Error analysis of structured Markov chains

Eleni Vatamidou

Error analysis of structured Markov chains

Eleni Vatamidou

Eindhoven University of Technology
Department of Mathematics
and Computer Science

Beta
Research School for Operations
Management and Logistics

Error analysis of structured Markov chains

Eleni Vatamidou

This work is part of the research programme *Error Bounds for Structured Markov Chains* (project nr. 613.001.006), which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).

Error analysis of structured Markov chains / by Eleni Vatamidou. –

Mathematics Subject Classification (2010): Primary 60K25, 91B30; Secondary 68M20, 60F10, 41A99.

A catalogue record is available from the Eindhoven University of Technology Library.
ISBN: 978-90-386-3878-2

This thesis is number D 192 of the thesis series of the Beta Research School for Operations Management and Logistics.

Kindly supported by a full scholarship from the legacy of K. Katseas.

Error analysis of structured Markov chains

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 maandag 22 juni 2015 om 16.00 uur

door

Eleni Vatamidou

geboren te Thessaloniki, Griekenland

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

voorzitter:	prof.dr. M.G.J. van den Brand
1e promotor:	prof.dr. A.P. Zwart
2e promotor:	prof.dr.ir. I.J.B.F. Adan
copromotor:	dr. M. Vlasiou
leden:	prof.dr. H. Albrecher (Universite de Lausanne)
	dr. B. Van Houdt (Universiteit Antwerpen)
	prof.dr.ir. G.J.J.A.N. van Houtum
	prof.dr. Z. Palmowski (University of Wrocław)

Acknowledgements

Once I was told that all you need to complete a PhD project is to withstand the disappointments that come along with research and have the guts to keep on going. The more than four years of research that was carried out during my PhD career – the outcome of which is this book – taught me that being surrounded by people that help, support, and guide you is also an essential factor that made this work possible. Thus, I would like to express my gratitude to all of you that were there when I needed you and kept me motivated.

First, I would like to thank my supervisor Maria Vlasiou for working closely with me all these years. I am very grateful for her advice, patience, recommendations, and support as a true friend. I am also deeply indebted to Bert Zwart who always guided me in the right direction and very generously shared his ideas with me. Ivo Adan with his inspiring enthusiasm, great intuition, and good sense of humour made our collaboration delightful and memorable. Thank you all for trusting me and providing me with the opportunity to improve my scientific skills under your supervision. It has been an honour to be your student.

Next, I would like to express my appreciation to the members of my doctoral committee: Hansjoerg Albrecher, Benny Van Houdt, Geert-Jan van Houtum, and Zbyszek Palmowski. Thank you for accepting our invitation, reading my dissertation with eagerness, and for your useful feedback. It was a pleasure interacting with all of you. Special thanks also to Zbyszek for inviting me to Wrocław and making my visit there really enjoyable.

A great many thanks to my professor in Thessaloniki, George Tsaklidis, who is the reason why I initiated this project. He was the person who recommended me for this position and provided me with valuable advice – both in life and research – throughout the years I know him. No matter how busy his schedule was, his door was always open for me, since I was a bachelor student, to discuss any matter that troubled me or simply to catch up on my news. A warm thank you is also owed to my professor Nikolaos Farmakis who most willingly offered me his help when I needed him.

One of my favourite quotes is that *friends are the family we choose*. This could not have been more accurate for Maximiliano Udenio whom I consider as my elder brother. In the most humorous way, he wrote recently for himself: "*During her stay in the Netherlands, Eleni met Maxi and her life was forever changed. Not only was he an inspiration for her research, but also in all matters of life. Changing light bulbs, fixing bikes, driving to far away lands... Nothing was too much for Maxi. His life affirming*

inspiration is the reason why you are reading these lines." Although these words were meant to amuse me, they are 100% true but incomplete! His constant encouragement, most precious advice, and genuine concern about me truly brightened my existence. I feel blessed meeting you and honoured to call you my friend. You are simply great.

Life brought it in a way that most of my beloved friends followed different paths and we are spread all over Europe. I always cherish the time we spent together and our long discussions about everything. Spyros, Vaso, Andreas, Duke, Paulos, Stauros, and Georgia, thank you for being so close to me, albeit so far away. In addition, I wish to express my indebtedness to a few friends that played an important role to my smooth settlement in the Netherlands. Elina, Erato, Amin, and Serban, thank you for being around and proved to me in various occasions that I can rely on you for everything.

Research on its own is already challenging, yet alone conducting it in a foreign country. Therefore, I owe a great deal to all those contributed, one way or another, to the realisation of this essay. I particularly wish to thank all the support staff in our department and EURANDOM that were always king enough to answer my questions. Whenever I entered the office of Wil Kortsmit with a new computer-related problem, he called it immediately "our problem" and showed a remarkable devotion in helping me. His assistance has been very fundamental in my numerical experiments. Marko, Jan-Pieter, Britt, and Jori were also very generous sharing with me their knowledge with computer coding when I sought their help.

This period of my life would not have been so amazing without the people I met on my way. I am so thankful to all EURANDOMers and other colleagues for the great experiences we shared. Many thanks to all of my lecturers in BETA and LNMB. Guido Janssen, I will never forget your kindness and respectable personality. Johan Hurink, thank you for your precious help during the LNMB meetings and the times we played "Mafia" together. Sindo Núñez Queija, thank you for all the fruitful discussions – not limited to math – that we had all these years. Finally, Onno Boxma, thank you for being such a great group leader and always very accessible to all of us.

From this list, my parents and brother could not be left out. My parents taught me that hard work and persistence is the key to succeed, not to quit on early difficulties, and always follow my dreams. Although separating from me was hard for them, they earnestly supported my decision to come to the Netherlands. One phone call was always enough for them to get on the first plane and be by my side. My brother was also very supportive. With his optimism and amazing humour, he managed to bridge our distance and hold a prominent place in my daily life. Μαμά, μπαμπά, Βαγγέλη, σας ευχαριστώ πολύ για όλη τη συμπαράσταση, τις συμβουλές σας, και που στέκεστε πάντα δίπλα μου όλα αυτά τα χρόνια. Σας αγαπώ πολύ.

Last but not least, I would like to thank my friend and long-time partner Stratos. I am extremely lucky having you in my life. Not only you inspired me to pursuit a PhD position abroad, but you adapted your own plans in order to follow me and offered your support by every means possible. Your belief in me and your unconditional love give me wings to fly.

Ελένη
May 2015

Contents

Acknowledgements **v**

1 Introduction **1**
1.1 Scope of the thesis . . . 1
1.2 Block-structured Markov processes . . . 4
1.2.1 Basic queueing models . . . 5
1.2.2 QBDs, GI/M/1–, and M/G/1–type processes . . . 6
1.3 Matrix-Analytic Methods . . . 8
1.3.1 Solutions for GI/M/1– and M/G/1–type processes . . . 8
1.3.2 Phase-type distributions . . . 10
1.3.3 Markovian Arrival Processes . . . 12
1.4 Truncations of the background state space . . . 14
1.4.1 Heavy-tailed models . . . 14
1.4.2 Networks of queues . . . 16
1.5 Beyond Matrix-Analytic Methods . . . 16
1.6 Our contribution . . . 18
1.7 Notation . . . 19
1.7.1 General symbols . . . 20
1.7.2 Distributions and random variables . . . 20
1.7.3 Matrices and vectors . . . 21

2 Spectral approximations **23**
2.1 Introduction . . . 23
2.2 Spectral approximation for the ruin probability . . . 25
2.2.1 Error bound for the ruin probability . . . 26
2.2.2 Completely monotone claim sizes . . . 28
2.3 Algorithm for the spectral approximation . . . 30
2.4 Heavy traffic and heavy tail approximations . . . 32
2.5 Numerical experiments . . . 33
2.5.1 Test distributions . . . 33

2.5.2 Numerical results . . . 35
2.6 Conclusions . . . 44

3 Corrected phase-type approximations **47**
3.1 Introduction . . . 47
3.2 Series expansions of the ruin probability . . . 49
3.3 Corrected phase-type approximations of the ruin probability . . . 54
3.3.1 Approximation errors . . . 55
3.3.2 Tail behaviour . . . 56
3.3.3 Relative error . . . 58
3.4 Numerical experiments . . . 61
3.4.1 Test distribution . . . 62
3.4.2 Numerical results . . . 64
3.5 Total loss and Value at Risk . . . 67

4 Corrected phase-type approximations in a Markovian environment **73**
4.1 Introduction . . . 73
4.2 Presentation of the model . . . 75
4.2.1 Preliminaries . . . 75
4.2.2 Construction of the corrected phase-type approximations . . . 87
4.3 Corrected replace approximation . . . 88
4.3.1 Replace base model . . . 88
4.3.2 Perturbation of the parameters of the replace base model . . . 92
4.3.3 Delay distribution of the perturbed model . . . 96
4.3.4 Corrected replace approximations . . . 109
4.4 Corrected discard approximation . . . 111
4.5 Numerical experiments . . . 117
4.6 Conclusions . . . 119

5 Truncated buffer approximations **121**
5.1 Introduction . . . 121
5.2 Presentation of the model . . . 123
5.3 State space truncation and error bounds . . . 124
5.3.1 Partition of the queue length probabilities . . . 125
5.3.2 Error bounds . . . 129
5.4 Exponential change of measure . . . 129
5.4.1 Random walk notation . . . 131
5.5 Asymptotic approximation for the maximum . . . 132
5.6 Asymptotics for the conditional mean return time . . . 136
5.6.1 Intuitive illustrations . . . 137
5.6.2 Rigorous proofs . . . 140
5.7 Numerical experiments . . . 151
5.7.1 Special case: single arrivals . . . 151
5.7.2 Numerical results . . . 152
5.8 Conclusions . . . 153

Appendix **157**
A.1 Subexponential distributions 157
A.2 Results on perturbation theory 157
A.3 Random walks and related results 162

Summary **167**

Bibliography **169**

Curriculum Vitae **183**

CHAPTER 1

Introduction

1.1 Scope of the thesis

In almost every aspect of our everyday life, we are involved with queues. Sometimes these queues are visible to us, for example when we stand in a supermarket line to pay for our shopping or when we stop at a red traffic light. Sometimes, however, we do not even realise that we are waiting in a queue, i.e. when we download a movie from the internet or when we perform a call. In all cases, there exist various factors that affect how fast a queue builds up or empties. In the end, these factors affect how long delays we experience in a queue until we achieve our objective – e.g. cross a red traffic light – or eventually abandon the queue – e.g. stop downloading the movie.

The field of mathematics that aims to study these phenomena is called *queueing theory*. In broad terms, queueing theory deals with models that involve a number of servers providing service to at least one queue of customers, where neither the customers nor the servers are necessarily individuals. For example, consider the to-do-lists most of us are familiar with, as a simple queueing model. In this model, a person has the role of the server that needs to complete all different obligations, which play the role of the customers. New customers (obligations) arrive one by one or in groups by following a specific pattern – including also the appearance of unexpected obligations. All arriving customers are first prioritised based on criteria such as the time they demand by the server in order to be completed or their significance. Afterwards, some of them are served only by one server, while others may require treatment by more servers simultaneously or consecutively. In addition, the server may not serve some customers e.g. because of overdue deadlines, lack of interest or perhaps because they chose to abandon the queue.

A main objective of queueing theory is to measure the performance of queueing systems. The performance of a queueing system can be expressed in terms of (mean) waiting times of arriving customers, (mean) queue lengths, traffic intensities (i.e. average occupancy of a server during a specified time period), or any other relevant

measure for the specific application we study. Traditionally, a major focus of queueing theory was to find *closed-form analytic solutions* for those performance measures, i.e. formulas involving the input parameters of the model such as the arrival process and service times. Nonetheless, finding only closed-form solutions for performance measures is not sufficient. The point is that by using these expressions one should be able to calculate exact numerical estimates for the performance measures.

In an ideal situation with deterministic input parameters, the performance measures of a queueing system are accurately computable. However, such deterministic behaviour is not customary in nature, where randomness is most often the case. Queueing models involving randomness are an example of stochastic models. In general, when analysing stochastic models it is hard or even impossible to find closed-form solutions for their performance measures. The more complicated the system, the harder it is to find such closed-form solutions. Therefore, simplifying assumptions are usually imposed to the input parameters so as to be able to (approximately) compute performance measures of interest.

A common simplifying assumption is that the inter-arrival times (the time between two successive arrivals of customers) and/or the service times are exponential. A basic property of the exponential distribution is the *memoryless* or *Markov* property, which is also transferred to stochastic processes involving exponential distributions. We say that a stochastic process has the Markov property if the conditional (on both past and present states) probability distribution of future states of the process depends only upon the present state, but not on the sequence of events that preceded it. As a consequence, exponential assumptions most often lead to models that can be analysed with the aid of Markov processes, the theory of which provides many closed-form expressions for performance measures.

Although exponential distributions are widely used in the analysis of queueing models, they are not always a realistic assumption. Two important restrictions of the exponential distribution are that it is not capable of modelling extremely long times (highly variable) and that it does not allow for correlations between arrivals of customers. Therefore, as a generalisation of the exponential distribution, for the analysis of more complicated queueing models, the use of mixtures and convolutions of exponentials has also been suggested. Models involving such constructions of exponentials can be represented by Markov processes that have a block structure. Although the Markov property suggests lack of memory and independence in the stochastic process, such constructions can approximate properties such as long-range dependence and high variability.

Typically, a block-structured Markov process is defined on a two-dimensional countable state space, where the first coordinate is called *level* and the second *phase*. An example is a tandem network with two queues, where customers join the waiting line of the second queue upon completing their service in the first queue. Here, the level corresponds to the number of customers in the second queue and the phase represents the number of customers in the first queue. Performance measures for this kind of block-structured models can be evaluated numerically in an algorithmic way with the aid of techniques that combine probability and matrix theory (Bini et al., 2005; Gail et al., 1996); sometimes closed-form expressions with probabilistic interpretations for performance measures can also be found. These techniques are widely known as *Matrix-Analytic Methods* (MAM) and they are very successful in the

numerical analysis of systems arising not only in queueing (Breuer and Baum, 2005), but also in applications like insurance (Badescu et al., 2009), telecommunications (Ost, 2001), supply chain management (He, 2014), and mathematical finance (Asmussen et al., 2004).

To analyse block-structured Markov chains with MAM, the phase space should be finite. However, there exist many practical situations with an infinite number of phases. For example, if in the above described tandem queueing network all arriving customers are admitted, then obviously the number of phases is unbounded. Another example with an infinite background state space is when heavy-tailed random variables are involved, where heavy-tailed distributions are ordinarily used to model long service times. In order to preserve the heavy-tailed property within the context of block-structured Markov processes, we should allow for an arbitrary number of phases.

Infinite phase spaces may be pragmatic because they usually reflect real-world dynamics, but the matrices of infinite dimensions appearing in the analysis of the corresponding systems do not allow for numerical investigation. MAM that work efficiently with systems of realistic size, are applicable to systems of infinite state space only under specific model restrictions. Truncation of the background state space overcomes the issue of infinitely many phases. From an application point of view, truncation of the background state space can be interpreted as approximating general distributions with (finite) phase-type distributions, and/or infinite waiting rooms (buffers) with finite ones. Truncation can also be done to overcome numerical issues with finite, but large state spaces.

Such approximations are customary in practice (Feldmann and Whitt, 1998; Heindl et al., 2004; Horváth and Telek, 2000; Nielsen, 2000), because truncations of the background state space lead to an approximate model that can be analysed exactly or numerically. However, they also introduce approximation errors. Most importantly, truncations typically are done in an empirical way that is deprived of a solid theoretical background. In other words, the truncation point may not be chosen in a constructive way to yield approximations of performance measures with a guaranteed accuracy.

The goal of this dissertation is to obtain a rigorous understanding of these types of errors for a number of practically relevant classes of stochastic systems. More precisely, we extend the applicability of MAM by establishing algorithms that yield provably accurate estimates for the performance of a wide class of systems, including systems with heavy tails, financial models, and queueing networks. Our main focus is on controlling the error of performance measures by strategically choosing the truncation point of the approximations and relating it to the error incurred.

Outline

The rest of this chapter is organised as follows. All the models discussed in this dissertation can be represented by Markov processes with a repetitive structure. Thus, in Section 1.2, we first present simple Markov processes that have a repetitive structure. Afterwards, we explain how a two-dimensional Markov process can retain a similar structure based on the way its states are ordered, and we give the basic terminology for block-structured Markov processes. Later, in Section 1.3, we focus on MAM. In Section 1.3.1, we give a brief overview on solution methods for the basic models with block structure. The approximations we derive in this dissertation rely heavily on

the characteristics of *phase-type* (PH) distributions and *Markovian Arrival Processes* (MArPs). Therefore, in Sections 1.3.2 and 1.3.3, we provide a detailed description for each of them, respectively.

In Section 1.4, we discuss truncations of the state space. Typical examples where MAM impose truncations are models involving heavy-tailed distributions and queueing networks with more than one queue. In this dissertation, we consider truncations of either type. Therefore, in Section 1.4.1, we provide an overview on the basic characteristics of heavy-tailed distributions and related results, while in Section 1.4.2, we do the same for queueing networks.

In Section 1.5, we provide background literature on alternative solution methods for Markov processes, which may have a more general structure. Furthermore, in Section 1.6, we explain the different models that we study and the derived results of this dissertation. Finally, in Section 1.7, we introduce some general rules with respect to the notation that we follow throughout the whole dissertation.

1.2 Block-structured Markov processes

In this dissertation, our focus is on infinite block-structured Markov processes. These kind of Markov processes (or chains) have a repetitive structure that ordinarily occurs when analysing queueing models via their *embedded Markov chains*. By embedding, we mean that we consider the Markov process only at the moments upon which the state of the system changes. In this section, we give an overview of the basic concepts and terminology for Markov chains with repetitive structures. Note that we do not use exclusively either *continuous time Markov chains* (CTMCs) or *discrete time Markov chains* (DTMCs) for this overview, but we choose the form that makes the presentation intuitively easier. Moreover, we present the aforementioned concepts in the context of queueing theory. Therefore, as a first step it is important to introduce some notation to describe the characteristics of a queueing system.

In queueing theory, the standard system that is used to describe and classify a queue is Kendall's notation (or simply Kendall notation). A three-factor notation was initially proposed by Kendall (1953) that was later extended to its current form:

$$A/S/s/c/p/D$$

where

A: stands for the description of the arrival process,

S: stands for the service time distribution,

s: stands for the number of servers in the system and can be any integer greater than or equal to 1 (including ∞),

c: stands for the capacity of the queue, i.e. the maximum number of customers that can be queued in the system ($c \geq 0$). If this argument is missing, the queue capacity is by default infinity.

p: stands for the system population, i.e. the maximum number of customers that can arrive in the queue. If this argument is missing, the system population is by default infinity.

D: stands for the queueing discipline, which can be FCFS (first come first served), LCFS (last come first served), or any other queueing discipline. If this argument is missing, then, by default, the queueing discipline is FCFS.

Some of the most commonly used symbols for the description of the inter-arrival and service time processes are the M (Markovian, memoryless or exponential), D (deterministic), and G (general), where in the case of independent arrivals of customers we use the notation GI. Also, for single server queues, the disciplines FCFS and LCFS are also noted as FIFO (first in first out) and LIFO (last in first out), respectively.

1.2.1 Basic queueing models

The simplest queueing system is the M/M/1 queue, where both arrival and service processes are Markovian and the customers are served by a single server with FIFO discipline. In this model, there is no restriction on the system population and queue capacity. Therefore, the number of customers in the queue is described by a *continuous time Markov chain* (CTMC) with state space the non-negative integers $\mathbb{S} = \{0, 1, \dots\}$ representing the number of customers in the system, i.e. both those waiting in line and the one receiving service. The upward transitions from n to $n+1$ occur at an exponential rate λ and describe customer arrivals, while transitions from n to n–1, for $n > 1$ occur at an exponential rate μ and describe completed services (departures) from the queue. The *infinitesimal generator* of the queue length process – i.e. the matrix with entries the rates at which the process jumps from state to state – is equal to

$$\mathbf{Q}_{M/M/1} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & \cdots \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ 0 & 0 & \mu & -(\lambda+\mu) & \lambda & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{1.1}$$

Since transitions can occur only at the nearest neighbours, this model is known as a pure *birth-death stochastic process*. The most important feature here is that $\mathbf{Q}$ is a tri-diagonal matrix: the elements of the upper diagonal are all equal, as are those of the lower diagonal and of the main diagonal, with the exception of the upper left corner.

Two other very basic queueing models are the GI/M/1 and the M/G/1 queues. To study the *steady state* or *limiting* behaviour of the queue length for these models one usually considers the embedded Markov chains at arrivals or departures of customers, respectively. Observe that we could attempt to find the steady-state behaviour by observing the queue length processes of these queues in continuous times, but the analysis is easier with embedded Markov chains. For example, the state of M/G/1 queue in continuous time can be described by a two-dimensional vector, where the first coordinate represents the number of customers in the system and the second coordinate corresponds to the elapsed service time of the customer in service. The one coordinate is discrete, but the other one continuous and this essentially perplexes the analysis. However, when we observe the system after departures, the state description simplifies to one-dimensional (only the number of customers in the system is required), because the elapsed service time of a new customer (if any) is zero.

The *probability transition matrix* of the GI/M/1 model – i.e. the matrix where each row represents the transition flow out of the corresponding state – takes the form

$$\mathbf{P}_{GI/M/1} = \begin{pmatrix} p_0 & \beta_0 & 0 & 0 & 0 & \cdots \\ p_1 & \beta_1 & \beta_0 & 0 & 0 & \cdots \\ p_2 & \beta_2 & \beta_1 & \beta_0 & 0 & \cdots \\ p_3 & \beta_3 & \beta_2 & \beta_1 & \beta_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{1.2}$$

where β_n denotes the probability of serving n customers during an inter-arrival time given that the server remains busy during this interval (thus there are more than n customers present) and $p_n = \sum_{i=n+1}^{\infty} \beta_i$. On the other hand, the transition probability matrix of the M/G/1 queue is equal to

$$\mathbf{P}_{M/G/1} = \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \cdots \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \cdots \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\ 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \cdots \\ 0 & 0 & 0 & \alpha_0 & \alpha_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{1.3}$$

where α_n denotes the probability that during a service time exactly n customers arrive. Observe that the transition probability matrix of the GI/M/1 queue is a *lower Hessenberg* matrix (Horn and Johnson, 1986), i.e. all the elements above the upper diagonal are equal to zero, while the transition probability matrix of the M/G/1 queue is an *upper Hessenberg* matrix, i.e. all the elements below the lower diagonal are equal to zero.

These three queueing models are the stepping stones of queueing theory, where the M/M/1 queue lies in the intersection of the other two. Consequently, these basic models have been treated extensively in the literature; for a comprehensive analysis of these models refer to Kleinrock (1976) and Asmussen (2003). However, performance measures for the M/G/1 queue are less easy to find. As we shall see later in Section 1.4.1, this is the case where the service times follow some heavy-tailed distribution. Therefore, despite its simplicity, the M/G/1 queue is an intriguing model, which we study in Chapters 2 and 3. The presentation of the GI/M/1 queue is included here for completeness.

1.2.2 QBDs, GI/M/1–, and M/G/1–type processes

In this section, we give an overview of block-structured Markov chains. In addition, we explain how the basic models of block-structured Markov chains relate to the basic queueing models we presented in Section 1.2.1.

As we mentioned in Section 1.1, a block-structured Markov chain is defined on a two-dimensional countable state space. As a result, its state space is of the form $\mathbb{S} = \{(i, j) : i \geq 0, 1 \leq j \leq m\}$, where m is the dimension of the phase space. To preserve some ordering among the states of such a chain, we partition its state space $\mathbb{S}$ as $\cup_{i\geq 0} l(i)$, where $l(i) = \{(i, 1), (i, 2), \ldots, (i, m)\}$, for $i \geq 0$. Except for the first coordinate, we also use the word *level* to denote the whole subset $l(i)$.

The most basic constructions of block-structured Markov chains are the GI/M/1– and M/GI/1–type processes. These, processes are widely known as *skip-free processes in one direction*. Specifically, the GI/M/1–type process is called *skip-free to the right* to indicate that the chain can move up only by one level at a time, although it may skip in the downward direction several levels in one transition. On the other hand, the M/G/1–type process is called *skip-free to the left*, because the chain can move down only by one level at a time, although it may skip in the upward direction several levels in one transition.

If we use the letters "L", "F", and "B" to denote the "local", "forward", and "backward" transition rates (Riska and Smirni, 2002b), the infinitesimal generator of the GI/M/1–type process is

$$\mathbf{Q_{GI/M/1}} = \begin{pmatrix} \widehat{\mathbf{L}} & \widehat{\mathbf{F}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \widehat{\mathbf{B}}^{(1)} & \mathbf{L} & \mathbf{F} & \mathbf{0} & \mathbf{0} & \cdots \\ \widehat{\mathbf{B}}^{(2)} & \mathbf{B}^{(1)} & \mathbf{L} & \mathbf{F} & \mathbf{0} & \cdots \\ \widehat{\mathbf{B}}^{(3)} & \mathbf{B}^{(2)} & \mathbf{B}^{(1)} & \mathbf{L} & \mathbf{F} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{1.4}$$

Observe that the above infinitesimal generator is a lower block Hessenberg type matrix and looks like the transition probability matrix of the simple GI/M/1 (see Eq. (1.2)); except for the fact that now, the elements of the matrix are themselves matrices. Therefore, the GI/M/1– type process is considered the matrix-equivalent of its corresponding simple model. Examples of GI/M/1–type process include systems that allow the customers to be served in groups and also systems that capture failure of service nodes (Grassmann and Stanford, 2000).

On the other hand, the infinitesimal generator of the M/G/1–type process is

$$\mathbf{Q_{M/G/1}} = \begin{pmatrix} \widehat{\mathbf{L}} & \widehat{\mathbf{F}}^{(1)} & \widehat{\mathbf{F}}^{(2)} & \widehat{\mathbf{F}}^{(3)} & \widehat{\mathbf{F}}^{(4)} & \cdots \\ \widehat{\mathbf{B}} & \mathbf{L} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \mathbf{F}^{(3)} & \cdots \\ \mathbf{0} & \mathbf{B} & \mathbf{L} & \mathbf{F}^{(1)} & \mathbf{F}^{(2)} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{B} & \mathbf{L} & \mathbf{F}^{(1)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{1.5}$$

Analogously, this infinitesimal generator is an upper block Hessenberg type matrix and the M/G/1–type process is considered the matrix equivalent of the simple M/G/1 queue (see Eq. (1.3)). Moreover, this type of processes usually characterise batch arrivals, i.e. simultaneous arrivals of customers in the queueing system (Grassmann and Stanford, 2000).

As the M/M/1 queue lies in the intersection of the GI/M/1 and the M/G/1 queues, at the intersection of the GI/M/1– and M/G/1–type models lie the *Quasi-Birth-Death processes* (QBDs). Therefore, QBDs are skip-free in both directions, which means that jumps between levels are exclusively to the nearest neighbours. Therefore, the

infinitesimal generator of QBDs takes the form

$$\mathbf{Q}_{QBD} = \begin{pmatrix} \widehat{\mathbf{L}} & \widehat{\mathbf{F}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \widehat{\mathbf{B}} & \mathbf{L} & \mathbf{F} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{B} & \mathbf{L} & \mathbf{F} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{B} & \mathbf{L} & \mathbf{F} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{1.6}$$

which is a tridiagonal block-matrix (see Eq. (1.1)).

In Section 1.3.1, we provide details on solution methods for these block-structured Markov chains. The benefit from their structure is that many algorithms from linear algebra require significantly less computational effort when applied to Hessenberg type matrices.

1.3 Matrix-Analytic Methods

In this dissertation, we construct algorithms for the numerical estimation of performance measures that cannot be found explicitly. As we shall see in Section 1.6 (also Chapters 2–4 in detail), MAM are on the basis of our derived approximations. Thus, we devote this section on the presentation of MAM.

MAM have been thriving since the work of Neuts (1989, 1994). Historically, they developed as two independent sets of techniques for the study the GI/M/1– and M/G/1–type Markov chains. By using a matrix formalism, they provide a framework that is widely used for the numerical analysis of various stochastic models described by block-structured Markov chains (Bini et al., 2005; Latouche and Ramaswami, 1999). In particular, MAM have been extensively applied to Markov chains on two-dimensional state spaces for which one-step transitions are allowed across several levels in one direction. Simple examples are queues with batch arrivals or group services.

The success of MAM is mainly attributed to the development of appealing algorithms that obtain numbers for systems of realistic size without having to resort to time-consuming simulations. A catalyst has been the ever-increasing ability of computers to perform numerical calculations. With a high-speed computer, elementary matrix operations can easily be programmed. In addition, software tools based on MAM have been developed (Bini et al., 2006; Van Velthoven et al., 2007; Pérez et al., 2008; Riska and Smirni, 2002a, 2007).

In Section 1.2.2, we presented the block structure of the GI/M/1– and M/G/1–type processes. This special structure is exploited by MAM to provide recursive and algorithmically tractable solutions for these types of processes. In Section 1.3.1, we briefly discuss such solution methods. Additionally, the theory of MAM is inextricably related to phase-type distributions and MArPs, which are usually perceived as the building blocks of MAM. Therefore, in Sections 1.3.2 and 1.3.3, we give an overview of phase-type distributions and MArPs, respectively.

1.3.1 Solutions for GI/M/1– and M/G/1–type processes

When estimating performance measures for the GI/M/1–type process, key to the MAM is the computation of an auxiliary matrix, traditionally denoted by $\mathbf{R}$. Similarly,

the analysis of the M/G/1–type process is related to an auxiliary matrix $\mathbf{G}$. The approach of MAM is to calculate the matrices $\mathbf{R}$ and $\mathbf{G}$ by iteratively solving matrix equations, where all the involved matrices have finite dimensions. In this section, we give an overview of the existing algorithms for the computation of these auxiliary matrices, their probabilistic interpretations, and other related results.

We start our analysis with the GI/M/1–type process and the auxiliary matrix $\mathbf{R}$. If $\mathbf{F}$, $\mathbf{L}$, and $\mathbf{B}^{(k)}$, $k = 1, 2, \ldots$, are according to Eq. (1.4), then matrix $\mathbf{R}$ is the unique non-negative solution to the matrix equation (Neuts, 1994; Riska and Smirni, 2002b)

$$\mathbf{F} + \mathbf{RL} + \sum_{k=1}^{\infty} \mathbf{R}^{k+1}\mathbf{B}^{(k)} = \mathbf{0}$$

and can be computed using iterative numerical algorithms (Ramaswami and Latouche, 1986). Matrix $\mathbf{R}$ has an important probabilistic interpretation: We denote as Δ^i the mean sojourn time in the state $(i-1, k)$ of $l(i-1)$ for $i \geq 2$. Then, the entry (k, n) of $\mathbf{R}$ is the expected time spent in the state (i, n) of $l(i)$, before the first visit into $l(i-1)$, expressed in time unit Δ^i, given the starting state in $l(i-1)$ is $(i-1, k)$ (Neuts, 1994, pages 30–35).

The matrix $\mathbf{R}$ is also called *geometric coefficient* because of the matrix-geometric relation that holds among the stationary probabilities of the sets $l(i)$. To make this clear, let $\boldsymbol{\pi}$ be the stationary probability vector of the GI/M/1–type process. If we consider its partition into the sub-vectors $\boldsymbol{\pi}_i$, $i \geq 0$, where $\boldsymbol{\pi}_i = (\pi_{ij})_{j=1,\ldots,m}$, then the following relation holds

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 \mathbf{R}^i, \qquad i = 1, 2, \ldots$$

This property leads to significant algebraic simplifications that result in the very elegant matrix-geometric solution that involves only the numerical evaluation of the matrix $\mathbf{R}$ and the stationary probabilities $\boldsymbol{\pi}_0$ (Neuts, 1994).

For M/G/1–type processes there is no geometric relation among the various probability vectors $\boldsymbol{\pi}_i$, $i = 1, 2, \ldots$, as in the case of GI/M/1–type processes (Riska and Smirni, 2002b). Therefore, although MAM have been proposed for the solution of the basic equation $\boldsymbol{\pi}\mathbf{Q}_{M/G/1} = \mathbf{0}$, they are significantly more complicated in this case because there is no explicit or simple solution to this system of equations (Bini et al., 2000; Meini, 1998; Neuts, 1989). Nonetheless, there exist recursive schemes based on $\mathbf{G}$ to compute the steady state probability vector (Ramaswami, 1988).

The algorithms for M/G/1–type processes involve the computation of the matrix $\mathbf{G}$, which is the unique solution to the matrix equation

$$\mathbf{B} + \mathbf{LG} + \sum_{k=1}^{\infty} \mathbf{F}^{(k)}\mathbf{G}^{k+1} = \mathbf{0},$$

where $\mathbf{B}$, $\mathbf{L}$, and $\mathbf{F}^{(k)}$, $k = 1, 2, \ldots$, are according to Eq. (1.5). The matrix $\mathbf{G}$ can be determined by using iterative algorithms (Latouche and Ramaswami, 1999; Meini, 1998) and has an important probabilistic interpretation: an entry (r, c) in $\mathbf{G}$ expresses the conditional probability that the process first enters the level $l(i-1)$ through the state $(i-1, c)$, given that it starts from state (i, r) of $l(i)$ (Neuts, 1989, page 81).

Finally, since QBDs are special cases of both GI/M/1– and M/G/1–type processes, its stationary probabilities can be found by using existing algorithms for either type of processes. However, because of its simplicity, the matrix-geometric solution is most preferable. Moreover, for QBDs both matrices $\mathbf{R}$ and $\mathbf{G}$ are defined and they are related through the fundamental equation (Latouche and Ramaswami, 1999, pages 137–8)

$$\mathbf{RB} = \mathbf{FG},$$

where $\mathbf{L}$ and $\mathbf{F}$ are according to Eq. (1.6). Different types of duality between $\mathbf{R}$ and $\mathbf{G}$, can also be found in Asmussen and Ramaswami (1990) and Ramaswami (1990). Due to the duality between the matrices $\mathbf{R}$ and $\mathbf{G}$, calculating only one of them is sufficient to have both. Ordinarily, the matrix $\mathbf{G}$ is computed first because it is easier to control the calculation errors (He, 2014).

1.3.2 Phase-type distributions

Phase-type distributions play a very prominent role in the construction of our approximations in Chapters 2–4. Thus, in this section, we give the definition of phase-type distributions, we introduce the related terminology, and we list their basic properties, which make the class of phase-type distributions a very popular class.

Phase-type distributions were first introduced by Neuts (1975) and they are characterised by a *finite* and *absorbing* Markov chain. The number of phases in a phase-type distribution is equal to the number of transient states in the associated (underlying) Markov chain. More precisely, a distribution B on $(0, \infty)$ is phase-type if B is the distribution of the time the Markov chain spends in the transient states until absorption. In addition, if $\boldsymbol{\alpha}$ is the initial probability vector for each of its transient states, $\boldsymbol{T}$ is the square matrix representing the transitions among the transient states, and E denotes the set of transient states, then B is said to be of phase-type with *representation* $(E, \boldsymbol{\alpha}, \boldsymbol{T})$ or simply $(\boldsymbol{\alpha}, \boldsymbol{T})$ (Asmussen, 2003). The infinitesimal generator of the underlying Markov chain of a phase-type distribution then takes the form

$$\mathbf{Q} = \begin{pmatrix} 0 & \mathbf{0} \\ \boldsymbol{t} & \boldsymbol{T} \end{pmatrix}, \tag{1.7}$$

where $\boldsymbol{t}$ is a column vector representing the transitions from the transient states to the absorption state and satisfies the relation $\boldsymbol{t} = -\boldsymbol{T}\mathbf{e}$. The basic characteristics of a phase-type distribution are

- the cumulative distribution function: $F(x) = 1 - \boldsymbol{\alpha} e^{\boldsymbol{T}} \mathbf{e}$,
- the density function: $f(x) = \boldsymbol{\alpha} e^{\boldsymbol{T}} \boldsymbol{t}$, and
- the nth moment: $M_n = (-1)^n n! \boldsymbol{\alpha} \boldsymbol{T}^{-1} \mathbf{e}$.

The simplest examples of phase-type distributions are mixtures and convolutions of exponential distributions; in particular Erlang distributions, defined as Gamma distributions with an integer shape parameter. More generally, the class of phase-type distributions comprises all series/parallel arrangements of exponential distributions, possibly with feedback (Fackrell, 2003). Since their first introduction, phase-type distributions became extremely popular, because they constitute a very versatile class

of distributions defined on the non-negative real numbers. Here, we explain the reasons why this class of distributions thrives in the area of applied probability.

Properties of phase-type distributions

First, phase-type distributions are a natural generalisation of the exponential distribution. Thus, stochastic models where the exponential distribution is used to model quantities (e.g. inter-arrival or service times) may admit extensions to phase-type distributions with some extra computational effort. In other words, often the underlying structure of a model can be preserved if the exponential distribution is simply replaced by a phase-type distribution. A typical example is the PH/PH/1 queue, which generalises the M/M/1 queue and can be analysed in an analogous manner.

Second, stochastic models involving phase-type distributions often lead to algorithmically tractable performance measures. The matrices involved in phase-type distributions consist entirely of real entries. Consequently, many performance measures, which are expressed in terms of the system parameters and exponentials of these matrices, can be implemented in algorithms and can be computed numerically with relative ease with the aid of a computer software. However, not only numerical performance measures can be calculated, but also qualitative performance measures can be established in stochastic models where phase-type distributions are used. For example, Takahashi (1981) showed that the waiting time (delay) distribution of a PH/PH/c queue has an exponential tail.

Third, the class of phase-type distributions is closed under a variety of operations (finite mixture, convolutions, superpositions). As a consequence, systems with phase-type inputs often generate phase-type outputs. This property is extremely useful especially when networks of systems are studied, where the output of one system is the input to another one. For example, Neuts (1994) showed that the stationary waiting time distribution in an M/PH/1 queue is phase-type, while Asmussen (1992a) extended this result to the G/PH/1 queue.

Fourth, the class of phase-type distributions is a proper subset of the class of distributions with rational *Laplace-Stieltjes transform* (LST) (Lipsky, 2009), which are widely known as matrix-exponential (ME) distributions (Asmussen and O'Cinneide, 2004; Bean et al., 2008; Fackrell, 2003, 2009). The ME distributions were first introduced by Cox (1955a,b). Since the LST – and also the *moment generating function* (m.g.f.) – of a ME distribution is expressed as a fraction of two polynomials, various performance measures have exact closed-form expressions (Asmussen and Bladt, 1997; Bladt and Neuts, 2003). Although ME distributions do not have simple probabilistic interpretations, performance measures involving ME distributions have ME representations. For example, the waiting time distribution of an G/ME/1 queue has a *zero-modified* (it has an atom at zero) ME representation (Asmussen and Bladt, 1997).

Finally, phase-type distributions are dense in the class of all distributions defined on the non-negative real numbers (Asmussen, 2000; Schassberger, 1973). In other words, any distribution on a positive support can be approximated arbitrarily close by a phase-type distribution in the sense of weak convergence. However, as Neuts (1994) remarked, there exist a number of simple distributions (e.g. the delayed exponential

distributions) where a prohibitive number of states is required to achieve a reasonable approximation by a phase-type distribution. On the other hand, the parameters of the underlying Markov chain that defines a phase-type distribution are very flexible, thus making the phase-type distributions exhibit a quite versatile behaviour (O'Cinneide, 1999).

Fitting techniques for phase-type distributions

Since phase-type distributions can approximate any distribution on a positive support arbitrarily close, many fitting techniques have been developed. These fitting techniques are either based on moment-matching (Johnson and Taaffe, 1989; Horváth and Telek, 2007) or on *maximum likelihood estimators* (MLEs) (Asmussen et al., 1996; Law et al., 1991). Moment matching techniques work efficiently, but they are not applicable in general. For example, Johnson and Taaffe (1989) found that there exist distributions that are impossible to approximate by a phase-type distribution so that the first three moments match exactly. Among the techniques that are based on MLEs, the most prevalent is the *expectation-maximisation* (EM) algorithm (Dempster et al., 1977). Other techniques based on MLEs methods to fit long-tailed distributions to special types of phase-type distributions like the Coxian (Horváth and Telek, 2000) and the hyperexponential (Feldmann and Whitt, 1998) distribution have also been suggested. Another example of an MLEs approach is a *divide-and-conquer technique* to fit data sets with non-monotone densities into a mixture of Erlang and hyperexponential distributions and also data sets with *completely monotone* densities into hyperexponential distributions (Riska et al., 2004). Finally, several tools have been developed to help with the task of fitting phase-type distributions to data. Some examples are: EMPHT (Asmussen et al., 1996), PhFit (Horváth and Telek, 2002), G-FIT (Thuümmler et al., 2006), and HyperStar (Reinecke et al., 2013).

1.3.3 Markovian Arrival Processes

The MArP was introduced by Neuts (1979). In broad terms, the idea of a MArP is to generalise the Poisson arrival process in a way to include non-exponential and/or dependent inter-arrival times, and correlated arrivals, but keep the tractability of the Poisson process. In Chapter 4, we consider approximations for the waiting time distribution of a MArP/G/1 queue. Therefore, in this section, we explain the mechanism of a MArP and we provide their properties and related results.

Similarly to phase-type distributions, a MArP is associated with a finite absorbing Markov chain. However, for its description it requires not only one initial probability vector $\boldsymbol{\alpha}$, but as many initial vectors as the number of transient states in the underlying Markov chain, i.e. one $\boldsymbol{\alpha}$ per transient state. To understand this, we explain here the mechanism of a MArP. When the Markov chain enters the absorption state a new customer arrives. Then, the process restarts from the transient part by remembering the last transient state that reached absorption. Consequently, a MArP is formally represented by two matrices $(\mathbf{D}_0, \mathbf{D}_1)$, where matrix $\mathbf{D}_0$ describes the interactions between the transient states of the underlying Markov chain and matrix $\mathbf{D}_1$ describes how the transient states of the underlying Markov chain are re-entered once absorption is reached.

For example, the (i, j) element of matrix $\mathbf{D}_0$ simply represents the rate at which the Markov chain moves from the transient state i to the transient state j. On the other hand, the same element in the $\mathbf{D}_1$ matrix represents the rate at which the Markov chain is absorbed from state i (where this absorption is accompanied by an arrival of a customer) and is instantaneously re-entered in the transient state j. All off-diagonal entries of $\mathbf{D}_0$ are non-negative, while matrix $\mathbf{D}_1$ has only non-negative entries. We also note that $\mathbf{D}_0 + \mathbf{D}_1$ is the infinitesimal generator describing the transitions of the transient states in the underlying Markov chain.

The class of MArPs is a very rich class of point processes, containing many well-known arrival processes in the applied probability literature as special cases. A special case of a MArP is the *Markov Modulated Poisson Process* (MMPP), which is a popular model for bursty arrivals (Fischer and Meier-Hellstern, 1993). Intuitively, a "burst" is a group of consecutive customers with shorter inter-arrival times than customers arriving before or after the burst. For MMPPs, matrix $\mathbf{D}_1$ has all entries zero, except for the diagonal ones. The class of MArPs contains also the class of phase-type renewal processes, i.e. renewal processes with phase-type inter-arrivals (Neuts, 1978). The phase-type distribution is a MArP where matrix $\mathbf{D}_1$ has equal rows. Moreover, the class of MArPs also allows for correlated renewal distributions and is closed under superposition, thinning, etc (Lucantoni, 1993).

A generalisation of the MArP is the *Batch Markovian Arrival Process* (BMArP), which was initially known under the name N-process; see (Neuts, 1979; Ramaswami, 1980). The BMArP extends the concept of MArP by allowing simultaneous batch arrivals of customers (batch absorptions) in the underlying Markov chain. Formally, a BMArP is represented by an infinite number (allowing for arbitrary large batch sizes) of matrices $\mathbf{D}_k$, $k \geq 0$, and its infinitesimal generator is $\sum_{k=0}^{\infty} \mathbf{D}_k$. The (i, j) element of matrix $\mathbf{D}_k$ has a similar meaning to the (i, j) element of matrix $\mathbf{D}_1$, where instead of an arrival of a single customer at absorption we have an arrival of a batch with size $k > 0$.

The introduction of BMArPs was motivated by the need to describe arrival processes that come from the superposition of many independent arrival streams of compound phase-type renewal processes, batch MMPPs, and batch Poisson processes. As the arrival streams are independent, it is natural to consider that the service time distributions of customers in each stream may differ from one another. Thus, BMArPs accommodate not only models with (correlated) batch arrivals (Lucantoni, 1991) but also models that bear dependencies between arrival and service processes (Takine and Hasegawa, 1994).

For the analysis of queueing models with MArP arrivals and i.i.d. service times, MAM have been a very popular tool (Lucantoni et al., 1990; Lucantoni, 1991; Ramaswami, 1980). If we substitute the Markovian arrival process in the M/G/1 queue to a MArP, then the MArP/G/1 queue lies in the context of M/G/1–type Markov processes (see Sections 1.2.2 and 1.3.1). The most important benefit by such a substitution is that the entire model behaves like a matrix generalisation of the M/G/1 queue. In fact, many expressions for performance measures of interest are natural matrix analogues of the corresponding expression in the M/G/1 queue. For example, it has been shown that the Laplace transform of the waiting time of a MArP/G/1 queue has a matrix expression analogous to the Pollazceck-Khinchine equation of an M/G/1 queue (Neuts, 1989; Ramaswami, 1980). Quite interestingly, many results that

hold for the MArP/G/1 queue hold true also for the BMArP/G/1 queue (Lucantoni, 1993).

Finally, stationary BMArPs are dense in the family of all stationary *Marked Point Processes* (MPPs) (Asmussen and Koole, 1993); overall, a MPP is a pure point process defined on the product space of points and marks. This shows that BMArPs can represent a wide range of behaviour although, from a practical point of view, the dimension of the matrices may be restrictive. Thus, many fitting techniques have developed for (B)MArPs, a brief overview of which we present below.

Fitting techniques for MArPs

Resembling the case of phase-type distributions, both moment-matching and MLE-based approaches exist for fitting data sets into (B)MArPs. The moment matching techniques on a large extend deal with two-state MMPPs (Gusella, 1991), while Meier-Hellstern (1987) and Rydén (1996) also proposed MLE algorithms for two-state MMPPs. Furthermore, the EM algorithm is used as an estimation procedure both for MArPs (Horváth and Okamura, 2013) and their more general form BMArPs (Breuer, 2000, 2002).

1.4 Truncations of the background state space

In the previous section, we gave an overview for MAM and their building blocks, phase-type distributions and MArPs. Particularly, in Section 1.3.1, we defined the auxiliary matrices $\mathbf{R}$ and $\mathbf{G}$, which are involved in the calculation of performance measures for block-structured Markov chains. Since the dimension of the phase space of a block-structured Markov chain determines the dimension of the matrices, the former should be finite. However, as we already mentioned in Section 1.1, there are a lot of practical situations where the phase space is infinite.

