\le\sup_{x\ge\widehat{x}_\rho}\frac{\lambda^2}{2}\frac{\mathbb{P}(B>x)}{(1-\rho)^2}\left[1+\frac{\lambda x\mathbb{P}(B>x)}{1-\rho}\right]\mathbb{E}[B^2].\end{aligned}\tag{5.35}$$

It follows that the second integral vanishes for all $x\ge\widehat{x}_\rho$ if both $\sup_{x\ge\widehat{x}_\rho}\frac{x\mathbb{P}(B>x)}{1-\rho}$ and $\sup_{x\ge\widehat{x}_\rho}\frac{\mathbb{P}(B>x)}{(1-\rho)^2}$ tend to zero as $\rho\uparrow 1$. These conditions are analysed by Potter's Theorem, which states that for any $0<\nu<\alpha-2$ there exists a constant $C_\nu>0$ such that $\mathbb{P}(B>x)\le C_\nu x^{-\alpha+\nu}$ for all x sufficiently large. In particular, we find

$$\sup_{x\ge 1/(1-\rho)}\frac{x\mathbb{P}(B>x)}{1-\rho}\le\sup_{x\ge 1/(1-\rho)}\frac{C_\nu x^{1-\alpha+\nu}}{1-\rho}\to 0,$$

and similarly $\sup_{x\geq 1/(1-\rho)} \frac{\mathbb{P}(B>x)}{(1-\rho)^2} \to 0$, implying that the theorem holds for $\widehat{x} \geq 1/(1-\rho)$. The proof of the theorem is completed after noting that $\frac{x\mathbb{P}(B>x)}{1-\rho} \leq \frac{\mathbb{P}(B>x)}{(1-\rho)^2}$ whenever $x \leq 1/(1-\rho)$.

5.7 Asymptotics of the first hitting time of level zero

This section is devoted to the proof of Theorem 5.3.5. We will validate expression (5.13), which considers the asymptotic behaviour of $\mathbb{P}(\tau > x \mid M_\tau > (1-\rho)x)$, and expression (5.15), which considers the asymptotic behaviour of the unconditional probability $\mathbb{P}(\tau > x)$ as $\rho \uparrow 1$. Expressions (5.13) and (5.15) together imply expression (5.14) through the inequality

$$\left|\mathbb{P}(Q \mid R) - 1\right| \leq \left|\mathbb{P}(R \mid Q) - 1\right| \times \left|\frac{\mathbb{P}(Q)}{\mathbb{P}(R)}\right| + \left|\frac{\mathbb{P}(Q)}{\mathbb{P}(R)} - 1\right| \tag{5.36}$$

for two events Q and R of non-zero probability.

Section 5.7.1 validates the asymptotic behaviour of $\mathbb{P}(\tau > x \mid M_\tau > (1-\rho)x)$. Thereafter, Section 5.7.2 proves the asymptotic behaviour of $\mathbb{P}(\tau > x)$ by means of a sample-path analysis that makes a distinction based on the supremum M_τ. The resulting events are then distinguished based on the number of jumps before τ or the first passage time of a specific level.

5.7.1 Asymptotics of conditional first hitting time

We first prove expression (5.13). Since the relation $\mathbb{P}(\tau > x \mid M_\tau > (1-\rho)x) \leq 1$ is always true, we only need to show that the relation $\sup_{x\geq x^*_\rho} \mathbb{P}(\tau > x \mid M_\tau > (1-\rho)x) - 1 \geq 0$ holds as $\rho \uparrow 1$.

Fix $p \in (1/2 + 1/k^*, 1)$ and define $h_u(x,\rho) := (1-\rho)x + g(x,\rho)$, where g is described by $g(x,\rho) := (1-\rho)^{2p-1}x^p$. The function $h_u(x,\rho)$ is an upper bound for, yet asymptotically equivalent to, $(1-\rho)x$. By conditioning on the event $\{M_\tau > h_u(x,\rho)\}$, the long term drift $-(1-\rho)$ of $X(t)$ implies that $\mathbb{P}(\tau > x \mid M_\tau > h_u(x,\rho))$ must tend to one.

To make this precise we follow the proof of Proposition 3.1 of Zwart [144]. Noting that $\{\sigma(y) < \tau\} = \{M_\tau > y\}$, the joint probability $\mathbb{P}(\tau > x; M_\tau > (1-\rho)x)$ is lower bounded as

$$\begin{aligned}\mathbb{P}(\tau > x; M_\tau > (1-\rho)x) &\geq \mathbb{P}(\tau > x; M_\tau > h_u(x,\rho)) \\ &\geq \mathbb{P}(\tau - \sigma(h_u(x,\rho)) > x \mid \sigma(h_u(x,\rho)) < \tau)\mathbb{P}(M_\tau > h_u(x,\rho)),\end{aligned}$$

where the conditional probability on the right-hand side can be represented as an

integral:

$$\begin{aligned}
&\mathbb{P}(\tau - \sigma(h_u(x,\rho)) > x \mid \sigma(h_u(x,\rho)) < \tau) \\
&\quad = \int_{h_u(x,\rho)}^{\infty} \mathbb{P}(\tau - \sigma(h_u(x,\rho)) > x \mid \sigma(h_u(x,\rho)) < \tau; X(\sigma(h_u(x,\rho))) = y) \\
&\qquad \times \mathrm{d}\mathbb{P}(X(\sigma(h_u(x,\rho))) \le y \mid \sigma(h_u(x,\rho)) < \tau) \\
&\quad \ge \int_{h_u(x,\rho)}^{\infty} \mathbb{P}(X(t) > 0; 0 \le t \le x \mid X(0) = y)\, \mathrm{d}\mathbb{P}(X(\sigma(h_u(x,\rho))) \le y \mid \sigma(h_u(x,\rho)) < \tau).
\end{aligned}$$

As the integrand is increasing in y, we obtain

$$\mathbb{P}(\tau > x; M_\tau > (1-\rho)x) \ge \mathbb{P}(X(t) > 0; 0 \le t \le x \mid X(0) = h_u(x,\rho))\mathbb{P}(M_\tau > h_u(x,\rho)). \quad (5.37)$$

Rewriting the first probability on the right-hand side of (5.37) yields

$$\begin{aligned}
\mathbb{P}(X(t) > 0; 0 \le t \le x \mid X(0) = h_u(x,\rho)) &= \mathbb{P}\left(\inf_{t\in[0,x]} \{X(t) - X(0)\} > -h_u(x,\rho) \right) \\
&\ge \mathbb{P}\left(\inf_{t\in[0,x]} \left\{ -\rho t + \sum_{i=1}^{N(t)} B_i \right\} > -g(x,\rho) \right) \\
&= 1 - \mathbb{P}\left(\sup_{t\in[0,x]} \left\{ \rho t - \sum_{i=1}^{N(t)} B_i \right\} \ge g(x,\rho) \right).
\end{aligned}$$

From Etemadi's inequality for Lévy processes [120, Lemma A.4], it then follows that

$$\begin{aligned}
\mathbb{P}(X(t) > 0; 0 \le t \le x \mid X(0) = h_u(x,\rho)) &\ge 1 - 3 \sup_{t\in[0,x]} \mathbb{P}\left(\rho t - \sum_{i=1}^{N(t)} B_i \ge g(x,\rho)/3 \right) \\
&\ge 1 - 3 \sup_{t\in[0,x]} \mathbb{P}\left(\left| \rho t - \sum_{i=1}^{N(t)} B_i \right| \ge g(x,\rho)/3 \right).
\end{aligned}$$

The variance of $\sum_{i=1}^{N(t)} B_i$ equals $\lambda \mathbb{E}[B^2]\, t$ and is dominated by $2\mathbb{E}[B^*]\, t$ for all $\rho \in [0,1]$. Therefore, noting that $3 \cdot 3^2 \cdot 2 = 54$, Chebyshev's inequality implies

$$\mathbb{P}(X(t) > 0; 0 \le t \le x \mid X(0) = h_u(x,\rho)) \ge 1 - \sup_{t\in[0,x]} \frac{54\mathbb{E}[B^*]\, t}{g(x,\rho)^2} = 1 - \frac{54\mathbb{E}[B^*]}{((1-\rho)^2 x)^{2p-1}} \to 1 \quad (5.38)$$

for all $x \ge x_\rho^*$ as $\rho \uparrow 1$. Let $\zeta(\rho) := 1 - 54\mathbb{E}[B^*](\log \frac{1}{1-\rho})^{-(2p-1)k^*}$. By relations (5.37) and (5.38) and Theorem 5.3.2, one readily finds

$$\begin{aligned}
&\sup_{x \ge x_\rho^*} \mathbb{P}(\tau > x \mid M_\tau > (1-\rho)x) \\
&\qquad \ge \sup_{x \ge x_\rho^*} \mathbb{P}(X(t) > 0; 0 \le t \le x \mid X(0) = h_u(x,\rho)) \, \frac{\mathbb{P}(M_\tau > h_u(x,\rho))}{\mathbb{P}(M_\tau > (1-\rho)x)} \\
&\qquad \gtrsim \zeta(\rho) \sup_{x \ge x_\rho^*} \frac{\mathbb{P}(B > h_u(x,\rho))}{\mathbb{P}(B > (1-\rho)x)} \sim \zeta(\rho) \left(1 + ((1-\rho)^2 x_\rho^*)^{p-1} \right)^{-\alpha},
\end{aligned}$$

which tends to one as $\rho \uparrow 1$. This validates expression (5.13) in Theorem 5.3.5.

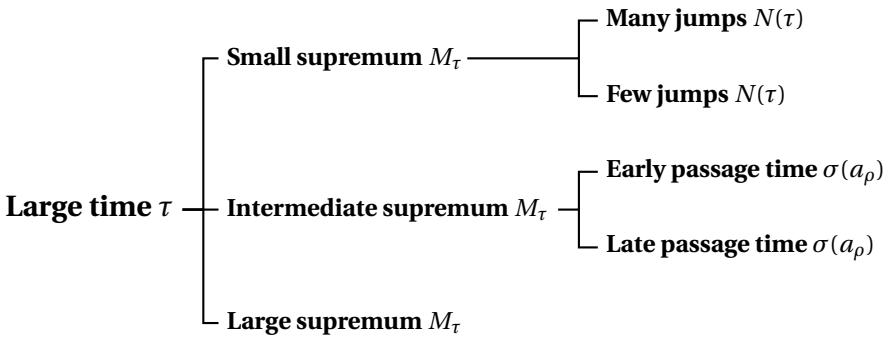


Figure 5.2: Visualization of proof structure. The event of a large time τ is analysed under three scenarios, depending on the size of the supremum M_τ. Two of these scenarios are again considered in more detail, where a distinction is based on the number of jumps before τ and the passage time of a high level a_ρ.

5.7.2 Asymptotics of unconditional first hitting time

This section validates expression (5.15). It follows from expression (5.13) that

$$\lim_{\rho\uparrow 1} \inf_{x\geq x_\rho^*} \frac{\mathbb{P}(\tau > x)}{\mathbb{P}(M_\tau > (1-\rho)x)} \geq 1. \tag{5.39}$$

Proving $\lim_{\rho\uparrow 1} \sup_{x\geq x_\rho^*} \frac{\mathbb{P}(\tau>x)}{\mathbb{P}(M_\tau>(1-\rho)x)} \leq 1$, however, requires far more work.

As noted at the beginning of this section, the event $\{\tau > x\}$ is analysed by distinguishing various scenarios. First, we specify scenarios $\{\tau > x, M_\tau \in \cdot\}$, where the supremum M_τ can be in three regions: small, intermediate and large. Then, the small and intermediate regions are shown to be negligible in Sections 5.7.2 through 5.7.2. Finally, the large M_τ region is shown to be asymptotically equivalent to $\mathbb{P}(\tau > x)$ in Section 5.7.2. The structure of the proof is visualised in Figure 5.2.

We now formalise the various scenarios. Fix $\varepsilon_\gamma > 0$ and $\varepsilon_\delta \in (0, (p-\frac{1}{2})k^* - 1)$, and define the functions $\gamma_\rho := \left(\log\frac{1}{1-\rho}\right)^{-\varepsilon_\gamma}$ and $\delta_\rho := \left(\log\frac{1}{1-\rho}\right)^{-(1+\varepsilon_\delta)}$. Similarly as before, the function $h_l(x,\rho) := (1-\gamma_\rho)(1-\rho)x - g(x,\rho)$, where $g(x,\rho) = (1-\rho)^{2p-1}x^p$, represents a lower bound for, yet is asymptotically equivalent to, $(1-\rho)x$. The three regions for M_τ are now given by

$$\begin{aligned}\mathbb{P}(\tau > x) &\leq \mathbb{P}(\tau > x; M_\tau \leq \delta_\rho(1-\rho)x) + \mathbb{P}(\tau > x; M_\tau \in (\delta_\rho(1-\rho)x, h_l(x,\rho)]) \\ &\quad + \mathbb{P}(M_\tau > h_l(x,\rho)) \\ &=: \widehat{\mathrm{I}}(x,\rho) + \widehat{\mathrm{II}}(x,\rho) + \widehat{\mathrm{III}}(x,\rho).\end{aligned}$$

The next paragraphs show that terms $\widehat{\mathrm{I}}$ and $\widehat{\mathrm{II}}$ both vanish faster than $\mathbb{P}(M_\tau > (1-\rho)x)$ for $x \geq x_\rho^*$ as $\rho \uparrow 1$. On the other hand, the final paragraph shows that term $\widehat{\mathrm{III}}$ asymptotically behaves as $\mathbb{P}(M_\tau > (1-\rho)x)$ in the same limiting regime.

Small supremum M_τ: term $\widehat{\mathrm{I}}$

Term $\widehat{\mathrm{I}}$ is the probability of a large first hitting time τ for which the corresponding process supremum M_τ is relatively small. First, we show that the number of jumps before τ is not much higher than the expected number of jumps. Then, we show that it is highly unlikely for a probable amount of small jumps to incur a large τ.