To overcome the issue with infinite phases, an appealing approach is to truncate the background state space and find approximate performance measures through the truncated model. Truncation of Markov chains, has already been considered before the appearance of MAM (Seneta, 1981). Seneta presents sufficient conditions for the convergence of the steady-state probability of the truncated Markov chain to the original Markov chain. He also provides an error bound expression, but this bound seems too cumbersome to be practical.

Two very classical examples with infinite phase spaces are models involving heavy-tailed distributions and queueing networks with more than one queue. In Chapters 2–4, we consider heavy-tailed models, while in Chapter 5 we deal with a tandem network of two queues. Thus, in Sections 1.4.1 and 1.4.2, we examine these two examples, respectively. For each of them, we give background literature and we explain how they relate to MAM.

1.4.1 Heavy-tailed models

In real-world applications, the input parameters – e.g. inter-arrival and service time distributions – are not explicitly known, but need to be estimated from data sets. There is abundance of evidence that the empirical distribution of many data sets

bears characteristics like *high variability*, i.e. its *coefficient of variation* is greater than the one of the exponential distribution (Bolotin, 1994; Duffy et al., 1994). Therefore, in such applications, the classical assumption of exponentially decaying probability distributions is not applicable (Embrechts et al., 1997).

An appropriate way to model highly-variable stochastic processes is by using heavy-tailed distributions (Sigman, 1999). Such distributions decay more slowly than any exponential function, which means that with such distributions there exists a non-trivial probability of an extremely large observation (Asmussen, 2003; Rolski et al., 1999). In the literature, the class of heavy-tailed distributions admits different definitions. For a detailed classification of heavy-tailed distributions, special cases, and their properties refer to Zwart (2001).

Data sets with heavy-tailed behaviour are commonly modelled by distributions such as Pareto, Weibull, and Lognormal. In general, heavy-tailed distributions have high coefficient of variation and sometimes infinite means. These characteristics increase the complexity of the analysis of queueing systems with such heavy-tailed input. Especially, the performance of queueing systems with heavy-tailed services is affected considerably, since measures like average queue length, average waiting times, and the corresponding distributions inherit the heavy-tailed behaviour (Feldmann and Whitt, 1998). Therefore, under the presence of heavy-tailed distributions, evaluations of performance measures become more challenging and sometimes even problematic (Ahn et al., 2012; Asmussen and Pihlsgård, 2005).

An attractive alternative to capture highly variable behaviour in data sets or distribution functions is by using phase-type distributions. Many approximation methods for heavy-tailed distributions have been proposed using special cases of phase-type distributions (Bladt et al., 2014; Feldmann and Whitt, 1998; Horváth and Telek, 2000; Sasaki et al., 2004; Starobinski and Sidi, 2000). Although queueing models become tractable and can be analysed with the aid of MAM, one disadvantage of phase-type approximations is that the accuracy of the performance measures cannot be pre-determined. Another drawback is that phase-type distributions require more parameters (in other words, phases) than Lognormal or Weibull distributions to capture the heavy-tailed behaviour, thus making estimation procedures more complex. On the contrary, when approximating some light-tailed distribution with a phase-type one, approximations for performance measures become highly accurate.

A different approach on evaluating performance measures for heavy-tailed models, is to study their asymptotic behaviour. When probability distributions belong to the class of *subexponential* distributions (Teugels, 1975), which is a special case of heavy-tailed distributions, asymptotic approximations are available (von Bahr, 1975; Borovkov and Foss, 1992; Embrechts and Veraverbeke, 1982; Foss et al., 2005; Olvera-Cravioto et al., 2011; Pakes, 1975). The main disadvantage of such approximations is that they provide a good fit only at the tail of the performance measure, especially in *heavy traffic*, i.e. when the occupation rate of the system is close to 1. Finally, results on error bounds (Kalashnikov, 2002) indicate that such bounds are rather pessimistic, especially in terms of relative errors, and in case of heavy traffic. There exist also bounds with the correct tail behaviour under subexponential claims (Kalashnikov and Norberg, 2002; Korshunov, 2011), but these bounds are only accurate at the tail.

A conclusion that can be safely drawn from all the above is that, although the literature is abundant with approximations for performance measures in the case of

light-tailed distributions, accurate approximations for performance measure in the case of heavy-tailed distributions are still an open topic. Therefore, in Chapters 2–4, our focus is on finding accurate approximations for performance measures of heavy-tailed models.

1.4.2 Networks of queues

Many important applications from manufacturing are involved with networks of systems, and thus the models describing these systems become too complex. For example, in Chapter 5, we study a simple network involving two queues in tandem. The state space includes the number of customers in front of each queue and is thus doubly-infinite. An approximation method is to truncate this state space by allowing only for a finite number of customers in front of the first or second queue, while any other customers who arrive and attempt to enter the queue are lost or block the system.

In general, infinite-buffered queueing networks seem to be more manageable (van Vuuren, 2007). This is due to the fact that the servers cannot get blocked and thus the departure processes from the stations are completely described by the arrival and service processes at the stations. In particular, there exist special cases of infinite-buffered queueing networks that admit product-form solutions (Latouche and Ramaswami, 1999). An example is a tandem queueing network with exponentially distributed service times in all stations, where customers arrive at the first queue according to a Poisson process, a model that is a special case of *Jackson networks* (Jackson, 1963).

Tandem queueing networks may involve general service time distributions and/or possibly finite buffers for different stations. Moreover, since traffic often exhibits correlations and burstiness, the arrival process should be able to capture these characteristics. However, queueing networks of this general type do not allow for an exact analysis. This is mainly due to concurrent non-exponential activities in combination with infinite state spaces.

A typical approach that may lead to a tractable model is to truncate the state space by considering finite buffers for some or all the stations. Although truncation seems a pragmatic approach, care is needed. For example, Kroese et al. (2004) analyse a simple two-node tandem Jackson network with exponential servers and Poisson arrivals directly and also approximate it via truncation techniques. They express the model as a QBD and they show that when the truncation threshold is converging to infinity, several quantities in the truncated system do not converge to the exact values. Moreover, they find that the decay rate of the truncated system may converge, but not necessarily to the decay rate of the original system.

1.5 Beyond Matrix-Analytic Methods

Although MAM are very powerful in the analysis of block-structured Markov chains, they are not the only techniques available for this class of Markov chains. Moreover, there exist more general Markov chains that MAM do not apply to. Therefore, in this section, we give a brief overview of other techniques.

As we mentioned in Section 1.3, technological advances were a catalyst to the popularity of MAM. Before the advent of fast computers, problems in stochastic modelling, particularly in queueing theory, were perceived as problems in analysis and their analysis relied on LST and complex analysis techniques; see for example Cohen (1982). Although these techniques are appropriate to produce numbers, they may not provide probabilistic insights to the system being analysed. Moreover, there exist situations where the evaluation of performance measures through LSTs is hard. An example is the LST of the waiting time distribution of the M/G/1 queue. Although this LST is given by the well-known *Pollaczek-Khinchine formula* (Asmussen, 2003; Asmussen and Albrecher, 2010), when the Laplace transform of the service times does not have a closed-form expression, then the Pollaczek-Khinchine formula cannot be used to find the waiting time distribution.

Generating functions have also been used for the analysis of queueing problems (Adan et al., 1993; Hofri, 1978). In particular, the equilibrium behaviour of two-dimensional Markov processes has been extensively studied by techniques that are based on generating functions. A classical example is the symmetric shortest queue problem, where Flatto and McKean (1977) and Kingman (1961) use uniformisation techniques to determine the generating function of the equilibrium of the joint queue length distribution. By using a more general approach, Cohen and Boxma (1983) and Fayolle and Iasnogorodski (1979) show that the analysis of the functional equation for the generating function can be reduced to that of a *Riemann-Hilbert boundary value problem*.

The analysis of the symmetric shortest queue problem initiated also the development of the *compensation approach* (Adan, 1994). The equilibrium probabilities of the shortest queue are well-known to satisfy two types of equations: the *equations in the interior points* and the *boundary conditions*. In broad, the compensation approach first finds the set of product form solutions that satisfy the equations in the interior points, and afterwards constructs a linear product-form solution from the latter set that satisfies also the boundary conditions. The name compensation is due to the fact that after introducing the first term, terms are added alternatively so as to compensate for errors on the two boundaries. The compensation approach leads to an explicit characterisation of the equilibrium probabilities, which can easily be exploited for the development of efficient algorithms, with the advantage of tight error bounds.

Queueing models with MArP arrivals and i.i.d. service times have also been analysed with MAM (Lucantoni et al., 1990; Lucantoni, 1991; Ramaswami, 1980). However, when service times of customers depend on the state of the underlying Markov chain upon arrival, MAM can no longer be used to analyse such a system. This is because the queue length and the state of the underlying Markov chain do not form a Markov chain. Therefore, alternative methods to obtain performance measures in queueing models with state-dependent service time distributions have been developed; e.g. by examining the *ladder height* distribution (Asmussen, 1991; Asmussen and Perry, 1992) or by a matrix-factorisation method that heavily relies on complex-plane methods (Regterschot and de Smit, 1986).

Perturbation analysis is another technique that has been used in the analysis of Markov chains (Altman et al., 2004). In broad terms, it involves the study of Markov chains where "small" changes in its *kernel* have taken place. Perturbation analysis has been used successfully to obtain error bounds for performance measures (Haviv and

Ritov, 1993; Heidergott et al., 2009) and puts emphasis on the analysis of *singular* perturbed Markov chains, i.e. the perturbation breaks down the class structure of the original chain (Avrachenkov and Lasserre, 1999). Approximations involving series expansions or performance measures have also been proposed. An example is the *Edgeworth series expansion* (Wallace, 1958), which is a refinement of the central limit theorem.

Another stream of research focuses on *corrected diffusion approximations* for stochastic processes (Blanchet and Zwart, 2010; Silvestrov, 2004). These approximations can be useful in applications where moments are computable, but the distribution is not. A numerically-oriented method that applies to fairly general exponential multidimensional queueing systems bas been developed by Hooghiemstra et al. (1988). This method is based on calculating a power-series expansions for the equilibrium probabilities as functions of the traffic intensity, and there is evidence that it works satisfactory for several queueing problems (Blanc, 1987, 1991, 1992).

Finally, a numerical approach that is widely used for the performance analysis of stochastic networks is *simulation*. In simulation models, the events that could occur while a system operates by following a sequence of steps, are generated by a computer program. The probabilistic nature of many events, such are arrivals of customers and service times, can be represented by sampling from a distribution that reflects the pattern with which the events occur. Thus, to capture the typical behaviour of a system, it is necessary to run the simulation model for a sufficiently long time, so that all events can occur a sufficiently large number of times. In principle, simulation models can describe whatever level of complexity is desired. However, simulation of heavy-tailed distributions for estimation of steady-state measures is not easy, as the simulation must run exceptionally long in order to capture the effect of the distribution tail, i.e., the rare events, which even with a small probability of occurrence can affect the system performance significantly.

1.6 Our contribution

This section gives an overview of the results in this dissertation, which is organised according to the different model that is studied. In Chapters 2 and 3, we consider the classical *Cramér-Lundberg risk model* (Asmussen and Albrecher, 2010; Prabhu, 1961). In this model, we have claims (for money) which arrive to an insurance company according to a Poisson process and the total income (premium) rate is 1. Given that the insurance company starts with some initial capital, we are interested in evaluating the ultimate ruin probability that the risk reserve ever drops below zero. More precisely, we focus on heavy-tailed *claim sizes*, which are known to make numerical evaluations of the ruin probability challenging. In each chapter, we follow a different approach to find numerical evaluations of the ruin probability under heavy-tailed claim sizes. Due to the duality between the probability of eventual ruin for an insurance company with an initial capital u and the stationary waiting probability $\mathbb{P}(W > u)$ of a G/G/1 queue, where service times in the queueing model correspond to the random claim sizes (Asmussen, 2003; Asmussen and Albrecher, 2010), our approximations are also valid for this model.

An attractive way to overcome the problem with heavy-tailed distributions is to approximate the claim sizes with a phase-type distribution. However, it is not

clear how many phases are enough in order to achieve a specific accuracy in the approximation of the ruin probability. In Chapter 2, we investigate the number of phases required so that we can achieve a pre-specified accuracy for the ruin probability and we provide error bounds. Also, in the special case of a *completely monotone* claim size distribution we develop an algorithm to estimate the ruin probability by approximating the *excess claim size distribution with* a hyperexponential one. Finally, we compare our approximation with two typical approximations for the ruin probability, i.e. the *heavy tail* and the *heavy traffic* approximations, by performing an extensive numerical study.

On the other hand, in Chapter 3, we provide approximations for the ruin probability by combining both phase-type distributions and asymptotic results. Motivated by statistical analysis, we describe how the claim sizes can be written as a mixture of a phase-type and a heavy-tailed distribution. From this representation of the claim size distribution, we derive with the aid of perturbation analysis a series expansion for the ruin probability. Our proposed approximations consist of the first two terms of this series expansion, where the first term has a phase-type representation. We refer to our approximations collectively as *corrected phase-type approximations* and we prove that they provide small absolute and relative errors. Finally, we show that the corrected phase-type approximations exhibit such a nice behaviour both in finite and infinite time horizon, and we check their accuracy through numerical experiments. The results of Chapters 2 and 3 are based on the research papers Vatamidou et al. (2014a) and Vatamidou et al. (2013b), respectively.

In Chapter 4, we investigate the applicability of the corrected phase-type approximations to a more involved queueing model. In particular, we consider a single server queue with FIFO discipline where customers arrive according to a MArP and their service times follow some general distribution. We explain how the general distribution can be written as a mixture of a phase-type and a heavy-tailed distribution, and with the aid of perturbation analysis we derive approximations for the queueing delay. We show that the developed approximations capture the exact tail behaviour and provide bounded relative errors. Moreover, we exhibit their performance with numerical examples. The results of this chapter are based on Vatamidou et al. (2013a, 2014b)

Finally in Chapter 5, we consider the $\mathrm{M}^X/\mathrm{M}/1 \rightarrow \bullet/\mathrm{M}/1$ tandem queueing network. Customers arrive in batches according to a Poisson stream and join the first queue, while the service times in each queue are exponential. A customer leaves the system after finishing service in both queues. In this model, the joint queue lengths can be represented by a QBD with an infinite phase space. However, MAM can be applied to find approximations for the joint queue length distribution only if the buffer size in front of either queue is finite. Therefore, we truncate the buffer size of the first queue and we find an asymptotic upper bound for our approximations. To derive the bound, we connect our two-dimensional queueing process with a two-dimensional random walk and with the aid of *large deviations theory* (LDT), we recognise three possible cases for the bound.

1.7 Notation

While each chapter focuses on a different model, we follow some general rules with respect to the notation throughout the whole dissertation, which we give below.

1.7.1 General symbols

The standard sets are denoted as follows:

$\mathbb{N} = \{0, 1, 2, \dots\}$	the natural numbers,
$\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$	the integer numbers,
$\mathbb{R} = (-\infty, \infty)$	the real numbers (or real line),
$\mathbb{C} = \{x + \mathrm{i}y : x, y \in \mathbb{R}\}$	the complex numbers.

We also introduce the notation $\mathbb{N}_n$ for the natural numbers up to the integer n, i.e. $\mathbb{N}_n = \{0, 1, \dots, n\}$.

For a real number x, we call *ceiling* the smallest integer that is greater than or equal to x and we denote it as $\lceil x \rceil$. On the other hand, we call the integer part of x *floor* and we denote it as $\lfloor x \rfloor$. Moreover, we define the function $(x)^+ := \max\{0, x\}$ and we use $\ln x$ for the natural logarithm of x.

Some other symbols that we use are $\mathbb{P}$ and $\mathbb{E}$ for the probability and the expectation, respectively. We also use the notation $\mathbb{1}$ for the indicator function and we denote the *Kronecker delta* as δ_{ij}, i.e. $\delta_{ij} = 0$ when $i \neq j$ and $\delta_{ij} = 1$ when $i = j$. Finally, the real part of a complex number s is denoted by $\Re(s)$.

1.7.2 Distributions and random variables

Consider a *cumulative distribution function* (c.d.f.) F. The total mass of the distribution is denoted as $\|F\|$ and it usually holds that $\|F\| = 1$. If $\|F\| < 1$, the distribution F is called *defective*. Its *complementary cumulative distribution function* (c.c.d.f.) is denoted by $\overline{F}$ and its nth convolution by F^{*n}. Furthermore, we denote as $\widetilde{F}$ its LST and as $\widehat{F}$ the approximation of F. In addition, if μ is the mean of the distribution F, we define its *stationary excess distribution* F^e as

$$F^e(x) = \frac{1}{\mu} \int_0^x \overline{F}(y) dy.$$

Some typical distributions are the exponential, the Erlang, the hyperexponential, and the geometric. Therefore, we denote as $E_k(\lambda)$ the Erlang distribution with k phases and rate parameter λ. For simplicity, we write $E(\lambda)$ for the exponential distribution with rate parameter λ. Furthermore, we use the notation H_k for a hyperexponential distribution with k phases. Finally, we write $G(p)$ for the geometric distribution with success probability $p \in (0, 1]$ and *probability distribution function* (p.d.f) defined as $f(k) = (1-p)^{k-1}p$, $k = 1, 2, \dots$ By convention, we use the notation $E(\lambda)$, $E_k(\lambda)$, H_k, and $G(p)$, for the random variable (r.v.) that corresponds to each respective distribution.

If two r.v.'s A, B are equal in distribution we write $A \stackrel{\mathfrak{D}}{=} B$. If the r.v. A follows the distribution D, we write $A \sim D$. The symbol $\sim$ is also used between two functions as $f(x) \sim g(x)$ to describe the relation $\lim_{x\to\infty} f(x)/g(x) = 1$; the meaning will always be clear from the context. Finally, the sup-norm distance between two distributions F_1 and F_2 on a positive support is defined as $\mathcal{D}(F_1, F_2) := \sup_x |F_1(x) - F_2(x)|$, $x \geq 0$. Also, the sup-norm distance between two random variables is defined to be the sup-norm distance between their distributions.

1.7.3 Matrices and vectors

We denote the matrices and vectors with boldface. In addition, we use the superscript T to denote the transpose of a vector or a matrix; e.g. $\mathbf{A}^T$ is the transpose of matrix $\mathbf{A}$. Moreover, we denote the rank of $\mathbf{A}$ as rank$\mathbf{A}$, while for a square matrix $\mathbf{A}$, we denote its determinant as $\det \mathbf{A}$.

Suppose now that $\mathbf{A}$ is a square matrix of dimension n and that $U, W \subset \{1, \ldots, n\}$, where the symbol "$\subset$" does not imply strict subsets. Then $\mathbf{A}_U^W$ is the sub-matrix of $\mathbf{A}$ if we keep the rows in U and the columns in W. Whenever the notation becomes very complicated, to avoid any confusion with the indices, we denote the ith column and row of matrix $\mathbf{A}$ with $\mathbf{A}_{\bullet i}$ and $\mathbf{A}_{i\bullet}$, respectively.

Two more matrix operations that we use are the Hadamard product and a new operator we introduce for the needs of Chapter 4. More precisely, we use the operator $\circ$ for the *Hadamard product* between two matrices of same dimensions; i.e. if $\mathbf{B} = (b_{ij})$ and $\mathbf{C} = (c_{ij})$ are $m \times n$ matrices, then the (i, j) element of the $m \times n$ matrix $\mathbf{B} \circ \mathbf{C}$ is equal to $b_{ij}c_{ij}$.

For the second operator, suppose that $\mathbf{A}$ and $\mathbf{B}$ are two square matrices of dimension n and $\Omega \subset \{1, \ldots, n\}$. If U and W are two disjoint sets such that $U \cup W \subset \Omega$, we use the notation $\mathbf{A}_\Omega^U \bowtie \mathbf{B}_\Omega^W$ for the matrix that has as columns the union of the columns U of matrix $\mathbf{A}$ and the columns W of matrix $\mathbf{B}$, ordered according to the index set $U \cup W$. For example, if $n = 6$, $\Omega = \{1, \ldots, 6\}$, $U = \{1, 2, 4\}$, and $W = \{3, 5\}$, then $\mathbf{A}_\Omega^{\{1,2,4\}} \bowtie \mathbf{B}_\Omega^{\{3,5\}} = (\mathbf{A}_{\bullet 1}, \mathbf{A}_{\bullet 2}, \mathbf{B}_{\bullet 3}, \mathbf{A}_{\bullet 4}, \mathbf{B}_{\bullet 5})$. Note that since the sets U and W are disjoint, there is a unique ordering of their union, which means that $\mathbf{A}_\Omega^U \bowtie \mathbf{B}_\Omega^W = \mathbf{B}_\Omega^W \bowtie \mathbf{A}_\Omega^U$.

Finally, the dimensions of the matrices (square or not) are clear from the context. Therefore, we do not use any indices to indicate their dimension. Some standard matrices and vectors are the following:

e: is the column vector with all elements equal to 1,

e$_i$: is a column vector with the element in position i equal to 1 and all other elements zero,

I: stands for the (square) identity matrix,

U: is the matrix with appropriate dimensions and all its elements equal to one, and

0: is the matrix with appropriate dimensions and all its elements equal to 0.

CHAPTER 2

Spectral approximations

2.1 Introduction

In this chapter, we consider the classical Cramér-Lundberg risk model, described in Section 1.6, and we find approximations for the ruin probability. Since we assume that the claim sizes arrive according to a Poisson process, the ruin probability can be found by using the well-known Pollaczek-Khinchine formula; see Eq. (2.1). This formula involves the convolutions of the excess claim size distribution, which cannot be easily computed, and thus one usually resorts to Laplace transforms. However, as we mentioned in Section 1.5, a major difficulty when analysing models with heavy-tailed distributions is that Laplace transforms of such distributions often do not have an analytic closed form. This is, in particular, the case for the Pareto and Weibull distributions. Thus, analytic methods, which use the Laplace transform of the claim sizes, are difficult (Abate and Whitt, 1999a) or even impossible to use in such cases.

As we explained in Section 1.4.1, a natural approach to provide approximations for the ruin probability is by approximating the claim size distribution with a phase-type one. We refer to these methods as *phase-type approximations*, because the approximate ruin probability has a phase-type representation (Asmussen, 1992a; Ramaswami, 1990). The main advantage of approximating a heavy-tailed claim size distribution with a phase-type distribution is that, in the latter case, the Laplace transform of the claim sizes has a closed form. However, in these cases, the exponential decay of phase-type approximations gives a big relative error at the tail and the evaluation of the ruin probability becomes more complicated.

Two approximations of the ruin probability based on the *safety loading*, which is defined as the relative amount by which the premium rate exceeds the average amount of claim per unit time, are the heavy traffic and light traffic. If, on average, the premiums exceed only slightly the expected claims then most appropriate for modelling is the heavy traffic approximation (Kalashnikov, 1997; Kingman, 1962). The drawback of this approximation though is that it requires finite first two moments for the claim

size distribution, a condition which may not be satisfied for several heavy-tailed distributions. On the other hand, when on average, the premiums are much larger than the expected claims, the light traffic approximation is used (Asmussen, 1992b; Bloomfield and Cox, 1972; Daley and Rolski, 1984, 1991; Sigman, 1992). However, in many applications, heavy traffic is most often argued to be the typical case rather than light traffic, which makes the light traffic approximation only of limited interest.

A particularly effective approach in handling distributions with infinite moments is the *Transform Approximation Method* (TAM). For some heavy-tailed distributions, such as the Pareto, (higher-order) moments may be infinite, thus making conventional moment-matching methods fail. The Laplace transform of a positive definite distribution, like the claim size distribution, exists always even if it does not have a closed analytic form. The TAM is based on the idea of approximating the Laplace transform of the claim sizes rather than directly their distribution (Harris et al., 2000; Harris and Marchal, 1998; Shortle et al., 2004). A drawback of this method though is that the accuracy of the approximation of the ruin probability cannot be predetermined. Finally, two known approximations for the ruin probability, which are based on the idea of matching the moments of a probability distribution, are the *Beekman-Bower's* (Beekman, 1969) and the *De Vylder's* (De Vylder, 1978) approximations.

In this chapter, we develop a new approach for approximating the ruin probability, when the claim sizes follow a heavy-tailed distribution. From the Pollaczek-Khinchine formula (see Section 2.2) it is clear that in order to evaluate the ruin probability, we only need to have a closed analytic form for the Laplace transform of the excess claim size distribution. For this reason, instead of approximating the claim size distribution, we approximate directly the stationary excess distribution with a hyperexponential distribution, a special case of a phase-type distribution. Since the Laplace transform of a hyperexponential distribution exists in a closed analytic form, we can numerically evaluate the ruin probability by inverting its Laplace transform.

An advantage of our approximation, which we call the *spectral approximation*, is that it has a predetermined accuracy. Thus, we first choose the accuracy we want to achieve in our approximation, and later on we determine the number of states for the hyperexponential distribution that are sufficient to guarantee this accuracy. Another interesting feature is that the bound that we guarantee is valid for the whole domain of the ruin probability and not only for a subset of it, contrary to other bounds that exist in the literature (Kalashnikov and Tsitsiashvili, 1999; Starobinski and Sidi, 2000).

Outline

In Section 2.2.1, we find bounds for the nth convolution of the excess claim size distribution. We prove that the bound for the convolution is linear with respect to the chosen accuracy for the excess claim size distribution. We also give the main result of this chapter, which is the error bound for the ruin probability.

Later, we focus on a class of heavy-tailed distributions that are in addition completely monotone, and we show that we can always approximate a completely monotone distribution with a hyperexponential one for any desired accuracy. We also prove that if the claim size distribution is completely monotone with finite mean, then the stationary excess distribution is also completely monotone. Furthermore, in Section 2.3, we present the steps of the *spectral approximation algorithm*, which approximates a com-

pletely monotone excess claim size distribution with a hyperexponential distribution for any desired accuracy.

Later on, we also compare the spectral approximation with the heavy traffic and the *heavy tail* approximations, where the latter is an asymptotic approximation for the ruin probability under subexponential claim sizes (Embrechts and Veraverbeke, 1982; Olvera-Cravioto et al., 2011). Thus, in Section 2.4, we give the basic characteristics of the latter two approximations and mention their advantages and disadvantages.

We devote Section 2.5 to numerical results. We do a series of experiments in order to compare the spectral approximation with the heavy traffic and the heavy tail approximations. As test distributions we use the Pareto, the Weibull, and a class of long-tail distributions introduced in Abate and Whitt (1999b). In addition, we extend a bound that is given in the literature (Brown, 1990) for the heavy traffic approximation to a specific case of the heavy traffic approximation that we use in our experiments. Finally, in Section 2.6, we discuss the results.

2.2 Spectral approximation for the ruin probability

In this section, we first introduce the mathematical framework and the basic terminology related to the classical Cramér-Lundberg risk model. We start our description with the *risk reserve process*. In broad terms, a risk reserve process $\{R(t)\}_{t\geq 0}$ is a model for the time evolution of the reserve of an insurance company, where the initial reserve is denoted by $u = R(0)$. In this model, claims arrive according to a Poisson process $\{N(t)\}_{t\geq 0}$ with rate λ. The claim sizes $U_1, U_2, \ldots$ are i.i.d. with common distribution G and independent of $\{N(t)\}$, and premiums flow in at a rate 1 per unit time. Putting all these together we see that

$$R(t) = u + t - \sum_{k=1}^{N(t)} U_k.$$

For mathematical purposes, it is frequently more convenient to work with the *claim surplus process* $\{S(t)\}_{t\geq 0}$, which is defined as $S(t) = u - R(t)$; as one can see from the expression above, this is merely a compound Poisson process with positive jumps and negative drift, a process well studied in the literature. The probability $\psi(u)$ of ultimate ruin is the probability that the reserve ever drops below zero, or equivalently the probability that the maximum $M = \sup_{0\leq t<\infty} S(t)$ ever exceeds u; i.e.

$$\psi(u) = \mathbb{P}(M > u).$$

Since we consider Poisson arrivals for the claims, for the evaluation of the ruin probability, the well-known Pollaczek-Khinchine formula (Asmussen, 2003; Asmussen and Albrecher, 2010) can be used:

$$1 - \psi(u) = (1-\rho)\sum_{n=0}^{\infty} \rho^n \left(G^e\right)^{*n}(u), \tag{2.1}$$

where $\rho < 1$ is the average amount of claims per unit time. Moreover, G^e is the stationary excess claim size distribution, which is defined as

$$G^e(u) = \frac{1}{\mathbb{E}U}\int_0^u \overline{G}(x)dx,$$

where $\mathbb{E}U$ is the (finite) mean of the claim sizes. The nth moment of the claim sizes is denoted by $\mathbb{E}U^n$.

For the evaluation of $\psi(u)$, Eq. (2.1) is not entirely satisfying because the infinite sum of convolutions at the right-hand side of the formula cannot be easily computed analytically and sometimes not even numerically. In order to overcome this difficulty we use Laplace transforms, which convert convolutions of distributions into powers of their Laplace transform. In terms of Laplace transforms, the Pollaczek-Khinchine formula can be written as

$$\widetilde{m}(s) := \mathbb{E}e^{-sM} = (1-\rho)\sum_{n=0}^{\infty}\rho^n\left(\widetilde{G}^e(s)\right)^n = \frac{1-\rho}{1-\rho\widetilde{G}^e(s)}. \tag{2.2}$$

From Eq. (2.2) it is clear why it is necessary to have a closed analytic form only for the Laplace transform of the *excess* claim size distribution, rather than the claim size distribution itself. Thus, the main idea of our algorithm is to approximate the excess claim size distribution with a phase-type distribution, which has a closed analytic Laplace transform, and apply Laplace inversion to evaluate the ruin probability.

Remark 2.1. In general, the premium rate of the Cramér-Lundberg risk model, say p, is not equal to 1. Here, we assumed that $p = 1$ in order to make our model comparable to the M/G/1 queue. However, when $p \neq 1$, we can normalise it to 1 by defining the risk reserve process $\check{R}(t) := R(t/p)$ and adapting the Poisson parameter. According to Asmussen and Albrecher (2010, Proposition I.1.3), the ultimate ruin probabilities in the original model and its time-scaled version coincide. Thus, without loss of generality, we may always assume that $p = 1$.

2.2.1 Error bound for the ruin probability

In this section, we provide a bound for the ruin probability when we approximate the excess claim size distribution with a known distribution, e.g. a phase-type distribution. If we approximate G^e with a known distribution (not necessarily a phase-type) then we can compute the ruin probability through the Pollaczek-Khinchine formula (2.1). From Eq. (2.1) and the triangular inequality, the error between the ruin probability and its approximation is then

$$\begin{aligned}\left|\psi(u) - \widehat{\psi}(u)\right| &= \left|\sum_{n=0}^{\infty}(1-\rho)\rho^n\left(\left(G^e\right)^{*n}(u) - \left(\widehat{G}^e\right)^{*n}(u)\right)\right| \\ &\leq \sum_{n=0}^{\infty}(1-\rho)\rho^n\left|\left(G^e\right)^{*n}(u) - \left(\widehat{G}^e\right)^{*n}(u)\right|,\end{aligned} \tag{2.3}$$

where $\widehat{\psi}$ is the exact result we obtain from the Pollaczek-Khinchine formula for the ruin probability when we use an approximate claim size distribution. From Eq. (2.3) we see that as a first step to find a bound for the ruin probability is to find a bound for the difference $\left|\left(G^e\right)^{*n}(u) - \left(\widehat{G}^e\right)^{*n}(u)\right|$. This is given in the following proposition.

Proposition 2.2. *If* $\sup_x\left|G^e(x) - \widehat{G}^e(x)\right| \leq \eta$ *for* $x \in [0,u]$, *then*

$$\left|\left(G^e\right)^{*n}(u) - \left(\widehat{G}^e\right)^{*n}(u)\right| \leq n\eta.$$

Proof. We prove this by induction. For $n = 2$,

$$\begin{aligned}\left|\left(G^e\right)^{*2}(u) - \left(\widehat{G}^e\right)^{*2}(u)\right| &= \left|G^e * G^e(u) - \widehat{G}^e * G^e(u) + \widehat{G}^e * G^e(u) - \widehat{G}^e * \widehat{G}^e(u)\right| \\ &\leq \left|\left(G^e - \widehat{G}^e\right) * G^e(u)\right| + \left|\left(G^e - \widehat{G}^e\right) * \widehat{G}^e(u)\right| \\ &\leq \int_0^u \underbrace{\left|\left(G^e - \widehat{G}^e\right)(u-x)\right|}_{\leq \eta} dG^e(x) + \int_0^u \underbrace{\left|\left(G^e - \widehat{G}^e\right)(u-x)\right|}_{\leq \eta} d\widehat{G}^e(x) \\ &\leq \eta G^e(u) + \eta \widehat{G}^e(u) \leq 2\eta.\end{aligned}$$

Assume now that the bound holds for a fixed n. We prove that it also holds for $n+1$.

$$\begin{aligned}&\left|\left(G^e\right)^{*(n+1)}(u) - \left(\widehat{G}^e\right)^{*(n+1)}(u)\right| \\ &\quad= \left|G^e * \left(G^e\right)^{*n}(u) \pm \widehat{G}^e * \left(G^e\right)^{*n}(u) - \widehat{G}^e * \left(\widehat{G}^e\right)^{*n}(u)\right| \\ &\quad\leq \left|\left(G^e - \widehat{G}^e\right) * \left(G^e\right)^{*n}(u)\right| + \left|\widehat{G}^e * \left((G^e)^{*n} - (\widehat{G}^e)^{*n}\right)(u)\right| \\ &\quad\leq \int_0^u \underbrace{\left|\left(G^e - \widehat{G}^e\right)(u-x)\right|}_{\leq \eta} d\left(G^e\right)^{*n}(u) \\ &\qquad + \int_0^u \underbrace{\left|\left((G^e)^{*n} - (\widehat{G}^e)^{*n}\right)(u-x)\right|}_{\leq n\eta} d\widehat{G}^e(u) \\ &\quad\leq \eta\left(G^e\right)^{*n}(u) + n\eta\left(\widehat{G}^e\right)^{*n}(u) \leq (n+1)\eta.\end{aligned}$$

□

In words, Proposition 2.2 says that if we bound the excess claim size distribution with some accuracy η, then a bound for its nth convolution is linear with respect to this accuracy η. Consequently, from Proposition 2.2, we have the following result.

Proposition 2.3. *If* $\sup_x \left|G^e(x) - \widehat{G}^e(x)\right| \leq \eta$ *for* $x \in [0, u]$, *then a bound for the ruin probability is*

$$\left|\psi(u) - \widehat{\psi}(u)\right| \leq \frac{\eta\rho}{1-\rho}.$$

Proof.

$$\begin{aligned}\left|\psi(u) - \widehat{\psi}(u)\right| &\leq \sum_{n=0}^{\infty}(1-\rho)\rho^n \left|\left(G^e\right)^{*n}(u) - \left(\widehat{G}^e\right)^{*n}(u)\right| \\ &\leq \sum_{n=0}^{\infty}(1-\rho)\rho^n n\eta = \eta\rho(1-\rho)\sum_{n=0}^{\infty} n\rho^{n-1} \\ &= \eta\rho(1-\rho)\cdot\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) = \eta\rho(1-\rho)\frac{1}{(1-\rho)^2} \\ &= \frac{\eta\rho}{1-\rho}.\end{aligned}$$

□

Notice that the bound in Proposition 2.3 is independent of u, for all $u \geq 0$. Thus, we conclude that the sup-norm distance between the ruin probability and its approximation is bounded as $\mathcal{D}(\psi, \widehat{\psi}) \leq \eta\rho/(1-\rho)$, whenever $\mathcal{D}(G^e, \widehat{G}^e) \leq \eta$. Observe that the term $1-\rho$ at the denominator has as consequence that, higher load ρ requires a more accurate approximation of the G^e to obtain tight bounds for the ruin probability.

To sum up, when the excess claim size distribution is approximated with some desired accuracy η, then a bound for the ruin probability, which is linear with respect to η, is guaranteed by Proposition 2.3. Thus, our next goal is to develop a way to approximate the excess claim size distribution with a hyperexponential one, a particular case of a phase-type distribution, with any desired accuracy. We complete this step in the next section.

2.2.2 Completely monotone claim sizes

We are interested in evaluating the ruin probability when the claim sizes follow a heavy-tailed distribution, such as Pareto or Weibull. These two distributions belong also to the class of *completely monotone* (c.m.) distributions, which is defined below.

Definition 2.4. A p.d.f. is said to be c.m. if all derivatives of f exist and if

$$(-1)^n f^{(n)}(u) \geq 0 \text{ for all } u > 0 \text{ and } n \geq 1.$$

Completely monotone distributions can be approximated arbitrarily close by hyperexponentials (Feldmann and Whitt, 1998). Here, we provide a method to approximate a completely monotone excess claim size distribution with a hyperexponential one in order to achieve any desired accuracy for the ruin probability. The following result is standard; see e.g. Feller (1971).

Theorem 2.5. *A p.d.f. f is called c.m. if and only if it is a mixture of exponential p.d.f.'s. That is,*

$$f(u) = \int_0^{+\infty} ye^{-yu} dA(y), \ u \geq 0,$$

for some proper c.d.f. A on a positive support. We call A the spectral *c.d.f. For the tail or the c.c.d.f. $\overline{F}$ of a c.m. distribution it holds that*

$$\overline{F}(u) = \int_u^{+\infty} f(x)dx = \int_0^{+\infty} \int_u^{+\infty} ye^{-yx} dx dA(y) = \int_0^{+\infty} e^{-yu} dA(y).$$

An alternative way to define a c.m. distribution is by using Laplace transforms. From Theorem 2.5, it is obvious that a p.d.f. f is c.m. if its tail can be written as the Laplace transform of some positive-definite distribution A. The following lemma is an immediate consequence.

Lemma 2.6. *If the claim size distribution G is c.m. then the excess claim size distribution G^e is c.m. too.*

Proof. If G is a completely monotone distribution, then for some spectral function A it holds that $\overline{G}(x) = \int_0^{+\infty} e^{-yx} dA(y)$. Thus,

$$\overline{G^e}(u) = \frac{1}{\mathbb{E}U} \int_u^{+\infty} \overline{G}(x)dx = \frac{1}{\mathbb{E}U} \int_u^{+\infty} \int_0^{+\infty} e^{-yx} dA(y)dx$$

$$= \frac{1}{\mathbb{E}U} \int_0^{+\infty} dA(y) \int_u^{+\infty} e^{-yx} dx = \int_0^{+\infty} e^{-yu} \frac{dA(y)}{y\mathbb{E}U} = \int_0^{+\infty} e^{-yu} dA^e(y),$$

where $dA^e(y) = \dfrac{dA(y)}{y\mathbb{E}U}$. □

In this chapter, we are interested in finding a bound for the excess claim size distribution. In order to achieve our goal, we approximate the spectral function of the excess claim size distribution by a step function with some fixed (and pre-determined) accuracy η and then calculate the error of the approximation for the excess claim size distribution itself.

Lemma 2.7. *Let A^e be the spectral function of the c.m. excess claim size distribution G^e and let the step function $\widehat{A}^e$ satisfy $\mathcal{D}(A^e, \widehat{A}^e) \leq \eta$. Then, $\mathcal{D}(G^e, \widehat{G}^e) \leq \eta$, where $\widehat{G}^e$ is the c.m. distribution with spectral function $\widehat{A}^e$.*

Proof. Since the spectral c.d.f. A^e is proper, we have by definition that it has no atom at 0 and that it is right continuous. Thus, $A^e(0) = 0$ and $A^e(+\infty) = 1 < \infty$. Then it holds that

$$\begin{aligned}\int_0^{+\infty} e^{-uy} dA^e(y) &= e^{-uy} A^e(y) \Big|_0^{+\infty} - \int_0^{+\infty} A^e(y) de^{-uy} \\ &= \int_0^{+\infty} u e^{-uy} A^e(y) dy.\end{aligned}$$

Suppose now that $\mathcal{D}(A^e, \widehat{A}^e) \leq \eta$. Then

$$\begin{aligned}\left|G^e(u) - \widehat{G}^e(u)\right| &= \left|\int_0^{+\infty} \left(A^e(y) - \widehat{A}^e(y)\right) u e^{-uy} dy\right| \\ &\leq \int_0^{+\infty} \underbrace{\left|A(y) - \widehat{A}^e(y)\right|}_{\leq \eta} u e^{-uy} dy \leq \eta,\end{aligned}$$

for all $u \geq 0$. So, $\mathcal{D}(G^e, \widehat{G}^e) \leq \eta$. □

By definition, a hyperexponential distribution with l phases is a c.m. distribution with spectral function a step function with l jumps. Summarising, if we want to approximate the claim size distribution with a hyperexponential with some fixed accuracy η, it is sufficient to approximate the spectral c.d.f. of the c.m. excess claim size distribution with a step function with the same accuracy. In Section 2.3, we present in detail our algorithm to approximate the ruin probability with guaranteed error bound δ by approximating the claim size distribution with accuracy of at most $\eta = \delta(1-\rho)/\rho$, a result which is a consequence of Proposition 2.3. The exact relation between the number of phases, the accuracy η, and the bound δ is given also in the same section.

2.3 Algorithm for the spectral approximation

We consider the compound Poisson model introduced in Section 2.2, where we assume that the claim size distribution G is c.m. with finite mean. We denote by A^e the spectral function (strictly increasing distribution) of the excess claim size distribution G^e. Here, we develop an algorithm to evaluate the ruin probability by approximating the excess claim size distribution with a hyperexponential one, where all phases have equal weights.

Before we present the spectral approximation algorithm, it is necessary to give an important property on which the Laplace inversion of the ruin probability will be based on. The Pollaczeck-Khinchine formula (2.2) can be written equivalently in the form

$$\widetilde{m}(s) = 1 - \rho + \rho \frac{(1-\rho)\widetilde{G}^e(s)}{1 - \rho\widetilde{G}^e(s)} = 1 - \rho + \rho\widetilde{m}_+(s),$$

where $\widetilde{m}_+(s)$ is the Laplace transform of M_+, which is defined by $\mathbb{P}\big(M_+ \in A\big) = \mathbb{P}\big(M \in A \;\big|\; M > 0\big)$ for $A \subset [0,\infty)$. We have the following lemma.

Lemma 2.8. *If G^e follows a hyperexponential distribution with l phases then M_+ follows a hyperexponential distribution with l phases as well (with different exponential rates and weights from the first one). In other words, $\widetilde{m}_+(s)$ can be written in the form $\sum_{i=1}^{l} c_i \frac{R_i}{R_i+s}$ for some $c_i, R_i, i = 1,\dots,l$.*

Proof. In Cohen (1982, Chapter II.5.10), it was proven that M_+ follows a hyperexponential distribution in the $G/K_l/1$ queue, where K_l denotes a matrix-exponential distribution with l phases. Since the $M/H_l/1$ is a special case of the $G/K_l/1$ queue, the result holds here too. □

After this, we present our algorithm:

Begin algorithm

1. Write $\overline{G}(u)$ as a mixture of exponentials.
2. By using Lemma 2.6, find the spectral function $A^e(y)$ of $\overline{G^e}(u)$.
3. Approximate $\overline{G^e}(u)$ by a hyperexponential distribution with l phases.
   (a) Choose the number of phases l.
   (b) Set the accuracy of the approximation $\eta = \dfrac{1}{l+1}$, such that $\left|G^e(u) - \widehat{G}^e(u)\right| \le \eta$.
   (c) Define l quantiles such that $A^e(\mu_i) = i\eta$, $i = 1,\dots,l$.
   (d) Approximate the spectral function by the step function

$$\widehat{A}^e(y) = \begin{cases} 0, & y \in [0,\mu_1), \\ \frac{i}{l}, & y \in [\mu_i,\mu_{i+1}), \\ 1, & y \ge \mu_l. \end{cases}$$

(e) Find the approximation of the excess claim size distribution as $\widehat{G}^e(u) = 1 - \frac{1}{l}\sum_{i=1}^{l} e^{-y\mu_i}$.

4. Calculate its Laplace transform $\widetilde{\widehat{G}^e}(s) = \frac{1}{l}\sum_{i=1}^{l} \frac{\mu_i}{\mu_i + s}$.

5. Choose ρ.

6. Calculate the approximation of the Laplace transform of M_+ through the formula $\widehat{\widetilde{m}}_+(s) = \frac{(1-\rho)\widetilde{\widehat{G}^e}(s)}{1 - \rho\widetilde{\widehat{G}^e}(s)}$.

7. Split $\widehat{\widetilde{m}}_+(s)$ into simple fractions. Afterwards, estimate their roots R_i and calculate also the coefficients c_i, by using Lemma 2.8,

8. Invert the Laplace transform of $\widehat{\widetilde{m}}(s) = 1 - \rho + \rho\widehat{\widetilde{m}}_+(s)$ and find that $\widehat{\psi}(u) = 1 - \rho + \rho\sum_{i=1}^{l} c_i(1 - e^{-R_i u})$, $u \geq 0$.

9. The accuracy for $\widehat{\psi}(u)$ is then $\delta \leq \eta\frac{\rho}{1-\rho}$.

End algorithm

Remark 2.9. At step (3b) of the algorithm, we approximate the spectral function A^e with a step function where the jumps occur at the quantiles μ_i and they are all of size $\eta + \eta^2/(1-\eta)$. It can be very easily verified that, by this choice of jumps, we avoid any atoms at 0 and we still achieve $\mathcal{D}(A^e, \widehat{A}^e) \leq \eta$.

Remark 2.10. Note that the algorithm was presented under the setting that we first fix the accuracy η for the approximation of the excess claim size distribution and then we evaluate the bound δ of the spectral approximation. With slight modifications, the algorithm can be presented by first fixing the desired accuracy δ for the approximation of the ruin probability. In this setting, we would have to set the number of required phases as $l = \lceil \rho/(1-\rho)\delta \rceil - 1$.

Remark 2.11. From the structure of the algorithm it is evident that we only need to write the c.m. claim size distribution as a mixture of exponentials. In comparison with the distributions used in the examples, there exist mixed distributions, such as the hyperexponential, that have a spectral function which is not strictly increasing and/or has jumps. In these cases, the algorithm cannot be applied as is and more attention needs to be paid. The problem appears at step (3c), when we invert the spectral function to find the quantiles. More precisely, for a non-strictly-increasing spectral function we might have $A^e(x) = i\eta$, for $x \in (a, b)$, with $a \neq b$, for some $i = 1, \ldots, l$. Therefore, since inversion will not give a unique value for the quantile μ_i, there must be a concrete way to define it. Also, when there are jumps, we might encounter the problem that $A^e(x) \neq i\eta$ for all $x \in (0, \infty)$. In this case, μ_i could take the value at which the jump occurred. All the above mentioned problems related to the determination of the quantiles can be overcome with small modifications to the algorithm.

2.4 Heavy traffic and heavy tail approximations

In this section, we present the heavy traffic (Kingman, 1962) and the heavy tail approximations (von Bahr, 1975; Borovkov and Foss, 1992; Embrechts and Veraverbeke, 1982; Pakes, 1975), which are most often used for the evaluation of the ruin probability. We first start with the heavy traffic approximation.

Heavy traffic approximation

If the claim size distribution G has a finite second moment, then as $\rho \to 1$, M, which was defined in Section 2.2, converges to an exponential random variable with mean $\mathbb{E}M$, i.e. $E(1/\mathbb{E}M)$. This result is known as the heavy traffic approximation (Kalashnikov, 1997). In other words,

$$\psi(u) \approx \psi_h(u) := e^{-u/\mathbb{E}M},$$

where $\mathbb{E}M = \rho\mathbb{E}U^2/2(1-\rho)\mathbb{E}U$. Although the heavy traffic approximation is given through a simple exponential, its biggest drawback is that it requires the first two moments of the claim size distribution to be finite, which is not always the case for heavy-tailed distributions, e.g. the Pareto.

Eq. (2.1) shows that M can be written as a geometric random sum with terms distributed according to G^e. Bounds for exponential approximations of geometric convolutions have been obtained by Brown (1990). Thus, we can acquire a bound for the ruin probability by applying Brown (1990, Theorem 2.1), which states that the sup-norm distance between M and an exponential random variable with the same mean, namely $E(1/\mathbb{E}M)$, is

$$\mathcal{D}(M, E(1/\mathbb{E}M)) = (1-\rho)\max(2\gamma, \gamma/\rho) = \begin{cases} 2(1-\rho)\gamma, & \text{if } \rho \geq \frac{1}{2} \\ (1-\rho)\gamma/\rho, & \text{if } 0 < \rho < \frac{1}{2}, \end{cases} \tag{2.4}$$

where $\gamma = 2\mathbb{E}U^3\mathbb{E}U/3(\mathbb{E}U^2)^2$. Thus, a finite *third* moment is required for the claim sizes in order to guarantee a bound for the heavy traffic approximation.

Heavy tail approximation

When the claim sizes belong to the subexponential class of distributions (Teugels, 1975), e.g. Weibull, lognormal, Pareto, etc., the heavy tail approximation can also be used. For $u \to \infty$, the heavy tail approximation is defined as

$$\psi(u) \approx \psi_t(u) := \frac{\rho}{1-\rho}\overline{G^e}(u).$$

This approximation is also given by a simple formula, which requires only the first moment of the claim size distribution to be finite. Its drawback though is that for values of ρ close to 1, or equivalently in the heavy traffic regime, the heavy tail approximation is useful only for extremely big values of u. For the heavy traffic setting, there exists a comparative analysis between the heavy traffic and the heavy tail approximations (Olvera-Cravioto et al., 2011) in which the point at which the heavy tail approximation becomes more suitable than the heavy traffic is examined.

In the following section, we compare the accuracy of the spectral approximation to the accuracy of the heavy traffic and the heavy tail approximations. An interesting

observation with respect to the spectral approximation is that, since it decays exponentially, it converges faster to zero than any heavy-tailed distribution. Thus, at the tail the spectral approximation is expected to underestimate the ruin probability. But an overestimation of the ruin probability for small values of u, compensates for the underestimation at the tail, as it will be apparent in Section 2.5.2.

2.5 Numerical experiments

In this section we implement our algorithm in order to check the accuracy of the spectral approximation. We test the spectral approximation in 3 different classes of c.m. heavy-tailed distributions: a class of long-tail distributions introduced in Abate and Whitt (1999b), the Weibull distribution, and the Pareto distribution.

2.5.1 Test distributions

First we present the three test distributions and thereafter we do a series of experiments to compare the accuracy of the spectral approximation with the accuracy of heavy tail approximation and when applicable with the heavy traffic approximation too.

Abate-Whitt distribution

Consider a claim size distribution with Laplace transform

$$\widetilde{G}(s) = 1 - \frac{s}{(\kappa + \sqrt{s})(1 + \sqrt{s})},$$

which has mean κ^{-1} and all higher moments infinite. The parameter κ can range over the positive values. This class of long-tailed distributions was introduced in Abate and Whitt (1999b), where it was also proven that the explicit formula for the ruin probability of the compound Poisson model with arrival rate for claims λ and $\rho = \lambda/\kappa < 1$ is

$$\psi(u) = \mathbb{P}(M > u) = \frac{\rho}{v_1 - v_2}\Big(v_1\zeta\left(v_2^2 u\right) - v_2\zeta\left(v_1^2 u\right)\Big)$$

where

$$\zeta(u) \equiv e^u \frac{2}{\sqrt{\pi}} \int_{\sqrt{u}}^{\infty} e^{-x^2} dx \quad \text{and} \quad v_{1,2} = \frac{1+\kappa}{2} \pm \sqrt{\left(\frac{1+\kappa}{2}\right)^2 - (1-\rho)\kappa}.$$

The existence of an exact formula for the ruin probability makes this distribution very interesting because we can compare the spectral approximation with the exact ruin probability and not with the outcome of a simulation.

For this model we have that the c.c.d.f. of the the claim size distribution is given by the formula

$$\overline{G}(u) = \left(\frac{1}{1-\kappa}\right)\Big(\zeta(u) - \kappa\zeta\left(\kappa^2 u\right)\Big).$$

With simple calculations we can verify that $\zeta(u)$ is c.m. since it can be written as a mixture of exponentials

$$\begin{aligned}
\zeta(u) &= e^u \frac{2}{\sqrt{\pi}} \int_{\sqrt{u}}^{\infty} e^{-x^2} dx \overset{z=x^2}{=} \frac{2e^u}{\sqrt{\pi}} \int_u^{+\infty} \frac{e^{-z}}{2\sqrt{z}} dz \\
&= \frac{1}{\sqrt{\pi}} \int_u^{+\infty} \frac{e^{-(z-u)}}{\sqrt{z}} dz \overset{t=z-u}{=} \frac{1}{\sqrt{\pi}} \int_0^{+\infty} \frac{e^{-t}}{\sqrt{t+u}} dt \\
&= \frac{1}{\sqrt{\pi}} \int_0^{+\infty} \frac{e^{-t}}{\sqrt{u}} \left(\frac{u}{t+u}\right)^{\frac{1}{2}} dt = \frac{1}{\sqrt{\pi}} \int_0^{+\infty} \frac{e^{-t}}{\sqrt{u}} \left(\frac{1}{\sqrt{\pi}} \int_0^{+\infty} \frac{\sqrt{u}}{\sqrt{y}} e^{-(u+t)y} dy \right) dt \\
&= \frac{1}{\pi} \int_0^{+\infty} \frac{e^{-uy}}{\sqrt{y}} \underbrace{\left(\int_0^{+\infty} e^{-(y+1)t} dt \right)}_{\frac{1}{y+1}} dy = \int_0^{+\infty} y e^{-uy} \frac{1}{\pi y^{3/2}(y+1)} dy.
\end{aligned}$$

The c.c.d.f. of the claim sizes is also c.m. That is,

$$\begin{aligned}
\overline{G}(u) &= \left(\frac{1}{1-\kappa}\right) \left(\zeta(u) - \kappa \zeta\left(\kappa^2 u\right) \right) \\
&= \frac{1}{1-\kappa} \int_0^{+\infty} y e^{-uy} \left[\frac{1}{\pi y^{3/2}(y+1)} - \frac{\kappa^2}{\pi y^{3/2}(y+\kappa^2)} \right] dy \\
&= \int_0^{+\infty} e^{-uy} \frac{\sqrt{y}(1+\kappa)}{\pi (y+1)(y+\kappa^2)} dy.
\end{aligned}$$

Note that for the heavy traffic approximation a finite second moment is required, which does not hold for this case. Therefore, for this distribution the heavy traffic approximation for the ruin probability cannot be evaluated. As a result, we compare the spectral approximation only with the heavy tail approximation.

Weibull

The c.c.d.f. of the Weibull(c,a) distribution with c and a the positive shape and scale parameters respectively is given by $\overline{G}(u) = e^{-(u/a)^c}$. It can be verified (Jewell, 1982) that the c.c.d.f. of the Weibull(0.5,a) distribution with fixed shape parameter 1/2 arises as a mixture of exponentials, where the *mixing measure* (measure of the spectral function) A is given by

$$dA(y) = \frac{a e^{-a^2/4y}}{2\sqrt{\pi y^3}} dy.$$

For this case we do not have an explicit formula for the ruin probability, thus we compare the spectral approximation to simulation results. Since the second moment of Weibull(c,a) is finite, namely $\mathbb{E}U^2 = 24a^2$, we can compare the spectral approximation with the heavy traffic approximation as well, contrary to the Abate-Whitt distribution, where only comparisons with the heavy tail approximation were possible.

Pareto

The third test function we use is the Pareto(a,b) distribution with shape parameter $a > 0$ and scale parameter $b > 0$. The Pareto(a,b) distribution with p.d.f. $g(u) =$

$ab/(1+bu)^{a+1}, u > 0$ is c.m. Its c.c.d.f. $\overline{G}(u) = (1+bu)^{-a}$ can be written as a mixture of exponentials in the form

$$(1+bu)^{-a} = \int_0^{+\infty} e^{-yu} e^{-y/b} \frac{\left(\frac{y}{b}\right)^{a-1}}{b\Gamma(a)} dy.$$

Also for this distribution the ruin probability does not exist in closed form. Therefore, we compare our approximation for this case to simulation results.

It is known that the nth moment of the Pareto distribution exists if and only if the shape parameter is greater than n. Since it would be interesting to compare the spectral approximation, not only with the heavy tail one, but with the heavy traffic too, it is necessary to have a finite second moment for the claim sizes. Moreover, as stated in Section 2.4, a bound for the heavy traffic approximation is guaranteed as long as the third moment of the distribution is finite. For these reasons, if we want to evaluate the heavy traffic approximation with a guaranteed bound for the Pareto(a,b), the shape parameter a must be chosen to be greater than 3.

2.5.2 Numerical results

The goal of this section is to implement our algorithm to check the accuracy of the spectral approximation and the tightness of its accompanying bound, which is given in Proposition 2.3.

Since the only restriction we have for the parameters of the three test distributions is that the shape parameter of the Pareto(a,b) must be greater than 3, we randomly select the parameters and thus we deal with the Abate-Whitt distribution with $\kappa = 2$, the Weibull(0.5,3) distribution, and the Pareto(4,3) distribution. We form and answer here the following questions:

Impact of phases. The bound of the spectral approximation is conversely proportional to the number of phases of the hyperexponential with which we approximate the excess claim size distribution (see Section 2.3). So, for a fixed claim rate ρ, the bound becomes tighter when the number of phases increases. Does this also mean that the spectral approximation becomes more accurate as the number of phases increases?
Experiment: We fix ρ and we compare three different spectral approximations with number of phases 10, 20, and 100 respectively, with the exact value of the ruin probability. For the Abate-Whitt distribution, we present the exact ruin probability with the three approximations in one graph; see Figure 2.1. For the Weibull and the Pareto distributions we compare the three approximations to the exact ruin probability that we obtain through simulation and display our results in Tables 2.1 and 2.2. As for all different values of ρ we get a similar results, we present our findings only for $\rho = 0.7$.
Answer: The conclusion is that, while the number of phases increases, a more accurate spectral approximation is achieved. This result is in line with our expectations, and we can safely conclude that for a fixed claim rate ρ more phases lead to a more accurate spectral approximation.

Quality of the bound. Is the bound strict or pessimistic? How far is the bound from the real error of the spectral approximation?
Experiment: We fix the bound of the spectral approximation to be equal to $\delta = 0.02$ and we evaluate the error functions (in absolute values) for the spectral approximation

when the claim rate ρ takes the values 0.1, 0.5, and 0.9. For these three cases we need 5, 49, and 449 phases respectively for the spectral approximation. We compare the guaranteed bound with the exact maximum error that is achieved; see Figures 2.2(a) to 2.2(c).

Also, for various combinations of the number of phases and the claim rate ρ, we calculate the ratios between the predicted bound of the spectral approximation and the achieved maximum error; see Table 2.3. We set out this experiment only for the Abate-Whitt distribution, because the existence of the exact ruin probability gives more accurate results.

Answer: An interesting observation that arises from Figure 2.2(a) is that the achieved maximum error of the spectral approximation seems to be almost half of the guaranteed bound. In order to verify that the bound is twice as big as the achieved maximum error we look at Table 2.3.

We first read the table horizontally, namely we fix the claim rate ρ. We observe that while we let the number of phases increase, the ratio between the predicted bound and the real maximum error becomes smaller and converges to 2. As it was mentioned earlier, the spectral approximation becomes more accurate when we increase the number of phases. Therefore, we conclude that the bound becomes tighter when for a fixed ρ we increase the number of phases.

We read now the table vertically, namely we fix the number of phases and we let the claim rate ρ increase. We observe that while we let ρ increase, both the predicted bound and the maximum error increase. Since the ratios between the bound and the maximum error increase too, we can conclude that the bound becomes less tight when the claim rate increases.

However, from Figures 2.2(b) and 2.2(c), we see that the achieved maximum error is not only 2 times smaller than the guaranteed bound but 4 times smaller! Gathering all the above together, we can conclude that the bound seems to be at least twice as big as the the achieved maximum error of the spectral approximation.

Comparison of Spectral, Heavy tail, Heavy traffic approximations. The accuracy of the spectral approximation can be predetermined through its bound. For a fixed range of u, which of the three approximations – spectral, heavy tail, and heavy traffic (when applicable) – is better than the others as $\rho \to 1$ or $\rho \to 0$, when the bound predicts accuracy of at most δ for the spectral approximation?

Experiment: We fix the bound of the spectral approximation to be equal to $\delta = 0.02$, and for $\rho = 0.1, 0.5$, and 0.9 we compare the spectral (with 5, 49, and 449 phases respectively), the heavy tail and the heavy traffic (when applicable) approximations. We present the distributions in a graph, where the displayed range of u is such that $\psi(u) > \delta$, because after this point the error is smaller than δ. The level δ is denoted on the graphs with a dashed horizontal line; see Figures 2.3(a) to 2.5(c).

Answer: We observe that the spectral approximation behaves nicely for all values of u. For small values of u, the spectral approximation is more accurate than the heavy tail approximation, where the second fails to provide us with a good estimation of the ruin probability, especially when $\rho \to 1$.

On the other hand, the heavy tail approximation is slightly more accurate than the spectral approximation at the tail. Although we cannot give an estimation for the point u^* at which the heavy tail approximation becomes more suitable than the spectral approximation, we observe that this point takes greater values as ρ increases and it

sometimes can be extremely big; i.e. see Figure 2.3(c).

Furthermore, according to our expectations, the spectral approximation overestimates the ruin probability for small values of u (this is more clear for small values of ρ) and underestimates it for large values of u. In all cases, the heavy traffic approximation is worse than the other two, since it exhibits a sharper behaviour than the spectral approximation. Namely, for small values of u it overestimates the ruin probability more than the spectral approximation, and for large values of u it underestimates the ruin probability more than the spectral approximation. Note also that, at the tail, the spectral approximation and the heavy traffic approximation are almost identical, which can be explained by the fact that both of them have an exponential decay.

Comparison between Spectral and Heavy traffic bounds. For the Weibull and the Pareto distributions, the heavy traffic approximation can be evaluated and it also has a guaranteed bound (Brown, 1990). So, is there a rule of thumb to help us choose between the spectral and the heavy traffic approximation, when they both guarantee the same bound?

Experiment: For various values of ρ, we compare the spectral approximation with the heavy traffic approximation when they both guarantee the same bound. More precisely, we fix ρ and determine the number of phases l^* of the spectral approximation for which both approximations guarantee the same bound. We calculate the two approximations and evaluate their maximum errors. We present our findings in a table, only for some values of ρ that the heavy traffic bound has a meaning, namely when it is smaller than 1; see Tables 2.4 and 2.5.

We can easily verify that for the Pareto(a,b) distribution, the heavy traffic bound depends on the shape parameter a, since $\gamma = (a-2)/(a-3)$. An interesting experiment that arises from this observation is to check whether we have a clearer picture on which of the spectral and heavy traffic approximations is the best in terms of accuracy, if we choose a big enough such that $\gamma \to 1$, namely if we make the heavy traffic bound tighter (for Pareto(4,3), $\gamma = 2$). For this reason, we repeat our last experiment for Pareto(15.6,2.7), which has $\gamma = 1.079$.

Answer: From Table 2.4, which gives the results for Weibull(0.5,3), we see that whenever the bounds are equal, the spectral approximation is more accurate than the heavy traffic approximation for all number of phases greater or equal than l^*. On the other hand, from Table 2.4, which gives the results for Pareto(4,3), we get a different picture. The conclusion that we draw from this table is that for a small number of phases (relatively smaller than 20) the heavy traffic approximation is better, while for a number of phases greater than 20 the conclusion reverses.

For Pareto(15.6,2.7), more phases were needed in the corresponding spectral approximation for the same values of ρ, because the heavy traffic bound is now tighter. The picture from Table 2.6 is not that clear. More precisely, even when the number of phases becomes relatively big we cannot draw a safe conclusion that the spectral approximation is better than the heavy traffic approximation.

At this point it is interesting to observe the following. The heavy traffic approximation as presented in Section 2.4 has no atoms. It is known (Asmussen and Albrecher, 2010) that the ruin probability has an atom of mass ρ at 0. Thus, the heavy traffic approximation is not very accurate for small values of u, especially when ρ takes relatively small values. For this reason, a more suitable heavy traffic approximation

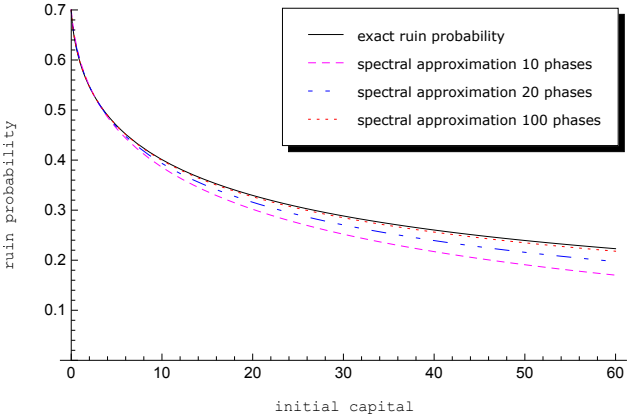


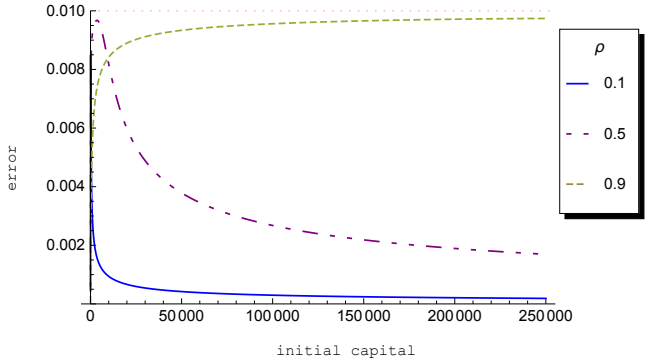
FIGURE 2.1: The spectral approximation for different number of phases, when $\rho = 0.7$ and the claims follow the Abate-Whitt distribution with $\kappa = 2$.

u	exact	sa 10 phases	sa 20 phases	sa 100 phases
0	0.70000	0.70000 (0.00000)	0.70000 (0.00000)	0.70000 (0.00000)
5	0.60745	0.61023 (0.00279)	0.60823 (0.00079)	0.60754 (0.00009)
10	0.54574	0.54696 (0.00122)	0.54569 (0.00005)	0.54527 (0.00047)
15	0.49580	0.49558 (0.00022)	0.49502 (0.00078)	0.49485 (0.00095)
20	0.45312	0.45172 (0.00139)	0.45181 (0.00130)	0.45189 (0.00122)
25	0.41603	0.41334 (0.00269)	0.41405 (0.00198)	0.41436 (0.00167)

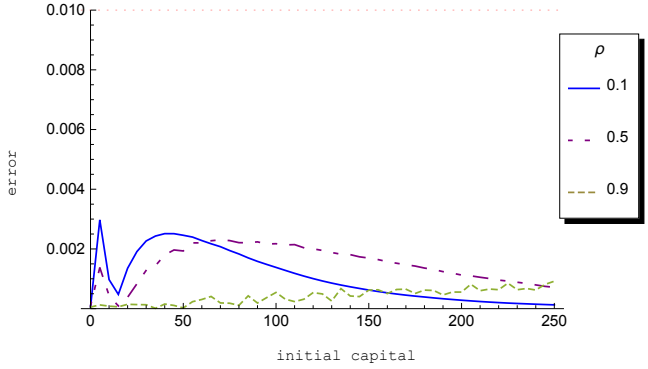
TABLE 2.1: The spectral approximation for different number of phases, when the claims follow the Weibull(0.5,3) distributions. The numbers in the brackets correspond to the absolute error of the exact ruin probability from its respective approximations.

u	exact	sa 10 phases	sa 20 phases	sa 100 phases
0.00	0.70000	0.70000 (0.00000)	0.70000 (0.00000)	0.70000 (0.00000)
0.10	0.54805	0.55012 (0.00207)	0.55008 (0.00203)	0.55005 (0.00200)
0.55	0.23572	0.22698 (0.00873)	0.23218 (0.00353)	0.23435 (0.00137)
1.00	0.11499	0.10194 (0.01305)	0.10851 (0.00648)	0.11146 (0.00352)
1.45	0.05983	0.04695 (0.01287)	0.05265 (0.00718)	0.05545 (0.00437)
1.90	0.03215	0.02187 (0.01028)	0.02609 (0.00606)	0.02838 (0.00377)

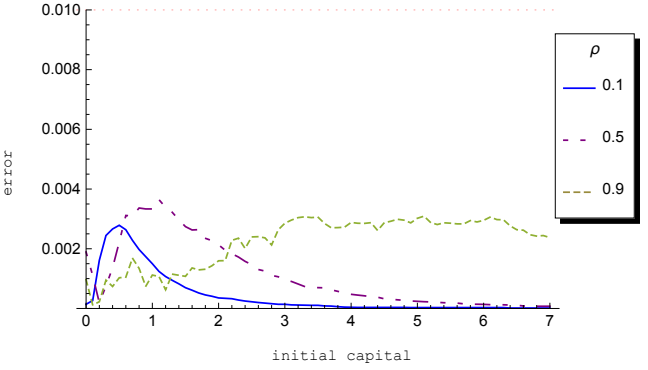
TABLE 2.2: The spectral approximation for different number of phases, when the claims follow the Pareto(4,3) distributions. The numbers in the brackets correspond to the absolute error of the exact ruin probability from its respective approximations.



(a) The claim size distribution is the Abate-Whitt distribution with $\kappa = 2$.

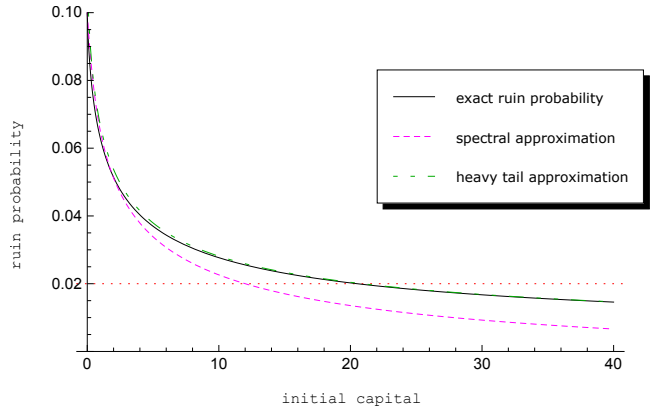


(b) The claim size distribution is Weibull$(0.5, 3)$.

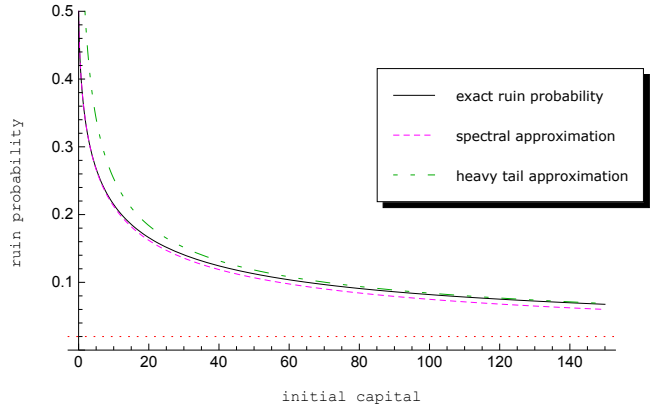


(c) The claim size distribution is Pareto$(4, 3)$.

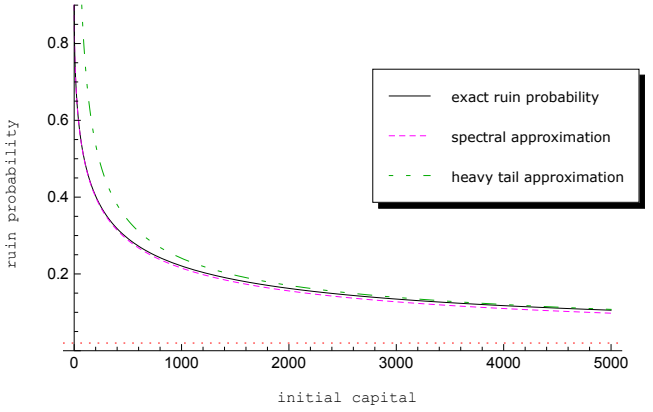
FIGURE 2.2: Error functions for the spectral approximation with guaranteed bound $\delta = 0.02$, when the claims follow each of the above distributions.



(a)

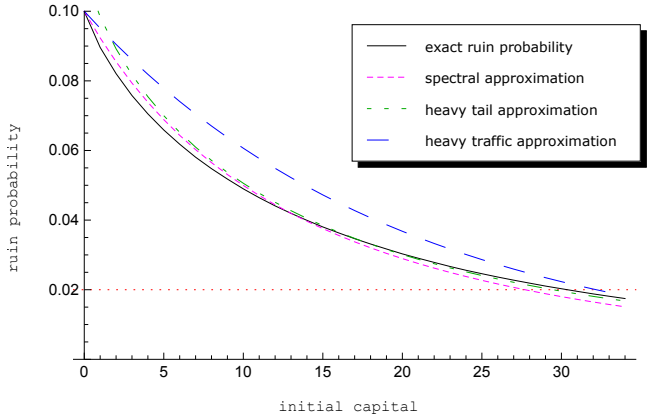


(b)

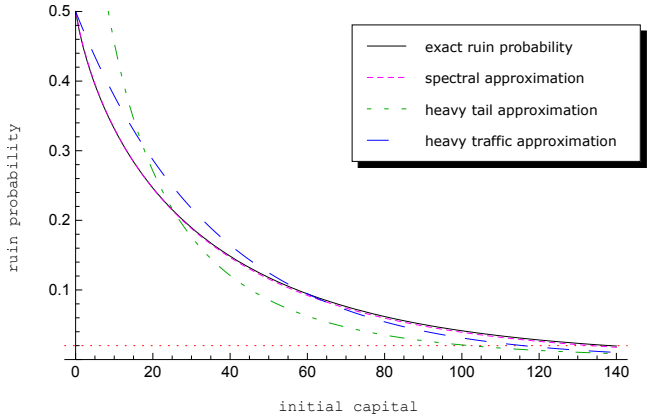


(c)

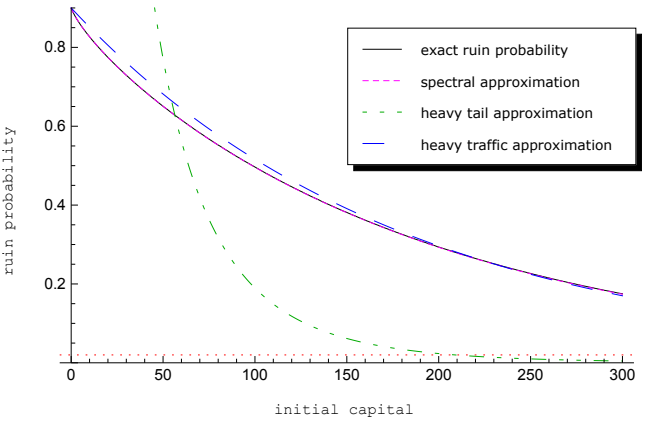
FIGURE 2.3: The spectral approximation for guaranteed bound $\delta = 0.02$, when the claims follow the Abate-Whitt distribution with $\kappa = 2$ and the average claim rate ρ is: (a) 0.1, (b) 0.5, and (c) 0.9.



(a)

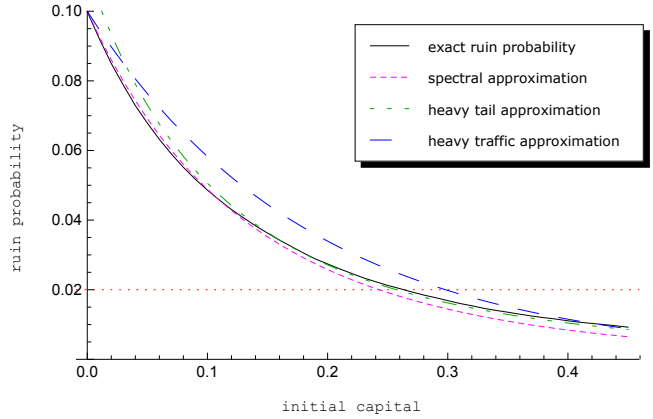


(b)

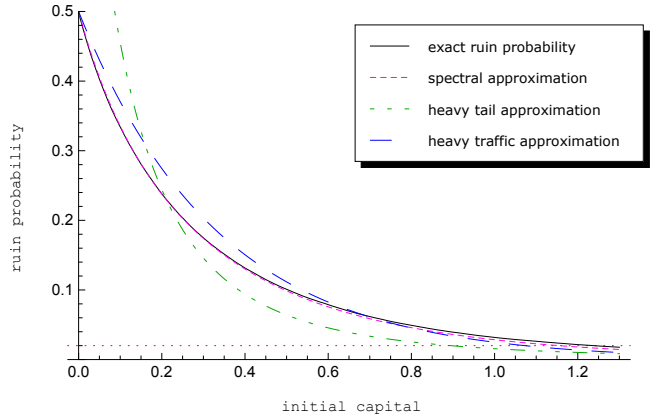


(c)

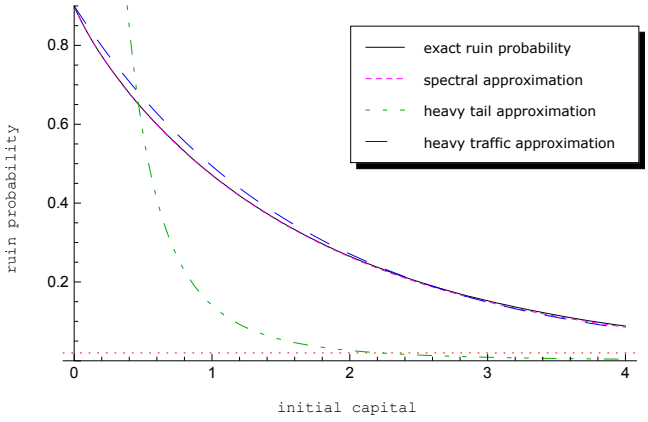
FIGURE 2.4: The spectral approximation for guaranteed bound $\delta = 0.02$, when the claims follow the Weibull(0.5,3) distribution and the average claim rate ρ is: (a) 0.1, (b) 0.5, and (c) 0.9.



(a)



(b)



(c)

FIGURE 2.5: The spectral approximation for guaranteed bound $\delta = 0.02$, when the claims follow the Pareto(4,3) distribution and the average claim rate ρ is: (a) 0.1, (b) 0.5, and (c) 0.9.

(a) 10 phases

ρ	bound	max error	ratio
0.1	0.010	0.0048	2.11
0.2	0.023	0.0106	2.13
0.3	0.039	0.0180	2.17
0.4	0.061	0.0275	2.21
0.5	0.091	0.0401	2.27
0.6	0.136	0.0580	2.35
0.7	0.212	0.0849	2.50
0.8	0.364	0.1299	2.80
0.9	0.818	0.2263	3.61

(b) 20 phases

ρ	bound	max error	ratio
0.1	0.005	0.0026	2.06
0.2	0.012	0.0057	2.08
0.3	0.020	0.0097	2.09
0.4	0.032	0.0150	2.12
0.5	0.048	0.0222	2.15
0.6	0.071	0.0326	2.19
0.7	0.111	0.0490	2.27
0.8	0.190	0.0787	2.42
0.9	0.429	0.1479	2.90

(c) 100 phases

ρ	bound	max error	ratio
0.1	0.001	0.0005	2.02
0.2	0.002	0.0012	2.02
0.3	0.004	0.0021	2.02
0.4	0.007	0.0033	2.03
0.5	0.010	0.0049	2.04
0.6	0.015	0.0073	2.05
0.7	0.023	0.0112	2.06
0.8	0.040	0.0189	2.10
0.9	0.089	0.0406	2.19

TABLE 2.3: Ratios between the guaranteed bound and the maximum error of the spectral approximation, when the claims follow the Abate-Whitt distribution with $\kappa = 2$.

ρ	HT bound	l^*	sp bound	max HT error	max sp error
0.82	0.78	5	0.76	0.0438	0.0312
0.85	0.65	8	0.63	0.0403	0.0253
0.88	0.52	13	0.52	0.0361	0.0196
0.91	0.39	25	0.39	0.0304	0.0139
0.94	0.26	59	0.26	0.0234	0.0081
0.97	0.13	248	0.13	0.0144	0.0013

TABLE 2.4: Comparison between the maximum heavy traffic and spectral errors, when the claims follow the Weibull(0.5,3) distribution.

ρ	HT bound	l^*	sp bound	max HT error	max sp error
0.82	0.90	4	0.91	0.0392	0.0453
0.85	0.75	7	0.71	0.0365	0.0387
0.88	0.60	11	0.61	0.0326	0.0330
0.91	0.45	21	0.46	0.0279	0.0261
0.94	0.30	51	0.30	0.0226	0.0166
0.97	0.15	215	0.15	0.0156	0.0074

TABLE 2.5: Comparison between the maximum heavy traffic and spectral errors, when the claims follow the Pareto(4,3) distribution.

ρ	HT bound	l^*	sp bound	max HT error	max sp error
0.82	0.568	7	0.569	0.0051	0.0068
0.85	0.473	11	0.472	0.0060	0.0066
0.88	0.379	18	0.386	0.0044	0.0044
0.91	0.284	35	0.281	0.0047	0.0026
0.94	0.190	82	0.189	0.0047	0.0014
0.97	0.095	340	0.095	0.0024	0.0025

TABLE 2.6: Comparison between the maximum heavy traffic and spectral errors, when the claims follow the Pareto(15.6,2.7) distribution.

(ψ_h) for our comparisons for all values of ρ seems to be

$$\psi(u) \approx \psi_h(u) := \rho e^{-\rho u/\mathbb{E}M}, \tag{2.5}$$

for which is easy to verify that it also has mean equal to $\mathbb{E}M$ and an atom of mass ρ at 0. Since we used a different heavy traffic approximation in all of our experiments than the one Brown (1990) compares the ruin probability with, we extended Brown's bound, given in Eq. (2.4), to this situation. Applying the triangular inequality to the sup-norm distance we get

$$\mathcal{D}(\psi, \psi_h) \leq \mathcal{D}(\psi, E(1/\mathbb{E}M)) + \mathcal{D}(E(1/\mathbb{E}M), \psi_h).$$

It is easy to verify that $\mathcal{D}(E(1/\mathbb{E}M), \psi_h) = 1 - \rho$, so the sup-norm distance between the ruin probability and the heavy traffic approximation we use for comparisons is

$$\mathcal{D}(\psi, \psi_h) \leq (1-\rho)\max(2\gamma, \gamma/\rho) + 1 - \rho = (1-\rho) \cdot \begin{cases} 2\gamma + 1, & \text{if } \rho \geq \frac{1}{2} \\ \gamma/\rho + 1, & \text{if } 0 < \rho < \frac{1}{2}, \end{cases} \tag{2.6}$$

where $\gamma = 2\mathbb{E}U^3\mathbb{E}U/3(\mathbb{E}U^2)^2$. When we referred to the heavy traffic approximation and its accompanying bound, in all of our experiments we meant those given from Eqs. (2.5) and (2.6), respectively.

2.6 Conclusions

In this chapter, we address the problem of how many phases are needed to approximate a heavy-tailed distribution with a phase-type distribution in such a way that one

can obtain a guaranteed bound on the approximation of the ruin probability (see Section 2.3). In doing so, we developed an explicit bound using the geometric random sum representation, which was combined with a spectral approximation of the excess claim size distribution.

The conclusions that we can draw, both for the spectral approximation and its bound, can be summarised as follows:

- The spectral approximation provides a good fit for all values of the initial capital u, especially for the small ones, where the heavy traffic and heavy tail approximations fail. Also, for small values of u the spectral approximation exhibits a behaviour of overestimating the ruin probability, while for larger values of u we have an underestimation of the ruin probability by the spectral approximation. Finally, for a fixed claim rate ρ, the more the phases we have for the approximate hyperexponential of the excess claim size distribution, the more accurate spectral approximation we achieve.

- The spectral bound, guaranteed by Proposition 2.3, becomes tighter when for a fixed claim rate ρ the number of phases is increased, while it becomes less tight when for a fixed number of phases the claim rate increases. Moreover, the bound seems to be at least twice as big as the achieved maximum error of the spectral approximation. But, based on the numerical examples we performed, we cannot conclude that this is the general rule.

- Based on existing analytical results and extensive experiments it is hard to draw a definitive conclusion on which approximation should be preferred: the heavy traffic approximation or the spectral approximation. We believe that obtaining more mathematical as well as experimental insights in this problem is an important topic for future research.

To sum up, the spectral approximation provides a good fit for all values of u and has a guaranteed accuracy, while it requires only a finite mean for the claim sizes.

CHAPTER 3

Corrected phase-type approximations

3.1 Introduction

In the previous chapter, we defined the spectral approximation for the ruin probability under heavy-tailed claim sizes, by approximating the claim size distribution with a hyperexponential one. We showed that the accuracy of our approximation is guaranteed by an upper bound and it can also be pre-determined by strategically choosing the number of phases of the hyperexponential distribution. However, being a phase-type approximation, the spectral approximation gives a big relative error at the tail of the ruin probability. Therefore, in this chapter, we develop a new method to construct accurate approximations for performance measures of heavy-tailed risk models that capture the correct tail behaviour. We use the classical risk model as a context and vehicle to demonstrate our key ideas, which we expect to have a much wider applicability in insurance.

The approximations we develop for ruin probabilities under heavy-tailed claims, combine desirable characteristics of the following three main approximation directions: phase-type approximations, asymptotic approximations, and error bounds. First, our approximations maintain the computational tractability of phase-type approximations. Additionally, they capture the correct tail behaviour, which so far could only be captured by asymptotic approximations, and have the advantage that finite higher-order moments are not required for the claim sizes. Last, they provide a provably small absolute error, independent of the initial capital, and a small relative error.

The idea of our approach stems from fitting procedures of the claim size distribution to data. Heavy-tailed statistical analysis suggests that for a sample with size n only a small fraction ($k_n/n \to 0$) of the upper-order statistics is relevant for estimating tail probabilities (Davis and Resnick, 1984; Hill, 1975; Resnick, 2007b). More information about the optimal choice of the k_nth upper-order statistic can be found in Haeusler and Teugels (1985). The remaining data set may be used to fit the bulk of the distribution. Since the class of phase-type distributions is dense in the class of all

positive definite probability distributions (Asmussen, 2003), a natural choice is to fit a phase-type distribution to the remaining data set (Asmussen et al., 1996). As a result, a mixture model for the claim size distribution is a natural assumption. Thus, our key idea is to use a mixture model for the claim size distribution in order to construct approximations of the ruin probability that combine the best elements of phase-type and asymptotic approximations.

We now sketch how to derive our approximations when the claim size distribution is a mixture of a phase-type distribution and a heavy-tailed one. Interpreting the heavy-tailed term of the claim size distribution in the mixture model as perturbation of the phase-type one and using perturbation theory, we can find the ruin probability as a complete series expansion. The first term of the expansion is the phase-type approximation of the ruin probability that occurs when we "remove" the heavy-tailed claim sizes from the system, either by discarding them or by replacing them with phase-type ones. We consider the model that appears when all heavy-tailed claims are removed as the "base" model. Due to the two different approaches of removing the heavy-tailed claim sizes, the ruin probability connects to two different base models and consequently to two different series expansions.

We show that adding the second term of the respective series expansions is sufficient to construct improved approximations, compared to their phase-type counterparts, the *discard* and the *replace approximations*, respectively. Since the second term of each series expansion works as a correction to its respective phase-type approximation, motivated by the terminology *corrected heavy traffic-approximations* (Asmussen, 2003), we refer to our approximations as *corrected phase-type approximations*. Therefore, in this chapter, we propose the *corrected discard approximation* and the *corrected replace approximation*. Both approximations have appealing properties: the corrected replace approximation tends to give better numerical estimates, while the corrected discard approximation is simpler and yields guaranteed upper and lower bounds.

Besides the ruin probability, we also find approximations for the *Value at Risk* (VaR), which is a very popular tool in real-world applications to measure the operational risk (McNeil et al., 2005). For a given portfolio, a VaR with a probability level α and fixed time horizon is defined as the threshold value such that the loss on the portfolio over the given time horizon exceeds this value with probability $1-\alpha$. It is of interest to quantify the operational risk through the statistical analysis of operational loss data (Embrechts and Samorodnitsky, 2004; Klugman et al., 2008) and to provide error bounds for the aggregate loss probability (Cox et al., 2008). Similarly to the ruin probability, things become more complicated under the presence of heavy-tailed data (Embrechts et al., 1997). Thus, in this chapter, we also provide the form of the *corrected phase-type approximations* for the aggregate loss over a fixed time period, and we show that they have the same appealing properties also for finite time.

Outline

The rest of the chapter is organised as follows. In Section 3.2, we introduce the model and we derive two series expansions for the ruin probability. From these series expansions we deduce approximations for the ruin probability, in Section 3.3, and we study their basic properties. In Section 3.4, we find the exact formula of the ruin probability for a specific mixture model and we study the extent of the achieved

improvement when we compare our approximations with phase-type approximations of their related base model through numerical experiments. In Section 3.5, we provide corrected phase-type approximations of the aggregate loss in finite time and we show through a numerical study that our approximations give excellent VaR estimates.

3.2 Series expansions of the ruin probability

As proof of concept, we apply our technique to the classical Cramér-Lundberg risk model, which was also introduced in Sections 1.6 and 2.2. Here, we assume that premiums flow in at a rate 1 per unit time and claims arrive according to a Poisson process $\{N_\epsilon(t)\}_{t\geq 0}$ with rate λ, where $\epsilon \in [0,1]$ is a parameter to be explained soon. The claim sizes $U_{\epsilon,k} \overset{\mathfrak{D}}{=} U_\epsilon$ are i.i.d. with common distribution G_ϵ and independent of $\{N_\epsilon(t)\}$. Motivated by statistical analysis, which proposes that only a small fraction of the upper-order statistics is relevant for estimating tail probabilities, we consider that an arbitrary claim size U_ϵ is phase-type (Neuts, 1994) with probability $1-\epsilon$ and heavy-tailed (Rolski et al., 1999) with probability ϵ, where $\epsilon \to 0$. In the forthcoming analysis, we use as general rule that all parameters depending on ϵ bear a subscript with the same letter. We assume that the phase-type claim sizes $B_k \overset{\mathfrak{D}}{=} B$ and the heavy-tailed claim sizes $C_k \overset{\mathfrak{D}}{=} C$ have both finite means, $\mathbb{E}B$ and $\mathbb{E}C$, respectively. If u is the initial capital, our risk reserve process takes the form

$$R_\epsilon(t) = u + t - \sum_{k=1}^{N_\epsilon(t)} U_{\epsilon,k}.$$

Using this model, we first examine in Sections 3.2–3.4 the ruin probability in infinite time horizon, and later on, in Section 3.5, we move to finite time horizon and we examine the aggregate loss.

For our model, we also define the claim surplus process $S_\epsilon(t) = u - R_\epsilon(t)$ and its maximum $M_\epsilon = \sup_{0\leq t<\infty} S_\epsilon(t)$. The probability $\psi_\epsilon(u)$ of ultimate ruin is then

$$\psi_\epsilon(u) = \mathbb{P}(M_\epsilon > u).$$

In addition, we assume that the average amount of claim per unit time $\rho_\epsilon = \lambda \mathbb{E}U_\epsilon$ is strictly smaller than 1 and thus the well-known Pollaczek-Khinchine formula (see Eq. (2.1)) can be used for the evaluation of the ruin probability

$$1 - \psi_\epsilon(u) = (1-\rho_\epsilon) \sum_{n=0}^{\infty} \rho_\epsilon^n (G_\epsilon^e)^{*n}(u), \tag{3.1}$$

where G_ϵ^e is the distribution of the stationary excess claim sizes U_ϵ^e. The infinite sum of convolutions at the right-hand side of (3.1) makes the evaluation of $\psi_\epsilon(u)$ difficult or even impossible for our mixture model. For this reason, one typically resorts to Laplace transforms. We use the notation $\widetilde{G}_\epsilon^e(s)$, $\widetilde{F}_p^e(s)$, and $\widetilde{F}_h^e(s)$ for the Laplace transforms of the stationary excess claim sizes $U_{\epsilon,k}^e \overset{\mathfrak{D}}{=} U_\epsilon^e$, $B_k^e \overset{\mathfrak{D}}{=} B^e$, and $C_k^e \overset{\mathfrak{D}}{=} C^e$, respectively. Moreover, we set $\rho_0 = \lambda \mathbb{E}B$ and $\rho_1 = \lambda \mathbb{E}C$, which means that the phase-type claims are responsible for average claim $(1-\epsilon)\rho_0$ per unit time and the heavy-

tailed claims are responsible for average claim $\epsilon\rho_1$ per unit time. Using this notation, we obtain $\rho_\epsilon = (1-\epsilon)\rho_0 + \epsilon\rho_1$. In terms of Laplace transforms, the Pollaczek-Khinchine formula can be written now as:

$$\widetilde{m}_\epsilon(s) := \mathbb{E}e^{-sM_\epsilon} = (1-\rho_\epsilon)\sum_{n=0}^{\infty}\rho_\epsilon^n\big(\widetilde{G}_\epsilon^e(s)\big)^n = \frac{1-\rho_\epsilon}{1-\rho_\epsilon\widetilde{G}_\epsilon^e(s)}$$
$$= \frac{1-(1-\epsilon)\rho_0-\epsilon\rho_1}{1-(1-\epsilon)\rho_0\widetilde{F}_p^e(s)-\epsilon\rho_1\widetilde{F}_h^e(s)}. \tag{3.2}$$

Applying Laplace inversion to Eq. (3.2) in order to find $\psi_\epsilon(u)$ is difficult (Abate and Whitt, 1999a) or even impossible, because the heavy-tailed component $\widetilde{F}_h^e(s)$ often does not have an analytic closed form. To overcome this difficulty, a phase-type approximation would suggest to "remove" the heavy-tailed claim sizes and find an explicit phase-type representation for the ruin probability of the resulting simpler model, which we use as base model for our analysis. In broad terms, we view the heavy-tailed claim sizes as perturbation of the phase-type claim sizes and we interpret ϵ as the perturbation parameter. With the aid of perturbation analysis, we find the ruin probability of our mixture model as a complete series expansion with first term the phase-type approximation that results from its base model.