We let $\lambda^* := (1+\eta_\rho)\lambda$, where $\eta_\rho := (1-\rho)/2$, and note that $\lambda^*\mathbb{E}[B] \in [0,1)$ whenever $\rho \in [0,1)$. Also, we introduce the i.i.d. random variables $B_{0,i}$ characterised by their common c.d.f. $\mathbb{P}(B_{0,i} \leq y) := \mathbb{P}(B \leq y \mid B \leq \delta_\rho(1-\rho)x)$.

Recall that $N(t)$ denotes the number of jumps during an interval of length t. In particular, $N(t)$ is Poisson distributed with mean λt. Let $N_0(t)$ be the number of jumps of size at most $\delta_\rho(1-\rho)x$ and $N_1(t)$ be the number of jumps of at least that size. Then $N_0(t)$ is Poisson distributed with mean $\lambda t\mathbb{P}(B < \delta_\rho(1-\rho)x)$, $N_1(t)$ is Poisson distributed with mean $\lambda t\mathbb{P}(B \geq \delta_\rho(1-\rho)x)$ and $N(t) = N_0(t) + N_1(t)$ for all $t \geq 0$.

We observe that if $\tau > x$, then all jumps before time x had a cumulative size of at least x. That is, if $\tau > x$ then it must be that $\sum_{i=0}^{N(x)} B_i > x$. Furthermore, it is easy to see that $M_\tau \geq B_\tau$. From these two observations we derive

$$\begin{aligned}
\widehat{\mathrm{I}}(x,\rho) &= \mathbb{P}\left(\tau > x, \sum_{i=0}^{N(x)} B_i > x, M_\tau \leq \delta_\rho(1-\rho)x\right) \\
&\leq \mathbb{P}\left(\sum_{i=0}^{N(x)} B_i > x, \bigvee_{i=0}^{N(x)} B_i \leq \delta_\rho(1-\rho)x\right) = \mathbb{P}\left(\sum_{i=0}^{N_0(x)} B_{0,i} > x, N_1(x) = 0\right) \\
&\leq \mathbb{P}\left(\sum_{i=0}^{N_0(x)} B_{0,i} > x\right) \leq \mathbb{P}(N_0(x) \geq \lambda^* x) + \mathbb{P}\left(\sum_{i=0}^{\lambda^* x} B_{0,i} > x\right) \\
&= \mathbb{P}(N(x) \geq \lambda^* x) + \mathbb{P}\left(\sum_{i=0}^{\lambda^* x} B_i > x \,\middle|\, \bigvee_{i=0}^{\lambda^* x} B_i \leq \delta_\rho(1-\rho)x\right) =: \widehat{\mathrm{Ia}}(x,\rho) + \widehat{\mathrm{Ib}}(x,\rho).
\end{aligned}$$

Here, term $\widehat{\mathrm{Ia}}$ corresponds to a system where the number of jumps greatly exceeds its expectation. Term $\widehat{\mathrm{Ib}}$ indicates a likely number of jumps, none of which has a size exceeding $\delta_\rho(1-\rho)x$.

Many jumps: term $\widehat{\mathrm{Ia}}$

From Markov's inequality, one can see that for all $s \geq 0$ we have

$$\widehat{\mathrm{Ia}}(x,\rho) = \mathbb{P}(e^{sN(x)} \geq e^{s\lambda^* x}) \leq e^{-s\lambda^* x}\mathbb{E}\left[e^{sN(x)}\right] = \exp[-\lambda x((1+\eta_\rho)s - e^s + 1)].$$

Taking the infimum over all $s \geq 0$ gives

$$\widehat{\mathrm{Ia}}(x,\rho) \leq \exp[-\lambda x \sup_{s\geq 0}((1+\eta_\rho)s - e^s + 1)] = \exp\left[-\lambda x\left((1+\eta_\rho)\log(1+\eta_\rho) - \eta_\rho\right)\right].$$

The bound $\log(1+\eta_\rho) \geq \frac{2\eta_\rho}{2+\eta_\rho}$ for $\eta_\rho > 0$ then yields $\widehat{\mathrm{Ia}}(x,\rho) \leq \exp\left[-\frac{\eta_\rho^2}{2+\eta_\rho}\lambda x\right]$. Dividing by $\mathbb{P}(M_\tau > (1-\rho)x)$, taking the supremum, applying Theorem 5.3.2 and using Potter's Theorem with $\nu > 0$ gives

$$\begin{aligned}
\sup_{x\geq x_\rho^*} \frac{\widehat{\mathrm{Ia}}(x,\rho)}{\mathbb{P}(M_\tau > (1-\rho)x)} &\lesssim \sup_{x\geq x_\rho^*} C\frac{1-\rho}{\rho\mathbb{P}(B > (1-\rho)x)} \exp\left[-\frac{\eta_\rho^2}{2+\eta_\rho}\lambda x\right] \\
&\leq \sup_{x\geq x_\rho^*} C\frac{1-\rho}{\rho} \exp\left[(\alpha+\nu)\log((1-\rho)x) - \frac{\eta_\rho^2}{2+\eta_\rho}\lambda x\right] \\
&\leq C\frac{1-\rho}{\rho}\exp\left[(\alpha+\nu)\log\frac{1}{1-\rho} - \frac{\lambda}{10}\log^{k^*}\frac{1}{1-\rho} + o\left(\log\frac{1}{1-\rho}\right)\right] \\
&\to 0
\end{aligned} \tag{5.40}$$

as $\rho \uparrow 1$.

Few jumps: term $\widehat{\mathrm{Ib}}$

Now consider term $\widehat{\mathrm{Ib}}$. The corresponding event is a large τ, caused by a probable amount of small jumps. The following theorem by Prokhorov [113] is used to show that this scenario is unlikely as ρ tends to 1.

Theorem 5.7.1 (Prokhorov [113], Theorem 1)**.** *Suppose that $\xi_i, i = 1,\ldots,n$ are independent, zero-mean random variables such that there exists a constant c for which $|\xi_i| \leq c$ for $i = 1,\ldots,n$, and $\sum_{i=1}^n \mathbb{V}\mathrm{ar}\{\xi_i\} < \infty$. Then*

$$\mathbb{P}\left(\sum_{i=1}^n \xi_i > y\right) \leq \exp\left[-\frac{y}{2c}\operatorname{arcsinh}\frac{yc}{2\sum_{i=1}^n \mathbb{V}\mathrm{ar}\{\xi_i\}}\right]. \tag{5.41}$$

Using the bound $\operatorname{arcsinh}(z) = \log(z+\sqrt{1+z^2}) \geq \log(2z)$, Prokhorov's inequality implies

$$\mathbb{P}\left(\sum_{i=1}^n \xi_i > y\right) \leq \left(\frac{cy}{\sum_{i=1}^n \mathbb{V}\mathrm{ar}\{\xi_i\}}\right)^{-\frac{y}{2c}}. \tag{5.42}$$

Define $Y_i := B_i - \mathbb{E}[B]$. Then

$$\widehat{\mathrm{Ib}}(x,\rho) = \mathbb{P}\left(\sum_{i=0}^{\lambda^* x} Y_i > (1-\lambda^*\mathbb{E}[B])x - \mathbb{E}[B] \,\middle|\, \bigvee_{i=0}^{\lambda^* x} Y_i \leq \delta_\rho(1-\rho)x - \mathbb{E}[B]\right).$$

Let $\sigma_{B_0}^2$ be the variance of B provided $B < \delta_\rho(1-\rho)x$. Then $\sigma_{B_0}^2 \le \sigma_B^2$ and hence, using (5.42),

$$\begin{aligned}\widehat{\mathrm{Ib}}(x,\rho) &\le \left(\frac{\delta_\rho(1-\rho)x-\mathbb{E}[B]}{\lambda^* x+1}\,\frac{(1-\lambda^*\mathbb{E}[B])x-\mathbb{E}[B]}{\sigma_B^2}\right)^{-\frac{(1-\lambda^*\mathbb{E}[B])x-\mathbb{E}[B]}{2\delta_\rho(1-\rho)x-2\mathbb{E}[B]}}\\ &= \exp\left[-(1+\phi_\rho^{(1)}(x))\frac{1-\lambda^*\mathbb{E}[B]}{2\delta_\rho(1-\rho)}\log\left(\frac{1-\phi_\rho^{(2)}(x)}{\lambda^*\sigma_B^2}(1-\lambda^*\mathbb{E}[B])(1-\rho)\delta_\rho x\right)\right],\end{aligned}$$

where the real-valued functions $\phi_\rho^{(i)}(x)$ are defined as

$$\phi_\rho^{(1)}(x) := \frac{1-\frac{\mathbb{E}[B]}{(1-\lambda^*\mathbb{E}[B])x}}{1-\frac{\mathbb{E}[B]}{\delta_\rho(1-\rho)x}}-1, \qquad \phi_\rho^{(2)}(x) := 1-\frac{\left(1-\frac{\mathbb{E}[B]}{\delta_\rho(1-\rho)x}\right)\left(1-\frac{\mathbb{E}[B]}{(1-\lambda^*\mathbb{E}[B])x}\right)}{1+\frac{1}{\lambda^* x}},$$

and satisfy $\phi_\rho^{(i)}(x)\to 0$ as $\rho\uparrow 1$ for all $x\ge x_\rho^*$. Additionally, the functions $\phi_\rho^{(i)}$ are non-negative and non-increasing for ρ sufficiently close to one. These properties imply that the inequality

$$\widehat{\mathrm{Ib}}(x,\rho) \le \exp\left[-\frac{1-\lambda^*\mathbb{E}[B]}{2\delta_\rho(1-\rho)}\log\left(\frac{1-\phi_\rho^{(2)}(x)}{\lambda^*\sigma_B^2}(1-\lambda^*\mathbb{E}[B])(1-\rho)\delta_\rho x\right)\right]$$

holds for ρ sufficiently close to one and $x\ge x_\rho^*$. Substitution of $\lambda^* = (1+\eta_\rho)\lambda = \frac{3-\rho}{2}\lambda$ subsequently gives

$$\widehat{\mathrm{Ib}}(x,\rho) \le \exp\left[-\frac{1}{4\delta_\rho}\log\left(\frac{1-\phi_\rho^{(2)}(x)}{3\lambda\sigma_B^2}(1-\rho)^2\delta_\rho x\right)\right].$$

Dividing the upper bound above by $\mathbb{P}(M_\tau > (1-\rho)x) \sim \frac{\rho}{1-\rho}\mathbb{P}(B>(1-\rho)x)$ and applying Potter's Theorem with $\nu>0$ yields

$$\begin{aligned}\frac{\widehat{\mathrm{Ib}}(x,\rho)}{\mathbb{P}(M_\tau>(1-\rho)x)} &\lesssim C\frac{1-\rho}{\rho}\exp\left[(\alpha+\nu)\log((1-\rho)x)-\frac{1}{4\delta_\rho}\log\left(\frac{1-\phi_\rho^{(2)}(x)}{3\lambda\sigma_B^2}(1-\rho)^2\delta_\rho x\right)\right]\\ &= C\frac{1-\rho}{\rho}\exp\left[\left(\alpha+\nu-\frac{1}{4\delta_\rho}\right)\log((1-\rho)x)-\frac{1}{4\delta_\rho}\log\left(\frac{1-\phi_\rho^{(2)}(x)}{3\lambda\sigma_B^2}(1-\rho)\delta_\rho\right)\right].\end{aligned}$$

The supremum over $x\ge x_\rho^*$ is attained in $x=x_\rho^*$ for ρ sufficiently close to one. That is,

$$\begin{aligned}\sup_{x\ge x_\rho^*}\frac{\widehat{\mathrm{Ib}}(x,\rho)}{\mathbb{P}(M_\tau>(1-\rho)x)} &\lesssim C\frac{1-\rho}{\rho}\exp\left[\left(\alpha+\nu-\frac{1}{4\delta_\rho}\right)\log\left(\frac{1}{1-\rho}\log^{k^*}\frac{1}{1-\rho}\right)\right.\\ &\qquad\left.-\frac{1}{4\delta_\rho}\log\left(\frac{1-\phi_\rho^{(2)}(x_\rho^*)}{3\lambda\sigma_B^2}(1-\rho)\delta_\rho\right)\right]\\ &= C\frac{1-\rho}{\rho}\exp\left[(\alpha+\nu)\log\left(\frac{1}{1-\rho}\log^{k^*}\frac{1}{1-\rho}\right)\right.\end{aligned}$$

$$-\frac{1}{4}\log^{1+\varepsilon_\delta}\left(\frac{1}{1-\rho}\right)\log\left(\frac{1-\phi_\rho^{(2)}(x_\rho^*)}{3\lambda\sigma_B^2}\log^{k^*-1-\varepsilon_\delta}\frac{1}{1-\rho}\right)\Bigg]$$
$$\to 0 \tag{5.43}$$

as $\rho\uparrow 1$. Together, (5.40) and (5.43) assure that term $\widehat{\mathrm{I}}$ is dominated by $\mathbb{P}(M_\tau>(1-\rho)x)$.