As mentioned in the introduction, we remove the heavy-tailed claims either by discarding them or by replacing them with phase-type ones. Therefore, the ruin probability $\psi_\epsilon(u)$ has two different series expansions, the discard and the replace expansions. We first derive the discard series expansion.

Discard series expansion for the ruin probability

From a mathematical point of view, when we discard the heavy-tailed claim sizes, we simply consider that $G_\epsilon(x) = (1-\epsilon)F_p(x) + \epsilon$, $x \geq 0$. This base model, for which the claim size distribution has an atom at zero, is equivalent to the compound Poisson risk model in which claims arrive with rate $(1-\epsilon)\lambda$ and follow the distribution of B. We denote by $M_\epsilon^\bullet$ the supremum of its corresponding claim surplus process. Thus, the Pollaczek-Khinchine formula for this base model takes the form

$$\widetilde{m}_\epsilon^\bullet(s) := \mathbb{E}e^{-sM_\epsilon^\bullet} = \frac{1-(1-\epsilon)\rho_0}{1-(1-\epsilon)\rho_0\widetilde{F}_p^e(s)}. \tag{3.3}$$

We denote by $\psi_\epsilon^\bullet(u)$ the discard phase-type approximation of $\psi_\epsilon(u)$ that appears when we apply Laplace inversion to the above formula. For this base model, the series expansion of $\psi_\epsilon(u)$ can be found in the following theorem.

Theorem 3.1. Discard expansion. *If $\psi_\epsilon^\bullet(u)$ is the phase-type approximation of the exact ruin probability $\psi_\epsilon(u)$ that occurs when we discard the heavy-tailed claim sizes and $M_{\epsilon,k}^\bullet \overset{\mathfrak{D}}{=} M_\epsilon^\bullet$, a series expansion of the exact ruin probability is given by*

$$\psi_\epsilon(u) = \psi_\epsilon^\bullet(u) + \sum_{n=1}^{\infty}\left(\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^n\big(L_{\epsilon,n}(u) - L_{\epsilon,n-1}(u)\big),$$

where $L_{\epsilon,n}(u) = \mathbb{P}(M_{\epsilon,0}^\bullet + M_{\epsilon,1}^\bullet + \cdots + M_{\epsilon,n}^\bullet + C_1^e + \cdots + C_n^e > u)$ and $L_{\epsilon,0}(u) = \mathbb{P}(M_{\epsilon,0}^\bullet > u) = \psi_\epsilon^\bullet(u)$. The discard series expansion convergences for all values of u and ϵ.

Proof. From Eqs. (3.2) and (3.3) we find

$$\begin{aligned}
\widetilde{m}_\epsilon(s) =& \frac{1-(1-\epsilon)\rho_0-\epsilon\rho_1}{1-(1-\epsilon)\rho_0\widetilde{F}_p^e(s)-\epsilon\rho_1\widetilde{F}_h^e(s)} = \frac{\big(1-(1-\epsilon)\rho_0\big)-\epsilon\rho_1}{\frac{1-(1-\epsilon)\rho_0}{\widetilde{m}_\epsilon^\bullet(s)}-\epsilon\rho_1\widetilde{F}_h^e(s)} \\
=& \frac{1-\frac{\epsilon\rho_1}{1-(1-\epsilon)\rho_0}}{\frac{1}{\widetilde{m}_\epsilon^\bullet(s)}-\frac{\epsilon\rho_1}{1-(1-\epsilon)\rho_0}\widetilde{F}_h^e(s)} \\
=& \widetilde{m}_\epsilon^\bullet(s)\left(1-\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)\frac{1}{1-\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\widetilde{m}_\epsilon^\bullet(s)\widetilde{F}_h^e(s)} \\
=& \widetilde{m}_\epsilon^\bullet(s)\left(1-\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)\sum_{n=0}^{\infty}\left(\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^n\big(\widetilde{m}_\epsilon^\bullet(s)\widetilde{F}_h^e(s)\big)^n \\
=& \widetilde{m}_\epsilon^\bullet(s)\sum_{n=0}^{\infty}\left(\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^n\big(\widetilde{m}_\epsilon^\bullet(s)\widetilde{F}_h^e(s)\big)^n \\
&-\sum_{n=0}^{\infty}\left(\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^{n+1}\big(\widetilde{m}_\epsilon^\bullet(s)\big)^{n+1}\big(\widetilde{F}_h^e(s)\big)^n \\
=& \widetilde{m}_\epsilon^\bullet(s) \\
&+\sum_{n=1}^{\infty}\left(\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^n\Big[\big(\widetilde{m}_\epsilon^\bullet(s)\big)^{n+1}\big(\widetilde{F}_h^e(s)\big)^n-\big(\widetilde{m}_\epsilon^\bullet(s)\big)^n\big(\widetilde{F}_h^e(s)\big)^{n-1}\Big].
\end{aligned}$$

By using Laplace inversion we obtain

$$\psi_\epsilon(u) = \psi_\epsilon^\bullet(u) + \sum_{n=1}^{\infty}\left(\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^n\big(L_{\epsilon,n}(u)-L_{\epsilon,n-1}(u)\big),$$

where $L_{\epsilon,n}(u) = \mathbb{P}(M_{\epsilon,0}^\bullet + M_{\epsilon,1}^\bullet + \cdots + M_{\epsilon,n}^\bullet + C_1^e + \cdots + C_n^e > u)$. Note that this power series expansion is valid if and only if

$$\left|\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\widetilde{m}_\epsilon^\bullet(s)\widetilde{F}_h^e(s)\right| < 1.$$

We know that $\left|\widetilde{m}_\epsilon^\bullet(s)\widetilde{F}_h^e(s)\right| \le 1$. Moreover, for stability reasons we assumed that $\rho_\epsilon < 1$. Consequently, $\epsilon\rho_1 < 1-\rho_0+\epsilon\rho_0$. Thus, the above condition is always satisfied and the series converges for all values of ϵ. □

Replace series expansion for the ruin probability

To find the replace series expansion, observe that the action of replacing the heavy-tailed claim sizes with phase-type ones translates into $\epsilon = 0$. For this base model, the Pollaczeck-Khinchine formula takes the form

$$\widetilde{m}_0(s) := \mathbb{E}e^{-sM_0} = \frac{1-\rho_0}{1-\rho_0\widetilde{F}_p^e(s)}, \tag{3.4}$$

where $M_0 = M_{\epsilon|\epsilon=0}$. Laplace inversion of $\widetilde{m}_0(s)$ gives the phase-type approximation $\psi_0(u)$ of the ruin probability $\psi_\epsilon(u)$. The series expansion of $\psi_\epsilon(u)$ in this case is given below.

Theorem 3.2. Replace expansion. *If $\psi_0(u)$ is the phase-type approximation of the exact ruin probability $\psi_\epsilon(u)$ that occurs when we replace the heavy-tailed claim sizes with phase type ones and $M_{0,k} \overset{\mathfrak{D}}{=} M_0$, then a series expansion of the exact ruin probability is*

$$\begin{aligned}\psi_\epsilon(u) =&\psi_0(u) + \rho_1 \sum_{n=1}^{\infty} \left(\frac{\epsilon}{1-\rho_0}\right)^n \sum_{k=0}^{n-1} \binom{n-1}{k} \rho_1^k (-\rho_0)^{n-1-k} \\ &\times \big(L_{n,k+1,n-1-k}(u) - L_{n-1,k,n-1-k}(u)\big) \\ &-\rho_0 \sum_{n=1}^{\infty} \left(\frac{\epsilon}{1-\rho_0}\right)^n \sum_{k=0}^{n-1} \binom{n-1}{k} \rho_1^k (-\rho_0)^{n-1-k} \\ &\times \big(L_{n,k,n-k}(u) - L_{n-1,k,n-1-k}(u)\big),\end{aligned}$$

where $L_{l,m,r}(u) = \mathbb{P}(M_{0,0} + M_{0,1} + \cdots + M_{0,l} + C_1^e + \cdots + C_m^e + B_1^e + \cdots + B_r^e > u)$ and $L_{0,0,0}(u) = \psi_0(u)$. A sufficient condition for the convergence of the replace series expansion for all values of u is $\epsilon < |1-\rho_0| \,/\, \max\{\rho_0, \rho_1\}$.

Proof. We set $D(s) = \rho_1 \widetilde{F}_h^e(s) - \rho_0 \widetilde{F}_p^e(s)$. By using (3.2) and (3.4) we find

$$\begin{aligned}\widetilde{m}_\epsilon(s) =&\frac{1-(1-\epsilon)\rho_0 - \epsilon\rho_1}{1-(1-\epsilon)\rho_0 \widetilde{F}_p^e(s) - \epsilon\rho_1 \widetilde{F}_h^e(s)} = \frac{(1-\rho_0) - \epsilon(\rho_1-\rho_0)}{1-\rho_0\widetilde{F}_p^e(s) - \epsilon\big(\rho_1\widetilde{F}_h^e(s) - \rho_0\widetilde{F}_p^e(s)\big)} \\ =&\frac{(1-\rho_0) - \epsilon(\rho_1-\rho_0)}{\frac{1-\rho_0}{\widetilde{m}_0(s)} - \epsilon D(s)} = \frac{1-\epsilon\frac{\rho_1-\rho_0}{1-\rho_0}}{\frac{1}{\widetilde{m}_0(s)} - \frac{\epsilon}{1-\rho_0} D(s)} \\ =&\widetilde{m}_0(s)\left(1-\epsilon\frac{\rho_1-\rho_0}{1-\rho_0}\right)\frac{1}{1-\frac{\epsilon}{1-\rho_0}\widetilde{m}_0(s)D(s)} \\ =&\widetilde{m}_0(s)\left(1-\epsilon\frac{\rho_1-\rho_0}{1-\rho_0}\right)\sum_{n=0}^{\infty}\left(\frac{\epsilon}{1-\rho_0}\right)^n \big(\widetilde{m}_0(s)D(s)\big)^n \\ =&\widetilde{m}_0(s)\sum_{n=0}^{\infty}\left(\frac{\epsilon}{1-\rho_0}\right)^n \big(\widetilde{m}_0(s)D(s)\big)^n \\ &-(\rho_1-\rho_0)\sum_{n=0}^{\infty}\left(\frac{\epsilon}{1-\rho_0}\right)^{n+1} \big(\widetilde{m}_0(s)\big)^{n+1}\big(D(s)\big)^n \\ =&\widetilde{m}_0(s) + \sum_{n=1}^{\infty}\left(\frac{\epsilon}{1-\rho_0}\right)^n \big(\widetilde{m}_0(s)\big)^n \Big[\widetilde{m}_0(s)\big(D(s)\big)^n - (\rho_1-\rho_0)\big(D(s)\big)^{n-1}\Big].\end{aligned}$$

But,

$$\begin{aligned}&\widetilde{m}_0(s)\big(D(s)\big)^n - (\rho_1-\rho_0)\big(D(s)\big)^{n-1} \\ &\quad=\widetilde{m}_0(s)\sum_{k=0}^{n}\binom{n}{k}\big(\rho_1\widetilde{F}_h^e(s)\big)^k\big(-\rho_0\widetilde{F}_p^e(s)\big)^{n-k} \\ &\qquad-(\rho_1-\rho_0)\sum_{k=0}^{n-1}\binom{n-1}{k}\big(\rho_1\widetilde{F}_h^e(s)\big)^k\big(-\rho_0\widetilde{F}_p^e(s)\big)^{n-1-k}\end{aligned}$$

$$
\begin{aligned}
=&\widetilde{m}_0(s)\Bigg[\sum_{k=1}^{n-1}\left(\binom{n-1}{k}+\binom{n-1}{k-1}\right)\big(\rho_1\widetilde{F}_h^e(s)\big)^k\big(-\rho_0\widetilde{F}_p^e(s)\big)^{n-k}\\
&+\big(\rho_1\widetilde{F}_h^e(s)\big)^n+\big(-\rho_0\widetilde{F}_p^e(s)\big)^n\Bigg]\\
&-(\rho_1-\rho_0)\sum_{k=0}^{n-1}\binom{n-1}{k}\big(\rho_1\widetilde{F}_h^e(s)\big)^k\big(-\rho_0\widetilde{F}_p^e(s)\big)^{n-1-k}\\
=&\rho_1\sum_{k=0}^{n-1}\binom{n-1}{k}\big(\widetilde{m}_0(s)\widetilde{F}_h^e(s)-1\big)\big(\rho_1\widetilde{F}_h^e(s)\big)^k\big(-\rho_0\widetilde{F}_p^e(s)\big)^{n-1-k}\\
&-\rho_0\sum_{k=0}^{n-1}\binom{n-1}{k}\big(\widetilde{m}_0(s)\widetilde{F}_p^e(s)-1\big)\big(\rho_1\widetilde{F}_h^e(s)\big)^k\big(-\rho_0\widetilde{F}_p^e(s)\big)^{n-1-k}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\widetilde{m}_\epsilon(s)=&\widetilde{m}_0(s)\\
&+\rho_1\sum_{n=1}^{\infty}\left(\frac{\epsilon}{1-\rho_0}\right)^n\Big(\big(\widetilde{m}_0(s)\big)^{n+1}\widetilde{F}_h^e(s)-\big(\widetilde{m}_0(s)\big)^n\Big)\\
&\times\sum_{k=0}^{n-1}\binom{n-1}{k}\big(\rho_1\widetilde{F}_h^e(s)\big)^k\big(-\rho_0\widetilde{F}_p^e(s)\big)^{n-1-k}\\
&-\rho_0\sum_{n=1}^{\infty}\left(\frac{\epsilon}{1-\rho_0}\right)^n\Big(\big(\widetilde{m}_0(s)\big)^{n+1}\widetilde{F}_p^e(s)-\big(\widetilde{m}_0(s)\big)^n\Big)\\
&\times\sum_{k=0}^{n-1}\binom{n-1}{k}\big(\rho_1\widetilde{F}_h^e(s)\big)^k\big(-\rho_0\widetilde{F}_p^e(s)\big)^{n-1-k}.
\end{aligned}
$$

By applying Laplace inversion we find

$$
\begin{aligned}
\psi_\epsilon(u)=&\psi_0(u)+\rho_1\sum_{n=1}^{\infty}\left(\frac{\epsilon}{1-\rho_0}\right)^n\sum_{k=0}^{n-1}\binom{n-1}{k}\rho_1^k(-\rho_0)^{n-1-k}\\
&\times\big(L_{n,k+1,n-1-k}(u)-L_{n-1,k,n-1-k}(u)\big)\\
&-\rho_0\sum_{n=1}^{\infty}\left(\frac{\epsilon}{1-\rho_0}\right)^n\sum_{k=0}^{n-1}\binom{n-1}{k}\rho_1^k(-\rho_0)^{n-1-k}\\
&\times\big(L_{n,k,n-k}(u)-L_{n-1,k,n-1-k}(u)\big),
\end{aligned}
$$

where $L_{l,m,r}(u)=\mathbb{P}(M_{0,0}+M_{0,1}+\cdots+M_{0,l}+C_1^e+\cdots+C_m^e+B_1^e+\cdots+B_r^e>u)$. Similarly to the discard expansion, the replace series converges for a given value of s if and only if

$$
\left|\frac{\epsilon}{1-\rho_0}\widetilde{m}_0(s)\big(\rho_1\widetilde{F}_h^e(s)-\rho_0\widetilde{F}_p^e(s)\big)\right|<1.
$$

If $\sigma=\max_s\left|\widetilde{m}_0(s)\big(\rho_1\widetilde{F}_h^e(s)-\rho_0\widetilde{F}_p^e(s)\big)\right|$, then a necessary and sufficient condition for the convergence of the replace series for all values of s is $\epsilon<|1-\rho_0|/\sigma$. However, we do

not have exact formulas for the Laplace transforms $\widetilde{m}_0(s)$, $\widetilde{F}_h^e(s)$, and $\widetilde{F}_p^e(s)$, and thus we can only find a sufficient condition for the convergence of the series. The Laplace transform of a distribution on a positive support, such as $\widetilde{F}_h^e(s)$ and $\widetilde{F}_p^e(s)$, is a positive function bounded by one. Therefore, it is immediate that $\max_{s\geq 0}\left|\rho_1\widetilde{F}_h^e(s)-\rho_0\widetilde{F}_p^e(s)\right| \leq \max\{\rho_0,\rho_1\}$. In addition, since $|\widetilde{m}_0(s)|\leq 1$, a sufficient condition for the convergence of the replace series is $\epsilon < |1-\rho_0| \,/\, \max\{\rho_0,\rho_1\}$. □

Note that Theorem 3.2 gives only a sufficient condition for the convergence of the replace series expansion. If all parameters involved are explicitly known, one can find a necessary condition in the way indicated in the proof of Theorem 3.2. In the next section, we propose two explicit approximations for the ruin probability based on these series expansions.

3.3 Corrected phase-type approximations of the ruin probability

The goal of this section is to provide approximations that maintain the numerical tractability but improve the accuracy of the phase-type approximations and that are able to capture the tail behaviour of the exact ruin probability. Large deviations theory suggests that a single catastrophic event, i.e. a heavy-tailed stationary claim size C^e, is sufficient to cause ruin (Embrechts et al., 1997). Observe that, for both the discard and replace series expansions, the second term contains a single appearance of C^e. For this reason, the proposed approximations for the ruin probability are constructed by the first two terms of their respective series expansions for the ruin probability (see Theorems 3.1 and 3.2), where the second term of each approximation is referred to as its correction term. We have the following definitions for the proposed approximations.

Definition 3.3. The corrected discard approximation of exact ruin probability $\psi_\epsilon(u)$ is defined as

$$\widehat{\psi}_{d,\epsilon}(u) := \psi_\epsilon^\bullet(u) + \frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\big(\mathbb{P}(M_{\epsilon,0}^\bullet + M_{\epsilon,1}^\bullet + C_1^e > u) - \mathbb{P}(M_{\epsilon,0}^\bullet > u)\big), \quad (3.5)$$

where $\psi_\epsilon^\bullet(u)$ is the discard phase-type approximation of $\psi_\epsilon(u)$.

In a similar manner, we define the approximation that connects to the replace expansion.

Definition 3.4. The corrected replace approximation of the exact ruin probability $\psi_\epsilon(u)$ is given by the formula

$$\begin{aligned}\widehat{\psi}_{r,\epsilon}(u) := \psi_0(u) &+ \frac{\epsilon\rho_1}{1-\rho_0}\big(\mathbb{P}(M_{0,0}+M_{0,1}+C_1^e>u)-\mathbb{P}(M_{0,0}>u)\big)\\ &- \frac{\epsilon\rho_0}{1-\rho_0}\big(\mathbb{P}(M_{0,0}+M_{0,1}+B_1^e>u)-\mathbb{P}(M_{0,0}>u)\big), \quad (3.6)\end{aligned}$$

where $\psi_0(u)$ is the replace phase-type approximation of $\psi_\epsilon(u)$.

In the following sections, we study characteristics of the corrected discard and the corrected replace approximations.

3.3.1 Approximation errors

Due to the construction of the two corrected phase-type approximations, the discard and the replace, their difference from the exact ruin probability is the sum of the remaining terms, namely the terms for $n \geq 2$. For the error of the corrected discard approximation, we have the following theorem.

Theorem 3.5. *The error of the corrected discard approximation is bounded from above and below as follows:*

$$\left(\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^2 \left(L_{\epsilon,2}(u) - L_{\epsilon,1}(u)\right) \leq \psi_\epsilon(u) - \widehat{\psi}_{d,\epsilon}(u) \leq \left(\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^2.$$

Proof. An interesting observation is that we can interpret the terms $L_{\epsilon,n}(u) - L_{\epsilon,n-1}(u)$ in Theorem 3.1 in terms of a renewal process $\{N_{D,\epsilon}(u), u \geq 0\}$ with a delayed first renewal $M^{\bullet}_{\epsilon,0}$. Consequently, $\mathbb{P}(N_{D,\epsilon}(u) = 0) = L_{\epsilon,0}(u)$ and $\mathbb{P}(N_{D,\epsilon}(u) = n) = L_{\epsilon,n}(u) - L_{\epsilon,n-1}(u)$, for $n \geq 1$. As a result,

$$\begin{aligned}
\psi_\epsilon(u) - \widehat{\psi}_{d,\epsilon}(u) &= \sum_{n=2}^{\infty} \left(\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^n \mathbb{P}(N_{D,\epsilon}(u) = n) \\
&= \left(\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^2 \mathbb{E}\left[\left(\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^{N_{D,\epsilon}(u)-2} \mathbb{1}\left(N_{D,\epsilon}(u) \geq 2\right)\right] \\
&\leq \left(\frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^2,
\end{aligned}$$

where the latter inequality holds because $\dfrac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0} < 1$. Thus, an upper bound for the approximation error is $\left(\dfrac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^2$. Due to the renewal argument, all terms in the discard series expansion are positive. Consequently, the corrected discard approximation always underestimates the exact ruin probability and the term $\left(\dfrac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\right)^2 \times \left(L_{\epsilon,2}(u) - L_{\epsilon,1}(u)\right)$ is a lower bound for the achieved error. □

Remark 3.6. Theorem 3.5 shows that the corrected discard approximation always underestimates the exact ruin probability, and its error is $O(\epsilon^2)$. Thus, the corrected discard approximation is a lower bound for the exact ruin probability.

As done in the proof of Theorem 3.5, similar probabilistic interpretations can also be given to the terms of the replace series expansion. However, due to the sign changes in the formula of the replace expansion (see Theorem 3.2), it is not immediate whether the corrected replace approximation underestimates or overestimates the exact ruin probability. This depends on the characteristics of the distributions involved. As we see in Section 3.4, both overestimation and underestimation are possible. Studying the areas of over- or underestimation of the ruin probability is beyond the scope of this chapter. In the sequel, we provide only absolute error bounds for the corrected replace approximation.

There are many possible ways to bound the error of the corrected replace approximation. For example, one could ignore all negative terms for $n \geq 2$ in the replace expansion and bound all positive terms. Of course, different techniques give different bounds. Among the different bounds we found, we present in Theorem 3.7 the one that is valid for the biggest range of the perturbation parameter ϵ.

Theorem 3.7. *When $\epsilon < |1-\rho_0|/(\rho_0+\rho_1)$, an upper bound for the absolute error that we achieve with the corrected replace approximation is*

$$\left|\psi_\epsilon(u) - \widehat{\psi}_{r,\epsilon}(u)\right| \leq \left(\frac{\epsilon}{1-\rho_0}\right)^2 (\rho_0+\rho_1)^2 \frac{1-\rho_0}{1-\rho_0-\epsilon(\rho_0+\rho_1)}.$$

Proof. Using the triangular inequality and the fact that the distance between two distributions is smaller than or equal to 1, we obtain

$$\begin{aligned}\left|\psi_\epsilon(u) - \widehat{\psi}_{r,\epsilon}(u)\right| &\leq (\rho_0+\rho_1)\sum_{n=2}^{\infty}\left(\frac{\epsilon}{1-\rho_0}\right)^n \sum_{k=0}^{n-1}\binom{n-1}{k}\rho_1^k\rho_0^{n-1-k} \\ &= (\rho_0+\rho_1)^2\left(\frac{\epsilon}{1-\rho_0}\right)^2 \sum_{n=2}^{\infty}\left(\frac{\epsilon}{1-\rho_0}(\rho_0+\rho_1)\right)^{n-2} \\ &= \left(\frac{\epsilon}{1-\rho_0}\right)^2 (\rho_0+\rho_1)^2 \frac{1-\rho_0}{1-\rho_0-\epsilon(\rho_0+\rho_1)},\end{aligned}$$

where the result holds only for $\epsilon(\rho_0+\rho_1)/|1-\rho_0| < 1$. □

Remark 3.8. Theorem 3.7 shows that the absolute error of the replace approximation is $O(\epsilon^2)$. Note that the expression

$$\left(\frac{\epsilon}{1-\rho_0}\right)^2 \sum_{k=0}^{1}\rho_1^k(-\rho_0)^{1-k}\Big[\rho_1\big(L_{2,k+1,1-k}(u) - L_{1,k,1-k}(u)\big) - \rho_0\big(L_{2,k,2-k}(u) - L_{1,k,1-k}(u)\big)\Big],$$

which corresponds to the term of the replace expansion (see Theorem 3.2) for $n=2$, is $O(\epsilon^2)$ and it could be used alternatively as an approximation of the real error.

An advantage of the corrected discard approximation over the corrected replace is the following. The fact that the corrected discard approximation underestimates the exact ruin probability gives a positive sign for its error, namely its difference from the exact ruin probability, which according to Theorem 3.5 is bounded from above and below. This information with respect to the nature of its error makes the corrected discard approximation much more controllable than the corrected replace approximation. In the next section, we study the tail behaviour of both corrected phase-type approximations.

3.3.2 Tail behaviour

To study the tail behaviour of the two approximations, we assume that the distribution of C^e belongs to the class of subexponential distributions $\mathcal{S}$; see Appendix A.1. Before

studying the tail behaviour of the approximations, we first give the tail behaviour of the exact ruin probability in the next theorem.

Theorem 3.9. *When $C^e \in \mathcal{S}$, the exact ruin probability $\psi_\epsilon(u)$ has the following tail behaviour:*

$$\psi_\epsilon(u) \sim \frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0-\epsilon\rho_1}\overline{F_h^e}(u).$$

Proof. When B has a phase-type distribution, then B^e has also a phase-type distribution (Asmussen and Albrecher, 2010), and consequently it has an exponential decay rate. Thus, by the definition of the stationary excess claim sizes U_ϵ^e (see Sections 3.2 and 1.7.2) and Property A.3, we have

$$\begin{aligned}\overline{G_\epsilon^e}(u) &= \frac{(1-\epsilon)\rho_0}{(1-\epsilon)\rho_0+\epsilon\rho_1}\overline{F_p^e}(u) + \frac{\epsilon\rho_1}{(1-\epsilon)\rho_0+\epsilon\rho_1}\overline{F_h^e}(u)\\ &\sim \frac{\epsilon\rho_1}{(1-\epsilon)\rho_0+\epsilon\rho_1}\overline{F_h^e}(u),\end{aligned} \tag{3.7}$$

which implies by Property A.2 that $U_\epsilon^e \in \mathcal{S}$. When $U_\epsilon^e \in \mathcal{S}$, it is known (Asmussen and Albrecher, 2010) that

$$\psi_\epsilon(u) \sim \frac{\rho_\epsilon}{1-\rho_\epsilon}\overline{G_\epsilon^e}(u), \tag{3.8}$$

where $\rho_\epsilon = (1-\epsilon)\rho_0 + \epsilon\rho_1 < 1$. Combining (3.7) and (3.8) yields the result. □

For the tail behaviour of the corrected discard approximation, the following result holds.

Theorem 3.10. *When $C^e \in \mathcal{S}$, we have for the corrected discard approximation the following tail behaviour:*

$$\widehat{\psi}_{d,\epsilon}(u) \sim \frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}\overline{F_h^e}(u).$$

Proof. The discard approximation $\psi_\epsilon^\bullet(u)$ has a phase-type representation; therefore, it is of $o\left(\overline{F_h^e}(u)\right)$. The same holds for the tail of the distribution of $M_{\epsilon,0}^\bullet + M_{\epsilon,1}^\bullet$. Moreover, since $C^e \in \mathcal{S}$, from Property A.3 we obtain $\mathbb{P}(M_{\epsilon,0}^\bullet + M_{\epsilon,1}^\bullet + C_1^e > u) \sim \overline{F_h^e}(u)$, which leads to the result by inserting these asymptotic estimates into (3.5). □

Theorem 3.10 shows that the corrected discard approximation captures the heavy-tailed behaviour of the exact ruin probability, but is off by a term $\epsilon\rho_1$ in the denominator. In fact, for all values of parameters, the tail of the discard approximation is always below the tail of the exact ruin probability, which is expected since the discard approximation gives an underestimation of the exact ruin probability.

On the other hand, for the tail behaviour of the corrected replace approximation of the ruin probability, the following result holds.

Theorem 3.11. *When $C^e \in \mathcal{S}$, we have for the corrected replace approximation of the ruin probability the following tail behaviour:*

$$\widehat{\psi}_{r,\epsilon}(u) \sim \frac{\epsilon\rho_1}{1-\rho_0}\overline{F_h^e}(u).$$

Proof. The class of phase-type distributions is closed under convolutions (Asmussen and Albrecher, 2010), which means that both $M_{0,0}+M_{0,1}$ and $M_{0,0}+M_{0,1}+B_1^e$ follow some phase-type distribution. Therefore, due to their exponential decay rate, $\psi_0(u)$, $\mathbb{P}(M_{0,0}+M_{0,1}>u)$ and $\mathbb{P}(M_{0,0}+M_{0,1}+B_1^e>u)$ are all of the order $o\left(\overline{F_h^e}(u)\right)$. In addition, since $C^e \in \mathcal{S}$, we obtain from Property A.3 that $\mathbb{P}(M_{0,0}+M_{0,1}+C_1^e>u) \sim \overline{F_h^e}(u)$. Inserting these asymptotic estimates into (3.6) leads to the result. □

Comparing the coefficients of $\overline{F_h^e}(u)$ in Theorems 3.10 and 3.11, we observe that the tail of the corrected replace approximation is always above the tail of the corrected discard approximation. To compare the tail behaviour of the corrected replace approximation to that of the exact ruin probability, we only need to compare the coefficients of $\overline{F_h^e}(u)$, and more precisely their denominators, as the expression with the largest denominator converges to zero faster. Therefore, the tails have the same behaviour when $\mathbb{E}B = \mathbb{E}C$, while the tail of the corrected replace approximation is above the tail of the exact ruin probability when $\mathbb{E}B > \mathbb{E}C$ and below when $\mathbb{E}B < \mathbb{E}C$.

3.3.3 Relative error

Following the results of Section 3.3.2, we show that the relative error at the tail for both approximations is $O(\epsilon)$.

Lemma 3.12. *When $C^e \in \mathcal{S}$, the relative error at the tail of the corrected discard approximation is*

$$R_{d,\epsilon}(u) = 1 - \frac{\widehat{\psi}_{d,\epsilon}(u)}{\psi_\epsilon(u)} \to \frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0}, \qquad as \quad u \to \infty.$$

Recall that for the corrected replace approximation, different values of parameters lead to both over- and under-estimation of the exact ruin probability. Thus, for this approximation it is more appropriate to evaluate the absolute relative error at its tail.

Lemma 3.13. *When $C^e \in \mathcal{S}$, the absolute relative error at the tail of the corrected replace approximation is*

$$|R_{r,\epsilon}(u)| = \left|1 - \frac{\widehat{\psi}_{r,\epsilon}(u)}{\psi_\epsilon(u)}\right| \to \left|\frac{\epsilon(\rho_1-\rho_0)}{1-\rho_0}\right|, \qquad as \quad u \to \infty,$$

and it goes asymptotically to zero when $\mathbb{E}B = \mathbb{E}C$.

Remark 3.14. Lemmas 3.12 and 3.13 indicate that the relative errors of both corrected phase-type approximations do not converge to 0 as $u \to \infty$. However, the approximations give the exact value of the ruin probability at the origin and have guaranteed bounds of the order $O(\epsilon^2)$ for all values of u. On the other hand, the asymptotic result of Theorem 3.9 has the correct tail behaviour but it gives relatively inaccurate estimates of the ruin probability for small values of u for some combinations of the involved parameters. In order to provide a compromise between our approximations and the asymptotic result of Theorem 3.9, one can simply change the coefficients $\epsilon\rho_1/(1-\rho_0+\epsilon\rho_0)$ and $\epsilon\rho_1/(1-\rho_0)$ of the correction terms (see Definitions 3.3 and 3.4, respectively) to $\epsilon\rho_1/(1-\rho_0+\epsilon\rho_0-\epsilon\rho_1)$, so that their tail

behaviour matches the correct tail behaviour. Of course, one should also multiply the first terms of the approximations with proper coefficients to obtain $\psi_\epsilon^\bullet(0) = \psi_0(0) = \psi_\epsilon(0)$. Such adjustments will lead to approximations with relative error at the tail that is asymptotically equal to zero. Moreover, the approximations work well for all values of the involved parameters, but may give worse results for small values of u when they are compared with the original corrected phase-type approximations.

The fact that the discard approximation always underestimates the ruin probability raises the question if it is possible to develop a result for its relative error for arbitrary values of u. The next theorem, which can be seen as the main technical contribution of this chapter, shows that this is indeed possible.

Theorem 3.15. *When $C^e \in \mathcal{S}$, there exists an $\eta > 0$, such that for all $\epsilon < \eta$, the relative error $R_{d,\epsilon}(u)$ of the discard approximation at the point u can be bounded by*

$$R_{d,\epsilon}(u) \leq \frac{\epsilon\rho_1}{1-\rho_0+\epsilon\rho_0} H_\epsilon(u) + \epsilon^2 K,$$

with $H_\epsilon(u) = \left(\dfrac{\mathbb{P}(M_{\epsilon,0}^\bullet + M_{\epsilon,1}^\bullet + M_{\epsilon,2}^\bullet + C_1^e + C_2^e > u)}{\mathbb{P}(M_{\epsilon,0}^\bullet + M_{\epsilon,1}^\bullet + C_1^e > u)} - 1 \right)$ *and K a finite constant.*

In order to prove Theorem 3.15, we first need the following lemma.

Lemma 3.16. *Let $X_{\epsilon,k}$ be an i.i.d. sequence such that $X_{\epsilon,k} \overset{\mathfrak{D}}{=} M_{\epsilon,0}^\bullet + M_{\epsilon,k}^\bullet + C_k^e$. There exists a constant K_0 independent of ϵ, such that*

$$\frac{\mathbb{P}(X_{\epsilon,1} + \cdots + X_{\epsilon,n} > u)}{\mathbb{P}(X_{\epsilon,1} > u)} \leq K_0^n,$$

for all u and for all n.

Proof. We follow a similar idea as the proof of Embrechts et al. (1997, Lemma 1.3.5), which is not directly applicable, as $X_{\epsilon,k}$ depends on ϵ. Let F be the distribution function of $X_{\epsilon,k}$. Since C_k^e is subexponential, and $M_{\epsilon,0}^\bullet$, $M_{\epsilon,k}^\bullet$ are light-tailed, according to Property A.3, $X_{\epsilon,k}$ is subexponential as well. We set $\alpha_n = \sup_u \overline{F^{*n}}(u)/\overline{F}(u)$. Observe that

$$\begin{aligned}\frac{\overline{F^{*(n+1)}}(u)}{\overline{F}(u)} &= 1 + \int_0^u \frac{\overline{F^{*n}}(u-x)}{\overline{F}(u)} dF(x) = 1 + \int_0^u \frac{\overline{F^{*n}}(u-x)}{\overline{F}(u-x)} \frac{\overline{F}(u-x)}{\overline{F}(u)} dF(x) \\ &\leq 1 + \frac{\alpha_n}{\overline{F}(u)} \left(\overline{F^{*2}}(u) - \overline{F}(u) \right) \leq 1 + \alpha_n(\alpha_2 - 1).\end{aligned}$$

Recursively, we find that

$$\alpha_{n+1} \leq \sum_{k=0}^{n-2} (\alpha_2 - 1)^k + \alpha_2(\alpha_2 - 1)^{n-1}.$$

From Definition A.1, we know that $\alpha_2 - 1 \geq 1$. So,

$$\alpha_{n+1} \leq \sum_{k=0}^{n-2} \alpha_2^k + \alpha_2^n \leq \sum_{k=0}^{n} \alpha_2^k \leq \alpha_2^{n+1},$$

therefore it suffices to show that α_2 is bounded in $\epsilon > 0$.

To this end, observe that $M^{\bullet}_{\epsilon,k}$ is stochastically decreasing in ϵ as it is the supremum of a compound Poisson process with arrival rate $\lambda(1-\epsilon)$. Therefore, the supremum that corresponds to the compound Poisson process with arrival rate λ ($\epsilon = 0$) is stochastically larger than all other suprema with $\epsilon > 0$ and we denote it by $M_{0,k}$. Letting $S = \sum_{k=1}^{4} M_{0,k}$, we see that

$$\begin{aligned}
\frac{\overline{F^{*2}}(u)}{\overline{F}(u)} =& \frac{\mathbb{P}(X_{\epsilon,1} + X_{\epsilon,2} > u)}{\mathbb{P}(X_{\epsilon,1} > u)} \leq \frac{\mathbb{P}(S + C^e_1 + C^e_2 > u)}{\mathbb{P}(C^e_1 > u)} \\
=& \frac{\mathbb{P}(S > u)}{\mathbb{P}(C^e_1 > u)} + \int_0^u \frac{\mathbb{P}(C^e_1 + C^e_2 > u - x)}{\mathbb{P}(C^e_1 > u - x)} \frac{\mathbb{P}(C^e_1 > u - x)}{\mathbb{P}(C^e_1 > u)} d\mathbb{P}(S \leq x) \\
\leq& \frac{\mathbb{P}(S > u)}{\mathbb{P}(C^e_1 > u)} \\
&+ \underbrace{\sup_{u>0} \frac{\mathbb{P}(C^e_1 + C^e_2 > u)}{\mathbb{P}(C^e_1 > u)}}_{>1} \frac{1}{\mathbb{P}(C^e_1 > u)} \int_0^u \mathbb{P}(C^e_1 > u - x) d\mathbb{P}(S \leq x) \\
&\leq \sup_{u>0} \frac{\mathbb{P}(C^e_1 + C^e_2 > u)}{\mathbb{P}(C^e_1 > u)} \sup_{u>0} \frac{\mathbb{P}(S + C^e_1 > u)}{\mathbb{P}(C^e_1 > u)}.
\end{aligned}$$

Both suprema are finite since C^e_1 is subexponential and S has a lighter tail than C^e_1. This completes the proof of the lemma. □

We now proceed with the proof of Theorem 3.15.

Proof of Theorem 3.15. Similarly to $X_{\epsilon,k}$, let $Y_{\epsilon,k}$ be an i.i.d. sequence such that $Y_{\epsilon,k} \overset{\mathfrak{D}}{=} M^{\bullet}_{\epsilon,k} + C^e_k$, and set $p_\epsilon = \dfrac{\epsilon\rho_1}{1 - \rho_0 + \epsilon\rho_0}$. Let η be such that $p_\eta K_0 = 1/2$ and suppose $\epsilon < \eta$. In addition, let N be a random variable such that $\mathbb{P}(N = n) = (1-p_\epsilon)p^n_\epsilon$ and observe that $M_\epsilon \overset{\mathfrak{D}}{=} M^{\bullet}_{\epsilon,0} + \sum_{k=1}^{N} Y_{\epsilon,k}$. For notational convenience, we assume that this equality holds almost surely through this proof. This enables us to write

$$\psi_\epsilon(u) - \widehat{\psi}_{d,\epsilon}(u) = \mathbb{P}(M_\epsilon > u; N \geq 2) - p^2_\epsilon \mathbb{P}(M^{\bullet}_{\epsilon,0} + Y_{\epsilon,1} > u),$$

so that

$$\begin{aligned}
R_{d,\epsilon}(u) &= \frac{\mathbb{P}(M_\epsilon > u; N \geq 2) - p^2_\epsilon \mathbb{P}(M^{\bullet}_{\epsilon,0} + Y_{\epsilon,1} > u)}{\mathbb{P}(M_\epsilon > u)} \\
&= \frac{\mathbb{P}(M_\epsilon > u; N \geq 2) - p^2_\epsilon \mathbb{P}(M^{\bullet}_{\epsilon,0} + Y_{\epsilon,1} > u)}{\mathbb{P}(M_\epsilon > u; N \geq 1)} \cdot \frac{\mathbb{P}(M_\epsilon > u; N \geq 1)}{\mathbb{P}(M_\epsilon > u)}. \qquad (3.9)
\end{aligned}$$

Note that $\mathbb{P}(M_\epsilon > u; N \geq 1)/\mathbb{P}(M_\epsilon > u) \leq 1$, where this ratio actually converges to 1 as $u \to \infty$. To analyse the other fraction of (3.9), from the memoryless property of N we obtain $\mathbb{P}(M_\epsilon > u; N \geq k) = p^k_\epsilon \mathbb{P}(M_\epsilon + Y_{\epsilon,1} + \cdots + Y_{\epsilon,k} > u)$ so

$$\begin{aligned}
\frac{\mathbb{P}(M_\epsilon > u; N \geq 2)}{\mathbb{P}(M_\epsilon > u; N \geq 1)} =& p_\epsilon \frac{\mathbb{P}(M_\epsilon + Y_{\epsilon,1} + Y_{\epsilon,2} > u)}{\mathbb{P}(M_\epsilon + Y_{\epsilon,1} > u)} \leq p_\epsilon \frac{\mathbb{P}(M_\epsilon + Y_{\epsilon,1} + Y_{\epsilon,2} > u)}{\mathbb{P}(X_{\epsilon,1} > u)} \\
\leq& p_\epsilon \mathbb{P}(N = 0) \frac{\mathbb{P}(X_{\epsilon,1} + Y_{\epsilon,2} > u)}{\mathbb{P}(X_{\epsilon,1} > u)}
\end{aligned}$$

$$
\begin{aligned}
&+ p_\epsilon \sum_{n=1}^{\infty} \mathbb{P}(N=n) \frac{\mathbb{P}(X_{\epsilon,1} + \cdots + X_{\epsilon,n+2} > u)}{\mathbb{P}(X_{\epsilon,1} > u)} \\
\leq & p_\epsilon \frac{\mathbb{P}(X_{\epsilon,1} + Y_{\epsilon,2} > u)}{\mathbb{P}(X_{\epsilon,1} > u)} + p_\epsilon^2 (1-p_\epsilon) K_0^3 \sum_{n=1}^{\infty} (p_\epsilon K_0)^{n-1} \\
\leq & \, p_\epsilon \frac{\mathbb{P}(X_{\epsilon,1} + Y_{\epsilon,2} > u)}{\mathbb{P}(X_{\epsilon,1} > u)} + p_\epsilon^2 2K_0^3.
\end{aligned}
$$

Finally, note that

$$
\frac{p_\epsilon^2 \mathbb{P}(M_{\epsilon,0}^{\bullet} + Y_{\epsilon,1} > u)}{\mathbb{P}(M_\epsilon > u; N \geq 1)} = p_\epsilon \frac{\mathbb{P}(M_{\epsilon,0}^{\bullet} + Y_{\epsilon,1} > u)}{\mathbb{P}(M_\epsilon + Y_{\epsilon,1} > u)}.
$$

As before, we can show there exists a constant K_1 such that $\dfrac{\mathbb{P}(M_\epsilon + Y_{\epsilon,1} > u)}{\mathbb{P}(M_{\epsilon,0}^{\bullet} + Y_{\epsilon,1} > u)} \leq 1 + p_\epsilon K_1$. Putting everything together, we conclude that

$$
\begin{aligned}
R_{d,\epsilon}(u) &\leq p_\epsilon \frac{\mathbb{P}(X_{\epsilon,1} + Y_{\epsilon,2} > u)}{\mathbb{P}(X_{\epsilon,1} > u)} + p_\epsilon^2 2K_0^3 - p_\epsilon \frac{1}{1 + p_\epsilon K_1} \\
&\leq p_\epsilon \left(\frac{\mathbb{P}(X_{\epsilon,1} + Y_{\epsilon,2} > u)}{\mathbb{P}(X_{\epsilon,1} > u)} - 1 \right) + p_\epsilon^2 K,
\end{aligned}
$$

for some constant K, completing the proof. □

The bound is sharp in the sense that $H_\epsilon(u) \to 1$ as $u \to \infty$, which recovers the relative error at the tail, up to a term $O(\epsilon^2)$. Moreover, $H_\epsilon(u)$ is uniformly bounded in u and ϵ.

3.4 Numerical experiments

In Section 3.2, we pointed out that the first terms of the discard and the replace expansions are phase-type approximations of $\psi_\epsilon(u)$. The goal of this section is to show numerically that adding the second term of these expansions leads to improved approximations (corrected discard and corrected replace approximations respectively) that are significantly more accurate than their phase-type counterparts. Moreover, the additional term has a great impact on the accuracy of the improved approximations even for small values of the perturbation parameter.

Therefore, in this section we check the accuracy of the corrected discard (see Definition 3.3) and the corrected replace approximations (see Definition 3.4) by comparing them with the exact ruin probability and their corresponding phase-type approximations. Since it is more meaningful to compare approximations with exact results than with simulation outcomes, we choose the general claim size distributions G_ϵ such that there exists an exact formula for the ruin probability $\psi_\epsilon(u)$.

In Section 3.4.1, we derive the exact formula for the ruin probability $\psi_\epsilon(u)$ for a specific choice of the claim size distribution. Using the latter claim size distribution, in Section 3.4.2, we perform our numerical experiments and we draw our conclusions.