Intermediate supremum M_τ: term $\widehat{\mathrm{II}}$

Term $\widehat{\mathrm{II}}$ corresponds to the event of a large τ that experience an intermediate supremum M_τ. Write

$$\begin{aligned}\sup_{x\geq x_\rho^*}\frac{\widehat{\mathrm{II}}(x,\rho)}{\mathbb{P}(M_\tau>(1-\rho)x)} &\leq \sup_{x\geq x_\rho^*}\frac{\mathbb{P}(M_\tau>\delta_\rho(1-\rho)x)}{\mathbb{P}(M_\tau>(1-\rho)x)}\\ &\quad\times\sup_{x\geq x_\rho^*}\mathbb{P}(\tau>x;M_\tau\leq h_l(x,\rho)\mid M_\tau>\delta_\rho(1-\rho)x)\\ &\sim\delta_\rho^{-\alpha}\sup_{x\geq x_\rho^*}\mathbb{P}(\tau>x;M_\tau\leq h_l(x,\rho)\mid M_\tau>\delta_\rho(1-\rho)x).\end{aligned}\tag{5.44}$$

Set $\kappa_\rho:=\left(\log\frac{1}{1-\rho}\right)^{-\varepsilon_\kappa}$ for some $\varepsilon_\kappa\geq\varepsilon_\gamma$, implying $\gamma_\rho-\kappa_\rho>0$. By considering the time $\sigma(a_\rho)$ when the process $X(t)$ first exceeds level $a_\rho:=\delta_\rho(1-\rho)x$, we can partition (5.44) into two events:

$$\begin{aligned}&\mathbb{P}(\tau>x;M_\tau\leq h_l(x,\rho)\mid M_\tau>\delta_\rho(1-\rho)x)\\ &\quad=\mathbb{P}(\tau>x;\sigma(a_\rho)\leq\kappa_\rho x;M_\tau\leq h_l(x,\rho)\mid\sigma(a_\rho)<\tau)\\ &\qquad+\mathbb{P}(\tau>x;\sigma(a_\rho)>\kappa_\rho x;M_\tau\leq h_l(x,\rho)\mid\sigma(a_\rho)<\tau)\\ &\quad\leq\mathbb{P}(\tau>(1-\kappa_\rho)x;M_\tau\leq h_l(x,\rho)\mid\sigma(a_\rho)<\tau)+\mathbb{P}(\sigma(a_\rho)>\kappa_\rho x\mid\sigma(a_\rho)<\tau)\\ &\quad=:\widehat{\mathrm{IIa}}(x,\rho)+\widehat{\mathrm{IIb}}(x,\rho).\end{aligned}$$

Term $\widehat{\mathrm{IIa}}$ is associated with sample paths that experiences an intermediate supremum and that may already hit zero after time $(1-\kappa_\rho)x$. Term $\widehat{\mathrm{IIb}}$ corresponds to a sample path where the process does not exceed level a_ρ before time $\kappa_\rho x$, provided that it will hit level a_ρ before it hits zero.

Early passage time: term $\widehat{\mathrm{IIa}}$

Term $\widehat{\mathrm{IIa}}$ is analysed along the lines of Section 5.7.1:

$$\begin{aligned}\widehat{\mathrm{IIa}}(x,\rho)&=\int_0^{h_l(x,\rho)}\mathbb{P}(X(t)>0;0\leq t\leq(1-\kappa_\rho)x;M_\tau\leq h_l(x,\rho)\mid\sigma(a_\rho)<\tau;X(0)=y)\\ &\qquad\times\mathrm{d}\mathbb{P}(X(0)\leq y\mid\sigma(a_\rho)<\tau)\\ &\leq\int_0^{h_l(x,\rho)}\mathbb{P}(X(t)>0;0\leq t\leq(1-\kappa_\rho)x\mid\sigma(a_\rho)<\tau;X(0)=y)\,\mathrm{d}\mathbb{P}(X(0)\leq y\mid\sigma(a_\rho)<\tau)\end{aligned}$$

$$\leq \mathbb{P}(X(t) > 0; 0 \leq t \leq (1-\kappa_\rho)x \mid X(0) = h_l(x,\rho))$$
$$\leq \mathbb{P}(X((1-\kappa_\rho)x) - X(0) > -h_l(x,\rho)).$$

Here, the second inequality holds as the integrand is increasing in y, and $a_\rho \leq h_l(x,\rho)$ for $x \geq x_\rho^*$ and ρ sufficiently close to one.

Define $A_0^\rho := 0$ and $A_i^\rho := \inf\{t \geq 0 : N(\sum_{j=0}^{i-1} A_j^\rho + t) \geq i\}$ for all $i \geq 1$. Then the A_i^ρ are i.i.d. exponentially distributed random variables with mean $1/\lambda$ and $\sum_{i=1}^{N(t)} A_i^\rho \leq t$ for all $t \geq 0$. We drop the superscript ρ for notational convenience. Now,

$$\begin{aligned}
\widehat{\mathrm{IIa}}(x,\rho) &\leq \mathbb{P}\left(-(1-\kappa_\rho)x + \sum_{i=1}^{N((1-\kappa_\rho)x)} B_i > -h_l(x,\rho)\right) \\
&= \mathbb{P}\left(-\rho(1-\kappa_\rho)x + \sum_{i=1}^{N((1-\kappa_\rho)x)} B_i > (\gamma_\rho - \kappa_\rho)(1-\rho)x + g(x,\rho)\right) \\
&\leq \mathbb{P}\left(\sum_{i=1}^{N((1-\kappa_\rho)x)} [B_i - \rho A_i] > (\gamma_\rho - \kappa_\rho)(1-\rho)x + g(x,\rho)\right).
\end{aligned}$$

Fix $q \in \left(\max\left\{2, \frac{(1+\varepsilon_\delta)\alpha}{(p-\frac{1}{2})k^*}\right\}, \alpha\right)$. By Chebyshev's inequality for general moments and Theorem 5.1 in Chapter 1 of Gut [68], there exists some constant C_q such that

$$\begin{aligned}
\widehat{\mathrm{IIa}}(x,\rho) &\leq \frac{\mathbb{E}\left[\left(\sum_{i=1}^{N((1-\kappa_\rho)x)} [B_i - \rho A_i]\right)^q\right]}{\left((\gamma_\rho - \kappa_\rho)(1-\rho)x + g(x,\rho)\right)^q} \leq \frac{C_q \mathbb{E}[|B_1 - \rho A_1|^q]\mathbb{E}[N((1-\kappa_\rho)x)^{q/2}]}{g(x,\rho)^q} \\
&\leq \frac{C_q \mathbb{E}[|B_1 - \rho A_1|^q]\mathbb{E}[N((1-\kappa_\rho)x)]^{q/2}}{g(x,\rho)^q},
\end{aligned}$$

where the last derivation is justified by Hölder's inequality. Subsequently, one may show from Jensen's inequality that $\mathbb{E}[|B-A|^q] \leq 2^{q-1}(\mathbb{E}[|A|^q] + \mathbb{E}[|B|^q])$ and therefore

$$\widehat{\mathrm{IIa}}(x,\rho) \leq \frac{C_q 2^{q-1}(\mathbb{E}[B_1^q] + \mathbb{E}[A_1^q])(\lambda(1-\kappa_\rho)x)^{q/2}}{(1-\rho)^{(2p-1)q} x^{pq}} \leq C\left((1-\rho)^2 x\right)^{-(p-\frac{1}{2})q}$$

for some constant C. By choice of p, q and ε_δ, we conclude that

$$\delta_\rho^{-\alpha} \sup_{x \geq x_\rho^*} \widehat{\mathrm{IIa}}(x,\rho) \leq C\left(\log\frac{1}{1-\rho}\right)^{(1+\varepsilon_\delta)\alpha - (p-\frac{1}{2})k^* q} \to 0 \tag{5.45}$$

as $\rho \uparrow 1$.

Late passage time: term $\widehat{\mathrm{IIb}}$

Term $\widehat{\mathrm{IIb}}$ is analysed with the following crucial lemma, and is proven in Section 5.8:

Lemma 5.7.2. *Suppose* $\mathbb{P}(B > x) = L(x)x^{-\alpha}$ *for some* $\alpha > 2$, $\alpha \neq 3$, *and* $L(x)$ *slowly varying. Define* $a_\rho^* := k^*\mu(\alpha-1)\frac{1}{1-\rho}\log\frac{1}{1-\rho}$ *for some* $k^* > 2$. *Then for any fixed* $y > 0$,

$$\sup_{a \geq a_\rho^*} \mathbb{E}[\sigma(a) \mid \sigma(a) < \tau; X_0 = y] = O\left(\frac{1}{1-\rho}\right) \tag{5.46}$$

as $\rho \uparrow 1$. *Similarly, without conditioning on the value of* X_0,

$$\sup_{a \geq a_\rho^*} \mathbb{E}[\sigma(a) \mid \sigma(a) < \tau] = O\left(\frac{1}{1-\rho}\right) \tag{5.47}$$

as $\rho \uparrow 1$.

Applying Markov's inequality and sequentially Lemma 5.7.2 to term $\widehat{\text{IIb}}$ yields

$$\begin{aligned}\delta_\rho^{-\alpha} \sup_{x \geq x_\rho^*} \widehat{\text{IIb}}(x,\rho) &\leq \sup_{x \geq x_\rho^*} \frac{\mathbb{E}[\sigma(\delta_\rho(1-\rho)x) \mid \sigma(\delta_\rho(1-\rho)x) < \tau]}{\delta_\rho^\alpha \kappa_\rho x} \\ &= O\left(\frac{1}{1-\rho}\right) \frac{(1-\rho)^2}{\left(\log\frac{1}{1-\rho}\right)^{k^*-(1+\varepsilon_\delta)\alpha-\varepsilon_\kappa}} \to 0\end{aligned} \tag{5.48}$$

as $\rho \uparrow 1$, thereby immediately completing the analysis of term $\widehat{\text{IIb}}$.

Large supremum M_τ: term $\widehat{\text{III}}$

Finally, we show that the probability of a large time τ is asymptotically equivalent to term $\widehat{\text{III}}$. Using Theorem 5.3.2, it directly follows that

$$\begin{aligned}\sup_{x \geq x_\rho^*} \frac{\widehat{\text{III}}(x,\rho)}{\mathbb{P}(M_\tau > (1-\rho)x)} &\lesssim \sup_{x \geq x_\rho^*} \frac{\mathbb{P}(B > h_l(x,\rho))}{\mathbb{P}(B > (1-\rho)x)} \\ &\sim \sup_{x \geq x_\rho^*} \left(\frac{(1-\gamma_\rho)(1-\rho)x - (1-\rho)^{2p-1}x^p}{(1-\rho)x}\right)^{-\alpha} \\ &= \left(1 - \left(\log\frac{1}{1-\rho}\right)^{-\varepsilon_\gamma} - \left(\log\frac{1}{1-\rho}\right)^{-(1-p)k^*}\right)^{-\alpha} \to 1\end{aligned}$$

as $\rho \uparrow 1$.

The proof of expression (5.15) and consequently the proof of Theorem 5.3.5 is completed once we validate Lemma 5.7.2, which is the subject of the next section.

5.8 Asymptotics of the conditional expectation of the passage time of level a

This section is dedicated to the proof of Lemma 5.7.2, which regards the expected first passage time of level a, $\sigma(a)$, provided that level a is reached before level 0: $\sigma(a) < \tau$. In particular, we consider high levels $a \geq a_\rho^* := k^*\mu(\alpha-1)\frac{1}{1-\rho}\log\frac{1}{1-\rho}$ for any $k^* > 2$.

The lemma considers two different scenarios. In the first scenario, we condition on the initial value $X(0) = y$. In the second scenario, the initial value $X(0)$ is a random variable with the same distribution as a general jump size B. The analysis for this latter scenario is based on the following decomposition:

$$\mathbb{E}[\sigma(a) \mid \sigma(a) < \tau] = \int_0^a \mathbb{E}[\sigma(a) \mid \sigma(a) < \tau; X(0) = y] \, \mathrm{d}\mathbb{P}(B \le y). \tag{5.49}$$

When analysing the integral in expression (5.49), a distinction is made between a "small" and a "large" random initial value; a precise definition of which is given at the end of these introductory paragraphs. The analysis of the first scenario of the lemma, where the initial value is fixed, is implicit in the analysis of a small random initial value. The proof of the first scenario is concluded at the end of Section 5.8.1.

The derivation of results in this section relies heavily on the theory of spectrally one-sided Lévy processes and q-scale functions, e.g. as documented by Kyprianou [90]. Our interest in q-scale functions $W_\rho^{(q)}$ originates from the close connection between the all-time supremum M_∞ and the 0-scale function $W_\rho(x) := W_\rho^{(0)}(x)$. Of particular importance is the relation

$$\mathbb{P}(M_\infty < x) = (1-\rho) W_\rho(x), \tag{5.50}$$

which can be derived from Corollary IX.3.4 in Asmussen [8] (e.g. as shown in [21]).

Define the Laplace exponent $\psi(\lambda) := \frac{1}{t} \log \mathbb{E}(e^{-\lambda X(t)})$ of $X(t)$ and its right-inverse $\varphi(q) := \sup\{\lambda \ge 0 : \psi(\lambda) = q\}$. Then, for every $q \ge 0$, the q-scale function $W_\rho^{(q)}(x) : \mathbb{R} \to [0, \infty)$ corresponding to the spectrally positive Lévy process $X(t)$ is defined on $x < 0$ as $W_\rho^{(q)}(x) = 0$, and on $x \ge 0$ by its Laplace transform:

$$\int_0^\infty e^{-\beta x} W_\rho^{(q)}(x) \, \mathrm{d}x = \frac{1}{\psi(\beta) - q} \quad \text{for } \beta > \varphi(q). \tag{5.51}$$

Additionally, Kyprianou gives a representation of $W_\rho^{(q)}(x)$ in terms of $W_\rho(x)$ in his relation (8.29):

$$W_\rho^{(q)}(x) = \sum_{k \ge 0} q^k W_\rho^{(k+1)\circledast}(x), \tag{5.52}$$

where the function $f^{1\circledast}(x)$ is identical to $f(x)$ and $f^{k\circledast}(x) := \int_0^x f^{(k-1)\circledast}(x-y) f(y) \, \mathrm{d}y$ denotes the k-fold convolution of f with itself.

An alternative representation of $W_\rho(x)$ is provided by expression (8.22) in Kyprianou [90], stating that there are a measure $n_\rho(\cdot)$ on the space of excursions of $X(t)$ from its previous minimum $\min\{X(s) : 0 \le s \le t\}$ and a random variable $\overline{\xi}_\rho$ associated with the height of an excursion, such that for all $b > x \ge 0$ we have

$$W_\rho(x) = W_\rho(b) \exp\left(- \int_x^b n_\rho(\overline{\xi}_\rho > t) \, \mathrm{d}t \right). \tag{5.53}$$

This representation will provide a useful property for the all-time supremum p.d.f. $f_{M_\infty}(x) := \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(M_\infty < y)\Big|_{y=x}$. Using the Pollaczek-Khintchine formula (cf. relation (5.20)), we write

$$f_{M_\infty}(x) = \sum_{n=1}^{\infty}(1-\rho)\rho^n \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(B_1^* + \ldots + B_n^* < y)\Big|_{y=x}$$

for $x > 0$. One may show by induction that $\frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(B_1^* + \ldots + B_n^* < y)$ is defined everywhere and is bounded by $1/\mathbb{E}[B]$ for all $n \geq 1$. As such, $f_{M_\infty}(x)$ is properly defined and bounded for all $x > 0$. Additionally, (5.53) implies that

$$\frac{f_{M_\infty}(x)}{\mathbb{P}(M_\infty < x)} = \frac{\mathrm{d}}{\mathrm{d}y}\log W_\rho(y)\Big|_{y=x} = n_\rho(\bar{\xi}_\rho > x) \tag{5.54}$$

is non-increasing in x.

For the remainder of this section, the subscripts ρ for $W_\rho(x)$ and $W_\rho^{(q)}(x)$ are discarded. We also introduce the short-hand notations $\mathbb{E}_y[\cdot]$ and $\mathbb{P}_y(\cdot)$ for the conditional expectation $\mathbb{E}[\cdot \mid X(0) = y]$ and conditional probability $\mathbb{P}(\cdot \mid X(0) = y)$, respectively.

Let $Z^{(q)}(x) := 1 + q\int_0^x W^{(q)}(y)\,\mathrm{d}y$. From (5.52) and the spectrally *positive* Lévy process interpretation of Theorem 8.1 in Kyprianou [90], it follows that the unconditional expectation $\mathbb{E}_y[\sigma(a)\mathbb{1}(\sigma(a) < \tau)]$ satisfies

$$\begin{aligned}
\mathbb{E}_y[\sigma(a)\mathbb{1}(\sigma(a) < \tau)] &= -\frac{\mathrm{d}}{\mathrm{d}q}\mathbb{E}_y[e^{-q\sigma(a)}\mathbb{1}(\sigma(a) < \tau)]\Big|_{q=0} \\
&= -\frac{\mathrm{d}}{\mathrm{d}q}Z^{(q)}(a-y) + \frac{W^{(q)}(a-y)}{W^{(q)}(a)}\frac{\mathrm{d}}{\mathrm{d}q}Z^{(q)}(a) \\
&\quad + Z^{(q)}(a)\frac{W^{(q)}(a)\frac{\mathrm{d}}{\mathrm{d}q}W^{(q)}(a-y) - W^{(q)}(a-y)\frac{\mathrm{d}}{\mathrm{d}q}W^{(q)}(a)}{(W^{(q)}(a))^2}\Big|_{q=0} \\
&= -\int_0^{a-y} W(t)\,\mathrm{d}t + \frac{W(a-y)}{W(a)}\int_0^a W(t)\,\mathrm{d}t + \frac{W(a)W^{2\circledast}(a-y) - W(a-y)W^{2\circledast}(a)}{(W(a))^2} \\
&= \frac{W(a-y)}{W(a)}\frac{\int_0^a (W(a) - W(a-t))W(t)\,\mathrm{d}t}{W(a)} - \frac{\int_0^{a-y}(W(a) - W(a-y-t))W(t)\,\mathrm{d}t}{W(a)}.
\end{aligned}$$

Now, from relation (8.12) in Kyprianou [90] one may deduce $\mathbb{P}_y(\sigma(a) < \tau) = \frac{W(a)-W(a-y)}{W(a)}$, which gives a representation of the conditional expectation $\mathbb{E}_y[\sigma(a) \mid \sigma(a) < \tau]$ in terms of the scale function:

$$\begin{aligned}
\mathbb{E}_y[\sigma(a) \mid \sigma(a) < \tau] = \frac{W(a-y)}{W(a)}&\frac{\int_0^a (W(a) - W(a-t))W(t)\,\mathrm{d}t}{W(a) - W(a-y)} \\
&- \frac{\int_0^{a-y}(W(a) - W(a-y-t))W(t)\,\mathrm{d}t}{W(a) - W(a-y)}.
\end{aligned}$$

Substitute (5.50) into the above expression to obtain

$$\begin{aligned}\mathbb{E}_y[\sigma(a) \mid \sigma(a) < \tau] &= \frac{\mathbb{P}(M_\infty < a-y)}{\mathbb{P}(M_\infty < a)} \frac{\int_0^a \mathbb{P}(M_\infty \in [a-t,a))\mathbb{P}(M_\infty < t)\,\mathrm{d}t}{(1-\rho)\mathbb{P}(M_\infty \in [a-y,a))} \\ &\quad - \frac{\int_0^{a-y} \mathbb{P}(M_\infty \in [a-y-t,a))\mathbb{P}(M_\infty < t)\,\mathrm{d}t}{(1-\rho)\mathbb{P}(M_\infty \in [a-y,a))} \\ &\leq \frac{\int_0^a \mathbb{P}(M_\infty \in [a-t,a))\mathbb{P}(M_\infty < t)\,\mathrm{d}t - \int_0^{a-y} \mathbb{P}(M_\infty \in [a-y-t,a))\mathbb{P}(M_\infty < t)\,\mathrm{d}t}{(1-\rho)\mathbb{P}(M_\infty \in [a-y,a))} \\ &=: \frac{K_{num}(y,a)}{K_{denom}(y,a)}. \end{aligned} \tag{5.55}$$

The analysis of this expression depends on the initial value y. We distinguish two categories of initial values: small and large values. Fix d such that $0 < d < 1 - \frac{2}{k^*} < 1$. Small values are of size at most $d \cdot a$, all other values are large values.

5.8.1 Small random initial value or fixed initial value

This section considers the process from a small initial value y, i.e. $y \leq da$. For any y-differentiable function $G(y,a)$, it is known that $G(y,a) = G(0,a) + \int_0^y \frac{\mathrm{d}}{\mathrm{d}s} G(s,a)\big|_{s=z}\,\mathrm{d}z$. This is now used to obtain an alternative representation of $K_{num}(y,a)$.

Let $M_\infty^{(i)}, i = 1,2$ be independent copies of M_∞. Taking the derivative of $K_{num}(s,a)$ with respect to s yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s} K_{num}(s,a) &= \mathbb{P}(M_\infty^{(2)} < a)\mathbb{P}(M_\infty^{(1)} < a-s) - \int_0^{a-s} \mathbb{P}(M_\infty^{(1)} < t)\,\mathrm{d}\mathbb{P}(M_\infty^{(2)} < a-s-t) \\ &= \mathbb{P}(M_\infty^{(2)} < a)\mathbb{P}(M_\infty^{(1)} < a-s) - \mathbb{P}(M_\infty^{(1)} + M_\infty^{(2)} < a-s) \\ &= \mathbb{P}(M_\infty^{(1)} < a-s) - \mathbb{P}(M_\infty^{(1)} + M_\infty^{(2)} < a-s) - \mathbb{P}(M_\infty^{(2)} \geq a)\mathbb{P}(M_\infty^{(1)} < a-s) \\ &= \mathbb{P}(M_\infty^{(1)} + M_\infty^{(2)} \geq a-s; M_\infty^{(1)} < a-s) \\ &\quad - \mathbb{P}(M_\infty^{(2)} \geq a-s)\mathbb{P}(M_\infty^{(1)} < a-s) + \mathbb{P}(M_\infty^{(2)} \in [a-s,a))\mathbb{P}(M_\infty^{(1)} < a-s) \\ &= \mathbb{P}(M_\infty^{(1)} + M_\infty^{(2)} \geq a-s; M_\infty^{(1)} < a-s; M_\infty^{(2)} < a-s) \\ &\quad + \mathbb{P}(M_\infty^{(2)} \in [a-s,a))\mathbb{P}(M_\infty^{(1)} < a-s), \end{aligned}$$

so that $K_{num}(0,a) = 0$ implies

$$\begin{aligned}\mathbb{E}_y[\sigma(a) \mid \sigma(a) < \tau] &\leq \frac{K_{num}(y,a)}{K_{denom}(y,a)} \\ &= \frac{\int_0^y \mathbb{P}(M_\infty^{(1)} + M_\infty^{(2)} \geq a-z; M_\infty^{(1)} < a-z; M_\infty^{(2)} < a-z)\,\mathrm{d}z}{(1-\rho)\mathbb{P}(M_\infty \in [a-y,a))} \\ &\quad + \frac{\int_0^y \mathbb{P}(M_\infty^{(2)} \in [a-z,a))\mathbb{P}(M_\infty^{(1)} < a-z)\,\mathrm{d}z}{(1-\rho)\mathbb{P}(M_\infty \in [a-y,a))} \\ &\leq \frac{\int_0^y \mathbb{P}(M_\infty^{(1)} + M_\infty^{(2)} \geq a-z; M_\infty^{(1)} < a-z; M_\infty^{(2)} < a-z)\,\mathrm{d}z}{(1-\rho)\mathbb{P}(M_\infty \in [a-y,a))} + \frac{y}{1-\rho}. \end{aligned}$$

By symmetry, we have

$$\begin{aligned}\mathbb{P}(M_\infty^{(1)} + M_\infty^{(2)} \geq u; M_\infty^{(1)} < u; M_\infty^{(2)} < u) &\leq 2\mathbb{P}(M_\infty^{(1)} + M_\infty^{(2)} \geq u; u/2 \leq M_\infty^{(1)} < u) \\ &\leq 2\mathbb{P}(M_\infty \in [u/2, u))\end{aligned}$$

and hence

$$\begin{aligned}\mathbb{E}_y[\sigma(a) \mid \sigma(a) < \tau] &\leq \frac{2\int_0^y \mathbb{P}\left(M_\infty \in \left[\frac{a-z}{2}, a-z\right)\right) \mathrm{d}z}{(1-\rho)\mathbb{P}(M_\infty \in [a-y, a))} + \frac{y}{1-\rho} \\ &\leq \frac{2y}{1-\rho}\left(1 + \frac{\mathbb{P}\left(M_\infty \in \left[\frac{a-y}{2}, a\right)\right)}{\mathbb{P}(M_\infty \in [a-y, a))}\right). \end{aligned} \tag{5.56}$$

Both local probabilities can be represented as a sum of local probabilities over an interval with fixed length. Subsequently, Theorem 5.3.1 is applied to bound the above ratio. Fix $y_{min} > 0$ and first consider (5.56) for $y_{min} \leq y \leq da$. For $S := y_{min}/2$, we have

$$\frac{\mathbb{P}\left(M_\infty \in \left[\frac{a-y}{2}, \frac{a}{2}\right)\right)}{\mathbb{P}\left(M_\infty \in \left[a-y, a\right)\right)} \leq \frac{\sum_{i=0}^{\lceil \frac{y}{2S}-1\rceil} \mathbb{P}\left(M_\infty \in \left[\frac{a-y}{2} + iS, \frac{a-y}{2} + (i+1)S\right)\right)}{\sum_{i=0}^{\lfloor \frac{y}{S}-1\rfloor} \mathbb{P}\left(M_\infty \in \left[a-y+iS, a-y+(i+1)S\right)\right)}.$$

We would now like to utilise Theorem 5.3.1. To this end, consider x_ρ as defined by Theorem 5.3.1 with parameter $(1-d)k^*/2 > 1$. Then for all $y \leq da$, we have $\frac{a-y}{2} \geq \frac{1-d}{2}a_\rho^* = x_\rho$. Hence, we observe that there exists a non-increasing function $\phi_\rho(\cdot) \downarrow 0$ for which the inequalities

$$1 - \phi_\rho\left(\frac{a-y}{2}\right) \leq \frac{\mathbb{P}\left(M_\infty \in \left[\frac{a-y}{2} + iS, \frac{a-y}{2} + (i+1)S\right)\right)}{\frac{\rho}{1-\rho}\mathbb{P}\left(B^* \in \left[\frac{a-y}{2} + iS, \frac{a-y}{2} + (i+1)S\right)\right)} \leq 1 + \phi_\rho\left(\frac{a-y}{2}\right) \tag{5.57}$$

both hold for all $y \leq da$ and $i \geq 0$. From $a - y \geq a - da \geq \frac{a}{k^*}$ one may subsequently conclude that the ratio of interest is bounded:

$$\begin{aligned}\frac{\mathbb{P}\left(M_\infty \in \left[\frac{a-y}{2}, \frac{a}{2}\right)\right)}{\mathbb{P}\left(M_\infty \in \left[a-y, a\right)\right)} &\leq \frac{1+\phi_\rho\left(\frac{a}{2k^*}\right)}{1-\phi_\rho\left(\frac{a}{k^*}\right)} \frac{\sum_{i=0}^{\lceil \frac{y}{2S}-1\rceil} \mathbb{P}\left(B^* \in \left[\frac{a-y}{2} + iS, \frac{a-y}{2} + (i+1)S\right)\right)}{\sum_{i=0}^{\lfloor \frac{y}{S}-1\rfloor} \mathbb{P}\left(B^* \in \left[a-y+iS, a-y+(i+1)S\right)\right)} \\ &\leq \frac{1+\phi_\rho\left(\frac{a}{2k^*}\right)}{1-\phi_\rho\left(\frac{a}{k^*}\right)} \frac{\frac{y}{2S}+1}{\frac{y}{S}-1} \frac{\mathbb{P}\left(B > \frac{a-y}{2}\right)}{\mathbb{P}\left(B > a\right)} \sim \frac{1+\frac{2S}{y}}{2-\frac{2S}{y}}\left(\frac{a-y}{2a}\right)^{-\alpha} \leq 2(2k^*)^\alpha.\end{aligned}$$