3.4.1 Test distribution

As claim size distribution we use a mixture of an exponential distribution with rate ν and a heavy-tailed one that belongs to a class of long-tailed distributions introduced in Abate and Whitt (1999b) (see also Section 2.5.1). The Laplace transform of the latter distribution is

$$\widetilde{F}_h(s) = 1 - \frac{s}{(\kappa + \sqrt{s})(1 + \sqrt{s})},$$

where $\mathbb{E}C = \kappa^{-1}$ and all higher moments are infinite. Furthermore, the Laplace transform of the stationary heavy-tailed excess claim size distribution is

$$\widetilde{F}_h^e(s) = \frac{\kappa}{(\kappa + \sqrt{s})(1 + \sqrt{s})},$$

which for $\kappa \neq 1$ can take the form

$$\widetilde{F}_h^e(s) = \left(\frac{\kappa}{1-\kappa}\right)\left(\frac{1}{\kappa + \sqrt{s}} - \frac{1}{1 + \sqrt{s}}\right).$$

For this combination of claim size distributions, the ruin probability can be found explicitly:

Theorem 3.17. *Assume that claims arrive according to a Poisson process with rate λ, the premium rate is 1, and the Laplace transform of the claim size distribution is*

$$\widetilde{G}_\epsilon(s) = (1-\epsilon)\frac{\nu}{s+\nu} + \epsilon\left(1 - \frac{s}{(\kappa + \sqrt{s})(1 + \sqrt{s})}\right),$$

with $\rho_\epsilon = \dfrac{\lambda}{\kappa\nu}\big(\kappa + \epsilon(\nu - \kappa)\big) < 1$. For this mixture model, the ruin probability is

$$\psi_\epsilon(u) = \frac{\lambda}{\kappa\nu}\Big(\kappa\nu - \lambda\big(\kappa + \epsilon(\nu-\kappa)\big)\Big)\sum_{i=1}^{4}\frac{a_i}{v_i(\epsilon)}\zeta\left(v_i^2(\epsilon)u\right),$$

where

$$\zeta(u) := e^u\frac{2}{\sqrt{\pi}}\int_{\sqrt{u}}^{\infty} e^{-x^2}dx,$$

and $-v_i(\epsilon)$, $i = 1, \ldots, 4$, are the roots of the polynomial

$$d(x) = x^4 + (\kappa+1)x^3 + (\kappa+\nu-\lambda)x^2 + (\kappa+1)(\nu-\lambda+\lambda\epsilon)x + \big(\kappa(\nu-\lambda) + \lambda\epsilon(\kappa-\nu)\big).$$

Finally, the coefficients a_i satisfy $a_i = \lim\limits_{x\to -v_i(\epsilon)}\dfrac{n(x)}{d(x)}\big(x + v_i(\epsilon)\big)$, $i = 1, \ldots, 4$, where

$$n(x) = (1-\epsilon)(\kappa + x)(1 + x) + \epsilon(x^2 + \nu).$$

Proof. The Laplace transform of the ruin probability $\mathcal{L}\{\psi_\epsilon(u)\}$ satisfies the equation

$$\mathcal{L}\{\psi_\epsilon(u)\} = \frac{\rho_\epsilon}{s}\left(1 - \frac{(1-\rho_\epsilon)\widetilde{G}_\epsilon^e(s)}{1 - \rho_\epsilon\widetilde{G}_\epsilon^e(s)}\right), \tag{3.10}$$

where $\rho_\epsilon = \frac{\lambda}{\kappa\nu}\big(\kappa + \epsilon(\nu-\kappa)\big)$ and

$$\begin{aligned}\widetilde{G}^e_\epsilon(s) &= \frac{1}{\mathbb{E}U_\epsilon}\big((1-\epsilon)\mathbb{E}B\widetilde{F}^e_p(s) + \epsilon\mathbb{E}C\widetilde{F}^e_h(s)\big)\\ &= \frac{\kappa\nu}{\kappa+\epsilon(\nu-\kappa)}\left((1-\epsilon)\frac{1}{\nu}\frac{\nu}{s+\nu} + \epsilon\frac{1}{\kappa}\frac{\kappa}{(\kappa+\sqrt{s})(1+\sqrt{s})}\right)\\ &= \frac{\kappa\nu}{\kappa+\epsilon(\nu-\kappa)}\cdot\frac{(1-\epsilon)(\kappa+\sqrt{s})(1+\sqrt{s})+\epsilon(s+\nu)}{(s+\nu)(\kappa+\sqrt{s})(1+\sqrt{s})}.\end{aligned}$$

If we set $w(s) = (1-\rho_\epsilon)\widetilde{G}^e_\epsilon(s)/\big(1-\rho_\epsilon\widetilde{G}^e_\epsilon(s)\big)$, then with simple calculations we find that

$$\begin{aligned}w(s) =& \frac{\kappa\nu-\lambda\big(\kappa+\epsilon(\nu-\kappa)\big)}{\kappa+\epsilon(\nu-\kappa)}\\ &\times\frac{(1-\epsilon)(\kappa+\sqrt{s})(1+\sqrt{s})+\epsilon(s+\nu)}{(s+\nu)(\kappa+\sqrt{s})(1+\sqrt{s})-\lambda(1-\epsilon)(\kappa+\sqrt{s})(1+\sqrt{s})-\lambda\epsilon(s+\nu)}.\end{aligned}$$

The denominator of $w(s)$,

$$d(\sqrt{s}) = s^2+(\kappa+1)s\sqrt{s}+(\kappa+\nu-\lambda)s+(\kappa+1)(\nu-\lambda+\lambda\epsilon)\sqrt{s}+\big(\kappa(\nu-\lambda)+\lambda\epsilon(\kappa-\nu)\big),$$

is a fourth degree polynomial with respect to $\sqrt{s}$. Let its roots be given by $-v_i(\epsilon)$, $i=1,\dots,4$, and let $n(s)$ denote the numerator of $w(s)$. Then,

$$\frac{n(s)}{d(s)} = \sum_{i=1}^{4}\frac{a_i}{\sqrt{s}+v_i(\epsilon)}. \tag{3.11}$$

Finally, the coefficients a_i are determined by the following equations

$$a_i = \lim_{\sqrt{s}\to -v_i(\epsilon)}\frac{n(s)}{d(s)}\big(\sqrt{s}+v_i(\epsilon)\big), \qquad i=1,\dots,4.$$

For $s=0$, we get from Eq. (3.11)

$$0 = n(0)-d(0)\sum_{i=1}^{4}\frac{a_i}{v_i(\epsilon)} = \kappa+\epsilon(\nu-\kappa)-\Big(\kappa\nu-\lambda\big(\kappa+\epsilon(\nu-\kappa)\big)\Big)\sum_{i=1}^{4}\frac{a_i}{v_i(\epsilon)}.$$

Substituting everything in Eq. (3.10), we find

$$\begin{aligned}\mathcal{L}\{\psi_\epsilon(u)\} =& \frac{1}{s}\frac{\lambda}{\kappa\nu}\big(\kappa+\epsilon(\nu-\kappa)\big)\left(1-\frac{\kappa\nu-\lambda\big(\kappa+\epsilon(\nu-\kappa)\big)}{\kappa+\epsilon(\nu-\kappa)}\sum_{i=1}^{4}\frac{a_i}{\sqrt{s}+v_i(\epsilon)}\right)\\ =& \frac{\lambda}{\kappa\nu}\left(\frac{\kappa+\epsilon(\nu-\kappa)}{s}-\Big(\kappa\nu-\lambda\big(\kappa+\epsilon(\nu-\kappa)\big)\Big)\sum_{i=1}^{4}\frac{a_i}{s\big(\sqrt{s}+v_i(\epsilon)\big)}\right)\\ =& \frac{\lambda}{\kappa\nu}\left(\frac{\kappa+\epsilon(\nu-\kappa)}{s}-\Big(\kappa\nu-\lambda\big(\kappa+\epsilon(\nu-\kappa)\big)\Big)\sum_{i=1}^{4}\frac{a_i}{v_i(\epsilon)s}\right)\end{aligned}$$

$$+\frac{\lambda}{\kappa\nu}\Big(\kappa\nu-\lambda\big(\kappa+\epsilon(\nu-\kappa)\big)\Big)\sum_{i=1}^{4}\frac{a_i}{v_i(\epsilon)}\frac{1}{\big(\sqrt{s}+v_i(\epsilon)\big)\sqrt{s}}$$

$$=\frac{\lambda}{\kappa\nu}\Big(\kappa\nu-\lambda\big(\kappa+\epsilon(\nu-\kappa)\big)\Big)\sum_{i=1}^{4}\frac{a_i}{v_i(\epsilon)}\frac{1}{\big(\sqrt{s}+v_i(\epsilon)\big)\sqrt{s}}.$$

Laplace inversion to $\mathcal{L}\{\psi_\epsilon(u)\}$ gives,

$$\psi_\epsilon(u)=\frac{\lambda}{\kappa\nu}\Big(\kappa\nu-\lambda\big(\kappa+\epsilon(\nu-\kappa)\big)\Big)\sum_{i=1}^{4}\frac{a_i}{v_i(\epsilon)}\zeta\left(v_i^2(\epsilon)u\right).$$

□

3.4.2 Numerical results

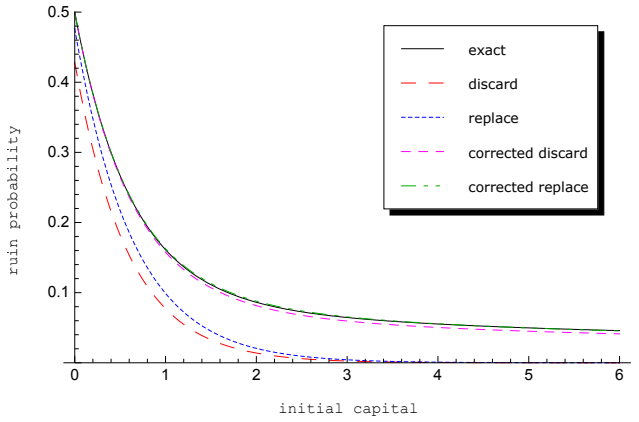
In this section, we fix values for the parameters of the mixture model described in the previous section and we perform our numerical experiments. Although we do not have any restrictions for the parameters of the involved claim size distributions, from a modelling point of view, it is counter-intuitive to fit a heavy-tailed claim size distribution with a mean smaller than the mean of the phase-type claim size distribution. For this reason, we select $\kappa = 2$ and $\nu = 3$.

For the perturbation parameter ϵ, the only restriction arises from the condition for the convergence of the replace series expansion (see Theorem 3.2). If we assume that $\rho_1 > \rho_0$, then the convergence condition for the replace expansion simplifies to $\epsilon < (1-\rho_0)/\rho_1$. A closer look at the formula reveals that, in the case of unequal means, for every value of ϵ there exists a value for the arrival rate λ such that the condition is satisfied. However, a logical constraint for the perturbation parameter is $\epsilon \leq 0.1$. The reason for this constraint is that in the case of phase-type approximations it is not natural to remove more than 10% of the data.

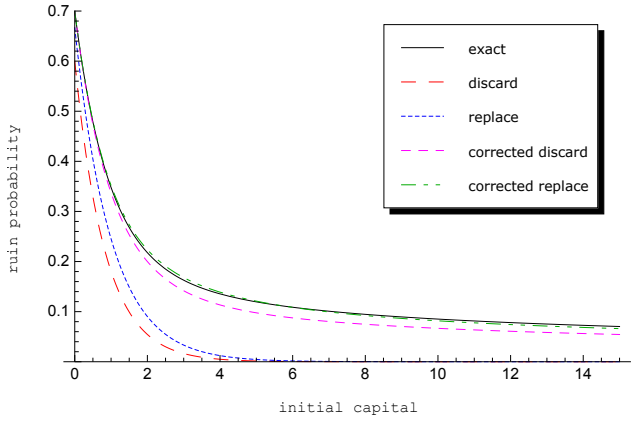
To start our experiments, we first choose the "worst case scenario" for the perturbation parameter, which is $\epsilon = 0.1$. It seems that this "worst case scenario" for the perturbation parameter is the "best case scenario" for the improvement we can achieve with the corrected phase-type approximations. When the perturbation parameter is big enough, a lot of information with respect to the tail behaviour of the ruin probability is missing from its phase-type approximations. So, it is quite natural to expect a great improvement when we add the second term of the respective series expansion, which contains a big part of this missing information. In this scenario, we compare the corrected phase-type approximations with their respective phase-type approximations when $\rho_{0.1}$ takes the values 0.5, 0.7, and 0.9.

From Figure 3.1, we conclude that the corrected discard and the corrected replace approximations manage to reduce the "gap" between their respective phase-type approximations and the exact ruin probability. Although the scale of the graphs is different, it is evident that the gap closes more efficiently for small values of ρ_ϵ, a conclusion that can be also supported theoretically by Section 3.3.2. Furthermore, the corrected replace approximation overestimates the ruin probability for small values of u and, as expected, it is better at the tail than the corrected discard approximation.

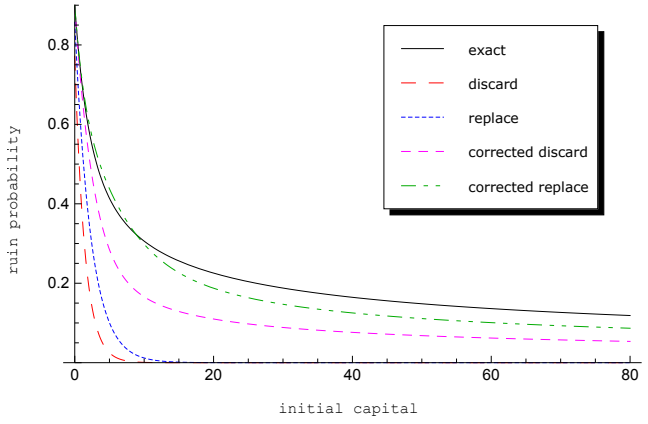
For small values of ρ_ϵ and small values of ϵ, one could argue that the gap between the exact ruin probability and its phase-type approximations is so small that the



(a)



(b)



(c)

FIGURE 3.1: Exact ruin probability with phase-type and corrected phase-type approximations for perturbation parameter 0.1 and average claim rate: (a) 0.5, (b) 0.7, and (c) 0.9.

corrected phase-type approximations do not improve on the accuracy of their phase-type counterparts. For this reason, we choose $\epsilon = 0.001$ and $\rho_{0.001} = 0.5$, and we compare all approximations with the exact ruin probability. We show that the improvement we achieve with the corrected phase-type approximations is still significant, even for this seemingly "bad scenario".

u	exact ruin pr.	discard	replace	cor. discard	cor. replace
0	0.50000000	0.49925037	0.49975012	0.50000000	0.50000000
1	0.11211000	0.11114757	0.11142576	0.11210955	0.11211017
2	0.02557910	0.02474466	0.02484381	0.02557847	0.02557930
3	0.00621454	0.00550887	0.00553925	0.00621386	0.00621466
4	0.00184042	0.00122643	0.00123504	0.00183975	0.00184047
5	0.00082276	0.00027304	0.00027536	0.00082212	0.00082275
6	0.00056334	0.00006078	0.00006139	0.00056273	0.00056329
7	0.00047969	0.00001353	0.00001368	0.00047910	0.00047962
8	0.00043993	3.01×10^{-6}	3.05×10^{-6}	0.00043937	0.00043985
9	0.00041336	6.70×10^{-7}	6.80×10^{-7}	0.00041284	0.00041329
10	0.00039235	1.49×10^{-7}	1.51×10^{-7}	0.00039183	0.00039225
a.r.e.	-	0.54664293	0.54471025	0.00067719	7.24×10^{-5}

TABLE 3.1: Exact ruin probability with phase-type and corrected phase-type approximations for perturbation parameter 0.001 and average claim rate 0.5. The last line corresponds to the average relative errors of each approximation, calculated from the provided values only.

From Table 3.1, we observe that even for this small value of ϵ the corrected discard and the corrected replace approximations yield significant improvements for their respective phase-type approximations. The difference between the exact ruin probability and the corrected phase-type approximations is $O(10^{-6})$, while for the phase-type approximations it is $O(10^{-3})$. In order to understand the magnitude of the improvement we achieve with the corrected phase-type approximations we need to look also at the relative errors of all the approximations involved. It is evident that the relative error of the phase-type approximations easily reaches values close to 1 (approximately after value 5 of the initial capital in this example), while the corrected phase-type approximations give a relative error $O(\epsilon)$.

An interesting observation is that the corrected replace approximation gives better numerical estimations than the corrected discard approximation, both in absolute and relative errors. This observation is also supported by the last line of Table 3.1, which provides the *average relative errors* (a.r.e.) of each approximation for the displayed values of the initial capital u. However, due to the sign changes in the formula of the replace expansion (see Theorem 3.2) it is difficult to find tight bounds for this approximation.

Finally, note that we performed extensive numerical experiments for various values of the perturbation parameter ϵ in the interval $[0.001, 0.1]$. We chose to present only the extreme cases, since the qualitative conclusions for the intermediate values of ϵ are similar to those of the extreme cases.

3.5 Total loss and Value at Risk

In this section, we give a brief overview of how our technique works when we calculate quantities in finite time horizon. As test example, we use the aggregate loss in a fixed period, and we provide the corrected phase-type approximations when the aggregate loss is a compound Poisson sum. Moreover, we extend our technique in case the aggregate loss is a compound mixed Poisson sum. Finally, we perform a small numerical experiment to compare the Value at Risk (VaR) for a given level α that we obtain from the original distribution, the corrected phase-type approximation and its corresponding phase-type approximation.

Suppose that we are interested in evaluating the aggregate loss in a fixed period $[0, t]$. The number $N_\epsilon(t)$ of claims U_ϵ over this fixed period follows a Poisson distribution with rate λt. Observe that $N_\epsilon(t)$ can be seen as a superposition of two independent Poisson processes $N_\epsilon^P(t)$ and $N_\epsilon^H(t)$, with rates $\lambda(1-\epsilon)t$ and $\lambda\epsilon t$ for the phase-type and the heavy-tailed claims sizes, respectively. Thus, we write

$$Loss_\epsilon(x,t) := \mathbb{P}\left(\sum_{k=1}^{N_\epsilon(t)} U_{\epsilon,k} > x\right) = \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + \sum_{k=1}^{N_\epsilon^H(t)} C_k > x\right).$$

To find $Loss_\epsilon(x,t)$, we condition on the number $N_\epsilon^H(t)$ of the heavy-tailed claim sizes and we get

$$\begin{aligned}
Loss_\epsilon(x,t) =& \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + \sum_{k=1}^{N_\epsilon^H(t)} C_k > x\right) \\
=& \sum_{n=0}^{\infty} \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + \sum_{k=1}^{n} C_k > x\right) \mathbb{P}\big(N_\epsilon^H(t) = n\big) \\
=& \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k > x\right) \mathbb{P}\big(N_\epsilon^H(t) = 0\big) \\
&+ \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + C > x\right) \mathbb{P}\big(N_\epsilon^H(t) = 1\big) \\
&+ \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + \sum_{k=1}^{N_\epsilon^H(t)} C_k > x \,\middle|\, N_\epsilon^H(t) \geq 2\right) \mathbb{P}\big(N_\epsilon^H(t) \geq 2\big) \\
=& \underbrace{\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k > x\right)}_{\text{PH-approximation}} e^{-\lambda\epsilon t} + \mathbb{P}\big(N_\epsilon^H(t) \geq 1\big) \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + C > x\right) \\
&+ \mathbb{P}\big(N_\epsilon^H(t) \geq 2\big) \left[\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + \sum_{k=1}^{N_\epsilon^H(t)} C_k > x \,\middle|\, N_\epsilon^H(t) \geq 2\right)\right. \\
&\left. - \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + C > x\right)\right].
\end{aligned}$$

As in the case of ruin probabilities, we define the corrected phase-type approximation by keeping only the terms that contain at most one appearance of the heavy-tailed claim sizes. Thus, we have the following definition.

Definition 3.18. The corrected discard approximation of the tail of the aggregated claim sizes in a fixed time interval $[0,t]$ is defined as

$$Loss_{d,\epsilon}(x,t) := e^{-\lambda\epsilon t}\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k > x\right) + (1-e^{-\lambda\epsilon t})\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + C > x\right).$$

Observe that the coefficient of the correction term in Definition 3.18 is equal to $\mathbb{P}\big(N_\epsilon^H(t) \geq 1\big)$ and not $\mathbb{P}\big(N_\epsilon^H(t) = 1\big)$ as one would expect. We used this modification in order to achieve more accurate estimates of the aggregate loss, without losing the main characteristic of the corrected discard approximation, which is the fact that it underestimates the exact distribution. According to the next theorem, the approximation error of $Loss_{d,\epsilon}(x,t)$ is of order $O(\epsilon^2)$. As it was the case for Theorem 3.7, there are many ways to find a lower bound for the error. In the next theorem, we present a bound yielding a simple expression.

Theorem 3.19. *The error of the corrected discard approximation $Loss_{d,\epsilon}(x,t)$ is bounded as follows:*

$$\begin{aligned}\mathbb{P}\big(N_\epsilon^H(t)\geq 2\big)&\left[\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + C_1 + C_2 > x\right) - \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + C_1 > x\right)\right]\\ &\leq Loss_\epsilon(x,t) - Loss_{d,\epsilon}(x,t) \leq \epsilon^2(\lambda t)^2.\end{aligned}$$

Proof. By using that conditional probabilities are less than or equal to 1, an upper bound for the error of the approximation $Loss_{d,\epsilon}(x,t)$ is found as

$$\begin{aligned}\mathbb{P}\big(N_\epsilon^H(t)\geq 2\big)&\left[\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + \sum_{k=1}^{N_\epsilon^H(t)} C_k > x \,\middle|\, N_\epsilon^H(t)\geq 2\right) - \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + C > x\right)\right]\\ &\leq \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + \sum_{k=1}^{N_\epsilon^H(t)} C_k > x \,\middle|\, N_\epsilon^H(t)\geq 2\right)\mathbb{P}\big(N_\epsilon^H(t)\geq 2\big)\\ &\leq \mathbb{P}\big(N_\epsilon^H(t)\geq 2\big) = \sum_{k=2}^{\infty}\frac{(\lambda t)^k}{k!}\epsilon^k e^{-\lambda\epsilon t} = \epsilon^2(\lambda t)^2\sum_{k=2}^{\infty}\frac{(\lambda t)^{k-2}}{k!}\epsilon^{k-2}e^{-\lambda\epsilon t}\\ &\leq \epsilon^2(\lambda t)^2\sum_{k=2}^{\infty}\frac{(\lambda t)^{k-2}}{(k-2)!}\epsilon^{k-2}e^{-\lambda\epsilon t} = \epsilon^2(\lambda t)^2.\end{aligned}$$

By using the obvious relation

$$\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + \sum_{k=1}^{N_\epsilon^H(t)} C_k > x \,\middle|\, N_\epsilon^H(t)\geq 2\right) \geq \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + C_1 + C_2 > x\right)$$

$$\geq \mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + C_1 > x\right),$$

it is easy to verify that the error is non-negative, which completes the proof. □

Remark 3.20. The corrected replace approximation can be constructed in a similar manner. However, special attention should be paid to the fact that we need to condition not only on the number $N_\epsilon^H(t)$ of heavy-tailed claims but also on the total number of claims, namely $N_\epsilon(t)$. This of course will lead to expressions with the same order of complexity with that of the approximation in Definition 3.4.

If the time t we are interested in is not fixed but a random variable, e.g. T, the total aggregate loss is a compound mixed Poisson r.v. The corrected discard approximation takes the form

$$\begin{aligned} Loss_{d,\epsilon}(x,T) = &\int_0^\infty e^{-\lambda\epsilon t}\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k > x\right) d\mathbb{P}(T \leq t) \\ &+ \int_0^\infty (1-e^{-\lambda\epsilon t})\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + C > x\right) d\mathbb{P}(T \leq t), \end{aligned}$$

and an upper bound for its error is $\epsilon^2\lambda^2\mathbb{E}T^2$. As a last result, we find a compact formula for the LST of the $Loss_{d,\epsilon}(x,T)$. We use the notation $\widetilde{F}_p(s)$, $\widetilde{F}_h(s)$, and $\tau(s)$ for the Laplace transforms of the phase-type claim sizes, the heavy-tailed claim sizes, and the r.v. T, respectively.

Theorem 3.21. *The LST of $Loss_{d,\epsilon}(x,T)$ is given by the formula*

$$\begin{aligned} \mathcal{L}\{Loss_{d,\epsilon}(x,T)\} = &\frac{1}{s} - \frac{1-\widetilde{F}_h(s)}{s}\tau\left(\lambda\big(1-(1-\epsilon)\widetilde{F}_p(s)\big)\right) \\ &- \frac{\widetilde{F}_h(s)}{s}\tau\left(\lambda(1-\epsilon)\big(1-\widetilde{F}_p(s)\big)\right). \end{aligned}$$

Proof. First, we define the LST of $\mathrm{S}_{N_\epsilon^P(t)} = \sum_{k=1}^{N_\epsilon^P(t)} B_k$ as

$$\begin{aligned} \widetilde{\mathrm{F}}_{\mathrm{S}_{N_\epsilon^P(t)}}(s) &= \int_0^\infty e^{-sx} d\mathbb{P}\big(\mathrm{S}_{N_\epsilon^P(t)} \leq x\big) = \sum_{k=0}^\infty \mathbb{P}\big(N_\epsilon^P(t)=k\big)\big(\widetilde{F}_p(s)\big)^k \\ &= \exp\{-\lambda(1-\epsilon)t\big(1-\widetilde{F}_p(s)\big)\}. \end{aligned}$$

Consequently, the LST of $Loss_{d,\epsilon}(x,T)$ satisfies

$$\begin{aligned} \mathcal{L}\{Loss_{d,\epsilon}(x,T)\} = &\int_{x=0}^\infty e^{-sx}\int_{t=0}^\infty e^{-\lambda\epsilon t} d\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k > x\right) d\mathbb{P}(T \leq t) \\ &+ \int_{x=0}^\infty e^{-sx}\int_{t=0}^\infty \big(1-e^{-\lambda\epsilon t}\big) d\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + C > x\right) d\mathbb{P}(T \leq t) \end{aligned}$$

t	simulation	discard	corrected discard
1	4.16	4.09	4.14
5	9.77	9.54	9.74
10	15.24	14.89	15.22
15	20.27	19.73	20.17
20	24.99	23.76	24.30

TABLE 3.2: Comparison of VaR values we obtain from simulation results and the phase-type and corrected phase-type approximations. The VaR level is $\alpha = 0.99$, the perturbation parameter 0.01, and the average claim rate 0.67.

$$
\begin{aligned}
&= \int_{t=0}^{\infty} e^{-\lambda\epsilon t} d\mathbb{P}(T \le t) \int_{x=0}^{\infty} e^{-sx} d\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k > x \right) \\
&\quad + \int_{t=0}^{\infty} \left(1 - e^{-\lambda\epsilon t}\right) d\mathbb{P}(T \le t) \int_{x=0}^{\infty} e^{-sx} d\mathbb{P}\left(\sum_{k=1}^{N_\epsilon^P(t)} B_k + C > x \right) \\
&= \int_{t=0}^{\infty} e^{-\lambda\epsilon t} \frac{1 - e^{-\lambda(1-\epsilon)t\left(1-\widetilde{F}_p(s)\right)}}{s} d\mathbb{P}(T \le t) \\
&\quad + \int_{t=0}^{\infty} \left(1 - e^{-\lambda\epsilon t}\right) \frac{1 - e^{-\lambda(1-\epsilon)t\left(1-\widetilde{F}_p(s)\right)}}{s} \widetilde{F}_h(s) d\mathbb{P}(T \le t) \\
&= \frac{1}{s} - \frac{1 - \widetilde{F}_h(s)}{s} \int_{t=0}^{\infty} e^{-\lambda\left(1-(1-\epsilon)\widetilde{F}_p(s)\right)t} d\mathbb{P}(T \le t) \\
&\quad - \frac{\widetilde{F}_h(s)}{s} \int_{t=0}^{\infty} e^{-\lambda(1-\epsilon)\left(1-\widetilde{F}_p(s)\right)t} d\mathbb{P}(T \le t) \\
&= \frac{1}{s} - \frac{1 - \widetilde{F}_h(s)}{s} \tau\left(\lambda\left(1 - (1-\epsilon)\widetilde{F}_p(s)\right)\right) \\
&\quad - \frac{\widetilde{F}_h(s)}{s} \tau\left(\lambda(1-\epsilon)\left(1 - \widetilde{F}_p(s)\right)\right),
\end{aligned}
$$

which completes the proof. □

One can find the corrected discard approximation analytically (or numerically) by applying Laplace inversion to $\mathcal{L}\{Loss_{d,\epsilon}(x,T)\}$.

A widely used risk measure that connects to the aggregate loss, is the Value at Risk (VaR), which is defined as the threshold value such that the probability of the aggregate loss to exceed this value is less than a given level α. In other words, the VaR is equal to the $(1-\alpha)$-quantile of $Loss_\epsilon(x,t)$. We show through a small numerical experiment that the VaR that is estimated with the corrected discard approximation is closer to the original VaR, than the one we obtain with the discard phase-type approximation. For our example, we choose the arrival rate $\lambda = 1$, the claim size distribution a mixture of an exponential distribution with rate 3/2 and a Pareto distribution with scale and shape parameters 1 and 2 respectively, and $\epsilon = 0.01$. We estimate the VaR values at level 0.99 for the interval $[0,t]$, for the values of $t = 1, 5, 10, 15, 20$. Note that we

simulated the system in order to estimate the exact VaR values. We summarise our results in Table 3.2.

We want to point out here that this numerical study differs from our previous examples. Although in all other examples we were comparing tail probabilities at given values, here we compare the values at which the original distribution and its approximations give us the same tail probability. This observation explains why the difference between the values in Table 3.2 are not of order $O(\epsilon^2)$.

CHAPTER 4

Corrected phase-type approximations in a Markovian environment

4.1 Introduction

In the previous chapter, we constructed the corrected phase-type approximations for heavy-tailed risk models. We showed that our approximations maintain the computational tractability of phase-type approximations, capture the correct tail behaviour, and provide provably small absolute and relative errors. Since our approximations combine such desirable characteristics, it is interesting to explore their applicability in more involved models. In many applications, significant correlations between arrivals of load-generating events make the numerical evaluation of performance measures a challenging problem. Therefore, in this chapter, we consider the MArP/G/1 queue with heavy-tailed service times and FIFO discipline and we develop the corrected phase-type approximations for the delay (waiting time) distribution. We find that our approximations capture the exact tail behaviour and provide bounded relative errors.

Recall from Section 1.3.3, that the Laplace transform of the delay of a MArP/G/1 queue has a matrix expression analogous to the Pollazceck-Khinchine equation of an M/G/1 queue (Neuts, 1989; Ramaswami, 1980). However, these closed-form expressions are only practical in case of phase-type service times (Asmussen, 2000, 2003). In particular, under phase-type service times it has been found that the delay distribution has a phase-type representation (Ramaswami, 1990; Sengupta, 1990) in a form which is explicit up to the solution of a matrix fixed point problem. Therefore, to construct our approximations, we exploit the latter property and we connect our model with a phase-type one.

Motivated by statistical analysis (see Section 3.1 for an explanation), we consider the service times as a mixture of a phase-type and a heavy-tailed distribution. As "base" model we use the model appearing when all heavy-tailed customers are removed, and we interpret the heavy-tailed term of the mixture model as perturbation of the

phase-type one. Using perturbation analysis, we find our approximations for the queueing delay in the mixture model as a sum of the delay of the base model, which itself is a phase-type approximation of the delay, and a heavy-tailed component that depends on the perturbation parameter. Large deviations theory suggests that a single catastrophic event, i.e. a stationary heavy-tailed service time, is sufficient to give a non-zero tail probability for the queueing delay (Embrechts et al., 1997). Since the heavy-tailed component contains such a catastrophic event, the second term of our approximation makes the phase-type approximation more robust so that the relative error at the tail does not explode. More details on the construction of our approximations are found in Section 4.2.2.

In Section 1.6, we noted that there exists duality between the stationary waiting probability $\mathbb{P}(W > u)$ of a G/G/1 queue and the probability of eventual ruin for an insurance company with an initial cash reserve u, where the claims in the risk model correspond to the service times of the queueing model (Asmussen, 2003; Asmussen and Albrecher, 2010). Thus, the corrected phase-type approximations can also be used to estimate the ruin probabilities in the risk model with arrival process of claims the time-reversed MArP of the MArP/G/1 queue. Finally, our technique can be applied to more general queueing models, i.e. queueing models with dependencies between inter-arrival and service times (Boxma and Perry, 2001; Smits et al., 2004), and also to models that allow for customers to arrive in batches (Lucantoni, 1991, 1993; Lucantoni et al., 1994).

A closely related work is Adan and Kulkarni (2003). They consider a single server queue, where the inter-arrival and service times depend on a common discrete Markov chain. In addition, they assume that a customer arrives in each phase transition and they find a closed form expression for the delay distribution under general service time distributions. However, by using the standard techniques of including *dummy* customers in a model, i.e. customers with zero service times, the arrivals of dummy customers in their model correspond to phase transitions not related to arrivals of customers in the typical MArP/G/1 queue (see Section 1.3.3). Thus, their results remain valid for the evaluation of the workload. In this chapter, we exploit this connection and, based on their results, we derive the corrected phase-type approximations for the delay of the MArP/G/1 queue.

Outline

The rest of the chapter is organised as follows. In Section 4.2, we introduce the model under consideration without assuming any special form for the service time distribution, and in Section 4.2.1, we find the general expressions for the Laplace transforms of the queueing delay a customer experiences upon arrival in each state.

In Section 4.2.2, we consider service time distributions that are a mixture of a phase-type distribution and a heavy-tailed one, and we explain the idea to construct our approximations. Later in Section 4.3.1, we specialise the results of Section 4.2.1 for phase-type service times. We use as a base model the phase-type model of Section 4.3.1, and we apply perturbation analysis to find in Section 4.3.2 the perturbed parameters and in Section 4.3.3 the desired Laplace transforms of the queueing delay in the mixture model. By using the latter results, we construct in Section 4.3.4 the approximations and we discuss their properties. In Section 4.4, we discuss an alternative way to

construct approximations for the queueing delay.

Furthermore, in Section 4.5, we use a specific mixture service time distribution for which the exact delay distribution can be calculated and we exhibit the accuracy of our approximations through numerical experiments. Due to the complexity of the formulas, we use a simple running example in order to explain the idea behind the calculations. Finally, the necessary theory on perturbation analysis and other related results can be found in Appendix A.2.

4.2 Presentation of the model

We consider a single server queue with FIFO discipline, where customers arrive according to a MArP. The arrivals are regulated by a Markov process $\{J_t\}_{t\geq 0}$ with a finite state space $\mathcal{N}$, say with N states. We assume that the service time distribution of a customer is independent of the state of $\{J_t\}$ upon his arrival. For this model, we are interested in finding accurate approximations for the delay distribution.

The intensity matrix $\mathbf{D}$ governing $\{J_t\}$ is denoted by the decomposition $\mathbf{D} = \mathbf{D}^{(1)} + \mathbf{D}^{(2)}$, where the matrix $\mathbf{D}^{(1)}$ is related to arrivals of *dummy* customers, while transitions in $\mathbf{D}^{(2)}$ are related to arrivals of *real* customers. Note that the diagonal elements of the matrix $\mathbf{D}^{(2)}$ may not be identically equal to zero. This means that if $d_{ii}^{(2)} > 0$, then a real customer arrives with rate $d_{ii}^{(2)}$ and we have a transition from state i to itself. However, phase transitions not associated with arrivals (dummy customers) from any state to itself are not allowed. Since the matrix $\mathbf{D}$ is an intensity matrix, its rows sum up to zero. Therefore, the diagonal elements of the matrix $\mathbf{D}^{(1)}$ are negative and they are defined as $d_{ii}^{(1)} = -\sum_{k\neq i} d_{ik}^{(1)} - \sum_{k=1}^{N} d_{ik}^{(2)}$.

For this model, we are interested in modelling heavy-tailed service times. As stated earlier, motivated by statistical analysis, we assume that the service time distribution of a real customer is a mixture of a phase-type distribution, $F_p(t)$, and a heavy-tailed one, $F_h(t)$. Namely, the service time distribution of a real customer has the form

$$G_\epsilon(t) = (1-\epsilon)F_p(t) + \epsilon F_h(t), \tag{4.1}$$

where ϵ is typically small.

Our goal is to find the delay distribution for this mixture model. Towards this direction, we present in the next section existing results (Adan and Kulkarni, 2003) for the evaluation of the delay distribution under the assumption of generally distributed service times. Ultimately, we wish to specialise these results to service times of the aforementioned form (4.1).

4.2.1 Preliminaries

Since the results of this section are valid for any service time distribution, we suppress the index ϵ and we use the notation $G(t)$ for the service time distribution of a real customer. We consider now the embedded Markov chain $\{Z_n\}_{n\geq 0}$ on the arrival epochs of customers (real and dummy) and we denote by $\mathbf{P}$ the transition probability matrix of the regulating Markov chain $\{Z_n\}$, which we assume to be irreducible. If λ_i

is the exponential exit rate from state i, i.e.

$$\lambda_i = \sum_{k \neq i} d_{ik}^{(1)} + \sum_{k=1}^{N} d_{ik}^{(2)}, \tag{4.2}$$

the transition probabilities can be calculated by

$$p_{ij} = \frac{d_{ij}^{(1)}(1-\delta_{ij}) + d_{ij}^{(2)}}{\lambda_i}.$$

In addition, an arriving customer at a transition from state i to state j is tagged i. If $p_{ij} > 0$, then we define the probability

$$q_{ij}^{(1)} = \frac{d_{ij}^{(1)}(1-\delta_{ij})}{d_{ij}^{(1)}(1-\delta_{ij}) + d_{ij}^{(2)}},$$

which is the probability of an arriving customer to be dummy conditioned on the event that there is a phase transition from state i to j. Similarly, conditioned on the event that there is a phase transition from i to j, the arriving customer is real with probability

$$q_{ij}^{(2)} = \frac{d_{ij}^{(2)}}{d_{ij}^{(1)}(1-\delta_{ij}) + d_{ij}^{(2)}}. \tag{4.3}$$

If $p_{ij} = 0$, then we define $q_{ij}^{(1)} = q_{ij}^{(2)} = 0$. Consequently, the conditional service time distribution of an arriving customer at a transition from i to j is $G_{ij}(t) = q_{ij}^{(1)} + q_{ij}^{(2)} G(t)$, and its LST is $\widetilde{G}_{ij}(s) = q_{ij}^{(1)} + q_{ij}^{(2)} \widetilde{G}(s)$, $i, j = 1, \ldots, N$, where $\widetilde{G}(s)$ is the LST of the service time distribution $G(t)$ of a real customer. In matrix form, the above quantities can be written as

$$\begin{aligned} \boldsymbol{\Lambda} &= \operatorname{diag}(\lambda_1, \ldots, \lambda_N), \\ \mathbf{Q}^{(1)} &= \big(q_{ij}^{(1)}\big), \\ \mathbf{Q}^{(2)} &= \big(q_{ij}^{(2)}\big), \\ \widetilde{\mathbf{G}}(s) &= \mathbf{Q}^{(1)} + \widetilde{G}(s)\mathbf{Q}^{(2)}. \end{aligned} \tag{4.4}$$

We also define the matrix

$$\mathbf{H}(s) = \widetilde{\mathbf{G}}(s) \circ \mathbf{P}\boldsymbol{\Lambda}, \tag{4.5}$$

which we will need later. Finally, let $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_N]$ be the stationary distribution of $\{Z_n\}_{n \geq 0}$, and μ be the mean of the service time distribution $G(t)$. Then the system is stable if the mean service time of a customer is less than the mean inter-arrival times between two consecutive customers in steady state. Namely,

$$\boldsymbol{\pi}\big(\boldsymbol{\Lambda}^{-\mathbf{1}} - \mathbf{M}\big)\mathbf{e} > 0, \tag{4.6}$$

where $\mathbf{M} = \mu \mathbf{Q}^{(2)} \circ \mathbf{P}$. Note that the (i, j) element of the matrix $\mathbf{Q}^{(2)} \circ \mathbf{P}$ is the unconditional probability that a real customer arrives at a transition from i to j.

From this point on, we use a simple running example so that we display the involved parameters and the derived formulas. The running example evolves progressively, which means that its parameters are introduced only once and the reader should consult a previous block of the example to recall the notation.

Running example. For our running example, we consider a MArP with Erlang-2 distributed inter-arrival times, where the exponential phases have both rate λ ($N = 2$). Therefore, the matrices $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ are given as follows:

$$\mathbf{D}^{(1)} = \begin{pmatrix} -\lambda & \lambda \\ 0 & -\lambda \end{pmatrix} \qquad \text{and} \qquad \mathbf{D}^{(2)} = \begin{pmatrix} 0 & 0 \\ \lambda & 0 \end{pmatrix}.$$

In this case, we have that $\lambda_1 = \lambda_2 = \lambda$, $p_{ij} = 1 - \delta_{ij}$, $q_{12}^{(1)} = q_{21}^{(2)} = 1$, and all other elements of the matrices $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$ are equal to zero. Observe that we only have transitions from state 1 to state 2 and from state 2 to state 1. Therefore, in state 1 we always have arrivals of dummy customers while in state 2 we only have arrivals of real customers. Thus, only the diagonal elements of the matrix $\widetilde{\mathbf{G}}(s)$ are not equal to zero, so that $\widetilde{G}_{11}(s) = 1$ and $\widetilde{G}_{22}(s) = \widetilde{G}(s)$. Finally, the stability condition takes its known form $\lambda\mu/2 < 1$. ■

Let now V denote the steady-state workload of the system just prior to an arrival of a customer. If the arriving customer is real, then the workload just prior to its arrival equals the delay or waiting time of the customer in the queue, which we denote by W. In terms of Laplace transforms, the steady-state workload of the system just prior to an arrival of a customer in state i is found as

$$\widetilde{\phi}_i(s) = \mathbb{E}(e^{-sV}; Z = i), \qquad \Re(s) \geq 0, \quad i = 1, \ldots, N,$$

where Z is the steady-state limit of Z_n. Gathering all the above Laplace transforms $\widetilde{\phi}_i(s)$, $i = 1, \ldots, N$, we construct the transform vector

$$\widetilde{\mathbf{\Phi}}(s) = \big(\widetilde{\phi}_1(s), \ldots, \widetilde{\phi}_N(s)\big).$$

We first provide some general theorems for the transform vector $\widetilde{\mathbf{\Phi}}(s)$ and we give its connection to $\widetilde{w}(s)$, which is defined as the Laplace transform of the queueing delay W of real customers. Later on we refine these results in order to provide more detailed information regarding the form of the elements $\widetilde{\phi}_i(s)$, $i = 1, \ldots, N$.

Theorem 4.1. *Provided that the stability condition* (4.6) *is satisfied, the transform vector* $\widetilde{\mathbf{\Phi}}(s)$ *satisfies*

$$\widetilde{\mathbf{\Phi}}(s)\big(\mathbf{H}(s) + s\mathbf{I} - \mathbf{\Lambda}\big) = s\mathbf{u}, \tag{4.7}$$

$$\widetilde{\mathbf{\Phi}}(0)\mathbf{e} = 1, \tag{4.8}$$

where $\mathbf{u} = [u_1, \ldots, u_N]$ *is a vector with* N *unknown parameters that needs to be determined.*

Note that the above theorem is similar to Adan and Kulkarni (2003, Theorem 3.1) and so is its proof. Therefore, we omit here the proof and we refer the reader to Adan and Kulkarni (2003, Theorem 3.1) for more details. Moreover, according to Takine and Hasegawa (1994), the ith element of the vector $\mathbf{u}$ represents the probability that, in steady state, the system is idle and the underlying Markov chain is in state i.

A real customer arrives in state i with probability $\sum_{j=1}^{N} p_{ij} q_{ij}^{(2)} = \sum_{j=1}^{N} d_{ij}^{(2)}/\lambda_i$, and consequently a real customer arrives in the system with probability $\sum_{i=1}^{N} \pi_i \cdot$

$\sum_{j=1}^{N} d_{ij}^{(2)}/\lambda_i$. Therefore, the following relation holds

$$\sum_{i=1}^{N} \pi_i \frac{\sum_{j=1}^{N} d_{ij}^{(2)}}{\lambda_i} \widetilde{w}(s) = \sum_{i=1}^{N} \frac{\sum_{j=1}^{N} d_{ij}^{(2)}}{\lambda_i} \widetilde{\phi}_i(s).$$

Thus, if $\boldsymbol{\omega}$ is a column vector of dimension N such that

$$\boldsymbol{\omega} = \frac{\boldsymbol{\Lambda}^{-1}\mathbf{D}^{(2)}\mathbf{e}}{\boldsymbol{\pi}\boldsymbol{\Lambda}^{-1}\mathbf{D}^{(2)}\mathbf{e}},$$

the Laplace transform of the queueing delay is found as

$$\widetilde{w}(s) = \widetilde{\boldsymbol{\Phi}}(s)\boldsymbol{\omega}, \qquad \Re(s) \geq 0. \tag{4.9}$$

If $\det\big(\mathbf{H}(s) + s\mathbf{I} - \boldsymbol{\Lambda}\big)$ denotes the determinant of the square matrix $\mathbf{H}(s) + s\mathbf{I} - \boldsymbol{\Lambda}$, then for the determination of the unknown vector $\mathbf{u}$, we have the following theorem.

Theorem 4.2. *The next two statements hold:*

1. *The equation* $\det\big(\mathbf{H}(s) + s\mathbf{I} - \boldsymbol{\Lambda}\big) = 0$ *has exactly* N *solutions* $s_1, \ldots, s_N$*, with* $s_1 = 0$ *and* $\Re(s_i) > 0$ *for* $i = 2, \ldots, N$.

2. *Suppose that the stability condition* (4.6) *is satisfied and that the above mentioned* $N - 1$ *solutions* $s_2, \ldots, s_N$ *are distinct. Let* $\mathbf{a}_i$ *be a non-zero column vector satisfying*

$$\big(\mathbf{H}(s_i) + s_i\mathbf{I} - \boldsymbol{\Lambda}\big)\mathbf{a}_i = 0, \quad i = 2, \ldots, N.$$

Then $\mathbf{u}$ *is given by the unique solution to the following* N *linear equations:*

$$\mathbf{u}\boldsymbol{\Lambda}^{-1}\mathbf{e} = \boldsymbol{\pi}\left(\boldsymbol{\Lambda}^{-1} - \mathbf{M}\right)\mathbf{e}, \tag{4.10}$$

$$\mathbf{u}\mathbf{a}_i = 0, \qquad i = 2, \ldots, N. \tag{4.11}$$

Again, Theorem 4.2 is similar to Adan and Kulkarni (2003, Theorems 3.2 and 3.3), and therefore, its proof is omitted. Alternatively, the vector $\mathbf{u}$ can be estimated by following the iterative approach presented in Takine and Hasegawa (1994).

Theorem 4.2 on the one hand provides us with an algorithm to calculate the vector $\mathbf{u}$ and on the other hand it guarantees that all elements of the transform vector $\widetilde{\boldsymbol{\Phi}}(s)$ are well-defined on the positive half-plane. To understand the latter remark observe the following. For simplicity, we set

$$\mathbf{E}(s) = \mathbf{H}(s) + s\mathbf{I} - \boldsymbol{\Lambda}. \tag{4.12}$$

Let $\mathcal{E}(s)$ be the adjoint matrix of $\mathbf{E}(s)$, so $\mathbf{E}(s)\mathcal{E}(s) = \det\mathbf{E}(s)\mathbf{I}$. Post-multiplying Eq. (4.7) with $\mathcal{E}(s)$, we have that $\widetilde{\boldsymbol{\Phi}}(s)\det\mathbf{E}(s) = s\mathbf{u}\mathcal{E}(s)$, and consequently

$$\widetilde{\boldsymbol{\Phi}}(s) = \frac{1}{\det\mathbf{E}(s)} s\mathbf{u}\mathcal{E}(s). \tag{4.13}$$

The first statement of Theorem 4.2 says that the determinant $\det\mathbf{E}(s)$ has the factors $s - s_i$, $i = 1, \ldots, N$, in its expression. This means that the transform vector $\widetilde{\boldsymbol{\Phi}}(s)$ has

N potential singularities on the positive half plane, since the determinant appears in the denominator. However, the second statement of Theorem 4.2 explains that the vector $\mathbf{u}$ is such that these problematic factors are canceled out.

Observe that Theorem 4.2 does not give us any information about the form of the elements of the transform vector $\widetilde{\mathbf{\Phi}}(s)$, which is the stepping stone for the construction of our approximations. For this reason, we proceed by finding an analytic expression for the aforementioned elements. It is apparent from Eq. (4.13) that for the evaluation of $\widetilde{\mathbf{\Phi}}(s)$, we only need $\det \mathbf{E}(s)$ and the adjoint matrix $\mathcal{E}(s)$. For the determination of these quantities, we introduce the following notation:

- As before, we denote the set of all states of the Markov process $\{J_t\}$ as $\mathcal{N} = \{1, \ldots, N\}$.
- If $S \subset \Omega$, for some set $\Omega \subset \mathcal{N}$, then S^c is the complementary set of S with respect to Ω. The number of elements in a set S is denoted as $|S|$.
- For a subset S of $\mathcal{N}$, we define $\boldsymbol{\lambda}^S = \prod_{i\in S} \lambda_i$ and $\zeta^S(s) = \prod_{i\in S}(s - \lambda_i)$. We also define $\boldsymbol{\lambda}^\emptyset = \zeta^\emptyset(s) = 1$ and $\det \mathbf{A}^\emptyset_\emptyset = 1$, where $\emptyset$ is the empty set.
- Suppose that S is a subset of Ω, for some set $\Omega \subset \mathcal{N}$, and that it follows some properties, i.e. "Property 1", etc. If we want to sum with respect to S, then we write under the symbol of summation first $S \subset \Omega$, followed by the properties. Namely, we write $\sum_{\substack{S\subset\Omega \\ \text{Property 1} \\ \text{etc}}}$. In some cases, to avoid lengthy expressions we will write instead of $\sum_{\substack{S\subset\Omega \\ \text{Properties of } S}} \sum_{\substack{R\subset\Omega_1 \\ \text{Properties of } R}}$ the double sum $\sum_{\substack{S\subset\Omega \\ \text{Properties of } S; \\ R\subset\Omega_1 \\ \text{Properties of } R}}$, where R is a subset of Ω_1, for some set $\Omega_1 \subset \mathcal{N}$. We apply the same rule also for multiple sums.

By using the above notation, we proceed with refining the desired quantities. More precisely, we first find $\det \mathbf{E}(s)$, then the adjoint matrix $\mathcal{E}(s)$, and finally the vector $s\mathbf{u}\mathcal{E}(s)$ that appears in the numerator of the transform vector $\widetilde{\mathbf{\Phi}}(s)$ (see Eq. (4.13)). Combining these results, one can easily derive $\widetilde{\mathbf{\Phi}}(s)$. We start by finding the determinant of the matrix $\mathbf{E}(s)$ (see Eq. (4.12)).

Theorem 4.3. *The determinant of the matrix* $\mathbf{E}(s)$ *can be explicitly calculated as follows:*

$$\det \mathbf{E}(s) = \sum_{S\subset\mathcal{N}} \boldsymbol{\lambda}^S \zeta^{S^c}(s) \det \left(\mathbf{Q}^{(1)} \circ \mathbf{P}\right)^S_S + \sum_{k=1}^{N} \widetilde{G}^k(s) \sum_{\substack{\Gamma\subset\mathcal{N} \\ |\Gamma|=k}} \sum_{\substack{S\subset\mathcal{N} \\ S\supset\Gamma}} \boldsymbol{\lambda}^S \zeta^{S^c}(s) \det \left(\left(\mathbf{Q}^{(1)} \circ \mathbf{P}\right)^{S\setminus\Gamma}_S \ltimes \left(\mathbf{Q}^{(2)} \circ \mathbf{P}\right)^{\Gamma}_S \right).$$

Proof. To prove the theorem, we need formulas that result from the properties of the determinants. We define the sets $F_i = \{1, \ldots, i\}$ and $L_i = \{i, \ldots, N\}$, where $F_0 = L_{N+1} = \emptyset$. Expansion by minors along the first row and the additive property of determinants give for $i \in \mathcal{N}$,

$$\det \mathbf{E}(s)^{L_i}_{L_i} = \widetilde{G}(s)\lambda_i \det \left(\left(\mathbf{Q}^{(2)} \circ \mathbf{P}\right)^{\{i\}}_{L_i}, \mathbf{E}(s)^{L_{i+1}}_{L_i} \right) + \lambda_i \det \left(\left(\mathbf{Q}^{(1)} \circ \mathbf{P}\right)^{\{i\}}_{L_i}, \mathbf{E}(s)^{L_{i+1}}_{L_i} \right)$$

$$+ (s-\lambda_i) \det \mathbf{E}(s)_{L_{i+1}}^{L_{i+1}}.$$

Suppose now that $V = \{i_1, \ldots, i_n\}$ and $W = \{j_1, \ldots, j_k\}$ are two non-overlapping ($V \cap W = \emptyset$) collections of n and k elements from $\mathcal{N}$, respectively, with $1 \leq n+k \leq N-1$. Furthermore, we choose j such that $j > \max\{l : l \in V \cup W\}$. Then, the determinant of the $(N+1-j+n+k)$-dimensional square matrix $\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)_{V\cup W\cup L_j}^{V} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)_{V\cup W\cup L_j}^{W}, \mathbf{E}(s)_{V\cup W\cup L_j}^{L_j} \Big)$ satisfies,

$$\begin{aligned}
&\det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)_{V\cup W\cup L_j}^{V} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)_{V\cup W\cup L_j}^{W}, \mathbf{E}(s)_{V\cup W\cup L_j}^{L_j} \Big) \\
&= \widetilde{G}(s)\lambda_j \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)_{V\cup W\cup L_j}^{V} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)_{V\cup W\cup L_j}^{W\cup\{j\}}, \mathbf{E}(s)_{V\cup W\cup L_j}^{L_{j+1}} \Big) \\
&\quad + \lambda_j \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)_{V\cup W\cup L_j}^{V\cup\{j\}} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)_{V\cup W\cup L_j}^{W}, \mathbf{E}(s)_{V\cup W\cup L_j}^{L_{j+1}} \Big) \\
&\quad + (s-\lambda_j) \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)_{V\cup W\cup L_{j+1}}^{V} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)_{V\cup W\cup L_{j+1}}^{W}, \mathbf{E}(s)_{V\cup W\cup L_{j+1}}^{L_{j+1}} \Big).
\end{aligned}$$

Note that $\det \mathbf{E}(s) = \det \mathbf{E}(s)_{L_1}^{L_1}$. The theorem is proven by applying recursively the above formulas. □

Observe that the determinant $\det \mathbf{E}(s)$ is an at most N degree polynomial with respect to the LST of the service time distribution $\widetilde{G}(s)$ of a real customer. Moreover, the coefficients of this polynomial are all polynomials with respect to s. Therefore, in case $\widetilde{G}(s)$ is a rational function in s, then $\det \mathbf{E}(s)$ is also a rational function in s and its eigenvalues can be easily calculated. Furthermore, all subsets Γ of $\mathcal{N}$ that appear in the second summand have at least one element, thus in the formula of $\det \mathbf{E}(s)$ it always holds that $\Gamma \neq \emptyset$.

Running example (continued). The matrix $\mathbf{E}(s)$ has elements $\mathbf{E}_{ii}(s) = s - \lambda_i$, $i = 1, 2$, $\mathbf{E}_{12}(s) = \lambda$, and $\mathbf{E}_{21}(s) = \lambda\widetilde{G}(s)$. We will calculate its determinant by using Theorem 4.3. It holds that $\det \big(\mathbf{Q}^{(1)} \circ \ \mathbf{P}\big)_{S}^{S} = 0$ for all subsets S of $\mathcal{N}$, except for $S = \emptyset$. Since $\Gamma \neq \emptyset$, it is evident that $\det\Big(\big(\mathbf{Q}^{(1)} \circ \ \mathbf{P}\big)_{S}^{S\setminus\Gamma} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)_{S}^{\Gamma} \Big) \neq 0$ only for $\Gamma = \{1\}$ and $S = \mathcal{N}$, because the 1st column of the matrix $\mathbf{Q}^{(1)}$ and the 2nd column of the matrix $\mathbf{Q}^{(2)}$ are zero. Combining all these we obtain

$$\begin{aligned}
\det \mathbf{E}(s) =& \boldsymbol{\lambda}^{\emptyset} \zeta^{\mathcal{N}}(s) \det \big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)_{\emptyset}^{\emptyset} \\
&+ \widetilde{G}(s) \boldsymbol{\lambda}^{\mathcal{N}} \zeta^{\emptyset}(s) \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)_{\mathcal{N}}^{\{2\}} \bowtie \big(\mathbf{Q}^{(2)} \circ \ \mathbf{P}\big)_{\mathcal{N}}^{\{1\}} \Big) \\
=& (s-\lambda)^2 - \lambda^2 \widetilde{G}(s).
\end{aligned}$$

■

In a similar manner, we find the explicit form of the adjoint matrix $\mathcal{E}(s)$ in the following theorem.

Theorem 4.4. *The adjoint matrix $\mathcal{E}(s)$ has elements*

$$\mathcal{E}_{ij}(s)=\begin{cases}\sum_{k=0}^{N-1}\widetilde{G}^k(s)\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i\}\\|\Gamma|=k;}}\boldsymbol{\lambda}^S\zeta^{S^c}(s)\\ \times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^{S\setminus\Gamma}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_S^{\Gamma}\Big), & i=j,\\ (-1)^{i+j}\sum_{k=1}^{N-1}\widetilde{G}^k(s)\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i,j\}\\|\Gamma|=k-1}}\sum_{\substack{S\subset\mathcal{N}\setminus\{i,j\}\\S\supset\Gamma;\\R\subset S\cap T_{ij}}}(-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{j\}}\zeta^{S^c}(s)\\ \times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{i\}}^{S\setminus\Gamma}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{i\}}^{\Gamma\cup\{j\}}\Big)\\ +(-1)^{i+j}\sum_{k=0}^{N-2}\widetilde{G}^k(s)\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i,j\}\\|\Gamma|=k}}\sum_{\substack{S\subset\mathcal{N}\setminus\{i,j\}\\S\supset\Gamma;\\R\subset S\cap T_{ij}}}(-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{j\}}\zeta^{S^c}(s)\\ \times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{i\}}^{(S\setminus\Gamma)\cup\{j\}}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{i\}}^{\Gamma}\Big), & i\neq j,\end{cases}$$

where $m_{ij}=\min\{i,j\}$, $M_{ij}=\max\{i,j\}$, and $T_{ij}=\{m_{ij}+1,\dots,M_{ij}-1\}$.

Proof. We use the definition $\mathcal{E}_{ij}(s)=(-1)^{i+j}\det\mathbf{E}(s)_{\mathcal{N}\setminus\{j\}}^{\mathcal{N}\setminus\{i\}}$ and we find recursive formulas for $\det\mathbf{E}(s)_{\mathcal{N}\setminus\{j\}}^{\mathcal{N}\setminus\{i\}}$, as in Theorem 4.3, by distinguishing between the cases $i=j$ and $i\neq j$. The case $i=j$ is merely an application of Theorem 4.3, where instead of the state space $\mathcal{N}$ we have $\mathcal{N}\setminus\{i\}$. Therefore,

$$\begin{aligned}\mathcal{E}_{ii}(s)=&\sum_{S\subset\mathcal{N}\setminus\{i\}}\boldsymbol{\lambda}^S\zeta^{S^c}(s)\det\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^S\\ &+\sum_{k=1}^{N-1}\widetilde{G}^k(s)\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i\}\\|\Gamma|=k}}\sum_{\substack{S\subset\mathcal{N}\setminus\{i\}\\S\supset\Gamma}}\boldsymbol{\lambda}^S\zeta^{S^c}(s)\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^{S\setminus\Gamma}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_S^{\Gamma}\Big).\end{aligned}$$

When $i\neq j$, we need to separate the two cases $i<j$ and $i>j$. We first deal with the case $i<j$. We then have,

$$\begin{aligned}\mathcal{E}_{ij}(s)=&(-1)^{i+j}\widetilde{G}(s)\lambda_j\det\Big(\mathbf{E}(s)_{\mathcal{N}\setminus\{j\}}^{F_{j-1}\setminus\{i\}},\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{\mathcal{N}\setminus\{j\}}^{\{j\}},\mathbf{E}(s)_{\mathcal{N}\setminus\{j\}}^{L_{j+1}}\Big)\\ &+(-1)^{i+j}\lambda_j\det\Big(\mathbf{E}(s)_{\mathcal{N}\setminus\{j\}}^{F_{j-1}\setminus\{i\}},\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{\mathcal{N}\setminus\{j\}}^{\{j\}},\mathbf{E}(s)_{\mathcal{N}\setminus\{j\}}^{L_{j+1}}\Big).\end{aligned}$$

We find $\mathcal{E}_{ij}(s)$ by expanding the determinants that appear above by minors along their first row. For this reason, it is important to know what is the position of the elements $\mathbf{E}_{nn}(s)=\widetilde{G}_{nn}(s)p_{nn}\lambda_n+s-\lambda_n$, $n\in\mathcal{N}\setminus\{i,j\}$, in the above reduced matrix. Note that the elements $\mathbf{E}_{nn}(s)$ with $n=i+1,\dots,j-1$, are on the diagonal of matrix $\mathbf{E}(s)$. However, when $j\neq i+1$ these elements drop to the lower-diagonal of the square matrices $\Big(\mathbf{E}(s)_{\mathcal{N}\setminus\{j\}}^{F_{j-1}\setminus\{i\}},\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{\mathcal{N}\setminus\{j\}}^{\{j\}},\mathbf{E}(s)_{\mathcal{N}\setminus\{j\}}^{L_{j+1}}\Big)$ and $\Big(\mathbf{E}(s)_{\mathcal{N}\setminus\{j\}}^{F_{j-1}\setminus\{i\}},\big(\mathbf{Q}^{(1)}\circ\ \mathbf{P}\big)_{\mathcal{N}\setminus\{j\}}^{\{j\}},\mathbf{E}(s)_{\mathcal{N}\setminus\{j\}}^{L_{j+1}}\Big)$.

It is immediately obvious that if this displacement takes place, it will result in a change of sign for the determinants. For this reason, we split the columns of the latter matrices in the subsets F_{i-1}, T, $\{j\}$, and L_{j+1}, where $T=\{i+1,\dots,j-1\}$. We fix some $m\in\mathcal{N}\setminus\{i,j\}$ and we separate the following cases:

1. $m \in F_{i-1}$. For every two non-overlapping collections of n and k elements from F_{m-1}, say $V = \{i_1, \ldots, i_n\}$ and $W = \{j_1, \ldots, j_k\}$, with $1 \leq n + k \leq m - 1$, it holds that

$$\begin{aligned}
&\det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{V}_{\Omega} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{W}_{\Omega}, \mathbf{E}(s)^{(L_m \cap F_{i-1}) \cup T}_{\Omega}, \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{\{j\}}_{\Omega}, \mathbf{E}(s)^{L_{j+1}}_{\Omega}\Big) \\
&= \widetilde{G}(s)\lambda_m \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{V}_{\Omega} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{W \cup \{m\}}_{\Omega}, \mathbf{E}(s)^{(L_{m+1} \cap F_{i-1}) \cup T}_{\Omega}, \\
&\quad \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{\{j\}}_{\Omega}, \mathbf{E}(s)^{L_{j+1}}_{\Omega}\Big) \\
&+ \lambda_m \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{V \cup \{m\}}_{\Omega} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{W}_{\Omega}, \mathbf{E}(s)^{(L_{m+1} \cap F_{i-1}) \cup T}_{\Omega}, \\
&\quad \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{\{j\}}_{\Omega}, \mathbf{E}(s)^{L_{j+1}}_{\Omega}\Big) \\
&+ (s - \lambda_m) \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{V}_{\Omega \setminus \{m\}} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{W}_{\Omega \setminus \{m\}}, \mathbf{E}(s)^{(L_{m+1} \cap F_{i-1}) \cup T}_{\Omega \setminus \{m\}}, \\
&\quad \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{\{j\}}_{\Omega \setminus \{m\}}, \mathbf{E}(s)^{L_{j+1}}_{\Omega \setminus \{m\}}\Big),
\end{aligned}$$

and,

$$\begin{aligned}
&\det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{V}_{\Omega} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{W}_{\Omega}, \mathbf{E}(s)^{(L_m \cap F_{i-1}) \cup T}_{\Omega} \big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{\{j\}}_{\Omega}, \mathbf{E}(s)^{L_{j+1}}_{\Omega}\Big) \\
&= \widetilde{G}(s)\lambda_m \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{V}_{\Omega} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{W \cup \{m\}}_{\Omega}, \mathbf{E}(s)^{(L_{m+1} \cap F_{i-1}) \cup T}_{\Omega}, \\
&\quad \big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{\{j\}}_{\Omega}, \mathbf{E}(s)^{L_{j+1}}_{\Omega}\Big) \\
&+ \lambda_m \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{V \cup \{m\}}_{\Omega} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{W}_{\Omega}, \mathbf{E}(s)^{(L_{m+1} \cap F_{i-1}) \cup T}_{\Omega}, \\
&\quad \big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{\{j\}}_{\Omega}, \mathbf{E}(s)^{L_{j+1}}_{\Omega}\Big) \\
&+ (s - \lambda_m) \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{V}_{\Omega \setminus \{m\}} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{W}_{\Omega \setminus \{m\}}, \mathbf{E}(s)^{(L_{m+1} \cap F_{i-1}) \cup T}_{\Omega \setminus \{m\}}, \\
&\quad \big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{\{j\}}_{\Omega \setminus \{m\}}, \mathbf{E}(s)^{L_{j+1}}_{\Omega \setminus \{m\}}\Big),
\end{aligned}$$

where $\Omega = V \cup W \cup (L_m \cap F_{i-1}) \cup \{i\} \cup T \cup L_{j+1}$.

2. $m \in T$ with $T \neq \emptyset$ (note that $T \neq \emptyset$ when $j \neq i + 1$). For every two non-overlapping collections of n and k elements from $F_{m-1} \setminus \{i\}$, say $V = \{i_1, \ldots, i_n\}$ and $W = \{j_1, \ldots, j_k\}$, with $1 \leq n + k \leq m - 2$, it holds that

$$\begin{aligned}
&\det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{V}_{\Omega} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{W}_{\Omega}, \mathbf{E}(s)^{L_m \cap T}_{\Omega}, \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{\{j\}}_{\Omega}, \mathbf{E}(s)^{L_{j+1}}_{\Omega}\Big) \\
&= \widetilde{G}(s)\lambda_m \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{V}_{\Omega} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{W \cup \{m\}}_{\Omega}, \mathbf{E}(s)^{L_{m+1} \cap T}_{\Omega}, \\
&\quad \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{\{j\}}_{\Omega}, \mathbf{E}(s)^{L_{j+1}}_{\Omega}\Big) \\
&+ \lambda_m \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{V \cup \{m\}}_{\Omega} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{W}_{\Omega}, \mathbf{E}(s)^{L_{m+1} \cap T}_{\Omega}, \\
&\quad \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{\{j\}}_{\Omega}, \mathbf{E}(s)^{L_{j+1}}_{\Omega}\Big) \\
&- (s - \lambda_m) \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)^{V}_{\Omega \setminus \{m\}} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)^{W}_{\Omega \setminus \{m\}}, \mathbf{E}(s)^{L_{m+1} \cap T}_{\Omega \setminus \{m\}},
\end{aligned}$$

$$\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{\{j\}}_{\Omega\setminus\{m\}},\mathbf{E}(s)^{L_{j+1}}_{\Omega\setminus\{m\}}\Big),$$

and

$$\begin{aligned}
&\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{V}_{\Omega}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{W}_{\Omega},\mathbf{E}(s)^{L_m\cap T}_{\Omega},\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{\{j\}}_{\Omega},\mathbf{E}(s)^{L_{j+1}}_{\Omega}\Big)\\
&=\widetilde{G}(s)\lambda_m\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{V}_{\Omega}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{W\cup\{m\}}_{\Omega},\mathbf{E}(s)^{L_{m+1}\cap T}_{\Omega},\\
&\quad\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{\{j\}}_{\Omega},\mathbf{E}(s)^{L_{j+1}}_{\Omega}\Big)\\
&+\lambda_m\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{V\cup\{m\}}_{\Omega}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{W}_{\Omega},\mathbf{E}(s)^{L_{m+1}\cap T}_{\Omega},\\
&\quad\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{\{j\}}_{\Omega},\mathbf{E}(s)^{L_{j+1}}_{\Omega}\Big)\\
&-(s-\lambda_m)\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{V}_{\Omega\setminus\{m\}}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{W}_{\Omega\setminus\{m\}},\mathbf{E}(s)^{L_{m+1}\cap T}_{\Omega\setminus\{m\}},\\
&\quad\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{\{j\}}_{\Omega\setminus\{m\}},\mathbf{E}(s)^{L_{j+1}}_{\Omega\setminus\{m\}}\Big),
\end{aligned}$$

where $\Omega=V\cup W\cup\{i\}\cup(L_m\cap T)\cup L_{j+1}$.

3. $m\in L_{j+1}$. For every two non-overlapping collections of n and k elements from $F_{m-1}\setminus\{i\}$, say $V=\{i_1,\dots,i_n\}$ and $W=\{j_1,\dots,j_k\}$, with $1\le n+k\le m-2$, it holds that

$$\begin{aligned}
&\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{V}_{\Omega}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{W}_{\Omega},\mathbf{E}(s)^{L_m}_{\Omega}\Big)\\
&=\widetilde{G}(s)\lambda_m\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{V}_{\Omega}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{W\cup\{m\}}_{\Omega},\mathbf{E}(s)^{L_{m+1}}_{\Omega}\Big)\\
&\quad+\lambda_m\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{V\cup\{m\}}_{\Omega}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{W}_{\Omega},\mathbf{E}(s)^{L_{m+1}}_{\Omega}\Big)\\
&\quad+(s-\lambda_m)\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{V}_{\Omega\setminus\{m\}}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{W}_{\Omega\setminus\{m\}},\mathbf{E}(s)^{L_{m+1}}_{\Omega\setminus\{m\}}\Big),
\end{aligned}$$

where $\Omega=V\cup W\cup L_m$.

By using the above formulas to evaluate all the involved determinants, we find that

$$\begin{aligned}
\mathcal{E}_{ij}(s)=&(-1)^{i+j}\widetilde{G}(s)\sum_{k=0}^{N-2}\widetilde{G}^k(s)\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i,j\}\\|\Gamma|=k;\\S\subset\mathcal{N}\setminus\{i,j\}\\S\supset\Gamma;\\R\subset S\cap T}}(-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{j\}}\zeta^{S^c}(s)\\
&\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{S\setminus\Gamma}_{S\cup\{i\}}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{\Gamma\cup\{j\}}_{S\cup\{i\}}\Big)\\
+(-1)^{i+j}&\sum_{k=0}^{N-2}\widetilde{G}^k(s)\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i,j\}\\|\Gamma|=k;\\S\subset\mathcal{N}\setminus\{i,j\}\\S\supset\Gamma;\\R\subset S\cap T}}(-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{j\}}\zeta^{S^c}(s)\\
&\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{(S\setminus\Gamma)\cup\{j\}}_{S\cup\{i\}}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{\Gamma}_{S\cup\{i\}}\Big),
\end{aligned}$$

which holds even when $T=\emptyset$.

We assume now that $i > j$, and we have to calculate

$$\begin{aligned}\mathcal{E}_{ij}(s) =&(-1)^{i+j}\widetilde{G}(s)\lambda_j \det\left(\mathbf{E}(s)^{F_{j-1}}_{\mathcal{N}\setminus\{j\}}, \left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)^{\{j\}}_{\mathcal{N}\setminus\{j\}}, \mathbf{E}(s)^{L_{j+1}\setminus\{i\}}_{\mathcal{N}\setminus\{j\}}\right)\\ &+(-1)^{i+j}\lambda_j \det\left(\mathbf{E}(s)^{F_{j-1}}_{\mathcal{N}\setminus\{j\}}, \left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)^{\{j\}}_{\mathcal{N}\setminus\{j\}}, \mathbf{E}(s)^{L_{j+1}\setminus\{i\}}_{\mathcal{N}\setminus\{j\}}\right).\end{aligned}$$

In this case, $T = \{j+1,\dots,i-1\}$. When $T \neq \emptyset$, for $n = j+1,\dots,i-1$, the elements $\mathbf{E}_{nn}(s) = \widetilde{G}_{nn}(s)p_{nn}\lambda_n + s - \lambda_n$, which are on the diagonal of matrix $\mathbf{E}(s)$, move to the upper-diagonal of the matrices $\left(\mathbf{E}(s)^{F_{j-1}}_{\mathcal{N}\setminus\{j\}}, \left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)^{\{j\}}_{\mathcal{N}\setminus\{j\}}, \mathbf{E}(s)^{L_{j+1}\setminus\{i\}}_{\mathcal{N}\setminus\{j\}}\right)$ and $\left(\mathbf{E}(s)^{F_{j-1}}_{\mathcal{N}\setminus\{j\}}, \left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)^{\{j\}}_{\mathcal{N}\setminus\{j\}}, \mathbf{E}(s)^{L_{j+1}\setminus\{i\}}_{\mathcal{N}\setminus\{j\}}\right)$.

The formula is exactly the same, with $T = \{i+1,\dots,j-1\}$. Thus, gathering all the above, for $i \neq j$

$$\begin{aligned}\mathcal{E}_{ij}(s) =&(-1)^{i+j}\sum_{k=1}^{N-1}\widetilde{G}^k(s)\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i,j\}\\ |\Gamma|=k-1;\\ S\subset\mathcal{N}\setminus\{i,j\}\\ S\supset\Gamma;\\ R\subset S\cap T_{ij}}}(-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{j\}}\zeta^{S^c}(s)\\ &\times\det\left(\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)^{S\setminus\Gamma}_{S\cup\{i\}}\bowtie\left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)^{\Gamma\cup\{j\}}_{S\cup\{i\}}\right)\\ &+(-1)^{i+j}\sum_{k=0}^{N-2}\widetilde{G}^k(s)\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i,j\}\\ |\Gamma|=k;\\ S\subset\mathcal{N}\setminus\{i,j\}\\ S\supset\Gamma;\\ R\subset S\cap T_{ij}}}(-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{j\}}\zeta^{S^c}(s)\\ &\times\det\left(\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)^{(S\setminus\Gamma)\cup\{j\}}_{S\cup\{i\}}\bowtie\left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)^{\Gamma}_{S\cup\{i\}}\right),\end{aligned}$$

where $m_{ij} = \min\{i,j\}$, $M_{ij} = \max\{i,j\}$, and $T_{ij} = \{m_{ij}+1,\dots,M_{ij}-1\}$. □

The adjoint matrix $\mathcal{E}(s)$ is equal to the transpose of the cofactor matrix of $\mathbf{E}(s)$. Therefore, similarly to $\det\mathbf{E}(s)$, each element of $\mathcal{E}(s)$ is an at most $N-1$ degree polynomial with respect to $\widetilde{G}(s)$. This observation explains also the similarity between the formula of $\det\mathbf{E}(s)$ and the diagonal elements of $\mathcal{E}(s)$.

Running example (continued). Recall that so far we have calculated $\det\mathbf{E}(s)$. By using the same arguments as for the evaluation of the determinant, from Theorem 4.4, we have for the adjoint matrix

$$\begin{aligned}\mathcal{E}_{ii}(s) =&\widetilde{G}^0(s)\boldsymbol{\lambda}^{\emptyset}\zeta^{\mathcal{N}\setminus\{i\}}(s)\det\left(\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)^{\emptyset}_{\emptyset}\bowtie\left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)^{\emptyset}_{\emptyset}\right) = s-\lambda, \qquad i=1,2,\\ \mathcal{E}_{12}(s) =&(-1)^{1+2}(-1)^{|\emptyset|}\boldsymbol{\lambda}^{\emptyset\cup\{2\}}\zeta^{\emptyset}(s)\det\left(\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)^{\{2\}}_{\{1\}}\bowtie\left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)^{\emptyset}_{\{1\}}\right) = -\lambda,\\ \mathcal{E}_{21}(s) =&(-1)^{2+1}\widetilde{G}(s)(-1)^{|\emptyset|}\boldsymbol{\lambda}^{\emptyset\cup\{1\}}\zeta^{\emptyset}(s)\det\left(\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)^{\emptyset}_{\{2\}}\bowtie\left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)^{\{1\}}_{\{2\}}\right)\\ =&-\lambda\widetilde{G}(s).\end{aligned}$$

■

For the evaluation of the Laplace transform $\widetilde{w}(s)$ of the queueing delay, it is only left to calculate $s\mathbf{u}\mathcal{E}(s)\boldsymbol{\omega}$ (see Eqs. (4.9) and (4.13)). Observe that the elements of the transform vector $\widetilde{\mathbf{\Phi}}(s)$ are defined as $\widetilde{\phi}_i(s) = s\mathbf{u}\mathcal{E}(s)\mathbf{e}_i/\det\mathbf{E}(s)$. The outcome of $s\mathbf{u}\mathcal{E}(s)\mathbf{e}_i$ is the inner product of the vector $s\mathbf{u}$ with the ith column of the matrix $\mathcal{E}(s)$. Therefore, as a first step we calculate the quantities $s\mathbf{u}\mathcal{E}(s)\mathbf{e}_i$, and we have the following theorem.

Theorem 4.5. *The numerator of the ith element of the transform vector* $\widetilde{\mathbf{\Phi}}(s)$ *takes the form*

$$\begin{aligned}
s\mathbf{u}\mathcal{E}(s)\mathbf{e}_i =& su_i \sum_{k=0}^{N-1} \widetilde{G}^k(s) \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i\}\\|\Gamma|=k;\\S\subset\mathcal{N}\setminus\{i\}\\S\supset\Gamma}} \boldsymbol{\lambda}^S \zeta^{S^c}(s) \det\left(\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)_S^{S\setminus\Gamma} \bowtie \left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)_S^{\Gamma}\right)\\
&+s\sum_{\substack{l=1\\l\neq i}}^{N} u_l(-1)^{l+i} \sum_{k=1}^{N-1} \widetilde{G}^k(s) \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\|\Gamma|=k-1}} \sum_{\substack{S\subset\mathcal{N}\setminus\{l,i\}\\S\supset\Gamma;\\R\subset S\cap T_{li}}} (-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s)\\
&\times \det\left(\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)_{S\cup\{l\}}^{S\setminus\Gamma} \bowtie \left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)_{S\cup\{l\}}^{\Gamma\cup\{i\}}\right)\\
&+s\sum_{\substack{l=1\\l\neq i}}^{N} u_l(-1)^{l+i} \sum_{k=0}^{N-2} \widetilde{G}^k(s) \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\|\Gamma|=k}} \sum_{\substack{S\subset\mathcal{N}\setminus\{l,i\}\\S\supset\Gamma;\\R\subset S\cap T_{li}}} (-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s)\\
&\times \det\left(\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)_{S\cup\{l\}}^{(S\setminus\Gamma)\cup\{i\}} \bowtie \left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)_{S\cup\{l\}}^{\Gamma}\right).
\end{aligned}$$

Proof. We calculate the inner product of the vector $s\mathbf{u}$ with the ith column of the matrix $\mathcal{E}(s)$, namely $s\mathbf{u}\mathcal{E}(s)\mathbf{e}_i = s\sum_{l=1}^{N} u_l\mathcal{E}_{li}(s) = s\sum_{\substack{l=1\\l\neq i}}^{N} u_l\mathcal{E}_{li}(s) + su_i\mathcal{E}_{ii}(s)$. By using the definition of $\mathcal{E}_{ij}(s)$, $\forall i,j\in\mathcal{N}$, and Theorem 4.4, the result is straightforward. □

By combining now the results of the Theorems 4.3 and 4.5 using Eq. (4.13), one can find the transform vector $\widetilde{\mathbf{\Phi}}(s)$.

Running example (continued). To find the transform vector $\widetilde{\mathbf{\Phi}}(s)$, we need to calculate the vector $\mathbf{u}\mathcal{E}(s)\mathbf{e}$. For each state we have

$$\begin{aligned}
\mathbf{u}\mathcal{E}(s)\mathbf{e}_1 =& u_1\mathcal{E}_{11}(s) + u_2\mathcal{E}_{21}(s) = u_1(s-\lambda) - u_2\lambda\widetilde{G}(s),\\
\mathbf{u}\mathcal{E}(s)\mathbf{e}_2 =& u_1\mathcal{E}_{12}(s) + u_2\mathcal{E}_{22}(s) = -u_1\lambda + u_2(s-\lambda).
\end{aligned}$$

The transform vector $\widetilde{\mathbf{\Phi}}(s)$ is then

$$\widetilde{\mathbf{\Phi}}(s) = \left(\frac{su_1(s-\lambda) - su_2\lambda\widetilde{G}(s)}{(s-\lambda)^2 - \lambda^2\widetilde{G}(s)}, \frac{-su_1\lambda + su_2(s-\lambda)}{(s-\lambda)^2 - \lambda^2\widetilde{G}(s)}\right).$$

■

The following remark connects the system of equations that is required for the evaluation of $\mathbf{u}$, which was introduced in Theorem 4.2, to the adjoint matrix $\mathcal{E}(s)$.

Remark 4.6. The second statement of Theorem 4.2 practically says that each s_i, $i = 2, \ldots, N$, is a simple eigenvalue of the matrix $\mathbf{H}(s) + s\mathbf{I} - \mathbf{\Lambda}$. Therefore, the column vector $\mathbf{a}_i$ belongs to the null space of the matrix $\mathbf{H}(s_i) + s_i\mathbf{I} - \mathbf{\Lambda}$. Based on some general results with respect to the form of the null space of a singular matrix (see Theorem A.4, Remark A.6, and Corollary A.7 for details), without loss of generality we can assume that the vector $\mathbf{a}_i$ is any non-zero column of the matrix $\mathcal{E}(s_i)$. Namely, if the mth column of $\mathcal{E}(s_i)$ is such a column, then

$$\mathbf{a}_i := \mathbf{a}_i(s_i) = \big(\mathcal{E}(s_i)\big)_{\mathcal{N}}^{\{m\}}, \qquad i = 2, \ldots, N. \tag{4.14}$$

This observation is very useful because it allows us to calculate in a straightforward way the desired system of equations and find closed form expressions for the vector $\mathbf{u}$. In addition, since the vectors $\mathbf{a}_i$, $i = 2, \ldots, N$, are matrix functions evaluated at the point $s = s_i$ we define the derivative of each $\mathbf{a}_i$ as

$$\mathbf{a}_i{}^{(1)} = \frac{d}{ds}\mathbf{a}_i(s)\bigg|_{s=s_i}, \qquad i = 2, \ldots, N.$$

The usefulness of the latter definition will be apparent in Section 4.3.2, where we provide an extension of Theorem 4.2 that helps us to calculate our approximations.

Running example (continued). The transform vector $\widetilde{\mathbf{\Phi}}(s)$ found in the previous block is expressed in terms of the vector $\mathbf{u}$. Here, we find an exact expression for $\mathbf{u}$. If s_2 is the only positive (and real) root of the equation $\det \mathbf{E}(s) = 0$, the vector $\mathbf{u}$ satisfies the system of Eqs. (4.10)–(4.11)

$$\begin{aligned} \frac{1}{\lambda}u_1 + \frac{1}{\lambda}u_2 &= \frac{1}{\lambda} - \frac{\mu}{2}, \\ -\lambda u_1 + (s_2 - \lambda)u_2 &= 0, \end{aligned}$$

where for the derivation of the second equation we used the second column of the matrix $\mathcal{E}(s)$. Namely, we used $\mathbf{a}_2 = \big(\mathcal{E}(s_2)\big)_{\mathcal{N}}^{\{2\}}$. It is easy to verify that the solution to the above system is given by

$$\mathbf{u} = \left(\left(1 - \frac{\lambda}{s_2}\right)\left(1 - \frac{\lambda\mu}{2}\right),\ \frac{\lambda}{s_2}\left(1 - \frac{\lambda\mu}{2}\right)\right).$$

■

Although Theorems 4.4 and 4.5 provide explicit expressions for the transform vector, they may not be practical in cases where the LST of the service time distribution of a real customer $\widetilde{G}(s)$, which is involved in the formulas, does not have a closed form; i.e. Pareto distribution. In such cases, one would have to either consort to a numerical evaluation of $\widetilde{G}(s)$ or approximate the transform vector $\widetilde{\mathbf{\Phi}}(s)$ in some other fashion. This chapter focuses on the latter approach, which we work out in detail in the following section by taking as starting point a mixture model for the service time distribution of a real customer.

4.2.2 Construction of the corrected phase-type approximations

We assume now that the service time distribution of a real customer is $G_\epsilon(t)$, which was defined in Eq. (4.1) as a mixture of a phase-type distribution and a heavy-tailed one. We will eventually show that the queueing delay can be written also as a mixture, in the sense that we can identify the queueing delay of a model with purely phase-type service times and some additional terms that involve the heavy-tailed service times. As a result, in order to derive our approximations, we first need to compute the delay in a MArP/PH/1 queue and afterwards use this as a base to further develop our approximations involving a heavy-tailed component. In the sequel, we give a more detailed description of our technique.

In terms of Laplace transforms, we get for our mixture service time distribution $\widetilde{G}_\epsilon(s) = (1-\epsilon)\widetilde{F}_p(s) + \epsilon\widetilde{F}_h(s)$. As observed in Section 4.2.1, when the service time distribution of a real customer is of phase type, then the determinant $\det \mathbf{E}(s)$ and the elements of the adjoint matrix $\mathcal{E}(s)$ are all rational functions in s. Therefore, after the cancelation of the problematic factors $s - s_i$, $i = 1, \ldots, N$, that appear in the denominator (see the analysis below Theorem 4.2), the elements of the transform vector $\widetilde{\mathbf{\Phi}}(s)$ are also rational functions in s and they can easily be inverted to find the delay distribution.

Note now that the LST of the service time distribution of a real customer $\widetilde{G}_\epsilon(s)$ can be written in the following two ways:

$$\widetilde{G}_\epsilon(s) = \widetilde{F}_p(s) + \epsilon\big(\widetilde{F}_h(s) - \widetilde{F}_p(s)\big) \quad \text{or} \quad \widetilde{G}_\epsilon(s) = (1-\epsilon)\widetilde{F}_p(s) + \epsilon + \epsilon\big(\widetilde{F}_h(s) - 1\big).$$

In both formulas, the LST of the service time distribution $\widetilde{G}_\epsilon(s)$ can be seen as perturbation of a phase-type distribution by a term that contains the heavy-tailed component $\widetilde{F}_h(s)$. The index ϵ is interpreted as the perturbation parameter and it is used for all the parameters of the system that depend on it. Next, we explain how these two different representations of the same formula can lead with the aid of perturbation analysis to two different approximations for the queueing delay.

We start our discussion with the first formula. We set $\widetilde{F}_h(s) \equiv \widetilde{F}_p(s)$ in the formula, or in other words, we assume that all of the customers come from the same phase-type distribution or equivalently that we replace all the heavy-tailed customers with phase-type ones. Therefore, one can find with $\widetilde{G}_\epsilon(s) = \widetilde{F}_p(s)$ the delay of a simpler MArP/PH/1 queue, by specialising the formulas of Section 4.2.1 to phase-type service times. As a next step, we find all the parameters of the mixture model as perturbation of the simpler phase-type model, which we use as base. Then, we write the queueing delay of the mixture model in a series expansion in ϵ, where the constant term is the delay of the MArP/PH/1 queue we used as base and all other terms contain the heavy-tailed service times.

We define our approximation by taking the first two terms of the aforementioned series, namely the up to ϵ-order terms. We call this approximation *corrected replace approximation*. The characterisation "corrected" comes from the fact that the ϵ-order term corrects the tail behaviour of the constant term, which as a phase-type approximation of the queueing delay is incapable of capturing the correct tail behaviour. Finally, the characterisation "replace" is due to the phase-type base model we used. This approximation is the extension of Approximation 3.4 to MArPs. We give analytically all the steps to derive the corrected replace approximation in Section 4.3.

In a similar manner, we construct the *corrected discard approximation* by using the second formula; see also Approximation 3.3. We first discard the heavy-tailed customers from the system by simply setting $\widetilde{F}_h(s) \equiv 1$. Afterwards, we derive the queueing delay of the phase-type base model with service time distribution $\widetilde{G}^{\bullet}_{\epsilon}(s) = (1-\epsilon)\widetilde{F}_p(s) + \epsilon$ for a real customer, which has an atom of size ϵ at zero. Throughout the chapter, we use $\widetilde{G}^{\bullet}_{\epsilon}(s)$ for the LST of the service time distribution of a real customer in the discard base model instead of $\widetilde{G}_{\epsilon}(s)$ to avoid confusion with the mixture model. We briefly discuss the details for the construction of the corrected discard approximation in Section 4.4.

In the next sections, we provide the steps to construct the corrected replace and the corrected discard approximations, which we call collectively *corrected phase-type approximations*.

4.3 Corrected replace approximation

In this section, we construct the corrected replace approximation. First, we calculate the queueing delay for the phase-type model that appears when we replace all the heavy-tailed customers with phase-type ones in Section 4.3.1, i.e. we specialise the results of Section 4.2.1 to phase-type service times. Later, in Section 4.3.2, we calculate the parameters of the mixture model with service time distribution $\widetilde{G}_{\epsilon}(s)$ given by Eq. (4.1) as perturbation of the parameters of the corresponding phase-type model, with perturbation parameter ϵ. In Section 4.3.3, we find a series expansion in ϵ of the queueing delay in the mixture model with constant term the queueing delay in the phase-type base model and all higher terms involving the heavy-tailed services. Finally, in Section 4.3.4, we construct the corrected replace approximation by keeping only the first two terms of the aforementioned series. We start in the next section with the analysis of the replace base model; i.e. the one containing only phase-type service times.

4.3.1 Replace base model

When we replace the heavy-tailed customers with phase-type ones, we consider the service time distribution $\widetilde{G}_{\epsilon}(s) = \widetilde{F}_p(s)$ for our phase-type base model. Observe that this service time distribution is independent of the parameter ϵ, and so will be all the other parameters of this simpler model. Thus, from a mathematical point of view, the action of replacing the heavy-tailed service times with phase-type ones is equivalent to setting $\epsilon = 0$ in the mixture model.

To avoid overloading the notation, we omit the subscript "0" (which is a consequence of the fact that $\epsilon = 0$) from the parameters of the replace phase-type model and we assume that the service time distribution of a real customer is some phase-type distribution with LST $\widetilde{G}(s) := \widetilde{F}_p(s) = q(s)/p(s)$, where $q(s)$ and $p(s)$ are appropriate polynomials without common roots. The degree of $p(s)$ is M, and without loss of generality, we choose the coefficient of its highest order term to be equal to 1. Finally,

the degree of the polynomial $q(s)$ is less than or equal to $M-1$. Define

$$K=\max_{k\neq 0}\left\{\max_{\Gamma\subset\mathcal{N}}\left\{\operatorname{rank}\left(\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)_{S}^{S\setminus\Gamma}\bowtie\left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)_{S}^{\Gamma}\right)\right\}:k=\mid\Gamma\mid,\text{ and }\Gamma\subset S\subset\mathcal{N}\right\}. \tag{4.15}$$

Then, the following result holds.

Proposition 4.7. *There exist x_j and y_j, with $\Re(x_j)>0$, $\Re(y_j)>0$, $j=1,\ldots,rM$, such that the Laplace transform $\widetilde{w}(s)$ of the queueing delay takes the form*

$$\widetilde{w}(s)=\frac{\mathbf{u}\boldsymbol{\omega}\prod_{j=1}^{rM}(s+y_j)}{\prod_{j=1}^{rM}(s+x_j)},$$

where the vector $\mathbf{u}$ is calculated according to Theorem 4.2 with the LST of the service times being equal to $\widetilde{F}_p(s)$, and r is some positive integer less than or equal to K defined by Eq. (4.15).

Proof. The main idea here is to find $\widetilde{w}(s)$ by taking $\widetilde{G}(s)=q(s)/p(s)$. In this case, the determinant $\det\mathbf{E}(s)$ (see Theorem 4.3) takes the form

$$\begin{aligned}\det\mathbf{E}(s)=&\sum_{S\subset\mathcal{N}}\boldsymbol{\lambda}^{S}\zeta^{S^c}(s)\det\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)_{S}^{S}\\&+\sum_{k=1}^{N}\left(\frac{q(s)}{p(s)}\right)^{k}\sum_{\substack{\Gamma\subset\mathcal{N}\\|\Gamma|=k}}\sum_{\substack{S\subset\mathcal{N}\\S\supset\Gamma}}\boldsymbol{\lambda}^{S}\zeta^{S^c}(s)\det\left(\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)_{S}^{S\setminus\Gamma}\bowtie\left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)_{S}^{\Gamma}\right),\end{aligned} \tag{4.16}$$

and the numerator of $\widetilde{w}(s)$ (see Eq. (4.9) and Theorem 4.5) becomes

$$\begin{aligned}s\mathbf{u}\mathcal{E}(s)\boldsymbol{\omega}=&s\sum_{i=1}^{N}u_i\omega_i\sum_{k=0}^{N-1}\left(\frac{q(s)}{p(s)}\right)^{k}\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i\}\\|\Gamma|=k;\\S\subset\mathcal{N}\setminus\{i\}\\S\supset\Gamma}}\boldsymbol{\lambda}^{S}\zeta^{S^c}(s)\\&\times\det\left(\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)_{S}^{S\setminus\Gamma}\bowtie\left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)_{S}^{\Gamma}\right)\\&+s\sum_{i=1}^{N}\omega_i\sum_{\substack{l=1\\l\neq i}}^{N}u_l(-1)^{l+i}\sum_{k=1}^{N-1}\left(\frac{q(s)}{p(s)}\right)^{k}\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\|\Gamma|=k-1;\\S\subset\mathcal{N}\setminus\{l,i\}\\S\supset\Gamma;\\R\subset S\cap T_{li}}}(-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s)\\&\times\det\left(\left(\mathbf{Q}^{(1)}\circ\mathbf{P}\right)_{S\cup\{l\}}^{S\setminus\Gamma}\bowtie\left(\mathbf{Q}^{(2)}\circ\mathbf{P}\right)_{S\cup\{l\}}^{\Gamma\cup\{i\}}\right)\\&+s\sum_{i=1}^{N}\omega_i\sum_{\substack{l=1\\l\neq i}}^{N}u_l(-1)^{l+i}\sum_{k=0}^{N-2}\left(\frac{q(s)}{p(s)}\right)^{k}\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\|\Gamma|=k;\\S\subset\mathcal{N}\setminus\{l,i\}\\S\supset\Gamma;\\R\subset S\cap T_{li}}}(-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s)\end{aligned}$$

$$\times \det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{(S\setminus\Gamma)\cup\{i\}} \bowtie \big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{\Gamma}\Big). \tag{4.17}$$

Observe that both the denominator (4.16) and the numerator (4.17) of $\widetilde{w}(s)$ are rational functions with denominators the polynomial $p(s)$ raised to some power. To simplify as much as possible the expression of $\widetilde{w}(s)$, we multiply Eqs. (4.16) and (4.17) with $\big(p(s)\big)^r$, where $r\in\mathcal{N}$ is the highest possible power of $p(s)$ that is involved in the formulas. It is immediately obvious that $r\le K$. Therefore, we multiply both Eqs. (4.16) and (4.17) with $\big(p(s)\big)^r$

When multiplied with $\big(p(s)\big)^r$, the denominator of $\widetilde{w}(s)$ becomes

$$\begin{aligned}\big(p(s)\big)^r \det\mathbf{E}(s) =& \big(p(s)\big)^r \sum_{S\subset\mathcal{N}} \boldsymbol{\lambda}^S\boldsymbol{\zeta}^{S^c}(s)\det\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^S\\ &+\sum_{k=1}^{N}\big(q(s)\big)^k\big(p(s)\big)^{r-k}\sum_{\substack{\Gamma\subset\mathcal{N}\\|\Gamma|=k}}\sum_{\substack{S\subset\mathcal{N}\\S\supset\Gamma}}\boldsymbol{\lambda}^S\boldsymbol{\zeta}^{S^c}(s)\\ &\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^{S\setminus\Gamma}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_S^{\Gamma}\Big).\end{aligned} \tag{4.18}$$

The term $\big(p(s)\big)^r\sum_{S\subset\mathcal{N}}\boldsymbol{\lambda}^S\boldsymbol{\zeta}^{S^c}(s)\det\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^S$ is a polynomial of degree $rM+N$. The coefficient of s^{rM+N} is found when we set $S=\emptyset$, and it is equal to 1.