Second, consider (5.56) for $0 < y < y_{min}$. Relation (5.54) implies

$$\begin{aligned}\frac{\mathbb{P}\left(M_\infty \in \left[\frac{a-y}{2}, \frac{a}{2}\right)\right)}{\mathbb{P}\left(M_\infty \in \left[a-y, a\right)\right)} &\leq \frac{y \sup_{z\in(0,y)} \frac{f_{M_\infty}\left(\frac{a-z}{2}\right)}{\mathbb{P}\left(M_\infty < \frac{a-z}{2}\right)} \mathbb{P}\left(M_\infty < \frac{a-z}{2}\right)}{2y \inf_{z\in(0,y)} \frac{f_{M_\infty}(a-z)}{\mathbb{P}(M_\infty < a-z)} \mathbb{P}(M_\infty < a-z)} \leq \frac{\frac{f_{M_\infty}\left(\frac{a-y}{2}\right)}{\mathbb{P}\left(M_\infty < \frac{a-y}{2}\right)} \mathbb{P}\left(M_\infty < \frac{a}{2}\right)}{2\frac{f_{M_\infty}(a)}{\mathbb{P}(M_\infty < a)} \mathbb{P}(M_\infty < a-y)} \\ &= \frac{f_{M_\infty}\left(\frac{a-y}{2}\right)}{2f_{M_\infty}(a)} \frac{\mathbb{P}\left(M_\infty < \frac{a}{2}\right)\mathbb{P}(M_\infty < a)}{\mathbb{P}\left(M_\infty < \frac{a-y}{2}\right)\mathbb{P}(M_\infty < a-y)} \sim \frac{f_{M_\infty}\left(\frac{a-y}{2}\right)}{2f_{M_\infty}(a)}\end{aligned}$$

as $a \to \infty$. We conclude that

$$\mathbb{E}_y[\sigma(a) \mid \sigma(a) < \tau] \lesssim C\frac{y}{1-\rho}\left(1 + \frac{f_{M_\infty}\left(\frac{a-y}{2}\right)}{f_{M_\infty}(a)}\mathbb{1}(y \le y_{min})\right). \tag{5.58}$$

The above relation explicitly shows the dependence of the asymptotic upper bound on y. This dependence is crucial in the analysis of the second part of the lemma, where we will integrate the upper bound over $\mathbb{P}(B < y)$. However, before addressing large initial values it should be noted that (5.58) also proves the first part of the lemma. There, y is fixed and the lemma follows directly after choosing $0 < y_{min} < y$.

5.8.2 Large random initial value

Complementary to the previous section, we now consider (5.55) for large initial values, i.e. $d\,a \le y < a$.

Let M_∞^* be a random variable with the excess distribution of M_∞ as its c.d.f. , that is, $\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(M_\infty^* < x) = \mathbb{P}(M_\infty \ge t)/\mathbb{E}[M_\infty]$. Using the equalities $\mathbb{P}(M_\infty < t) = 1 - \mathbb{P}(M_\infty \ge t)$ and $\int_0^a \mathbb{P}(M_\infty \in [a-t, a))\,\mathrm{d}t = \mathbb{E}[M_\infty \mathbb{1}(M_\infty < a)]$, we find

$$\begin{aligned}
K_{num}(y,a) &= \int_0^a \mathbb{P}(M_\infty \in [a-t,a))\mathbb{P}(M_\infty < t)\,\mathrm{d}t \\
&\quad - \int_0^{a-y} \mathbb{P}(M_\infty \in [a-y-t,a))\mathbb{P}(M_\infty < t)\,\mathrm{d}t \\
&= \mathbb{E}[M_\infty \mathbb{1}(M_\infty < a)] - \mathbb{E}[M_\infty]\int_0^a \mathbb{P}(M_\infty \in [a-t,a))\,\mathrm{d}\mathbb{P}(M_\infty^* < t) \\
&\quad - \mathbb{E}[M_\infty \mathbb{1}(M_\infty < a-y)] + \mathbb{E}[M_\infty]\int_0^{a-y} \mathbb{P}(M_\infty \in [a-y-t,a))\,\mathrm{d}\mathbb{P}(M_\infty^* < t) \\
&\le \mathbb{E}[M_\infty \mathbb{1}(M_\infty \in [a-y,a))] + \mathbb{E}[M_\infty]\mathbb{P}(M_\infty \in [a-y-M_\infty^*, a); M_\infty^* < a-y) \\
&\le a\mathbb{P}(M_\infty \in [a-y,a)) + \mathbb{E}[M_\infty].
\end{aligned}$$

It therefore follows that

$$\mathbb{E}_y[\sigma(a) \mid \sigma(a) < \tau] - \frac{a}{1-\rho} \le \frac{\mathbb{E}[M_\infty]}{(1-\rho)\mathbb{P}(M_\infty \in [a-y,a))} \le \frac{\mathbb{E}[M_\infty]}{(1-\rho)\mathbb{P}(M_\infty \in [(1-d)\,a,a))},$$

where $\mathbb{E}[M_\infty] = \frac{\rho}{1-\rho}\frac{\mathbb{E}[B^2]}{2\mathbb{E}[B]}$.

Similar to the analysis of small initial values, Theorem 5.3.1 invokes

$$\mathbb{E}_y[\sigma(a) \mid \sigma(a) < \tau] \lesssim \frac{a}{1-\rho} + \frac{C}{(1-\rho)\,d\,a\mathbb{P}(B > a)} \tag{5.59}$$

and consequently completes the analysis of the conditional expectation for large initial values.

5.8.3 Synthesis of small and large random initial value

From relations (5.49), (5.58) and (5.59), one may deduce that

$$\begin{aligned}
&\sup_{a\geq a_\rho^*}\mathbb{E}[\sigma(a)\mid\sigma(a)<\tau]\\
&\quad\leq\sup_{a\geq a_\rho^*}\int_0^{da}\mathbb{E}_y[\sigma(a)\mid\sigma(a)<\tau]\,\mathrm{d}\mathbb{P}(B<y)+\sup_{a\geq a_\rho^*}\int_{da}^{a}\mathbb{E}_y[\sigma(a)\mid\sigma(a)<\tau]\,\mathrm{d}\mathbb{P}(B<y)\\
&\quad\lesssim\frac{C}{1-\rho}\sup_{a\geq a_\rho^*}\int_0^{da}y\,\mathrm{d}\mathbb{P}(B<y)+\frac{C}{1-\rho}\sup_{a\geq a_\rho^*}\int_0^{y_{min}}y\cdot\frac{f_{M_\infty}\left(\frac{a-y}{2}\right)}{f_{M_\infty}(a)}\,\mathrm{d}\mathbb{P}(B<y)\\
&\qquad+\sup_{a\geq a_\rho^*}\mathbb{P}(B\geq da)\sup_{y\in[da,a)}\mathbb{E}_y[\sigma(a)\mid\sigma(a)<\tau]\\
&\quad\lesssim\frac{C\mathbb{E}[B]}{1-\rho}+\frac{Cy_{min}}{1-\rho}\sup_{a\geq a_\rho^*}\int_0^{y_{min}}\frac{f_{M_\infty}\left(\frac{a-y}{2}\right)}{f_{M_\infty}(a)}\,\mathrm{d}\mathbb{P}(B<y)\\
&\qquad+\sup_{a\geq a_\rho^*}\frac{a}{1-\rho}\mathbb{P}(B\geq da)+\sup_{a\geq a_\rho^*}\frac{C}{(1-\rho)a}\frac{\mathbb{P}(B\geq da)}{\mathbb{P}(B\geq a)}.
\end{aligned}\tag{5.60}$$

The third term is dominated by its Markov's bound $\frac{\mathbb{E}[B]}{(1-\rho)d}$. Also, the integral in the second term is ultimately bounded by a constant. This follows from the fact that $\frac{f_{M_\infty}(x)}{\mathbb{P}(M_\infty\leq x)}$ is non-increasing and application of Theorem 5.3.1 as before:

$$\begin{aligned}
\int_0^{y_{min}}\frac{f_{M_\infty}\left(\frac{a-y}{2}\right)}{f_{M_\infty}(a)}\,\mathrm{d}\mathbb{P}(B<y)&\leq\frac{\mathbb{P}(M_\infty\leq\frac{a}{2})}{f_{M_\infty}(a)}\int_0^{y_{min}}\frac{f_{M_\infty}\left(\frac{a-y}{2}\right)}{\mathbb{P}(M_\infty\leq\frac{a-y}{2})}\,\mathrm{d}\mathbb{P}(B<y)\\
&\leq\mathbb{P}(B<y_{min})\frac{\mathbb{P}(M_\infty\leq\frac{a}{2})}{\mathbb{P}(M_\infty\leq a)}\frac{\mathbb{P}(M_\infty\leq a)}{f_{M_\infty}(a)}\frac{f_{M_\infty}\left(\frac{a-y_{min}}{2}\right)}{\mathbb{P}(M_\infty\leq\frac{a-y_{min}}{2})}\\
&=C\frac{\mathbb{P}(M_\infty\leq\frac{a}{2})}{\mathbb{P}(M_\infty\leq a)}\inf_{y\in(0,y_{min})}\frac{\mathbb{P}(M_\infty\leq a+y)}{f_{M_\infty}(a+y)}\inf_{y\in(0,y_{min})}\frac{f_{M_\infty}\left(\frac{a+y-2y_{min}}{2}\right)}{\mathbb{P}(M_\infty\leq\frac{a+y-2y_{min}}{2})}\\
&\leq C\frac{\mathbb{P}(M_\infty\leq\frac{a}{2})}{\mathbb{P}(M_\infty\leq a)}\frac{\mathbb{P}(M_\infty\leq a+y_{min})}{\mathbb{P}(M_\infty\leq\frac{a-y_{min}}{2})}\frac{\inf_{y\in(0,y_{min})}f_{M_\infty}\left(\frac{a+y-2y_{min}}{2}\right)}{\sup_{y\in(0,y_{min})}f_{M_\infty}(a+y)}\\
&\lesssim C\frac{\int_0^{y_{min}}f_{M_\infty}\left(\frac{a+y-2y_{min}}{2}\right)\mathrm{d}y}{\int_0^{y_{min}}f_{M_\infty}(a+y)\,\mathrm{d}y}=C\frac{\mathbb{P}\left(M_\infty\in\left(\frac{a-2y_{min}}{2},\frac{a-y_{min}}{2}\right)\right)}{\mathbb{P}(M_\infty\in(a,a+y_{min}))}\\
&\lesssim C\frac{\mathbb{P}\left(B>\frac{a-2y_{min}}{2}\right)}{\mathbb{P}(B>a+y_{min})}\sim C\left(1-\frac{3y_{min}}{a+y_{min}}\right)^{-\alpha}
\end{aligned}$$

as $a\to\infty$. Substituting this into (5.60) gives

$$\sup_{a\geq a_\rho^*}\mathbb{E}[\sigma(a)\mid\sigma(a)<\tau]\lesssim\frac{C}{1-\rho}+\frac{Cy_{min}}{1-\rho}\sup_{a\geq a_\rho^*}\left(1-\frac{3y_{min}}{a+y_{min}}\right)^{-\alpha}+\sup_{a\geq a_\rho^*}\frac{C}{(1-\rho)a}d^{-\alpha}.\tag{5.61}$$

Since all suprema are obtained in $a = a_\rho^*$ as $\rho \uparrow 1$, the above expressions can be written in terms of $1/(1-\rho)$:

$$\sup_{a \geq a_\rho^*} \mathbb{E}[\sigma(a) \mid \sigma(a) < \tau] \lesssim \frac{C}{1-\rho} + \frac{C}{\log \frac{1}{1-\rho}} = O\left(\frac{1}{1-\rho}\right).$$

This completes the proof of Lemma 5.7.2.

5.9 Tightness of bounds – proofs

This section presents the proofs of Lemma 5.3.6, Corollary 5.3.7 and Lemma 5.3.8, respectively.

5.9.1 Local Kingman heavy-traffic approximation

Complete monotonicity of the p.d.f. $f_{M_\infty}(\cdot)$ follows from Corollary 3.2 in Keilson [78]. As $f_{M_\infty}(\cdot)$ is non-increasing, it follows that the random variable V with c.d.f. $F_V(0) := 0$, $F_V(x) := 1 - \frac{f_{M_\infty}(x)}{f_{M_\infty}(0+)}$, $x > 0$, is well-defined. Relation (5.16) is now derived by analysing the Laplace-Stieltjes transform of V.

Let $\widetilde{M}_\infty(\cdot)$ and $\widetilde{B^*}(\cdot)$ denote the Laplace-Stieltjes transforms of M_∞ and B^*, respectively. On the one hand, we have [6, relation (7.9)]

$$\int_{0+}^{\infty} e^{-st} f_{M_\infty}(t)\,\mathrm{d}t = \widetilde{M}_\infty(s) - \mathbb{P}(M_\infty = 0) = \frac{1-\rho}{1-\rho\widetilde{B^*}(s)} - (1-\rho) = \frac{\rho(1-\rho)\widetilde{B^*}(s)}{1-\rho\widetilde{B^*}(s)}. \quad (5.62)$$

On the other hand, integration by parts yields

$$\int_{0+}^{\infty} e^{-st} f_{M_\infty}(t)\,\mathrm{d}t = \frac{1}{s} f_{M_\infty}(0+) + \frac{1}{s}\int_{0+}^{\infty} e^{-st}\,\mathrm{d}f_{M_\infty}(t). \quad (5.63)$$

Combining (5.62) and (5.63) gives

$$\mathbb{E}[e^{sV}] = -\int_0^\infty e^{-st}\,\mathrm{d}\frac{f_{M_\infty}(t)}{f_{M_\infty}(0+)} = 1 - \frac{\rho(1-\rho)s\widetilde{B^*}(s)}{f_{M_\infty}(0+)(1-\rho\widetilde{B^*}(s))} = 1 - \frac{\mathbb{E}[B]s\widetilde{B^*}(s)}{1-\rho\widetilde{B^*}(s)},$$

since $f_{M_\infty}(0+) = (1-\rho)\lambda$ (cf. relation (5.20)).