Let now n be the degree of the polynomial $q(s)$. Therefore, the second term of the right hand side of Eq. (4.18) is a polynomial of degree at most $n+(r-1)M+N-1$ (the highest order of s is found when $|S|=1$). Since $n\le M-1$, it is immediately obvious that $\big(p(s)\big)^r\det\mathbf{E}(s)$ is a polynomial of degree $N+rM$, thus it has exactly $N+rM$ roots. From Theorem 4.2, we know that exactly $N-1$ of its roots have positive real part and that zero is also a root. We denote these roots as $s_1=0$ and $s_2,\dots,s_N$, and we assume them to be simple. We denote the remaining rM roots with negative real part as $-x_j$, $j=1,\dots,rM$. Consequently, the denominator of $\widetilde{w}(s)$ is written as

$$\big(p(s)\big)^r\det\mathbf{E}(s) = s\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+x_j). \tag{4.19}$$

Similarly, the numerator of $\widetilde{w}(s)$ becomes

$$\begin{aligned}\big(p(s)\big)^r s\mathbf{u}\mathcal{E}(s)\boldsymbol{\omega} =& s\sum_{i=1}^{N}u_i\omega_i\big(p(s)\big)^r\sum_{S\subset\mathcal{N}\setminus\{i\}}\boldsymbol{\lambda}^S\boldsymbol{\zeta}^{S^c}(s)\det\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^S\\ &+s\sum_{i=1}^{N}u_i\omega_i\sum_{k=1}^{N-1}\big(q(s)\big)^k\big(p(s)\big)^{r-k}\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i\}\\|\Gamma|=k;\\S\subset\mathcal{N}\setminus\{i\}\\S\supset\Gamma}}\boldsymbol{\lambda}^S\boldsymbol{\zeta}^{S^c}(s)\\ &\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^{S\setminus\Gamma}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_S^{\Gamma}\Big)\end{aligned}$$

$$+s\sum_{i=1}^{N}\omega_i\sum_{\substack{l=1\\l\neq i}}^{N}u_l(-1)^{l+i}\sum_{k=1}^{N-1}\big(q(s)\big)^k\big(p(s)\big)^{r-k}\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\|\Gamma|=k-1;\\S\subset\mathcal{N}\setminus\{l,i\}\\S\supset\Gamma;\\R\subset S\cap T_{li}}}(-1)^{|R|}$$
$$\times\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s)\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{S\setminus\Gamma}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{\Gamma\cup\{i\}}\Big)$$
$$+s\sum_{i=1}^{N}\omega_i\sum_{\substack{l=1\\l\neq i}}^{N}u_l(-1)^{l+i}\sum_{k=0}^{N-2}\big(q(s)\big)^k\big(p(s)\big)^{r-k}\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\|\Gamma|=k;\\S\subset\mathcal{N}\setminus\{l,i\}\\S\supset\Gamma;\\R\subset S\cap T_{li}}}(-1)^{|R|}$$
$$\times\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s)\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{(S\setminus\Gamma)\cup\{i\}}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{\Gamma}\Big).$$

It is easy to verify that $\big(p(s)\big)^r s\mathbf{u}\mathcal{E}(s)\boldsymbol{\omega}$ is also a polynomial of degree $rM+N$. The coefficient of s^{rM+N} is equal to $\mathbf{u}\boldsymbol{\omega}$ and it is determined by the term $s\sum_{i=1}^{N}u_i\omega_i\big(p(s)\big)^r\sum_{S\subset\mathcal{N}\setminus\{i\}}\boldsymbol{\lambda}^S\zeta^{S^c}(s)\det\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^S$ for $S=\emptyset$. We know from Theorem 4.2, that the vector $\mathbf{u}$ is such that the numbers s_k, $k\in\mathcal{N}$, are also roots of the numerator of $\widetilde{w}(s)$. We denote the rest rM roots of the numerator as $-y_j$, $j=1,\ldots,rM$. Therefore, the numerator of $\widetilde{w}(s)$ is written as

$$\big(p(s)\big)^r s\mathbf{u}\mathcal{E}(s)\boldsymbol{\omega}=\mathbf{u}\boldsymbol{\omega}s\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+y_j).\tag{4.20}$$

Combining Eqs. (4.19) and (4.20), the result is immediate. □

The formula of $\widetilde{w}(s)$ is a rational function that corresponds to a phase-type distribution. Applying Laplace inversion to $\widetilde{w}(s)$, we can find the exact tail probabilities of the queueing delay; namely we can find $\mathbb{P}(W>t)$.

Running example (continued). Given that we have already calculated the transform vector $\widetilde{\boldsymbol{\Phi}}(s)$, we can now calculate the Laplace transform $\widetilde{w}(s)$ of the queueing delay for phase-type customers, by using Proposition 4.7. In our example, $K=1$ and consequently $r=1$. In addition, $\boldsymbol{\omega}^T=(0,2)$. Thus, $\widetilde{w}(s)$ under phase-type service times is

$$\widetilde{w}(s)=2\widetilde{\phi}_2(s)=2\frac{s^2u_2-s\lambda(u_1+u_2)}{(s-\lambda)^2-\lambda^2\widetilde{F}_p(s)}=2\frac{s^2p(s)u_2-sp(s)\lambda(u_1+u_2)}{(s-\lambda)^2p(s)-\lambda^2q(s)}.$$

Observe that both the numerator and the denominator of $\widetilde{w}(s)$ are polynomials of degree $M+2$. Moreover, Theorem 4.2 guarantees that 0 and s_2 are common roots of them. If $-y_j$ and $-x_j$, $j=1,\ldots,M$, $\Re(x_j),\Re(y_j)>0$, are the remaining roots of the numerator and the denominator, respectively, the Laplace transform of the queueing delay can be written as

$$\widetilde{w}(s)=\frac{2u_2s(s-s_2)\prod_{j=1}^{M}(s+y_j)}{s(s-s_2)\prod_{j=1}^{M}(s+x_j)}=\frac{2u_2\prod_{j=1}^{M}(s+y_j)}{\prod_{j=1}^{M}(s+x_j)}.$$

■

As pointed out in Section 4.2.2, the LST of the service time distribution $\widetilde{G}_\epsilon(s)$ (see Eq. (4.1)) can be seen as perturbation of $\widetilde{F}_p(s)$ by the term $\epsilon\big(\widetilde{F}_h(s) - \widetilde{F}_p(s)\big)$. In the next section we write the parameters of the mixture model as perturbation of the parameters of the replace base model.

4.3.2 Perturbation of the parameters of the replace base model

In order to find the queueing delay in the mixture model as a series expansion in ϵ with constant term the queueing delay in the replace base model, we apply perturbation analysis to the parameters of the mixture model that depend on ϵ. Thus, we first check which of the parameters in the mixture model depend on ϵ and then we represent them as perturbation of the parameters of the replace base model.

Since the matrices $\mathbf{P}$, $\mathbf{Q}^{(1)}$, $\mathbf{Q}^{(2)}$, and $\mathbf{\Lambda}$ (see Section 4.2.1) depend only on the arrival process, they are invariant under any perturbation of the service time distribution. However, the matrix $\widetilde{\mathbf{G}}_\epsilon(s)$, and consequently $\mathbf{H}_\epsilon(s)$ change, and so does the stability condition (see Eqs. (4.4)–(4.6)). Let now $\widetilde{F}^e_p(s)$ and $\widetilde{F}^e_h(s)$ be the LSTs of the stationary-excess service time distributions $F_p(t)$ and $F_h(t)$, and μ_p and μ_h be the finite means of the phase-type and heavy-tailed service times, respectively. Then, we obtain

$$\widetilde{\mathbf{G}}_\epsilon(s) = \widetilde{\mathbf{G}}(s) + \epsilon s\big(\mu_p \widetilde{F}^e_p(s) - \mu_h \widetilde{F}^e_h(s)\big)\mathbf{Q}^{(2)},$$

and

$$\mathbf{H}_\epsilon(s) = \widetilde{\mathbf{G}}_\epsilon(s) \circ \mathbf{P}\mathbf{\Lambda} = \mathbf{H}(s) + \epsilon s\big(\mu_p \widetilde{F}^e_p(s) - \mu_h \widetilde{F}^e_h(s)\big)\mathbf{Q}^{(2)} \circ \mathbf{\Lambda}.$$

Finally, the stability condition takes the form

$$\boldsymbol{\pi}(\mathbf{\Lambda}^{-\mathbf{1}} - \mathbf{M}_\epsilon)\mathbf{e} > 0, \tag{4.21}$$

where $\mathbf{M}_\epsilon = \mathbf{M} + \epsilon s\big(\mu_h - \mu_p\big)\mathbf{Q}^{(2)} \circ \mathbf{P}$.

Under the stability condition (4.21), Theorem 4.1 holds for the transform vector $\widetilde{\mathbf{\Phi}}_\epsilon(s)$, for some row vector $\mathbf{u}_\epsilon$. More precisely, there exists a unique vector $\mathbf{u}_\epsilon$ such that the transform vector $\widetilde{\mathbf{\Phi}}_\epsilon(s)$ satisfies the system of equations:

$$\widetilde{\mathbf{\Phi}}_\epsilon(s)\big(\mathbf{H}_\epsilon(s) + s\mathbf{I} - \mathbf{\Lambda}\big) = s\mathbf{u}_\epsilon, \tag{4.22}$$

$$\widetilde{\mathbf{\Phi}}_\epsilon(0)\mathbf{e} = 1, \tag{4.23}$$

where the vector $\mathbf{u}_\epsilon$ is calculated according to Theorem 4.2.

Recall that the evaluation of $\mathbf{u}_\epsilon$ goes through the evaluation of the positive eigenvalues of the matrix

$$\mathbf{E}_\epsilon(s) = \mathbf{H}_\epsilon(s) + s\mathbf{I} - \mathbf{\Lambda} = \mathbf{E}(s) + \epsilon s\big(\mu_p \widetilde{F}^e_p(s) - \mu_h \widetilde{F}^e_h(s)\big)\mathbf{Q}^{(2)} \circ \mathbf{\Lambda}. \tag{4.24}$$

Observe that the above representation of the matrix $\mathbf{E}_\epsilon(s)$ is a linear perturbation in ϵ of the matrix $\mathbf{E}(s)$ of the base model. Thus, according to results on perturbation of analytic matrix functions (De Terán, 2011; Lancaster et al., 2003), we have that the positive eigenvalues of the matrix $\mathbf{E}_\epsilon(s)$ and their corresponding eigenvectors are

analytic functions in ϵ. Consequently, one can find a series representation in ϵ for all the involved quantities that are necessary for the evaluation of the vector $\mathbf{u}_\epsilon$ (see Theorem 4.2). By using these parameters, we can find a complete series representation for the transform vector $\widetilde{\mathbf{\Phi}}_\epsilon(s)$ and by applying Laplace inversion to each term of this series we can find a formal expression for the queueing delay that is a series expansion in ϵ. As we stated earlier, we only need the first two terms of the latter series to define the corrected replace approximation. Therefore, in our analysis, we keep only the terms up to order ϵ of each involved perturbed parameter.

In the next theorem, we provide an algorithm to calculate the first order approximation in ϵ of the vector $\mathbf{u}_\epsilon$, given that we have already calculated the vector $\mathbf{u}$ of the replace base model, by specialising Theorem 4.2 to phase-type service times.

Theorem 4.8. *Let* $\mathbf{u}$ *be the unique solution to the Eqs.* (4.10)–(4.11) *for the replace base model. If the roots* $s_2, \dots, s_N$ *of* $\det\big(\mathbf{H}(s) + s\mathbf{I} - \mathbf{\Lambda}\big) = 0$ *with positive real part are simple, then*

1. *the equation* $\det\big(\mathbf{H}_\epsilon(s) + s\mathbf{I} - \mathbf{\Lambda}\big) = 0$ *has exactly* N *non-negative solutions* $s_{\epsilon,1}, \dots, s_{\epsilon,N}$, *with* $s_{\epsilon,1} = 0$ *and* $s_{\epsilon,i} = s_i - \epsilon\delta_i + O(\epsilon^2)$ *for* $i = 2, \dots, N$, *where*

$$\delta_i := \delta(s_i) = \frac{\sum_{j=1}^{N} \det\big(\mathbf{E}(s_i)_{\bullet 1}, \dots, \mathbf{K}(s_i)_{\bullet j}, \dots, \mathbf{E}(s_i)_{\bullet N}\big)}{\sum_{j=1}^{N} \det\big(\mathbf{E}(s_i)_{\bullet 1}, \dots, \mathbf{E}^{(1)}(s_i)_{\bullet j}, \dots, \mathbf{E}(s_i)_{\bullet N}\big)},$$

and $\mathbf{K}(s) = s\big(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s)\big)\mathbf{Q}^{(2)} \circ \mathbf{\Lambda}$.

2. *We assume that the stability condition* (4.21) *is satisfied and we set* $\mathbf{A} = \big(\mathbf{\Lambda^{-1}e}, \mathbf{a}_2, \dots, \mathbf{a}_N\big)$ *(see Eq.* (4.14)*) and* $\mathbf{c} = \big(\boldsymbol{\pi}(\mathbf{\Lambda^{-1}} - \mathbf{M})\mathbf{e}, 0, \dots, 0\big)$. *Then, the vector* $\mathbf{u}_\epsilon$ *is the unique solution to the system of* N *linear equations*

$$\mathbf{u}_\epsilon\big(\mathbf{A} - \epsilon\mathbf{B} + O(\epsilon^2\mathbf{U})\big) = \mathbf{c} + \epsilon\mathbf{d}, \tag{4.25}$$

where $\mathbf{B} = \big(\mathbf{0}, \delta_2 \mathbf{a}_2{}^{(1)} - \mathbf{k}_2, \dots, \delta_N \mathbf{a}_N{}^{(1)} - \mathbf{k}_N\big)$ *and* $\mathbf{d} = \big((\mu_p - \mu_h)\boldsymbol{\pi}\mathbf{Q}^{(2)} \circ \mathbf{Pe}, 0, \dots, 0\big)$, *with* $\mathbf{k}_i$, $i = 2, \dots, N$, *being a column vector with coordinates*

$$k_{i,j} = (-1)^{m+j} \sum_{k=1}^{N-1} \det\Bigg(\Big(\mathbf{E}(s_i)^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\Big)_{\bullet 1}, \dots, \Big(\mathbf{K}(s_i)^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\Big)_{\bullet k}, \dots, \Big(\mathbf{E}(s_i)^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\Big)_{\bullet N-1}\Bigg), \qquad j \in \mathcal{N},$$

and the choice of m *explained in Remark 4.6.*

Proof. For the first part, we write $\det\big(\mathbf{H}_\epsilon(s) + s\mathbf{I} - \mathbf{\Lambda}\big)$ as a perturbed function of $\det\big(\mathbf{H}(s) + s\mathbf{I} - \mathbf{\Lambda}\big)$ and we apply perturbation analysis to show that the eigenvalues with positive real part of the first determinant are perturbation of the latter's eigenvalues with positive real part. Namely, since $\mathbf{K}(0)$ is an $N \times N$ zero matrix, it is evident that $s_{\epsilon,1} = 0$ is an eigenvalue of the matrix $\mathbf{H}_\epsilon(s) + s\mathbf{I} - \mathbf{\Lambda}$ (see Eq. (4.24)). According to Corollary A.11, the numbers $s_{\epsilon,i}$, $i = 2, \dots, N$, are also simple eigenvalues of this matrix. Thus, according to Theorem 4.2, there are no other roots of the equation $\det\big(\mathbf{E}(s) + \epsilon\mathbf{K}(s)\big) = 0$ with non-negative real part besides the values $s_{\epsilon,i}$, $i \in \mathcal{N}$.

For the second part of the proof, by using Remark 4.6 and perturbation analysis, we find the form of the right eigenvectors of $\det\big(\mathbf{H}_\epsilon(s)+s\mathbf{I}-\mathbf{\Lambda}\big)$ that correspond to its eigenvalues with positive real part. Then, we follow closely the proof of Theorem 4.2 to find the system of equations that the vector $\mathbf{u}_\epsilon$ satisfies. More precisely, by using Theorem A.12, we can evaluate $N-1$ column vectors $\mathbf{w}_{\epsilon,i}$ such that

$$\big(\mathbf{H}_\epsilon(s_{\epsilon,i})+s_{\epsilon,i}\mathbf{I}-\mathbf{\Lambda}\big)\mathbf{w}_{\epsilon,i}=0, \qquad i=2,\dots,N.$$

Since $s_{\epsilon,i}\neq 0$, $i=2,\dots,N$, post-multiplying Eq. (4.22) with $s=s_{\epsilon,i}$ by $\mathbf{w}_{\epsilon,i}$, we obtain

$$\mathbf{u}_\epsilon\mathbf{w}_{\epsilon,i}=0, \qquad i=2,\dots,N.$$

To derive the remaining equation, we take the derivative of Eq. (4.22) with respect to s, yielding

$$\widetilde{\mathbf{\Phi}}_\epsilon(s)\big(\mathbf{H}_\epsilon^{(1)}(s)+\mathbf{I}\big)+\widetilde{\mathbf{\Phi}}_\epsilon^{(1)}(s)\big(\mathbf{H}_\epsilon(s)+s\mathbf{I}-\mathbf{\Lambda}\big)=\mathbf{u}_\epsilon.$$

By setting $s=0$, we get

$$\widetilde{\mathbf{\Phi}}_\epsilon(0)\big(\mathbf{H}_\epsilon^{(1)}(0)+\mathbf{I}\big)+\widetilde{\mathbf{\Phi}}_\epsilon^{(1)}(0)\big(\mathbf{P}-\mathbf{I}\big)\mathbf{\Lambda}=\mathbf{u}_\epsilon.$$

Post-multiplying by $\mathbf{\Lambda^{-1}e}$ gives

$$\widetilde{\mathbf{\Phi}}_\epsilon(0)\big(\mathbf{H}_\epsilon^{(1)}(0)+\mathbf{I}\big)\mathbf{\Lambda^{-1}e}+\widetilde{\mathbf{\Phi}}_\epsilon^{(1)}(0)\big(\mathbf{P}-\mathbf{I}\big)\mathbf{\Lambda\Lambda^{-1}e}=\mathbf{u}_\epsilon\mathbf{\Lambda^{-1}e}.$$

Finally, by using $(\mathbf{P}-\mathbf{I})\mathbf{e}=0$, $\mathbf{H}_\epsilon^{(1)}(0)=-\mathbf{M\Lambda}+\epsilon\big(\mu_p-\mu_h\big)\mathbf{Q}^{(2)}\circ\mathbf{P\Lambda}$, and $\widetilde{\mathbf{\Phi}}_\epsilon(0)=\boldsymbol{\pi}$ (where the latter follows from Eq. (4.22) with $s=0$ and the normalisation Eq. (4.23)), the above can be simplified to

$$\boldsymbol{\pi}\left(\mathbf{\Lambda^{-1}}-\mathbf{M}\right)\mathbf{e}+\epsilon(\mu_p-\mu_h)\boldsymbol{\pi}\mathbf{Q}^{(2)}\circ\mathbf{Pe}=\mathbf{u}_\epsilon\mathbf{\Lambda^{-1}e}.$$

The uniqueness of the solution follows from the general theory of Markov chains that under the condition of stability, there is a unique stationary distribution and thus also a unique solution $\widetilde{\mathbf{\Phi}}_\epsilon(s)$ to the Eqs. (4.22) and (4.23). This completes the proof. □

Remark 4.9. When the number of states is $N=2$, the column vector $\mathbf{k}_2$ of Theorem 4.8 is equal to

$$\mathbf{k}_2=\big(\mathbf{K}_{22}(s_2),-\mathbf{K}_{21}(s_2)\big)^T \qquad \text{or} \qquad \mathbf{k}_2=\big(-\mathbf{K}_{12}(s_2),\mathbf{K}_{11}(s_2)\big)^T,$$

depending on whether $m=1$ or $m=2$, respectively. The case $N=1$ is merely the M/G/1 queue, which was treated in Chapter 2 (due to the duality between the two models).

Running example (continued). In order to evaluate vector $\mathbf{u}_\epsilon$, we first need to calculate the perturbed root $s_{\epsilon,2}$, and more precisely the term δ_2. Observe that in our case only element $\mathbf{K}_{21}(s)=s\lambda\big(\mu_p\widetilde{F}_p^e(s)-\mu_h\widetilde{F}_h^e(s)\big)$ of matrix $\mathbf{K}(s)$ is not equal to zero. Then, the numerator of δ_2 becomes

$$\det\big(\mathbf{E}(s_2)_{\mathcal{N}}^{\{1\}},\mathbf{K}(s_2)_{\mathcal{N}}^{\{2\}}\big)+\det\big(\mathbf{K}(s_2)_{\mathcal{N}}^{\{1\}},\mathbf{E}(s_2)_{\mathcal{N}}^{\{2\}}\big)=-s_2\lambda^2\big(\mu_p\widetilde{F}_p^e(s_2)-\mu_h\widetilde{F}_h^e(s_2)\big),$$

and its denominator takes the form

$$\det\big(\mathbf{E}(s_2)_{\mathcal{N}}^{\{1\}}, \mathbf{E}^{(1)}(s_2)_{\mathcal{N}}^{\{2\}}\big) + \det\big(\mathbf{E}^{(1)}(s_2)_{\mathcal{N}}^{\{1\}}, \mathbf{E}(s_2)_{\mathcal{N}}^{\{2\}}\big) = 2(s_2-\lambda) - \lambda^2 \widetilde{F}_p^{(1)}(s_2),$$

because the first derivative of matrix $\mathbf{E}(s)$ is

$$\mathbf{E}^{(1)}(s) = \begin{pmatrix} 1 & 0 \\ \lambda \widetilde{F}_p^{(1)}(s) & 1 \end{pmatrix}.$$

Combining the above, we have

$$\delta_2 = \frac{-s_2\lambda^2\big(\mu_p \widetilde{F}_p^e(s_2) - \mu_h \widetilde{F}_h^e(s_2)\big)}{2(s_2-\lambda) - \lambda^2 \widetilde{F}_p^{(1)}(s_2)}.$$

Recall that for the determination of the vector $\mathbf{a}_2$ we had used the second column of the adjoint matrix, namely we had chosen $m = 2$. Thus, according to Remark 4.9 vector $\mathbf{k}_2$ is a zero column vector of dimension 2. Since ${\mathbf{a}_2}^{(1)}$ is the second column of matrix $\mathbf{E}^{(1)}(s)$, it holds that $\mathbf{B}_{22} = \delta_2$ and all other elements of $\mathbf{B}$ are equal to zero. Finally, $\mathbf{d} = \big(\frac{1}{2}(\mu_p - \mu_h), 0\big)$. ■

By matching the coefficients of ϵ on the left and right side of Eq. (4.25), we can write the vector of unknown parameters $\mathbf{u}_\epsilon$ as $\mathbf{u}_\epsilon = \mathbf{u} + \epsilon\mathbf{z} + O(\epsilon^2\mathbf{e})$. The exact form of vector $\mathbf{z}$ is given in the following lemma, which we give without proof.

Lemma 4.10. *Vector* $\mathbf{u}_\epsilon$ *can be written in the form*

$$\mathbf{u}_\epsilon = \mathbf{u} + \epsilon\mathbf{z} + O(\epsilon^2\mathbf{e}),$$

where

$$\mathbf{z} = \big(\mathbf{c}\mathbf{A}^{-1}\mathbf{B} + \mathbf{d}\big)\,\mathbf{A}^{-1}.$$

Running example (continued). For the evaluation of $\mathbf{z}$ we need to find the inverse of matrix $\mathbf{A}$, namely we need

$$\mathbf{A}^{-1} = \frac{\lambda}{s_2}\begin{pmatrix} s_2-\lambda & \lambda \\ -\frac{1}{\lambda} & \frac{1}{\lambda} \end{pmatrix}.$$

By observing that $\mathbf{c}\mathbf{A}^{-1} = \mathbf{u}$ and following the calculations of Lemma 4.10 we obtain

$$\mathbf{z} = \frac{\lambda}{s_2}\left[\frac{1}{2}(\mu_p-\mu_h)(s_2-\lambda) - \frac{1}{s_2}\Big(1-\frac{\lambda\mu_p}{2}\Big)\delta_2, \frac{\lambda}{2}(\mu_p-\mu_h) + \frac{1}{s_2}\Big(1-\frac{\lambda\mu_p}{2}\Big)\delta_2\right].$$

■

In our analysis, we used first order perturbation with respect to the parameter ϵ. The exact same procedure can be followed if higher order terms of ϵ are desired. However, this would result to the increase of the complexity of the formulas. In the next section, we provide the formulas for the evaluation of the perturbed transform vector $\widetilde{\mathbf{\Phi}}_\epsilon(s)$ and the Laplace transform $\widetilde{w}_\epsilon(s)$ of the queueing delay.

4.3.3 Delay distribution of the perturbed model

If $\mathcal{E}_\epsilon(s)$ is the adjoint matrix of $\mathbf{E}_\epsilon(s)$ (see Eq. (4.24)), then the ith element of the transform vector $\widetilde{\mathbf{\Phi}}_\epsilon(s)$ is defined as

$$\widetilde{\phi}_{\epsilon,i}(s) = \frac{s\mathbf{u}_\epsilon \mathcal{E}_\epsilon(s)\mathbf{e}_i}{\det \mathbf{E}_\epsilon(s)}.$$

Therefore, to find the exact formula of $\widetilde{\phi}_{\epsilon,i}(s)$ we need to find $\det \mathbf{E}_\epsilon(s)$ and $s\mathbf{u}_\epsilon\mathcal{E}_\epsilon(s)\mathbf{e}_i$. By using the binomial identity and omitting higher order powers of ϵ, we have that $\Big(\widetilde{F}_p(s)+\epsilon s\big(\mu_p\widetilde{F}^e_p(s)-\mu_h\widetilde{F}^e_h(s)\big)\Big)^k = \big(\widetilde{F}_p(s)\big)^k + \epsilon k\big(\widetilde{F}_p(s)\big)^{k-1} s\big(\mu_p\widetilde{F}^e_p(s)-\mu_h\widetilde{F}^e_h(s)\big) + O(\epsilon^2)$. We give the following lemmas without proof. The first one gives the formula for the evaluation of the denominator of the desired quantity.

Lemma 4.11. *If* $\det \mathbf{E}(s)$ *is evaluated according to Theorem 4.3 with* $\widetilde{G}(s)=\widetilde{F}_p(s)$, *then* $\det \mathbf{E}_\epsilon(s)$ *can be written as perturbation of* $\det \mathbf{E}(s)$ *as follows*

$$\begin{aligned}\det \mathbf{E}_\epsilon(s) =& \det \mathbf{E}(s) + \epsilon s\big(\mu_p\widetilde{F}^e_p(s)-\mu_h\widetilde{F}^e_h(s)\big)\sum_{k=1}^{N} k\big(\widetilde{F}_p(s)\big)^{k-1}\sum_{\substack{\Gamma\subset\mathcal{N}\\|\Gamma|=k}}\sum_{\substack{S\subset\mathcal{N}\\S\supset\Gamma}}\boldsymbol{\lambda}^S\zeta^{S^c}(s)\\ &\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{S\setminus\Gamma}_{S}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{\Gamma}_{S}\Big)+O(\epsilon^2).\end{aligned}$$

Running example (continued). To find the perturbed determinant $\det \mathbf{E}_\epsilon(s)$, observe that only the combination $k=1$ with $\Gamma=\{1\}$ and $S=\mathcal{N}$ gives a non-zero coefficient for ϵ. Therefore,

$$\begin{aligned}\det \mathbf{E}_\epsilon(s) =& \det \mathbf{E}(s) + \epsilon s\big(\mu_p\widetilde{F}^e_p(s)-\mu_h\widetilde{F}^e_h(s)\big)\boldsymbol{\lambda}^{\mathcal{N}}\zeta^{\emptyset}(s)\\ &\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{\{2\}}_{\mathcal{N}}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{\{1\}}_{\mathcal{N}}\Big)\\ =&\det \mathbf{E}(s)-\epsilon\lambda^2 s\big(\mu_p\widetilde{F}^e_p(s)-\mu_h\widetilde{F}^e_h(s)\big).\end{aligned}$$

■

The next lemma gives the numerator of each $\widetilde{\phi}_{\epsilon,i}(s)$, $i\in\mathcal{N}$.

Lemma 4.12. *If* $s\mathbf{u}\mathcal{E}(s)\mathbf{e}_i$ *is evaluated according to Theorem 4.5 with* $\widetilde{G}(s)=\widetilde{F}_p(s)$, *then* $s\mathbf{u}_\epsilon\mathcal{E}_\epsilon(s)\mathbf{e}_i$ *can be written as perturbation of* $s\mathbf{u}\mathcal{E}(s)\mathbf{e}_i$ *as follows*

$$\begin{aligned}s\mathbf{u}_\epsilon\mathcal{E}_\epsilon(s)\mathbf{e}_i =& s\mathbf{u}\mathcal{E}(s)\mathbf{e}_i\\ &+\epsilon s\Bigg[z_i\sum_{k=1}^{N-1}\big(\widetilde{F}_p(s)\big)^k\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i\}\\|\Gamma|=k;\\S\subset\mathcal{N}\setminus\{i\}\\S\supset\Gamma}}\boldsymbol{\lambda}^S\zeta^{S^c}(s)\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{S\setminus\Gamma}_{S}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{\Gamma}_{S}\Big)\\ &+z_i\sum_{S\subset\mathcal{N}\setminus\{i\}}\boldsymbol{\lambda}^S\zeta^{S^c}(s)\det\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{S}_{S}\end{aligned}$$

$$
\begin{aligned}
&+\sum_{\substack{l=1\\l\neq i}}^{N} z_l(-1)^{l+i}\sum_{k=1}^{N-1}\big(\widetilde{F}_p(s)\big)^k \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\|\Gamma|=k-1;\\S\subset\mathcal{N}\setminus\{l,i\}\\S\supset\Gamma;\\R\subset S\cap T_{li}}} (-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s)\\
&\quad\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{S\setminus\Gamma}_{S\cup\{l\}}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{\Gamma\cup\{i\}}_{S\cup\{l\}}\Big)\\
&+\sum_{\substack{l=1\\l\neq i}}^{N} z_l(-1)^{l+i}\sum_{k=0}^{N-2}\big(\widetilde{F}_p(s)\big)^k \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\|\Gamma|=k}}\sum_{\substack{S\subset\mathcal{N}\setminus\{l,i\}\\S\supset\Gamma;\\R\subset S\cap T_{li}}} (-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s)\\
&\quad\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{(S\setminus\Gamma)\cup\{i\}}_{S\cup\{l\}}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{\Gamma}_{S\cup\{l\}}\Big)\\
&+s\big(\mu_p\widetilde{F}^e_p(s)-\mu_h\widetilde{F}^e_h(s)\big)\Bigg(u_i\sum_{k=1}^{N-1}k\big(\widetilde{F}_p(s)\big)^{k-1}\\
&\quad\times\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i\}\\|\Gamma|=k}}\sum_{\substack{S\subset\mathcal{N}\setminus\{i\}\\S\supset\Gamma}}\boldsymbol{\lambda}^S\zeta^{S^c}(s)\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{S\setminus\Gamma}_{S}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{\Gamma}_{S}\Big)\\
&+\sum_{\substack{l=1\\l\neq i}}^{N} u_l(-1)^{l+i}\sum_{k=1}^{N-1}k\big(\widetilde{F}_p(s)\big)^{k-1}\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\|\Gamma|=k-1}}\sum_{\substack{S\subset\mathcal{N}\setminus\{l,i\}\\S\supset\Gamma;\\R\subset S\cap T_{li}}} (-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s)\\
&\quad\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{S\setminus\Gamma}_{S\cup\{l\}}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{\Gamma\cup\{i\}}_{S\cup\{l\}}\Big)\\
&+\sum_{\substack{l=1\\l\neq i}}^{N} u_l(-1)^{l+i}\sum_{k=1}^{N-2}k\big(\widetilde{F}_p(s)\big)^{k-1}\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\|\Gamma|=k}}\sum_{\substack{S\subset\mathcal{N}\setminus\{l,i\}\\S\supset\Gamma;\\R\subset S\cap T_{li}}} (-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s)\\
&\quad\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{(S\setminus\Gamma)\cup\{i\}}_{S\cup\{l\}}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{\Gamma}_{S\cup\{l\}}\Big)\Bigg)\Bigg]+O(\epsilon^2),
\end{aligned}
$$

where z_i, $i\in\mathcal{N}$, are the coordinates of the vector $\mathbf{z}$ given in Lemma 4.10.

Running example (continued). By doing the calculations for each state without taking into account terms that are equal to zero, we obtain:

$$
\begin{aligned}
s\mathbf{u}_\epsilon\mathcal{E}_\epsilon(s)\mathbf{e}_1 =&s\mathbf{u}\mathcal{E}(s)\mathbf{e}_1+\epsilon s\Bigg[z_1\boldsymbol{\lambda}^{\emptyset}\zeta^{\{2\}}(s)\det\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{\emptyset}_{\emptyset}\\
&+s\big(\mu_p\widetilde{F}^e_p(s)-\mu_h\widetilde{F}^e_h(s)\big)\Big(u_2(-1)^{2+1}(-1)^{|\emptyset|}\boldsymbol{\lambda}^{\emptyset\cup\{1\}}\zeta^{\emptyset}(s)q^{(2)}_{21}p_{21}\Big)\\
&+z_2(-1)^{2+1}\widetilde{F}_p(s)(-1)^{|\emptyset|}\boldsymbol{\lambda}^{\emptyset\cup\{1\}}\zeta^{\emptyset}(s)q^{(2)}_{21}p_{21}\Bigg]+O(\epsilon^2)\\
=&s\mathbf{u}\mathcal{E}(s)\mathbf{e}_1+\epsilon s\Big(z_1(s-\lambda)-z_2\lambda\widetilde{F}_p(s)+s\big(\mu_p\widetilde{F}^e_p(s)-\mu_h\widetilde{F}^e_h(s)\big)(-\lambda u_2)\Big)\\
&+O(\epsilon^2),
\end{aligned}
$$

and

$$\begin{aligned} s\mathbf{u}_\epsilon\mathcal{E}_\epsilon(s)\mathbf{e}_2 =& s\mathbf{u}\mathcal{E}(s)\mathbf{e}_2 + \epsilon s\Bigg[z_2\boldsymbol{\lambda}^{\emptyset}\zeta^{\{1\}}(s)\det\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{\emptyset}_{\emptyset} \\ &+ z_1(-1)^{1+2}(-1)^{|\emptyset|}\boldsymbol{\lambda}^{\emptyset\cup\{2\}}\zeta^{\emptyset}(s)q^{(1)}_{12}p_{12}\Bigg] + O(\epsilon^2) \\ =& s\mathbf{u}\mathcal{E}(s)\mathbf{e}_2 + \epsilon s\big(-z_1\lambda + z_2(s-\lambda)\big) + O(\epsilon^2). \end{aligned}$$

■

By combining the results of Lemmas 4.11–4.12, we have the following proposition for the Laplace transform $\widetilde{w}_\epsilon(s)$ of the queueing delay.

Proposition 4.13. *If $\widetilde{w}(s)$ is calculated according to Proposition 4.7 for the replace base model, then there exist unique coefficients β, γ, α_k, β_k, γ_k, $k=2,\dots,N$, and $\alpha''_{j,l}$, $\beta''_{j,l}$, $\gamma''_{j,l}$, $j=1,\dots,\sigma$, $l=1,\dots,r_j$, such that the Laplace transform $\widetilde{w}_\epsilon(s)$ of the queueing delay of the mixture model satisfies*

$$\begin{aligned} \widetilde{w}_\epsilon(s) =& \widetilde{w}(s) + \epsilon\frac{1}{\mathbf{u}\boldsymbol{\omega}}\widetilde{w}(s)\Bigg[\Bigg(\mathbf{z}\boldsymbol{\omega} + \sum_{k=2}^{N}\frac{\alpha_k}{s-s_k} + \sum_{j=1}^{\sigma}\sum_{l=1}^{r_j}\frac{\alpha''_{j,l}\cdot(y_j)^{r_j-l+1}}{(s+y_j)^{r_j-l+1}}\Bigg) \\ &+ \big(\mu_p\widetilde{F}^e_p(s) - \mu_h\widetilde{F}^e_h(s)\big)\Bigg(\beta + \sum_{k=2}^{N}\frac{\beta_k}{s-s_k} + \sum_{j=1}^{\sigma}\sum_{l=1}^{r_j}\frac{\beta''_{j,l}\cdot(y_j)^{r_j-l+1}}{(s+y_j)^{r_j-l+1}}\Bigg) \\ &- \big(\mu_p\widetilde{F}^e_p(s) - \mu_h\widetilde{F}^e_h(s)\big)\widetilde{w}(s)\Bigg(\gamma + \sum_{k=2}^{N}\frac{\gamma_k}{s-s_k} + \sum_{j=1}^{\sigma}\sum_{l=1}^{r_j}\frac{\gamma''_{j,l}\cdot(y_j)^{r_j-l+1}}{(s+y_j)^{r_j-l+1}}\Bigg)\Bigg] \\ &+ O(\epsilon^2), \end{aligned}$$

where the vector $\mathbf{z}$ given in Lemma 4.10.

Proof. First, we show how we can factorise the numerator and the denominator of $\widetilde{w}_\epsilon(s)$ so that we recognise $\widetilde{w}(s)$ as a factor of $\widetilde{w}_\epsilon(s)$. The remaining factors of $\widetilde{w}_\epsilon(s)$ form a series expansion in ϵ, from which we keep only the ϵ-order terms. The latter terms involve the Laplace transforms of rational functions to which we apply simple fraction decomposition to complete the proof.

Recall now that r is the maximum power of $p(s)$ that appears in the formulas. Therefore, to use perturbation analysis, we multiply both $\det\mathbf{E}_\epsilon(s)$ and $s\mathbf{u}_\epsilon\mathcal{E}_\epsilon(s)\boldsymbol{\omega}$ with $\big(p(s)\big)^r$. So, if we set

$$\begin{aligned} \xi_{rM+N-1}(s) =& \sum_{k=1}^{N} k\big(q(s)\big)^{k-1}\big(p(s)\big)^{r-k+1}\sum_{\substack{\Gamma\subset\mathcal{N}\\|\Gamma|=k;\\S\subset\mathcal{N}\\S\supset\Gamma}}\boldsymbol{\lambda}^S\zeta^{S^c}(s) \\ &\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)^{S\setminus\Gamma}_{S}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)^{\Gamma}_{S}\Big), \end{aligned} \tag{4.26}$$

then,

$$\big(p(s)\big)^r\det\mathbf{E}_\epsilon(s) = \big(p(s)\big)^r\det\mathbf{E}(s) + \epsilon s\big(\mu_p\widetilde{F}^e_p(s) - \mu_h\widetilde{F}^e_h(s)\big)\xi_{rM+N-1}(s) + O(\epsilon^2).$$

Note that the polynomial $\xi_{rM+N-1}(s)$ is of degree at most $rM+N-1$, and the coefficient of s^{rM+N-1} is equal to $\gamma = \sum_{i=1}^{N} \lambda_i \det\big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)_{\{i\}}^{\{i\}} = \sum_{i=1}^{N} \lambda_i q_{ii}^{(2)} p_{ij}$. Theorem 4.8 guarantees that the function $\big(p(s)\big)^r \det \mathbf{E}_\epsilon(s)$ has exactly $N-1$ roots with positive real part and it also has $s_{\epsilon,1} = 0$. The roots with positive real part are of the form $s_{\epsilon,k} = s_k - \epsilon\delta_k + O(\epsilon^2)$, $k = 2, \dots, N$, where

$$\begin{aligned}\delta_k &= \frac{\big(\mu_p \widetilde{F}_p^e(s_k) - \mu_h \widetilde{F}_h^e(s_k)\big)\xi_{rM+N-1}(s_k)}{\prod_{\substack{l=2\\ l\neq k}}^{N}(s_k - s_l)\prod_{j=1}^{rM}(s_k + x_j)} \\ &= \frac{\big(\mu_p \widetilde{F}_p^e(s_k) - \mu_h \widetilde{F}_h^e(s_k)\big)\xi_{rM+N-1}(s_k)\widetilde{w}(s_k)}{\mathbf{u}\boldsymbol{\omega}\prod_{\substack{l=2\\ l\neq k}}^{N}(s_k - s_l)\prod_{j=1}^{rM}(s_k + y_j)}.\end{aligned} \tag{4.27}$$

Thus, if we set

$$d(s) = \frac{\big(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s)\big)\xi_{rM+N-1}(s)\widetilde{w}(s)}{\mathbf{u}\boldsymbol{\omega}\prod_{k=2}^{N}(s - s_k)\prod_{j=1}^{rM}(s + y_j)} - \sum_{k=2}^{N}\frac{\delta_k}{s - s_k}, \tag{4.28}$$

the denominator of $\widetilde{w}_\epsilon(s)$ multiplied by $\big(p(s)\big)^r$ can be written as

$$\big(p(s)\big)^r \det \mathbf{E}_\epsilon(s) = s\prod_{j=1}^{rM}(s + x_j)\prod_{k=2}^{N}(s - s_k + \epsilon\delta_k + O(\epsilon^2))\big(1 + \epsilon d(s) + O(\epsilon^2)\big). \tag{4.29}$$

Note that the function $d(s)$ is well defined in the positive half plane due to the definition (4.27) of δ_k, $k = 2, \dots, N$. Similarly, if we set

$$\begin{aligned}\xi_{i,l,rM+N-2}(s) =& \mathbb{1}_{\{l=i\}}\sum_{k=1}^{N-1} k\big(q(s)\big)^{k-1}\big(p(s)\big)^{r-k+1}\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i\}\\ |\Gamma|=k;\\ S\subset\mathcal{N}\setminus\{i\}\\ S\supset\Gamma}}\boldsymbol{\lambda}^S\zeta^{S^c}(s) \\ &\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^{S\setminus\Gamma}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_S^{\Gamma}\Big) \\ &+\mathbb{1}_{\{l\neq i\}}\Bigg[(-1)^{l+i}\sum_{k=1}^{N-1} k\big(q(s)\big)^{k-1}\big(p(s)\big)^{r-k+1}\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\ |\Gamma|=k-1}}\sum_{\substack{S\subset\mathcal{N}\setminus\{l,i\}\\ S\supset\Gamma;\\ R\subset S\cap T_{li}}}(-1)^{|R|} \\ &\times\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s)\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{S\setminus\Gamma}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{\Gamma\cup\{i\}}\Big) \\ &+(-1)^{l+i}\sum_{k=1}^{N-2} k\big(q(s)\big)^{k-1}\big(p(s)\big)^{r-k+1}\sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\ |\Gamma|=k}}\sum_{\substack{S\subset\mathcal{N}\setminus\{l,i\}\\ S\supset\Gamma;\\ R\subset S\cap T_{li}}}(-1)^{|R|} \\ &\times\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s)\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{(S\setminus\Gamma)\cup\{i\}}\bowtie\big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{\Gamma}\Big)\Bigg],\end{aligned} \tag{4.30}$$

and

$$\begin{aligned}\xi'_{i,l,rM+N-1}(s) =& \mathbb{1}_{\{l=i\}}\Bigg[\big(p(s)\big)^r \sum_{S\subset\mathcal{N}\setminus\{i\}} \boldsymbol{\lambda}^S \zeta^{S^c}(s) \det\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^S \\ &+ \sum_{k=1}^{N-1} \big(q(s)\big)^k \big(p(s)\big)^{r-k} \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i\}\\|\Gamma|=k}} \sum_{\substack{S\subset\mathcal{N}\setminus\{i\}\\S\supset\Gamma}} \boldsymbol{\lambda}^S \zeta^{S^c}(s) \\ &\times \det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^{S\setminus\Gamma} \bowtie \big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_S^{\Gamma}\Big)\Bigg] \\ &+\mathbb{1}_{\{l\neq i\}}\Bigg[(-1)^{l+i}\sum_{k=1}^{N-1} \big(q(s)\big)^k \big(p(s)\big)^{r-k} \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\|\Gamma|=k-1}} \sum_{\substack{S\subset\mathcal{N}\setminus\{l,i\}\\S\supset\Gamma;\\R\subset S\cap T_{li}}} (-1)^{|R|} \\ &\times \boldsymbol{\lambda}^{S\cup\{i\}} \zeta^{S^c}(s) \det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{S\setminus\Gamma} \bowtie \big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{\Gamma\cup\{i\}}\Big) \\ &+(-1)^{l+i}\sum_{k=0}^{N-2} \big(q(s)\big)^k \big(p(s)\big)^{r-k} \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\|\Gamma|=k}} \sum_{\substack{S\subset\mathcal{N}\setminus\{l,i\}\\S\supset\Gamma;\\R\subset S\cap T_{li}}} (-1)^{|R|} \boldsymbol{\lambda}^{S\cup\{i\}} \zeta^{S^c}(s) \\ &\times \det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{(S\setminus\Gamma)\cup\{i\}} \bowtie \big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{\Gamma}\Big)\Bigg], \end{aligned} \tag{4.31}$$

then

$$\begin{aligned}\big(p(s)\big)^r s\mathbf{u}_\epsilon\mathcal{E}_\epsilon(s)\boldsymbol{\omega} =& \big(p(s)\big)^r s\mathbf{u}\mathcal{E}(s)\boldsymbol{\omega} + \epsilon s\Bigg[\sum_{i=1}^N \omega_i \sum_{l=1}^N z_l \xi'_{i,l,rM+N-1}(s) \\ &+ s\big(\mu_p\widetilde{F}_p^e(s) - \mu_h\widetilde{F}_h^e(s)\big)\sum_{l=1}^N \omega_i \sum_{l=1}^N u_l \xi_{i,l,rM+N-2}(s)\Bigg] + O(\epsilon^2).\end{aligned}$$