From the above, we deduce

$$\begin{aligned}\mathbb{E}[e^{(1-\rho)sV}] &= 1 - \frac{\mathbb{E}[B](1-\rho)s\widetilde{B^*}((1-\rho)s)}{1-\rho\widetilde{B^*}((1-\rho)s)} = 1 - \frac{\mathbb{E}[B](1-\rho)s\widetilde{B^*}((1-\rho)s)}{1-\rho\left(1-\mathbb{E}[B^*](1-\rho)s + o(1-\rho)\right)} \\ &\to 1 - \frac{\mathbb{E}[B]s}{1+\mathbb{E}[B^*]s}\end{aligned}$$

as $\rho \uparrow 1$. Inverting this expression and applying the Continuity Theorem [54, Section XIII.1, Theorem 2a] gives

$$\mathbb{P}((1-\rho)V \leq x) \to 1 - \frac{\mathbb{E}[B]}{\mathbb{E}[B^*]} e^{-\frac{x}{\mathbb{E}[B^*]}}, \quad (5.64)$$

provided $\mathbb{E}[B^*] \geq \mathbb{E}[B]$. Under this assumption, the lemma statement follows from the definition of $F_V(x)$. The proof is therefore concluded once we verify that all completely monotone densities $f_B(\cdot)$ satisfy $\mathbb{E}[B^*] \geq \mathbb{E}[B]$.

Bernstein's theorem [23] states that any completely monotone function can be represented as mixture of exponential functions. In particular, there exists a non-decreasing function $\mu(\cdot)$ such that

$$f_B(x) = \int_0^\infty e^{-tx}\,\mathrm{d}\mu(t). \tag{5.65}$$

From this representation, one may derive $1 = \int_0^\infty \frac{1}{t}\,\mathrm{d}\mu(t)$, $\mathbb{E}[B] = \int_0^\infty \frac{1}{t^2}\,\mathrm{d}\mu(t)$ and $\mathbb{E}[B^2] = \int_0^\infty \frac{2}{t^3}\,\mathrm{d}\mu(t)$. A straightforward computation yields

$$\mathbb{E}[B^2] - 2\mathbb{E}[B]^2 = \int_0^\infty \int_0^\infty \frac{1}{st}\left(\frac{1}{s} - \frac{1}{t}\right)^2 \mathrm{d}\mu(s)\,\mathrm{d}\mu(t) \geq 0.$$

The claimed property follows from $\mathbb{E}[B^*] - \mathbb{E}[B] = (\mathbb{E}[B^2] - 2\mathbb{E}[B]^2)/(2\mathbb{E}[B]) \geq 0$.

5.9.2 Lower bound of the function x_ρ

Since the p.d.f.'s of both M_∞ and B^* are well-defined and non-increasing, one can see that

$$\begin{aligned}\frac{\mathbb{P}(M_\infty \in \frac{y}{1-\rho} + \Delta)}{\frac{\rho}{1-\rho}\mathbb{P}(B^* \in \frac{y}{1-\rho} + \Delta)} &= \frac{\int_{y/(1-\rho)}^{y/(1-\rho)+T} f_{M_\infty}(t)\,\mathrm{d}t}{\frac{\rho}{1-\rho}\int_{y/(1-\rho)}^{y/(1-\rho)+T} \mathbb{P}(B > t)/\mathbb{E}[B]\,\mathrm{d}t} \geq \frac{f_{M_\infty}\left(\frac{y}{1-\rho} + T\right)}{\frac{\lambda}{1-\rho}\mathbb{P}\left(B > \frac{y}{1-\rho}\right)} \\ &\geq \frac{\frac{1}{1-\rho} f_{M_\infty}\left(\frac{y+T}{1-\rho}\right)}{\frac{\lambda}{(1-\rho)^2}\mathbb{P}\left(B > \frac{y}{1-\rho}\right)}.\end{aligned}$$

Fix $0 < \nu < \alpha - 2$. According to Potter's Theorem there exists a constant $C > 0$ such that $\mathbb{P}(B > x) \leq Cx^{-\alpha+\nu}$ for x sufficiently large. Hence, by Lemma 5.3.6,

$$\lim_{\rho\uparrow 1} \frac{\mathbb{P}\left(M_\infty \in \frac{y}{1-\rho} + \Delta\right)}{\frac{\rho}{1-\rho}\mathbb{P}\left(B^* \in \frac{y}{1-\rho} + \Delta\right)} \geq \lim_{\rho\uparrow 1} \frac{\frac{1}{1-\rho} f_{M_\infty}\left(\frac{y+T}{1-\rho}\right)}{\lambda C(1-\rho)^{\alpha-2-\nu} y^{-\alpha+\nu}} = \infty.$$

5.9.3 Lower bound of the function x^*_ρ

Theorem 1 in Abate and Whitt [5] states that $\frac{1}{1-\rho}\mathbb{P}\left(\tau > \frac{t}{(1-\rho)^2}\right)$ converges to a function $f_{\mathrm{R}}(t)$ as $\rho \uparrow 1$ for all $t > 0$. Thus, since F_B is regularly varying with index $-\alpha < -2$, we find

$$\lim_{\rho\uparrow 1} \frac{\mathbb{P}(\tau > y^*_\rho)}{\frac{\rho}{1-\rho}\mathbb{P}(B > (1-\rho)y^*_\rho)} = \lim_{\rho\uparrow 1} \frac{\frac{1}{1-\rho}\mathbb{P}\left(\tau > \frac{y}{(1-\rho)^2}\right)}{\frac{\rho}{(1-\rho)^2}\mathbb{P}\left(B > \frac{y}{1-\rho}\right)} = f_R(y) \lim_{\rho\uparrow 1} \frac{1}{\frac{\rho}{(1-\rho)^2}\mathbb{P}\left(B > \frac{y}{1-\rho}\right)} = \infty.$$

5.A Inequalities

This appendix is dedicated to the proof of Lemma 5.5.1. Takács [133, Section 29] and Cohen [38] have independently shown that

$$\mathbb{P}(M_\tau > x) = \frac{1}{\lambda}\frac{\mathrm{d}}{\mathrm{d}y}\log\mathbb{P}(M_\infty < y)\big|_{y=x}, \tag{5.66}$$

of which we analyse the right-hand side by means of scale functions.

The definition and some properties of scale functions were provided in Section 5.8; for this appendix we recall that $\mathbb{P}(M_\infty < x) = (1-\rho)W_\rho(x)$ and that $\frac{\mathrm{d}}{\mathrm{d}y}\log W_\rho(y)$ is non-increasing and positive (cf. relation (5.54)). This latter property implies that $\log W_\rho(x)$ is concave.

Rewriting (5.66) in terms of the scale function $W_\rho(x)$, exploiting its concavity and using $\log x \le x-1$ gives

$$\begin{aligned}\mathbb{P}(M_\tau > x) &= \frac{1}{\lambda}\frac{\mathrm{d}}{\mathrm{d}y}\log W_\rho(y)\big|_{y=x} \le \frac{1}{\lambda}\left[\log W_\rho(x) - \log W_\rho(x-1)\right] \\ &= \frac{1}{\lambda}\log\frac{W_\rho(x)}{W_\rho(x-1)} \le \frac{1}{\lambda}\left[\frac{W_\rho(x)}{W_\rho(x-1)} - 1\right] = \frac{1}{\lambda}\frac{\mathbb{P}(M_\infty \in [x-1,x))}{\mathbb{P}(M_\infty < x-1)}\end{aligned}$$

for all $x > 1$, which concludes the upper bound. Using the inequality $\log x \ge 1 - \frac{1}{x}$ for all $x > 0$ and slightly altering the above analysis yields the lower bound.

Bibliography

Self-references

[S1] Bansal, N., Kamphorst, B. & Zwart, B. (2018). Achievable performance of blind policies in heavy traffic. *Mathematics of Operations Research, Articles in Advance*.

[S2] Kamphorst, B. & Zwart, B. (2018a). Heavy-traffic analysis of sojourn time under the foreground-background scheduling policy. Manuscript submitted for publication.

[S3] Kamphorst, B. & Zwart, B. (2018b). Uniform asymptotics for compound Poisson processes with regularly varying jumps and vanishing drift. *Stochastic Processes and their Applications, In Press*.

References

[4] Aalto, S. & Ayesta, U. (2006). Mean delay analysis of multi level processor sharing disciplines. In *Proceedings IEEE INFOCOM 2006. 25th IEEE international conference on computer communications* (pp. 1–11).

[5] Abate, J. & Whitt, W. (1995). Limits and approximations for the busy-period distribution in single-server queues. *Probability in the Engineering and Informational Sciences*, *9*(4), 581–602.

[6] Adan, I. & Resing, J. (2002). Queueing theory. Eindhoven University of Technology, Eindhoven.

[7] Asmussen, S. (1998). Subexponential asymptotics for stochastic processes: Extremal behavior, stationary distributions and first passage probabilities. *Annals of Applied Probability*, *8*(2), 354–374.

[8] Asmussen, S. (2003). *Applied probability and queues*. Springer.

[9] Asmussen, S. & Klüppelberg, C. (1996). Large deviations results for subexponential tails, with applications to insurance risk. *Stochastic Processes and their Applications*, *64*(1), 103–125.

[10] Avrahami, N. & Azar, Y. (2003). Minimizing total flow time and total completion time with immediate dispatching. In *Proceedings of the fifteenth annual ACM symposium on parallel algorithms and architectures* (pp. 11–18). ACM.

[11] Avram, F., Palmowski, Z. & Pistorius, M. R. (2007). On the optimal dividend problem for a spectrally negative Lévy process. *Annals of Applied Probability*, *17*(1), 156–180.

[12] Awerbuch, B., Azar, Y., Leonardi, S. & Regev, O. (1999). Minimizing the flow time without migration. In *Proceedings of the thirty-first annual ACM symposium on theory of computing* (pp. 198–205). STOC '99. ACM.

[13] Baltrūnas, A., Daley, D. J. & Klüppelberg, C. (2004). Tail behaviour of the busy period of a GI/GI/1 queue with subexponential service times. *Stochastic Processes and their Applications*, *111*(2), 237–258.

[14] Bansal, N. (2005). On the average sojourn time under M/M/1/SRPT. *Operations Research Letters*, *33*(2), 195–200.

[15] Bansal, N. & Gamarnik, D. (2006). Handling load with less stress. *Queueing Systems*, *54*(1), 45–54.

[16] Bansal, N. & Harchol-Balter, M. (2001). Analysis of SRPT scheduling: Investigating unfairness. In *Proceedings of the 2001 ACM SIGMETRICS international conference on measurement and modeling of computer systems* (pp. 279–290). SIGMETRICS '01. ACM.

[17] Bateman, H. (1954). Tables of integral transforms [volumes I & II]. McGraw-Hill Book Company.

[18] Beasley, J. E., Sonander, J. & Havelock, P. (2001). Scheduling aircraft landings at london heathrow using a population heuristic. *The Journal of the Operational Research Society*, *52*(5), 483–493.

[19] Becchetti, L. & Leonardi, S. (2004). Nonclairvoyant scheduling to minimize the total flow time on single and parallel machines. *Journal of the ACM*, *51*(4), 517–539.

[20] Beirlant, J., Broniatowski, M., Teugels, J. L. & Vynckier, P. (1995). The mean residual life function at great age: Applications to tail estimation. *Journal of Statistical Planning and Inference*, *45*(1), 21–48.

[21] Bekker, R., Boxma, O. J. & Resing, J. A. C. (2009). Lévy processes with adaptable exponent. *Advances in Applied Probability*, *41*(1), 117–205.

[22] Beneš, V. E. (1957). On queues with Poisson arrivals. *The Annals of Mathematical Statistics*, *28*(3), 670–677.

[23] Bernstein, S. (1929). Sur les fonctions absolument monotones. *Acta Mathematica*, *52*(1), 1–66.

[24] Bingham, N. H., Goldie, C. M. & Teugels, J. L. (1989). *Regular variation*. Cambridge University Press.

[25] Blanchet, J. & Lam, H. (2013). Uniform large deviations for heavy-tailed queues under heavy traffic. *Boletin de la Sociedad Matemática Mexicana*, *19*(3).

[26] Błażewicz, J., Ecker, K. H., Pesch, E., Schmidt, G. & Weglarz, J. (2013). *Scheduling computer and manufacturing processes*. Springer science & Business media.

[27] Borodin, A. & El-Yaniv, R. (1998). *Online computation and competitive analysis*. Cambridge University Press.

[28] Borovkov, A. A. (1970). Factorization identities and properties of the distribution of the supremum of sequential sums. *Theory of Probability and Its Applications*, *15*(3), 359–402.

[29] Borovkov, A. A. & Borovkov, K. A. (2008). *Asymptotic analysis of random walks: Heavy-tailed distributions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press.

[30] Borst, S. C., Boxma, O. J., Núñez-Queija, R. & Zwart, A. P. (2003). The impact of the service discipline on delay asymptotics. *Performance Evaluation*, *54*(2), 175–206. Modelling Techniques and Tools for Computer Performance Evaluation.

[31] Boxma, O. J. (1978). On the longest service time in a busy period of the M/G/1 queue. *Stochastic Processes and their Applications*, *8*(1), 93–100.

[32] Boxma, O. & Zwart, B. (2007). Tails in scheduling. *ACM SIGMETRICS Performance Evaluation Review*, *34*(4), 13–20.

[33] Braverman, A., Dai, J. G. & Miyazawa, M. (2017). Heavy traffic approximation for the stationary distribution of a generalized jackson network: The BAR approach. *Stochastic Systems*, *7*(1), 143–196.

[34] Brockmeyer, E., Halstrm, H. L. & Jensen, A. (1948). *The life and works of A.K. Erlang*. Akademiet for de Tekniske Videnskaber.

[35] C, L. & Iyer, S. A. (2013). Application of queueing theory in health care: A literature review. *Operations Research for Health Care*, *2*(1–2), 25–39.

[36] Chekuri, C., Khanna, S. & Zhu, A. (2001). Algorithms for minimizing weighted flow time. In *Proceedings of the thirty-third annual ACM symposium on theory of computing* (pp. 84–93). STOC '01. ACM.

[37] Chen, H. & Yao, D. D. (2013). *Fundamentals of queueing networks: Performance, asymptotics, and optimization*. Springer Science & Business Media.

[38] Cohen, J. W. (1968). Extreme value distribution for the M/G/1 and the G/M/1 queueing systems. *Annales de l'institut Henri Poincaré, section B*, *4*(1), 83–98.

[39] Cohen, J. W. (1982). *The single server queue*. North-Holland Amsterdam.

[40] Conway, R. W., Maxwell, W. L. & Miller, L. W. (1967). *Theory of scheduling*. Addison-Wesley Publishing Company, Inc.

[41] Corbató, F. J., Merwin-Daggett, M. & Daley, R. C. (1962). An experimental time-sharing system. In *Proceedings of the May 1–3, 1962, Spring Joint Computer Conference* (pp. 335–344). AIEE-IRE '62 (Spring). ACM.

[42] Cuyt, A., Petersen, V. B., Verdonk, B., Waadeland, H. & Jones, W. B. (2008). *Handbook of continued fractions for special functions*. Springer.

[43] Dębicki, K. & Mandjes, M. (2012). Lévy-driven queues. *Surveys in Operations Research and Management Science*, *17*(1), 15–37.

[44] Dembo, A. & Zeitouni, O. (2010). *Large deviations techniques and applications* (2nd ed.). Stochastic Modelling and Applied Probability. Springer-Verlag.

[45] Denisov, D. & Kugler, J. (2014). Heavy traffic and heavy tails for subexponential distributions. *arXiv:1403.7325v2 [math.PR]*.

[46] Denisov, D., Dieker, A. B. & Shneer, V. (2008). Large deviations for random walks under subexponentiality: The big-jump domain. *The Annals of Probability*, *36*(5), 1946–1991.

[47] Du, J., Leung, J. Y.-T. & Young, G. H. (1990). Minimizing mean flow time with release time constraint. *Theoretical Computer Science*, *75*(3), 347–355.

[48] Durrett, R. (1980). Conditioned limit theorems for random walks with negative drift. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *52*(3), 277–287.

[49] Edmonds, J. (1965). Paths, trees, and flowers. *Canadian Journal of Mathematics*, *17*(3), 449–467.

[50] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling extremal events: For insurance and finance*. Springer.

[51] Embrechts, P. & Veraverbeke, N. (1982). Estimates for the probability of ruin with special emphasis on the possibility of large claims. *Insurance: Mathematics and Economics*, *1*(1), 55–72.

[52] Erlang, A. K. (1909). Sandsynlighedsregning og telefonsamtaler. *Nyt Tidsskrift for Matematik B*, *20*(6), 33–39.

[53] Erlang, A. K. (1917). Løsning af nogle problemer fra sandsynlighedsregningen af betydning for de automatiske telefoncentraler. *Elektroteknikeren*, *13*.

[54] Feller, W. (1971). *An introduction to probability theory and its applications: Volume II*. John Wiley & Sons.

[55] Feuerstein, E., Marchetti-Spaccamela, A., Schalekamp, F., Sitters, R., van der Ster, S., Stougie, L. & van Zuylen, A. (2017). Minimizing worst-case and average-case makespan over scenarios. *Journal of Scheduling*, *20*(6), 545–555.

[56] Fiat, A. & Woeginger, G. J. (1998). Online algorithms: The state of the art. *LNCS, Springer-Verlag, 1442*.

[57] Foss, S. & Konstantopoulos, T. (2004). An overview of some stochastic stability methods. *Journal of the Operations Research Society of Japan*, *47*(4), 275–303.

[58] Foss, S., Konstantopoulos, T. & Zachary, S. (2007). Discrete and continuous time modulated random walks with heavy-tailed increments. *Journal of Theoretical Probability*, *20*(3), 581–612.

[59] Foss, S., Korshunov, D. & Zachary, S. (2013). *An introduction to heavy-tailed and subexponential distributions*. Springer.

[60] Gamarnik, D. & Zeevi, A. (2006). Validity of heavy traffic steady-state approximation in generalized Jackson networks. *Annals of Applied Probability, 16*(1), 56–90.

[61] Ganesh, A. J., O'Connell, N. & Wischik, D. J. (2004). *Big queues*. Springer.

[62] Garey, M. R. & Johnson, D. S. (1979). *Computers and intractibility*. WH Freeman Co.

[63] Glynn, P. W. (1990). Diffusion approximations. In *Stochastic models* (Vol. 2, pp. 145–198). Handbooks in Operations Research and Management Science. Elsevier.

[64] Goldie, C. M. & Klüppelberg, C. (1998). Subexponential distributions. *A practical guide to heavy tails: Statistical techniques and applications*, 435–459.

[65] Graham, R. L. (1966). Bounds for certain multiprocessing anomalies. *Bell Labs Technical Journal, 45*(9), 1563–1581.

[66] Graham, R. L., Lawler, E. L., Lenstra, J. K. & Rinnooy Kan, A. H. G. (1979). Optimization and approximation in deterministic sequencing and scheduling: A survey. *Annals of Discrete Mathematics, 5*, 287–326.

[67] Gupta, D. & Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions, 40*(9), 800–819.

[68] Gut, A. (1988). *Stopped random walks*. Springer.

[69] de Haan, L. (1974). Equivalence classes of regularly varying functions. *Stochastic Processes and their Applications, 2*(3), 243–259.

[70] Haque, L. & Armstrong, M. J. (2007). A survey of the machine interference problem. *European Journal of Operational Research, 179*(2), 469–482.

[71] Harchol-Balter, M. (2013). *Performance modeling and design of computer systems: Queueing theory in action*. Cambridge University Press.

[72] Harchol-Balter, M., Bansal, N., Schroeder, B. & Agrawal, M. (2001). SRPT scheduling for web servers. *Lecture notes in computer science, 2221*, 11–20.

[73] Hubalek, F. & Kyprianou, E. (2011). Old and new examples of scale functions for spectrally negative Lévy processes. In *Seminar on stochastic analysis, random fields and applications VI* (pp. 119–145). Springer.

[74] Jackson, J. R. (1955). *Scheduling a production line to minimize maximum tardiness*. California University Los Angeles.

[75] Johnson, S. M. (1954). Optimal two-and three-stage production schedules with setup times included. *Naval Research Logistics (NRL), 1*(1), 61–68.

[76] Kalyanasundaram, B. & Pruhs, K. R. (2003). Minimizing flow time nonclairvoyantly. *Journal of the ACM, 50*(4), 551–567.

[77] Kasperski, A., Kurpisz, A. & Zieliński, P. (2012). Parallel machine scheduling under uncertainty. In *International conference on information processing and management of uncertainty in knowledge-based systems* (pp. 74–83). Springer.

[78] Keilson, J. (1978). Exponential spectra as a tool for the study of server-systems with several classes of customers. *Journal of Applied Probability*, *15*(1), 162–170.

[79] Kelly, F. & Yudovina, E. (2014). *Stochastic networks*. Cambridge University Press.

[80] Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *Annals of Mathematical Statistics*, *24*(3), 338–354.

[81] Khintchine, A. Ya. (1932). The mathematical theory of a stationary queue. *Matematicheskii Sbornik*, *39*(4), 73–84.

[82] Kiefer, J. & Wolfowitz, J. (1955). On the theory of queues with many servers. *Transactions of the AMS*, *78*(1), 1–18.

[83] Kingman, J. F. C. (1961). The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society*, *57*, 902–904.

[84] Kingman, J. F. C. (1962). On queues in heavy traffic. *Journal of the Royal Statistical Society, Series B*, *24*(2), 383–392.

[85] Kleinrock, L. (1975). *Queueing Systems I: Theory*. John Wiley & Sons.

[86] Kleinrock, L. (1976). *Queueing Systems II: Computer theory*. John Wiley & Sons.

[87] Klüppelberg, C., Kyprianou, A. E. & Maller, R. A. (2004). Ruin probabilities and overshoots for general Lévy insurance risk processes. *Annals of Applied Probability*, *14*(4), 1766–1801.

[88] Kugler, J. & Wachtel, V. (2013). Upper bounds for the maximum of a random walk with negative drift. *Journal of Applied Probability*, *50*(4), 1131–1146.

[89] Kuznetsov, A., Kyprianou, A. E. & Rivero, V. (2013). The theory of scale functions for spectrally negative Lévy processes. In *Lévy matters II* (pp. 97–186). Springer.

[90] Kyprianou, A. E. (2014). *Introductory lectures on fluctuations of Lévy processes with applications* (2nd ed.). Springer.

[91] Lawler, E. L., Lenstra, J. K. & Rinnooy Kan, A. H. G. (1982). Recent developments in deterministic sequencing and scheduling: A survey. In *Deterministic and stochastic scheduling* (pp. 35–73). Springer.

[92] Lee, A. M. (1966). *Applied queueing theory*. Macmillan.

[93] Lenstra, J. K., Rinnooy Kan, A. & Brucker, P. (1977). Complexity of machine scheduling problems. In *Annals of discrete mathematics* (Vol. 1, pp. 343–362). Elsevier.

[94] Leonardi, S. & Raz, D. (1997). Approximating total flow time on parallel machines. In *Proceedings of the twenty-ninth annual ACM symposium on theory of computing* (pp. 110–119). STOC '97. ACM.

[95] Lin, M., Wierman, A. & Zwart, B. (2011). Heavy-traffic analysis of mean response time under shortest remaining processing time. *Performance Evaluation*, *68*(10), 955–966.

[96] Lindley, D. V. (1952). The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*, *48*(2), 277–289.

[97] Lotov, V. I. (2002). Inequalities for the moments and distribution of the ladder height of a random walk. *Siberian Mathematical Journal*, *43*(4), 655–660.

[98] Lu, D., Sheng, H. & Dinda, P. (2004). Size-based scheduling policies with inaccurate scheduling information. In *Modeling, analysis, and simulation of computer and telecommunications systems, 2004. (MASCOTS 2004). Proceedings. the IEEE computer society's 12th annual international symposium on* (pp. 31–38).

[99] Mandjes, M. & Zwart, B. (2006). Large deviations of sojourn times in processor sharing queues. *Queueing Systems*, *52*(4), 237–250.

[100] Maulik, K. & Zwart, B. (2006). Tail asymptotics for exponential functionals of Lévy processes. *Stochastic Processes and their Applications*, *116*(2), 156–177.

[101] de Meyer, A. & Teugels, J. L. (1980). On the asymptotic behaviour of the distributions of the busy period and service time in M/G/1. *Journal of Applied Probability*, *17*(3), 802–813.

[102] Motwani, R., Phillips, S. & Torng, E. (1994). Nonclairvoyant scheduling. *Theoretical Computer Science*, *130*(1), 17–47.

[103] Nair, J., Wierman, A. & Zwart, B. (2010). Tail-robust scheduling via limited processor sharing. *Performance Evaluation*, *67*(11), 978–995.

[104] von Neumann, J. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, *100*(1), 295–320.

[105] Nuyens, M. & Wierman, A. (2008). The foreground–background queue: A survey. *Performance Evaluation*, *65*(3), 286–307.

[106] Nuyens, M., Wierman, A. & Zwart, B. (2008). Preventing large sojourn times using SMART scheduling. *Operations Research*, *56*(1), 88–101.

[107] Olvera-Cravioto, M., Blanchet, J. & Glynn, P. W. (2011). On the transition from heavy traffic to heavy tails for the M/G/1 queue: The regularly varying case. *Annals of Applied Probability*, *21*(2), 645–668.

[108] Olvera-Cravioto, M. & Glynn, P. W. (2011). Uniform approximations for the M/G/1 queue with subexponential processing times. *Queueing Systems*, *68*(1), 1–50.

[109] Pinedo, M. (2012). *Scheduling* (4th ed.). Springer.

[110] Pollaczek, F. (1930a). Über eine Aufgabe der Wahrscheinlichkeitstheorie. I. *Mathematische Zeitschrift*, *32*(1), 64–100.

[111] Pollaczek, F. (1930b). Über eine Aufgabe der Wahrscheinlichkeitstheorie. II. *Mathematische Zeitschrift*, *32*(1), 729–750.

[112] Potts, C. N. & Strusevich, V. A. (2009). Fifty years of scheduling: A survey of milestones. *Journal of the Operational Research Society*, *60*(S1), S41–S68.

[113] Prokhorov, Yu. V. (1959). An extremal problem in probability theory. *Theory of Probability and Its Applications*, *4*(2), 201–203.

[114] Prokhorov, Yu. V. (1963). Transition phenomena in queueing processes. I. *Litovskiĭ Matematicheskiĭ Sbornik*, *3*(1), 199–205.

[115] Pruhs, K. (2007). Competitive online scheduling for server systems. *ACM SIGMETRICS Performance Evaluation Review*, *34*(4), 52–58.

[116] Pruhs, K., Sgall, J. & Torng, E. (2004). Online scheduling. In *Handbook of scheduling: Algorithms, models, and performance analysis*.

[117] Puha, A. L. et al. (2015). Diffusion limits for shortest remaining processing time queues under nonstandard spatial scaling. *Annals of Applied Probability*, *25*(6), 3381–3404.

[118] Remerova, M., Foss, S. & Zwart, B. (2014). Random fluid limit of an overloaded polling model. *Advances in Applied Probability*, *46*(1), 76–101.

[119] Resnick, S. I. (1987). *Extreme values, regular variation and point processes*. Springer.

[120] Rhee, C.-H., Blanchet, J. & Zwart, B. (2016). Sample path large deviations for heavy-tailed Lévy processes and random walks. *arXiv:1606.02795v3 [math.PR]*.

[121] Righter, R. & Shanthikumar, J. G. (1989). Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probability in the Engineering and Informational Sciences*, *3*(3), 323–333.

[122] Righter, R., Shanthikumar, J. G. & Yamazaki, G. (1990). On extremal service disciplines in single-stage queueing systems. *Journal of Applied Probability*, *27*(2), 409–416.

[123] Salveson, M. E. (1952). On a quantitative method in production planning and scheduling. *Econometrica*, *20*(4), 554–590.

[124] Schrage, L. E. (1967). The queue M/G/1 with feedback to lower priority queues. *Management Science*, *13*(7), 466–474.

[125] Schrage, L. E. (1968). Letter to the editor — a proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, *16*(3), 687–690.

[126] Schroeder, B. & Harchol-Balter, M. (2006). Web servers under overload: How scheduling can help. *ACM Transactions on Internet Technology*, *6*(1), 20–52.

[127] Scully, Z., Harchol-Balter, M. & Scheller-Wolf, A. (2018). SOAP: One clean analysis of all age-based scheduling policies. *arXiv:1712.00790v2 [math.PR]*.

[128] Sgall, J. (1998). On-line scheduling. In *Online algorithms* (pp. 196–231). Springer.

[129] Stolyar, A. L. & Ramanan, K. (2001). Largest weighted delay first scheduling: Large deviations and optimality. *Annals of Applied Probability*, *11*(1), 1–48.

[130] Syski, R. (1961). Review of *Mathematical Methods in the Theory of Queueing*, by A. Ya. Khintchine. *The Incorporated Statistician, 11*(1), 57–61.

[131] Taha, H. A. (1971). *Operations research: An introduction*. Macmillan.

[132] Takács, L. (1962). The time dependence of a single-server queue with Poisson input and general service times. *The Annals of Mathematical Statistics, 33*(4), 1340–1348.

[133] Takács, L. (1967). *Combinatorial methods in the theory of stochastic processes*. New York, NY, USA: John Wiley & Sons.

[134] Vazirani, V. V. (2013). *Approximation algorithms*. Springer Science & Business Media.

[135] Whitt, W. (1974). Heavy traffic limit theorems for queues: A survey. In *Mathematical methods in queueing theory: Proceedings of a conference at Western Michigan University, May 10–12, 1973* (pp. 307–350). Springer.

[136] Whitt, W. (2002). *Stochastic-process limits: An introduction to stochastic-process limits and their application to queues*. Springer Science & Business Media.

[137] Wierman, A., Harchol-Balter, M. & Osogami, T. (2005). Nearly insensitive bounds on SMART scheduling. In *Proceedings of the 2005 ACM SIGMETRICS international conference on measurement and modeling of computer systems* (pp. 205–216). SIGMETRICS '05. ACM.

[138] Wierman, A. & Nuyens, M. (2008). Scheduling despite inexact job-size information. *ACM SIGMETRICS Performance Evaluation Review, 36*(1), 25–36.

[139] Wierman, A. & Zwart, B. (2012). Is tail-optimal scheduling possible? *Operations Research, 60*(5), 1249–1257.

[140] Williamson, D. P. & Shmoys, D. B. (2011). *The design of approximation algorithms*. Cambridge University Press.

[141] Wolff, R. W. (1989). *Stochastic modeling and the theory of queues*. Pearson College Division.

[142] Yao, A. C.-C. (1977). Probabilistic computations: Toward a unified measure of complexity. In *2013 IEEE 54th annual symposium on foundations of computer science* (pp. 222–227). IEEE.

[143] Zhang, J. & Zwart, B. (2008). Steady state approximations of limited processor sharing queues in heavy traffic. *Queueing Systems, 60*(3), 227–246.

[144] Zwart, A. P. (2001). Tail asymptotics for the busy period in the GI/G/1 queue. *Mathematics of Operations Research, 26*(3), 485–493.

[145] Zwart, A. P. & Boxma, O. J. (2000). Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Systems, 35*(1), 141–166.

Summary

Heavy-traffic behaviour of scheduling policies in queues

This dissertation lies in the intersection of two mathematical areas: queueing theory and scheduling theory. Queueing theory concerns itself with the analysis and control of congestion phenomena. It provides a better understanding of the performance of call centres, data communication networks, hospital emergency departments and many more applications. Scheduling theory is devoted to the study of decision-making processes that deal with the allocation of resources to tasks. Its insights are, among others, utilised in computer operating systems, manufacturing processes and surgery planning in hospitals.

Both queueing theory and scheduling theory consider a number of customers (tasks) that all need to be served (processed) by one or more servers (machines). Researchers from both areas design policies (algorithms) that assign the customers to the servers, and assess their performance. The fact that many of these policies are applicable in both communities emphasises their resemblance.

The main differences between the two areas are the characteristics of the customers and the choice of performance metrics. In queueing models, one assumes that there is a some degree of uncertainty. In particular, is often unknown when the next customer arrives or how much service he or she requires, although the corresponding probability distributions might be available. Additionally, it is assumed that the inflow of customers never ceases. It is for these reasons that researchers analyse performance metrics like the expected waiting time of a customer, or the probability that a customer has to wait for at least five minutes. Scheduling theory, on the other hand, historically focuses on models where the probability distributions are generally unknown or non-existing. One instead assumes that there is a limited number of customers, and that these customers potentially enter the system at the most inconvenient times. Researchers then aim to give guarantees on the behaviour of policies in these worst-case scenarios.

As illustrated above, some queueing models are closely related to scheduling models and vice versa. At the same time it seems that researchers of either community are often unaware of the results and the techniques used in the other area. This dissertation

contributes in bridging this gap by presenting (1) a novel combination of techniques from both worlds and (2) several analyses of policies in queueing models. Most of our results apply to the "heavy-traffic regime", which is a specification to queueing models where the amount of requested service approaches the capacity of the server. This regime reflects the growing demand that we observe in many applications and simultaneously reduces the complexity of the analyses.

We begin this dissertation with an extended version of the above introduction into the areas of queueing and scheduling theory, and discuss a selection of relevant literature. Then, in Chapter 2, we investigate the performance of "blind" scheduling policies in the GI/GI/1 queueing model. Blind policies do not base their decisions on the service requirement of a customer, which may pose an advantage if this information is inaccurate or unavailable. We show that blind policies can not perform as good as their information-dependent counterparts for general models and derive a lower bound on this sub-optimality. Subsequently, we show that there exists a blind policy that actually achieves this lower bound, implying that this is in some sense the best blind policy possible.

The above results are derived under the assumption that there is only a single server; however, the techniques that are used seem to be applicable to more general models. In particular, one may attempt to follow a similar approach to analyse queueing models with multiple servers. Chapter 3 shows that one such approach is not quite suitable to obtain the desired extension and discusses alternative approaches.

Chapter 4 focuses on the Foreground-Background policy in the M/GI/1 model. Several performance metrics of the Foreground-Background policy depend greatly on the stochastic properties of the service requirements. Among these metrics is the sojourn time: the time that a customer resides in the system. In this chapter we quantify the heavy-traffic behaviour of the mean sojourn time and the probability of a long sojourn time under a broad range of service-requirement distributions.

Our fifth and final chapter also considers the probability that customers experience a sojourn time of at least x time units in M/GI/1 queueing models. We obtain results for both the First In First Out and the Last In First Out policy under the assumption that the service-time distribution is heavy-tailed. More specifically, we show how large x needs to be in order for these results to remain valid in heavy traffic.

Samenvatting

Gedrag van planningsalgoritmes in zwaar belaste wachtrijen

Dit proefschrift beschouwt het raakvlak tussen twee mathematische disciplines: wachtrijtheorie en planningstheorie. Wachtrijtheorie concentreert zich op de analyse en beïnvloeding van opstoppingsfenomenen. De kennis uit deze discipline wordt benut bij het effectief beheren van telefooncentrales, datanetwerken, spoedeisende hulpposten en nog vele andere toepassingen. Planningstheorie richt zich op de beslissingsprocessen bij het toewijzen van benodigdheden aan taken. De inzichten uit deze discipline worden onder andere toegepast op besturingssystemen, fabrieksprocessen en het inplannen van operaties in ziekenhuizen.

Zowel wachtrijtheorie als planningstheorie gaat uit van een aantal klanten (taken) die allemaal bediend (verwerkt) willen worden door één of meerdere medewerkers (machines). Onderzoekers vanuit beide disciplines ontwerpen hiertoe algoritmes die de klanten toewijzen aan de medewerkers, en bestuderen vervolgens hoe goed de algoritmes werken. Het feit dat veel van deze algoritmes toepasbaar zijn in beide disciplines benadrukt hun gelijkenis.

De grootste verschillen tussen de disciplines zijn de eigenschappen van de klanten en de prestatie indicatoren. In wachtrijmodellen wordt aangenomen dat er een bepaalde mate van onzekerheid is. Zo is het vaak onbekend wanneer de volgende klant arriveert of hoelang een klant geholpen moet worden, hoewel men soms beschikt over de bijbehorende kansverdelingen. Daarnaast wordt aangenomen dat de toestroom van klanten nooit stopt. Vanwege deze eigenschappen zijn onderzoekers vaak geïnteresseerd in de gemiddelde wachttijd van een klant, of de kans dat een klant meer dan vijf minuten moet wachten. Aan de andere kant richt planningstheorie zich op modellen waar de kansverdelingen onbekend zijn of niet eens bestaan. In plaats daarvan gaat men er vanuit dat er een beperkt aantal klanten is, en dat deze klanten mogelijk op de meest ongunstige tijden arriveren. Onderzoekers tonen vervolgens garanties aan met betrekking tot de prestatie van algoritmes in deze doemscenario's.

Zoals hierboven is beschreven zijn sommige wachtrijmodellen nauw verwant aan planningsmodellen en omgekeerd. Tegelijkertijd lijkt het vaak zo te zijn dat de onderzoe-

kers van de ene gemeenschap zich niet bewust zijn van de resultaten en technieken uit de andere gemeenschap. Dit proefschrift draagt bij aan het overbruggen van deze kloof door (1) een nieuwe combinatie van technieken uit de twee disciplines te presenteren en (2) verscheidene algoritmes te analyseren in wachtrijmodellen. De meeste resultaten zijn gericht op zwaar belaste wachtrijen, waarmee wordt bedoeld dat er dusdanig veel werk in het model is dat de medewerker het nog net bij kan houden. Deze specificatie reflecteert de groeiende vraag die we in veel toepassingsgebieden waarnemen en vereenvoudigt tegelijkertijd de doorgaans complexe analyses.

We beginnen dit proefschrift met een uitgebreidere versie van de bovenstaande introductie in de wachtrij- en planningstheorie en bespreken een selectie van de relevante literatuur. Daarna bestuderen we in Hoofdstuk 2 de prestatie van "blinde" algoritmes in het GI/GI/1 wachtrijmodel. Blinde algoritmes baseren hun keuzes niet op de benodigde bedieningsduur per klant, wat een voordeel kan zijn als deze informatie onnauwkeurig of niet beschikbaar is. We laten zien dat, in algemene wachtrijmodellen, blinde algoritmes niet altijd zo goed kunnen presteren als hun informatie-afhankelijke tegenpolen en leiden een ondergrens af voor deze inefficiëntie. Vervolgens laten we zien dat er een blind algoritme bestaat dat deze ondergrens daadwerkelijk bereikt, wat impliceert dat dit in zekere zin het optimale blinde algoritme is.

De bovenstaande resultaten zijn afgeleid onder de aanname dat er slechts één medewerker is. De gebruikte technieken lijken echter ook toepasbaar te zijn in algemenere modellen. Men zou dus kunnen proberen om wachtrijmodellen met meerdere medewerkers volgens een soortgelijke methode te analyseren. In Hoofdstuk 3 laten we van een specifieke methode zien dat deze ongeschikt is voor de gewenste modeluitbreiding, waarna we een korte uiteenzetting van alternatieve methodes geven.

Hoofdstuk 4 richt zich op het Foreground-Background algoritme in het M/GI/1 wachtrijmodel. Een aantal prestatie indicatoren hangt sterk af van de stochastische eigenschappen van de benodigde bedieningsduur per klant. Tot deze indicatoren behoort ook de verblijftijd van een klant: de tijdspanne waarin de klant in het systeem is. In dit hoofdstuk kwantificeren we, voor een brede selectie van bedieningsduurverdelingen, de groei van de gemiddelde verblijftijd en de kans op een lange verblijftijd als het systeem zwaar belast is.

Ook ons vijfde en laatste hoofdstuk bestudeert de kans dat een klant een verblijftijd van minstens x tijdseenheden heeft in een M/GI/1 wachtrijmodel. We presenteren resultaten voor de First In First Out en Last In First Out algoritmes onder de aanname dat de bedieningsduurverdelingen zwaarstaartig zijn. Specifiek laten we zien hoe groot x moet zijn om zeker te weten dat de resultaten geldig blijven als het systeem zwaar belast raakt.

About the author

Bart Kamphorst was born in Rockanje, municipality Westvoorne, the Netherlands, on the 6th of August, 1990. After obtaining his atheneum diploma at the bilingual PENTA college CSG Jacob van Liesveldt in Hellevoetsluis in 2008, Bart studied mathematics at Leiden University where he finished the bachelor's program in 2011. Bart subsequently completed the master's track Applied Mathematics with honours in 2013 after an eight-month internship at the independent research organisation TNO, where he investigated cascading failures in high-voltage power grids for his master's thesis.

In October 2013, Bart started in the PhD project titled "Bridging deterministic and stochastic analysis of scheduling policies" in the Stochastics group of Centrum Wiskunde & Informatica in Amsterdam under daily supervision of prof. dr. A.P. Zwart. The project was initiated to obtain new insights at the interface of applied probability and theoretical computer science. The corresponding results are presented in this dissertation, which will be defended on Thursday the 31st of May, 2018.