Note that the polynomial $\sum_{l=1}^N \omega_i \sum_{l=1}^N z_l \xi'_{i,l,rM+N-1}(s)$ is of degree $rM+N-1$, and the coefficient of s^{rM+N-1} is $\mathbf{z}\boldsymbol{\omega}$. Analogously, the polynomial $s\sum_{l=1}^N \omega_i \sum_{l=1}^N u_l \times \xi_{i,l,rM+N-2}(s)$ is of degree at most $rM+N-1$, and the coefficient of s^{rM+N-1} is equal to $\beta = \sum_{l=1}^N \omega_i \sum_{l=1}^N u_l \times \Big(\mathbb{1}_{\{l=i\}} \sum_{\substack{j=1\\j\neq i}}^N \lambda_j q_{jj}^{(2)} p_{jj} + \mathbb{1}_{\{l\neq i\}}(-1)^{l+i}\lambda_i q_{li}^{(2)} p_{li}\Big)$. The first part is for $S=\Gamma=\{j\}$, and the second part for $S=\Gamma=\emptyset$. Theorem 4.8 guarantees that the roots $s_{\epsilon,k}$, $k\in\mathcal{N}$, are also roots of the numerator of $\widetilde{w}_\epsilon(s)$. Therefore, applying perturbation analysis to $\big(p(s)\big)^r s\mathbf{u}_\epsilon\mathcal{E}_\epsilon(s)\boldsymbol{\omega}$ results in an equivalent definition for each δ_k, $k=2,\ldots,N$, as

$$\delta_k = \frac{\big(\mu_p\widetilde{F}_p^e(s_k) - \mu_h\widetilde{F}_h^e(s_k)\big)s_k\sum_{i,l=1}^N \omega_i u_l \xi_{i,l,rM+N-2}(s_k)}{\mathbf{u}\boldsymbol{\omega}\prod_{\substack{l=2\\l\neq k}}^N (s_k - s_l)\prod_{j=1}^{rM}(s_k+y_j)}$$

$$+\frac{\sum_{i=1}^{N}\omega_i\sum_{l=1}^{N}z_l\xi'_{i,l,rM+N-1}(s_k)}{\mathbf{u\omega}\prod_{\substack{l=2\\l\neq k}}^{N}(s_k-s_l)\prod_{j=1}^{rM}(s_k+y_j)}. \tag{4.32}$$

Now, if we set

$$\begin{aligned} n(s)=&\frac{\big(\mu_p\widetilde{F}_p^e(s)-\mu_h\widetilde{F}_h^e(s)\big)s\sum_{i,l=1}^{N}\omega_i u_l\xi_{i,l,rM+N-2}(s)}{\mathbf{u\omega}\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+y_j)}\\ &+\frac{\sum_{i=1}^{N}\omega_i\sum_{l=1}^{N}z_l\xi'_{i,l,rM+N-1}(s)}{\mathbf{u\omega}\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+y_j)}-\sum_{k=2}^{N}\frac{\delta_k}{s-s_k}, \end{aligned} \tag{4.33}$$

the numerator of $\widetilde{w}_\epsilon(s)$ multiplied by $\big(p(s)\big)^r$ can be written as

$$\big(p(s)\big)^r s\mathbf{u}_\epsilon\mathcal{E}_\epsilon(s)\boldsymbol{\omega}=\mathbf{u\omega}s\prod_{j=1}^{rM}(s+y_j)\prod_{k=2}^{N}\big(s-s_k+\epsilon\delta_k+O(\epsilon^2)\big)\times\big(1+\epsilon n(s)+O(\epsilon^2)\big). \tag{4.34}$$

Note that the function $n(s)$ is well defined in the positive half plane due to the definition (4.32) of δ_k, $k=2,\dots,N$. Combining (4.29) and (4.34), we obtain

$$\begin{aligned} \widetilde{w}_\epsilon(s)=&\frac{\mathbf{u\omega}\prod_{j=1}^{rM}(s+y_j)}{\prod_{j=1}^{rM}(s+x_j)}\cdot\frac{1+\epsilon n(s)+O(\epsilon^2)}{1+\epsilon d(s)+O(\epsilon^2)}\\ =&\widetilde{w}(s)\big(1+\epsilon n(s)+O(\epsilon^2)\big)\big(1-\epsilon d(s)+O(\epsilon^2)\big)\\ =&\widetilde{w}(s)+\epsilon\widetilde{w}(s)\big(n(s)-d(s)\big)+O(\epsilon^2)\\ =&\widetilde{w}(s)+\epsilon\frac{1}{\mathbf{u\omega}}\widetilde{w}(s)\Bigg(\frac{\sum_{i=1}^{N}\omega_i\sum_{l=1}^{N}z_l\xi'_{i,l,rM+N-1}(s)}{\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+y_j)}\\ &+\big(\mu_p\widetilde{F}_p^e(s)-\mu_h\widetilde{F}_h^e(s)\big)\frac{s\sum_{i=1}^{N}\omega_i\sum_{l=1}^{N}u_l\xi_{i,l,rM+N-2}(s)}{\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+y_j)}\\ &-\big(\mu_p\widetilde{F}_p^e(s)-\mu_h\widetilde{F}_h^e(s)\big)\widetilde{w}(s)\frac{\xi_{rM+N-1}(s)}{\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+y_j)}\Bigg)+O(\epsilon^2)\\ =&\widetilde{w}(s)+\epsilon\frac{1}{\mathbf{u\omega}}\widetilde{w}(s)\Bigg[\Bigg(\mathbf{z\omega}+\sum_{k=2}^{N}\frac{\alpha_k}{s-s_k}+\sum_{j=1}^{rM}\frac{\alpha'_j\cdot y_j}{s+y_j}\Bigg)\\ &+\big(\mu_p\widetilde{F}_p^e(s)-\mu_h\widetilde{F}_h^e(s)\big)\Bigg(\beta+\sum_{k=2}^{N}\frac{\beta_k}{s-s_k}+\sum_{j=1}^{rM}\frac{\beta'_j\cdot y_j}{s+y_j}\Bigg)\\ &-\big(\mu_p\widetilde{F}_p^e(s)-\mu_h\widetilde{F}_h^e(s)\big)\widetilde{w}(s)\Bigg(\gamma+\sum_{k=2}^{N}\frac{\gamma_k}{s-s_k}+\sum_{j=1}^{rM}\frac{\gamma'_j\cdot y_j}{s+y_j}\Bigg)\Bigg]+O(\epsilon^2), \end{aligned}$$

where the last equality comes from simple fraction decomposition under the assumption that the roots $-y_j$, $j=1,\dots,rM$, are simple. The coefficients $\alpha_k,\beta_k,\gamma_k$, $k=2,\dots,N$,

and $\alpha'_j, \beta'_j, \gamma'_j$, $j = 1, \ldots, rM$, are as follows

$$\alpha_k = \frac{\sum_{i=1}^N \omega_i \sum_{l=1}^N z_l \xi'_{i,l,rM+N-1}(s_k)}{\prod_{\substack{l=2 \\ l\neq k}}^N (s_k - s_l) \prod_{j=1}^{rM} (s_k + y_j)}, \tag{4.35}$$

$$\beta_k = \frac{s_k \sum_{i=1}^N \omega_i \sum_{l=1}^N u_l \xi_{i,l,rM+N-2}(s_k)}{\prod_{\substack{l=2 \\ l\neq k}}^N (s_k - s_l) \prod_{j=1}^{rM} (s_k + y_j)}, \tag{4.36}$$

$$\gamma_k = \frac{\xi_{rM+N-1}(s_k)}{\prod_{\substack{l=2 \\ l\neq k}}^N (s_k - s_l) \prod_{j=1}^{rM} (s_k + y_j)}, \tag{4.37}$$

$$\alpha'_j = \frac{\sum_{i=1}^N \omega_i \sum_{l=1}^N z_l \xi'_{i,l,rM+N-1}(-y_j)}{y_{i,j} \prod_{k=2}^N (-y_j - s_k) \prod_{\substack{l=1 \\ l\neq j}}^{rM} (-y_j + y_l)},$$

$$\beta'_j = \frac{-\sum_{i=1}^N \omega_i \sum_{l=1}^N u_l \xi_{i,l,rM+N-2}(-y_j)}{\prod_{k=2}^N (-y_j - s_k) \prod_{\substack{l=1 \\ l\neq j}}^{rM} (-y_j + y_l)},$$

$$\gamma'_j = \frac{\xi_{rM+N-1}(-y_j)}{y_j \prod_{k=2}^N (-y_j - s_k) \prod_{\substack{l=1 \\ l\neq j}}^{rM} (-y_j + y_l)}.$$

The above results hold when all roots $-y_j$, $j = 1, \ldots, rM$, are simple. Suppose now that only σ of the roots are distinct and that the multiplicity of root $-y_j$, $j = 1, \ldots, \sigma$, is r_j, such that $\sum_{j=1}^{\sigma} r_j = rM$. In this case,

$$\begin{aligned}
\widetilde{w}_\epsilon(s) =& \widetilde{w}(s) + \epsilon \frac{1}{\mathbf{u\omega}} \widetilde{w}(s) \Bigg[\left(\mathbf{z\omega} + \sum_{k=2}^N \frac{\alpha_k}{s - s_k} + \sum_{j=1}^{\sigma} \sum_{l=1}^{r_j} \frac{\alpha''_{j,l} \cdot (y_j)^{r_j - l + 1}}{(s + y_j)^{r_j - l + 1}} \right) \\
&+ \left(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s) \right) \left(\beta + \sum_{k=2}^N \frac{\beta_k}{s - s_k} + \sum_{j=1}^{\sigma} \sum_{l=1}^{r_j} \frac{\beta''_{j,l} \cdot (y_j)^{r_j - l + 1}}{(s + y_{i,j})^{r_j - l + 1}} \right) \\
&- \left(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s) \right) \widetilde{w}(s) \left(\gamma + \sum_{k=2}^N \frac{\gamma_k}{s - s_k} + \sum_{j=1}^{\sigma} \sum_{l=1}^{r_j} \frac{\gamma''_{j,l} \cdot (y_j)^{r_j - l + 1}}{(s + y_j)^{r_j - l + 1}} \right) \Bigg] \\
&+ O(\epsilon^2),
\end{aligned} \tag{4.38}$$

where α_k, β_k, and γ_k, $k = 2, \ldots, N$, are defined through (4.35)–(4.37). For each $j = 1, \ldots, \sigma$, the coefficients $\alpha''_{j,p}$, $p = 1, \ldots, r_j$, are the unique solution to the following linear system of r_j equations

$$\begin{aligned}
&\frac{d}{ds^n} \left[\sum_{i=1}^N \omega_i \sum_{l=1}^N z_l \xi'_{i,l,rM+N-1}(s) \right] \Bigg|_{s=-y_j} \\
&\quad = \frac{d}{ds^n} \left[\prod_{k=2}^N (s - s_k) \prod_{\substack{l=1 \\ l\neq j}}^{\sigma} (s + y_l)^{r_l} \sum_{p=1}^{r_j} \alpha''_{j,p} \, (y_j)^{r_j - p + 1} (s + y_j)^{p-1} \right] \Bigg|_{s=-y_j},
\end{aligned}$$

for $n = 0, \ldots, r_j$. Similarly, for each $j = 1, \ldots, \sigma$, the coefficients $\beta''_{j,p}$ and $\gamma''_{j,p}$, $p = 1, \ldots, r_j$, are the respective unique solutions to the following two linear system of

r_j equations

$$\frac{d}{ds^n}\left[s\sum_{i=1}^{N}\omega_i\sum_{l=1}^{N}u_l\xi_{i,l,rM+N-2}(s)\right]\Bigg|_{s=-y_j}$$
$$=\frac{d}{ds^n}\left[\prod_{k=2}^{N}(s-s_k)\prod_{\substack{l=1\\l\neq j}}^{\sigma}(s+y_l)^{r_l}\sum_{p=1}^{r_j}\beta''_{j,p}(y_j)^{r_j-p+1}(s+y_j)^{p-1}\right]\Bigg|_{s=-y_j},$$
$$\frac{d}{ds^n}\left[\xi_{rM+N-1}(s)\right]\Bigg|_{s=-y_j}$$
$$=\frac{d}{ds^n}\left[\prod_{k=2}^{N}(s-s_k)\prod_{\substack{l=1\\l\neq j}}^{\sigma}(s+y_l)^{r_l}\sum_{p=1}^{r_j}\gamma''_{j,p}(y_j)^{r_j-p+1}(s+y_j)^{p-1}\right]\Bigg|_{s=-y_j},$$

for $n=0,\ldots,r_j$. □

Before we evaluate $\widetilde{w}_\epsilon(s)$ in our running example, we apply Laplace inversion to the coefficient of ϵ in the series expansion of $\widetilde{w}_\epsilon(s)$. Finally, let B^e and C^e be the generic stationary excess phase-type and heavy-tailed service times, respectively.

Theorem 4.14. *If $\widetilde{\theta}(s)$ is the coefficient of ϵ in the series expansion of $\widetilde{w}_\epsilon(s)$ in Proposition 4.13, its Laplace inversion $\Theta(t)=\mathcal{L}^{-1}\{\widetilde{\theta}(s)\}$ is equal to the expression $\Theta(t)=\frac{1}{\mathbf{u\omega}}\Big(\Theta_1(t)+\Theta_2(t)\Big)$, where $\Theta_1(t),\Theta_2(t)$ are given as follows*

$$\begin{aligned}
\Theta_1(t)=&\Big(\mathbf{z\omega}-\sum_{k=2}^{N}\frac{\alpha_k}{s_k}\Big)\mathbb{P}(W>t)\\
&+\Big(\beta-\sum_{k=2}^{N}\frac{\beta_k}{s_k}\Big)\Big(\mu_p\mathbb{P}(W+B^e>t)-\mu_h\mathbb{P}(W+C^e>t)\Big)\\
&-\Big(\gamma-\sum_{k=2}^{N}\frac{\gamma_k}{s_k}\Big)\Big(\mu_p\mathbb{P}(W+W'+B^e>t)-\mu_h\mathbb{P}(W+W'+C^e>t)\Big)\\
&-\sum_{j=1}^{\sigma}\sum_{l=1}^{r_j}\Bigg(\gamma''_{j,l}\Big(\mu_p\mathbb{P}\big(W+W'+B^e+E_{r_j-l+1}(y_j)>t\big)\\
&\quad-\mu_h\mathbb{P}\big(W+W'+C^e+E_{r_j-l+1}(y_j)>t\big)\Big)\\
&\quad-\beta''_{j,l}\Big(\mu_p\mathbb{P}\big(W+B^e+E_{r_j-l+1}(y_j)>t\big)-\mu_h\mathbb{P}\big(W+C^e+E_{r_j-l+1}(y_j)>t\big)\Big)\\
&\quad-\alpha''_{j,l}\mathbb{P}\big(W+E_{r_j-l+1}(y_j)>t\big)\Bigg),\\
\Theta_2(t)=&-\sum_{k=2}^{N}\frac{1}{s_k}\Bigg(\gamma_k\Big(\mu_p\mathbb{P}\big(t<W+W'+B^e<t+E(s_k)\big)\\
&\quad-\mu_h\mathbb{P}\big(t<W+W'+C^e<t+E(s_k)\big)\Big)
\end{aligned}$$

$$- \beta_k \Big(\mu_p \mathbb{P}\big(t < W + B^e < t + E(s_k)\big) - \mu_h \mathbb{P}\big(t < W + C^e < t + E(s_k)\big) \Big)$$
$$- \alpha_k \mathbb{P}\big(t < W < t + E(s_k)\big) \Bigg),$$

and $W^{'}$ is independent and follows the same distribution of W.

Proof. Here, we follow the notation we introduced in Proposition 4.13. We denote by $\widetilde{\theta}(s)$ the correction term (the coefficient of ϵ) in the expression of $\widetilde{w}_\epsilon(s)$. In order to apply Laplace inversion to $\widetilde{\theta}(s)$, we first reorder the involved terms (see Eq. (4.38)) as

$$\begin{aligned}
\widetilde{\theta}(s) =& \frac{1}{\mathbf{u}\boldsymbol{\omega}} \widetilde{w}(s) \Bigg[\Bigg(\mathbf{z}\boldsymbol{\omega} + \beta\big(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s)\big) - \gamma\big(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s)\big)\widetilde{w}(s) \Bigg) \\
&+ \sum_{k=2}^{N} \frac{1}{s - s_k} \Bigg(\alpha_k + \beta_k\big(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s)\big) - \gamma_k\big(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s)\big)\widetilde{w}(s) \Bigg) \\
&+ \sum_{j=1}^{\sigma} \sum_{l=1}^{r_j} \frac{1}{(s + y_j)^{r_j - l + 1}} \Bigg(\alpha''_{j,l} + \beta''_{j,l}\big(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s)\big) - \gamma''_{j,l}\big(\mu_p \widetilde{F}_p^e(s) \\
&- \mu_h \widetilde{F}_h^e(s)\big)\widetilde{w}(s) \Bigg) \Bigg]. \qquad (4.39)
\end{aligned}$$

From the above formula it is evident that only the terms in the middle bracket cannot be inverted directly as they are, because of the singularities they seem to have in the positive half plane; i.e. they contain the factors $(s - s_k)^{-1}$, $k = 2, \ldots, N$. Thus, we treat them separately in the next lines. From the two equivalent definitions (4.27) and (4.32) of the perturbation terms δ_k, $k = 2, \ldots, N$, and the relations (4.35)–(4.37) we obtain that

$$\alpha_k \widetilde{w}(s_k) + \beta_k\big(\mu_p \widetilde{F}_p^e(s_k) - \mu_h \widetilde{F}_h^e(s_k)\big)\widetilde{w}(s_k) - \gamma_k\big(\mu_p \widetilde{F}_p^e(s_k) - \mu_h \widetilde{F}_h^e(s_k)\big)\big(\widetilde{w}(s_k)\big)^2 = 0,$$

for $k = 2, \ldots, N$. With the aid of the above equations, we show that eventually all the problematic factors $(s - s_k)^{-1}$, $k = 2, \ldots, N$ cancel and the simplified terms we obtain from this cancelation are well defined on the positive half. The above equations are equivalent to

$$\begin{aligned}
0 =& \alpha_k \int_{x=0}^{\infty} e^{-s_k x} d\mathbb{P}(W \le x) + \beta_{i,k} \Big(\mu_p \int_{x=0}^{\infty} e^{-s_k x} d\mathbb{P}(W + B^e \le x) \\
&- \mu_h \int_{x=0}^{\infty} e^{-s_k x} d\mathbb{P}(W + C^e \le x) \Big) - \gamma_k \Big(\mu_p \int_{x=0}^{\infty} e^{-s_k x} d\mathbb{P}(W + W^{'} + B^e \le x) \\
&- \mu_h \int_{x=0}^{\infty} e^{-s_k x} d\mathbb{P}(W + W^{'} + C^e \le x) \Big), \qquad (4.40)
\end{aligned}$$

$k = 2, \ldots, N$. We first show that

$$\mathcal{L}^{-1} \Bigg(\sum_{k=2}^{N} \frac{1}{s - s_k} \Big(\alpha_k \widetilde{w}(s) + \beta_k\big(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s)\big)\widetilde{w}(s)$$

$$
\begin{aligned}
&- \gamma_k\big(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s)\big)\big(\widetilde{w}(s)\big)^2\Big)\Big) \\
=&\sum_{k=2}^{N}\Bigg[\gamma_k\Big(\mu_p\int_{y=x}^{\infty} e^{s_k(x-y)}d\mathbb{P}(W+W^{'}+B^e\le y) \\
&- \mu_h\int_{y=x}^{\infty} e^{s_k(x-y)}d\mathbb{P}(W+W^{'}+C^e\le y)\Big) - \alpha_k\int_{y=x}^{\infty} e^{s_k(x-y)}d\mathbb{P}(W\le y) \\
&- \beta_k\Big(\mu_p\int_{y=x}^{\infty} e^{s_k(x-y)}d\mathbb{P}(W+B^e\le y) - \mu_h\int_{y=x}^{\infty} e^{s_k(x-y)}d\mathbb{P}(W+C^e\le y)\Big)\Bigg].
\end{aligned}
\tag{4.41}
$$

Since Laplace transforms turn convolutions of functions into their product, using the property $\int_{y=0}^{\infty} f(y)dy = \int_{y=0}^{x} f(y)dy + \int_{y=x}^{\infty} f(y)dy$ and the Eqs. (4.40) we obtain

$$
\begin{aligned}
&\mathcal{L}\Bigg\{\sum_{k=2}^{N}\Bigg[\gamma_k\Bigg(\mu_p\int_{y=x}^{\infty} e^{s_k(x-y)}d\mathbb{P}(W+W^{'}+B^e\le y) \\
&- \mu_h\int_{y=x}^{\infty} e^{s_k(x-y)}d\mathbb{P}(W+W^{'}+C^e\le y)\Bigg) \\
&- \beta_k\Bigg(\mu_p\int_{y=x}^{\infty} e^{s_k(x-y)}d\mathbb{P}(W+B^e\le y) - \mu_h\int_{y=x}^{\infty} e^{s_k(x-y)}d\mathbb{P}(W+C^e\le y)\Bigg) \\
&- \alpha_k\int_{y=x}^{\infty} e^{s_k(x-y)}d\mathbb{P}(W\le y)\Bigg]\Bigg\} \\
=&\mathcal{L}\Bigg\{\sum_{k=2}^{N}\Bigg[-\gamma_k\Bigg(\mu_p\int_{y=0}^{x} e^{s_k(x-y)}d\mathbb{P}(W+W^{'}+B^e\le y) \\
&- \mu_h\int_{y=0}^{x} e^{s_k(x-y)}d\mathbb{P}(W+W^{'}+C^e\le y)\Bigg) \\
&+ \beta_k\Bigg(\mu_p\int_{y=0}^{x} e^{s_k(x-y)}d\mathbb{P}(W+B^e\le y) - \mu_h\int_{y=0}^{x} e^{s_k(x-y)}d\mathbb{P}(W+C^e\le y)\Bigg)\Bigg] \\
&+ \alpha_k\int_{y=0}^{x} e^{s_k(x-y)}d\mathbb{P}(W\le y)\Bigg\} \\
=&\sum_{k=2}^{N}\frac{1}{s-s_k}\Big(\alpha_k\widetilde{w}(s) + \beta_k\big(\mu_p\widetilde{F}_p^e(s) - \mu_h\widetilde{F}_h^e(s)\big)\widetilde{w}(s) \\
&- \gamma_k\big(\mu_p\widetilde{F}_p^e(s) - \mu_h\widetilde{F}_h^e(s)\big)\big(\widetilde{w}(s)\big)^2\Big)\Big),
\end{aligned}
$$

which proves (4.41).

To find the tail probabilities that correspond to the terms in the middle bracket of Eq. (4.39), we integrate the inverted Laplace transform in Eq. (4.41) from t to ∞,

and we obtain

$$
\begin{aligned}
&\sum_{k=2}^{N}\Bigg[\gamma_k\Bigg(\mu_p\int_{x=t}^{\infty}\int_{y=x}^{\infty}e^{s_k(x-y)}d\mathbb{P}(W+W^{'}+B^e\le y)dx\\
&-\mu_h\int_{x=t}^{\infty}\int_{y=x}^{\infty}e^{s_k(x-y)}d\mathbb{P}(W+W^{'}+C^e\le y)dx\Bigg)\\
&-\beta_k\Bigg(\mu_p\int_{x=t}^{\infty}\int_{y=x}^{\infty}e^{s_k(x-y)}d\mathbb{P}(W+B^e\le y)dx\\
&-\mu_h\int_{x=t}^{\infty}\int_{y=x}^{\infty}e^{s_k(x-y)}d\mathbb{P}(W+C^e\le y)dx\Bigg)-\alpha_k\int_{x=t}^{\infty}\int_{y=x}^{\infty}e^{s_k(x-y)}d\mathbb{P}(W\le y)dx\Bigg]\\
=&\sum_{k=2}^{N}\Bigg[\gamma_k\Bigg(\mu_p\int_{y=t}^{\infty}e^{-s_ky}d\mathbb{P}(W+W^{'}+B^e\le y)\int_{x=t}^{y}e^{s_kx}dx\\
&-\mu_h\int_{y=t}^{\infty}e^{-s_ky}d\mathbb{P}(W+W^{'}+C^e\le y)\int_{x=t}^{y}e^{s_kx}dx\Bigg)\\
&-\beta_k\Bigg(\mu_p\int_{y=t}^{\infty}e^{-s_ky}d\mathbb{P}(W+B^e\le y)\int_{x=t}^{y}e^{s_kx}dx\\
&-\mu_h\int_{y=t}^{\infty}e^{-s_ky}d\mathbb{P}(W+C^e\le y)\int_{x=t}^{y}e^{s_kx}dx\Bigg)\\
&-\alpha_k\int_{y=t}^{\infty}e^{-s_ky}d\mathbb{P}(W\le y)\int_{x=t}^{y}e^{s_kx}dx\Bigg]\\
=&\sum_{k=2}^{N}\Bigg[\frac{\gamma_k}{s_k}\Bigg(\mu_p\int_{y=t}^{\infty}d\mathbb{P}(W+W^{'}+B^e\le y)-\mu_h\int_{y=t}^{\infty}d\mathbb{P}(W+W^{'}+C^e\le y)\Bigg)\\
&-\frac{\beta_k}{s_k}\Bigg(\mu_p\int_{y=t}^{\infty}d\mathbb{P}(W+B^e\le y)-\mu_h\int_{y=t}^{\infty}d\mathbb{P}(W+C^e\le y)\Bigg)\\
&-\frac{\gamma_k}{s_k}\Bigg(\mu_p\int_{y=t}^{\infty}e^{-s_k(y-t)}d\mathbb{P}(W+W^{'}+B^e\le y)\\
&-\mu_h\int_{y=t}^{\infty}e^{-s_k(y-t)}d\mathbb{P}(W+W^{'}+C^e\le y)\Bigg)\\
&+\frac{\beta_k}{s_k}\Bigg(\mu_p\int_{y=t}^{\infty}e^{-s_k(y-t)}d\mathbb{P}(W+B^e\le y)-\mu_h\int_{y=t}^{\infty}e^{-s_k(y-t)}d\mathbb{P}(W+C^e\le y)\Bigg)\\
&+\frac{\alpha_k}{s_k}\int_{y=t}^{\infty}e^{-s_k(y-t)}d\mathbb{P}(W\le y)-\frac{\alpha_k}{s_k}\int_{y=t}^{\infty}d\mathbb{P}(W\le y)\Bigg]\\
=&\sum_{k=2}^{N}\frac{1}{s_k}\Bigg[-\gamma_k\Bigg(\mu_p\mathbb{P}\big(t<W+W^{'}+B^e<t+E(s_k)\big)
\end{aligned}
$$

$$
\begin{aligned}
&- \mu_h \mathbb{P}\big(t < W + W^{'} + C^e < t + E(s_k)\big)\Big) \\
&+ \beta_k \Big(\mu_p \mathbb{P}\big(t < W + B^e < t + E(s_k)\big) - \mu_h \mathbb{P}\big(t < W + C^e < t + E(s_k)\big)\Big) \\
&+ \alpha_k \mathbb{P}\big(t < W < t + E(s_k)\big)\Big] \\
&+ \sum_{k=2}^{N} \frac{1}{s_k}\Big[\gamma_k \Big(\mu_p \mathbb{P}(W + W^{'} + B^e > t) - \mu_h \mathbb{P}(W + W^{'} + C^e > t)\Big) \\
&- \beta_k \Big(\mu_p \mathbb{P}(W + B^e > t) - \mu_h \mathbb{P}(W + C^e > t)\Big) - \alpha_k \mathbb{P}(W > t)\Big].
\end{aligned}
$$

By using now the property $\mathcal{L}^{-1}\{a^{n+1}/(s+a)^{n+1}\} = \dfrac{1}{n!}a^{n+1}t^n \times e^{-at}$, $t \geq 0$, of the inverse Laplace transform, we see that the terms $\dfrac{(y_j)^{r_j-l+1}}{(s+y_j)^{r_j-l+1}}$ in Eq. (4.39) correspond to the Laplace transform of an $E_{r_j-l+1}(y_j)$ r.v. Combining all the above, the result in immediate, which completes the proof of the theorem. □

Remark 4.15. Note that an $E_k(\lambda)$ distribution ($k \geq 1$) is defined for a non-negative real valued rate λ. To state Theorem 4.14, we assumed that all the roots s_k, $k = 2, \ldots, N$, and $-y_j$, $j = 1, \ldots, rM$, are real-valued. In most systems, this assumption in not always true. Recall that the previously mentioned roots are roots of a polynomial with real coefficients (see also the analysis above Eq. (4.20)). Therefore, from the Complex Conjugate Root Theorem it holds that if e.g. s_2 is complex, then its complex conjugate $\overline{s_2}$ is also a root. Thus, we write $E_{\Re(s_2)}$ instead of E_{s_2} and $E_{\overline{s_2}}$, because every parameter or function that depends on $\overline{s_2}$ appears as a complex conjugate of the corresponding quantity that depends on s_2, and their imaginary parts cancel out. The same result holds for all other roots.

Running example (continued). For the evaluation of the Laplace transform $\widetilde{w}_\epsilon(s) = \widetilde{\mathbf{\Phi}}_\epsilon(s)\boldsymbol{\omega}$ of the queueing delay W_ϵ, we follow the steps in the proof of Proposition 4.13. Here we show how the expressions of the proof are simplified for our running example.

Recall that in our example, $r = 1$, and assume that only σ of the roots $-y_j$ are distinct and that the multiplicity of each of them is r_j, such that $\sum_{j=1}^{\sigma} r_j = M$. Therefore, we first find $p(s)\det \mathbf{E}_\epsilon(s)$ and $p(s)s\mathbf{u}_\epsilon\mathcal{E}_\epsilon(s)\boldsymbol{\omega}$. If we set $\xi(s) = -\lambda^2 p(s)$, $\xi_1'(s) = -2\lambda p(s)$, and $\xi_2'(s) = 2(s-\lambda)p(s)$, then we obtain

$$
\begin{aligned}
p(s)\det \mathbf{E}_\epsilon(s) &= p(s)\det \mathbf{E}(s) + \epsilon s\big(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s)\big)\xi(s) + O(\epsilon^2), \\
p(s)s\mathbf{u}_\epsilon\mathcal{E}_\epsilon(s)\boldsymbol{\omega} &= p(s)s\mathbf{u}\mathcal{E}(s)\boldsymbol{\omega} + \epsilon s \sum_{l=1}^{2} z_l \xi_l'(s) + O(\epsilon^2).
\end{aligned}
$$

We define the functions $d(s)$ and $n(s)$ (see Eqs. (4.28) and (4.33) respectively) as

$$d(s) = \frac{\left(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s)\right)\xi(s)\widetilde{w}(s)}{\mathbf{u}\boldsymbol{\omega}(s-s_2)\prod_{j=1}^{\sigma}(s+y_j)^{r_j}} - \frac{\delta_2}{s-s_2},$$

$$n(s) = \frac{\sum_{l=1}^{2} z_l \xi_l'(s)}{\mathbf{u}\boldsymbol{\omega}(s-s_2)\prod_{j=1}^{\sigma}(s+y_j)^{r_j}} - \frac{\delta_2}{s-s_2},$$

where the two equivalent definitions of δ_2 (see Eqs. (4.27) and (4.32)) take the form

$$\delta_2 = \frac{\left(\mu_p \widetilde{F}_p^e(s_2) - \mu_h \widetilde{F}_h^e(s_2)\right)\xi(s_2)\widetilde{w}(s_2)}{\mathbf{u}\boldsymbol{\omega}\prod_{j=1}^{\sigma}(s_2+y_j)^{r_j}} = \frac{\sum_{l=1}^{2} z_l \xi_l'(s_2)}{\mathbf{u}\boldsymbol{\omega}\prod_{j=1}^{\sigma}(s_2+y_j)^{r_j}}.$$

Following the calculations after Eq. (53) we get that

$$\begin{aligned}\widetilde{w}_\epsilon(s) = \widetilde{w}(s) + \epsilon\frac{1}{\mathbf{u}\boldsymbol{\omega}}\widetilde{w}(s)\Bigg(&\frac{\sum_{l=1}^{2} z_l \xi_l'(s)}{(s-s_2)\prod_{j=1}^{\sigma}(s+y_j)^{r_j}} \\ &- \left(\mu_p \widetilde{F}_p^e(s) - \mu_h \widetilde{F}_h^e(s)\right)\widetilde{w}(s)\frac{\xi(s)}{(s-s_2)\prod_{j=1}^{\sigma}(s+y_j)^{r_j}}\Bigg) \\ &+ O(\epsilon^2). \end{aligned} \tag{4.42}$$

Now, we apply simple fraction decomposition to the rational functions

$$\frac{\sum_{l=1}^{2} z_l \xi_l'(s)}{(s-s_2)\prod_{j=1}^{\sigma}(s+y_j)^{r_j}}, \qquad \frac{\xi(s)}{(s-s_2)\prod_{j=1}^{\sigma}(s+y_j)^{r_j}}.$$

Thus, we calculate

$$\alpha_2 = \frac{\sum_{l=1}^{2} z_l \xi_l'(s_2)}{\prod_{j=1}^{\sigma}(s_2+y_j)^{r_j}}, \qquad \gamma_2 = \frac{\xi(s_2)}{\prod_{j=1}^{\sigma}(s_2+y_j)^{r_j}},$$

and for $j = 1, \ldots, \sigma$, $p = 1, \ldots, r_j$, the coefficients $\alpha''_{j,p}$ and $\gamma''_{j,p}$, are respectively the unique solutions to the following two linear systems of r_j equations

$$\begin{aligned}\frac{d}{ds^n}\left[\sum_{l=1}^{2} z_l \xi_l'(s)\right]\Bigg|_{s=-y_j} &= \frac{d}{ds^n}\left[(s-s_2)\prod_{\substack{l=1\\ l\neq j}}^{\sigma}(s+y_l)^{r_l}\sum_{p=1}^{r_j}\alpha''_{j,p}(y_j)^{r_j-p+1}(s+y_j)^{p-1}\right]\Bigg|_{s=-y_j}, \\ \frac{d}{ds^n}\left[\xi(s)\right]\Bigg|_{s=-y_j} &= \frac{d}{ds^n}\left[(s-s_2)\prod_{\substack{l=1\\ l\neq j}}^{\sigma}(s+y_l)^{r_l}\sum_{p=1}^{r_j}\gamma''_{j,p}(y_j)^{r_j-p+1}(s+y_j)^{p-1}\right]\Bigg|_{s=-y_j},\end{aligned}$$

$n = 0, \ldots, r_j$. In addition, the polynomial $\xi(s)$ is of degree M and the polynomial $\sum_{l=1}^{2} z_l \xi_l'(s)$ is of degree $M+1$, with the coefficient of s^{M+1} equal to $2z_2$. Combining all these, we write Eq. (4.42) as

$$\begin{aligned}\widetilde{w}_\epsilon(s) =& \widetilde{w}(s) + \epsilon \frac{1}{2u_2}\widetilde{w}(s)\Bigg[\left(2z_2 + \frac{\alpha_2}{s-s_2} + \sum_{j=1}^{\sigma}\sum_{l=1}^{r_j}\frac{\alpha''_{j,l}\cdot(y_j)^{r_j-l+1}}{(s+y_j)^{r_j-l+1}}\right)\\ &-\left(\mu_p\widetilde{F}_p^e(s) - \mu_h\widetilde{F}_h^e(s)\right)\widetilde{w}(s)\left(\frac{\gamma_2}{s-s_2} + \sum_{j=1}^{\sigma}\sum_{l=1}^{r_j}\frac{\gamma''_{j,l}\cdot(y_j)^{r_j-l+1}}{(s+y_j)^{r_j-l+1}}\right)\Bigg] + O(\epsilon^2).\end{aligned}$$

Observe that in this case $\gamma = 0$ and all β coefficients are also equal to zero. Thus, if $\widetilde{\theta}(s)$ is the coefficient of ϵ in the series expansion of $\widetilde{w}_\epsilon(s)$, we apply Theorem 4.14 to find its Laplace inversion as

$$\begin{aligned}\Theta(t) =& \frac{1}{2u_2}\Bigg[\left(2z_2 - \frac{\alpha_2}{s_2}\right)\mathbb{P}(W > t)\\ &+\frac{\gamma_2}{s_2}\Big(\mu_p\mathbb{P}(W + W' + B^e > t) - \mu_h\mathbb{P}(W + W' + C^e > t)\Big)\\ &-\frac{1}{s_2}\Bigg(\gamma_2\Big(\mu_p\mathbb{P}\big(t < W + W' + B^e < t + E(s_2)\big)\\ &\quad - \mu_h\mathbb{P}\big(t < W + W' + C^e < t + E(s_2)\big)\Big) - \alpha_2\mathbb{P}\big(t < W < t + E(s_2)\big)\Bigg)\\ &-\sum_{j=1}^{\sigma}\sum_{l=1}^{r_j}\Bigg(\gamma''_{j,l}\Big(\mu_p\mathbb{P}\big(W + W' + B^e + E_{r_j-l+1}(y_j) > t\big)\\ &\quad - \mu_h\mathbb{P}\big(W + W' + C^e + E_{r_j-l+1}(y_j) > t\big)\Big) - \alpha''_{j,l}\mathbb{P}\big(W + E_{r_j-l+1}(y_j) > t\big)\Bigg)\Bigg],\end{aligned}$$

where W' is independent and follows the same distribution of W. ■

By applying Laplace inversion to the first two terms of the series expansion in ϵ of the queueing delay, we obtain that the first term is a phase-type approximation of the queueing delay that results from the replace base model (see Section 4.3.1). In addition, the second term, which we refer to as correction term and is found explicitly in Theorem 4.14, involves linear combinations of terms that have a probabilistic interpretation. More precisely, these terms are either tail probabilities of convoluted r.v. or probabilities for some of the aforementioned convoluted r.v. that lie between a fixed value t and the same value t shifted by an exponential time. Finally, observe that these convoluted r.v. involve the heavy-tailed stationary-excess service time r.v. C^e at most once. Combining the results of Proposition 4.13 and Theorem 4.14, in the next section we define our approximations.

4.3.4 Corrected replace approximations

The goal of this section is to provide approximations that maintain the numerical tractability of the phase-type approximations, but improve their accuracy, and that

are able to capture the tail behaviour of the exact delay distribution. As we pointed out in Section 3.3, a single appearance of a stationary excess heavy-tailed service time C^e is sufficient to capture the correct tail behaviour of the exact queueing delay. As we observed in Section 4.3.3, the correction term contains terms with single appearances of C^e. For this reason, the proposed approximation for the queueing delay is constructed by the first two terms of its respective series expansion for the queueing delay. We propose the following approximation:

Approximation 4.16. The corrected replace approximation of the survival function $\mathbb{P}(W_\epsilon > t)$ of the exact queueing delay is defined as

$$\widehat{\varphi}_{r,\epsilon}(t) := \mathbb{P}(W > t) + \epsilon \frac{1}{\mathbf{u}\boldsymbol{\omega}} \Big(\Theta_1(t) + \Theta_2(t) \Big),$$

where $\mathbb{P}(W > t)$ is the replace phase-type approximation of $\mathbb{P}(W_\epsilon > t)$, and $\Theta_1(t), \Theta_2(t)$ are given in Theorem 4.14.

The following result shows that the corrected replace approximation makes sense rigorously.

Proposition 4.17. *If $\mathbb{P}(W > t)$ is the replace approximation of the exact queueing delay $\mathbb{P}(W_\epsilon > t)$, then as $\epsilon \to 0$, it holds that*

$$\frac{\mathbb{P}(W_\epsilon > t) - \mathbb{P}(W > t)}{\epsilon} \to \Theta(t),$$

where $\Theta(t)$ is given in Theorem 4.14.

Proof. In Proposition 4.13, we found that

$$\widetilde{w}_\epsilon(s) = \widetilde{w}(s) + \epsilon \widetilde{\theta}(s) + O(\epsilon^2),$$

where $\widetilde{\theta}(s)$ is the LST of the signed measure $\Theta(t)$ introduced in Proposition 4.14. The above equation implies that

$$\frac{\widetilde{w}_\epsilon(s) - \widetilde{w}(s)}{\epsilon} = \widetilde{\theta}(s) + o(1). \tag{4.43}$$

We set $n = \dfrac{1}{\epsilon}$ and we define the sequence of functions

$$\widetilde{\chi}_n(s) := \frac{1}{\epsilon} \big(\widetilde{w}_\epsilon(s) - \widetilde{w}(s) \big),$$

where $\widetilde{\chi}_n(s)$ is the LST of the measure $X_n(t) = \big(\mathbb{P}(W_\epsilon > t) - \mathbb{P}(W > t) \big)/\epsilon$. By using Eq. (4.43), we obtain that $\widetilde{\chi}_n(s) \to \widetilde{\theta}(s)$, for all $s > 0$ as $n \to \infty$ (or equivalently $\epsilon \to 0$). Thus, it follows from the *Extended Continuity Theorem* (Feller, 1971, Theorem XIII.2) that

$$\frac{\mathbb{P}(W_\epsilon > t) - \mathbb{P}(W > t)}{\epsilon} \to \Theta(t),$$

which completes the proof. □

Although Approximation 4.16 gives an approximation of the queueing delay that can be calculated explicitly and is computationally tractable, it involves the evaluation of many terms. Therefore, to simplify the formula of the approximation, it makes sense to ignore terms that do not contribute significantly to the accuracy of the corrected replace approximation. Such terms seem to be the probabilities in $\Theta_2(t)$, which is defined in Theorem 4.14. Therefore, we define the *simplified corrected replace approximation* as follows.

Approximation 4.18. The simplified corrected replace approximation of the survival function $\mathbb{P}(W_\epsilon > t)$ of the exact delay is defined as

$$\widehat{\varphi}_{sr,\epsilon}(t) := \mathbb{P}(W > t) + \epsilon \frac{1}{\mathbf{u}\boldsymbol{\omega}} \Theta_1(t),$$

where $\mathbb{P}(W > t)$ is the replace phase-type approximation of $\mathbb{P}(W_\epsilon > t)$ and $\Theta_1(t)$ is given in Theorem 4.14.

4.4 Corrected discard approximation

In this section, we construct the corrected discard approximation. There are two different approaches to obtain this approximation. In the first one, we follow the same steps as in the construction of the corrected replace approximation. Namely, we first calculate the queueing delay for the simpler phase-type model when we discard the heavy-tailed customers and then we write the queueing delay of the mixture model as perturbation of the queueing delay in the discard base model. However, here we use an alternative approach that connects the discard base model with the replace base model.

As we mentioned in Section 4.2.2, when we discard the heavy-tailed customers we simply consider that

$$\widetilde{G}_\epsilon^\bullet(s) = (1-\epsilon)\widetilde{F}_p(s) + \epsilon = \widetilde{F}_p(s) + \epsilon\big(1 - \widetilde{F}_p(s)\big) = \widetilde{F}_p(s) + \epsilon s \mu_p \widetilde{F}_p^e(s).$$

Although the service time distribution $\widetilde{G}_\epsilon^\bullet(s)$ has an atom at zero, the resulting delay distribution has a phase-type representation and consequently it can be directly calculated through Laplace inversion of its LST $\widetilde{w}_\epsilon^\bullet(s)$. However, it is difficult to apply perturbation analysis to find the connection between $\widetilde{w}_\epsilon^\bullet(s)$ and $\widetilde{w}_\epsilon(s)$, because both of them depend on ϵ.

Observe that $\widetilde{G}_\epsilon^\bullet(s)$ can be expressed as perturbation of $\widetilde{F}_p(s)$ by the term $\epsilon s \mu_p \widetilde{F}_p^e(s)$. Therefore, we can apply perturbation analysis to find a connection between $\widetilde{w}_\epsilon^\bullet(s)$ and $\widetilde{w}(s)$, which is the Laplace transform of the queueing delay in the replace base model, and then use the connection of $\widetilde{w}(s)$ with $\widetilde{w}_\epsilon(s)$ to establish a connection between $\widetilde{w}_\epsilon(s)$ and $\widetilde{w}_\epsilon^\bullet(s)$. Thus, as a first step we express the matrices in the discard base model as perturbation of the ones in the replace base model, by setting $\widetilde{F}_h(s) \equiv 1$ in the results of Section 4.3.2. So, we define the matrices

$$\begin{aligned}
\widetilde{\mathbf{G}}_\epsilon^\bullet(s) =& \widetilde{\mathbf{G}}(s) + \epsilon s \mu_p \widetilde{F}_p^e(s) \mathbf{Q}^{(2)}, \\
\mathbf{H}_\epsilon^\bullet(s) =& \mathbf{H}(s) + \epsilon s \mu_p \widetilde{F}_p^e(s) \mathbf{Q}^{(2)} \circ \mathbf{P}\boldsymbol{\Lambda}, \\
\mathbf{E}_\epsilon^\bullet(s) =& \mathbf{E}(s) + \epsilon s \mu_p \widetilde{F}_p^e(s) \mathbf{Q}^{(2)} \circ \mathbf{P}\boldsymbol{\Lambda},
\end{aligned}$$

$$\mathbf{M}_\epsilon^\bullet = \mathbf{M} - \epsilon s \mu_p \mathbf{Q}^{(2)} \circ \mathbf{P}.$$

Now, we provide a series of results, which occur as corollaries of their corresponding results in Sections 4.3.2 and 4.3.3, for the evaluation of the Laplace transform $\widetilde{w}_\epsilon^\bullet(s) = s\mathbf{u}_\epsilon^\bullet \mathcal{E}_\epsilon^\bullet(s)\boldsymbol{\omega}/\det \mathbf{E}_\epsilon^\bullet(s)$. The first two corollaries are for the evaluation of the vector $\mathbf{u}_\epsilon^\bullet$ of unknown parameters.

Corollary 4.19. *Let* $\mathbf{u}$ *be the unique solution to the Eqs.* (4.10)–(4.11) *for the replace base model. If the roots* $s_2, \ldots, s_N$ *of* $\det\big(\mathbf{H}(s) + s\mathbf{I} - \mathbf{\Lambda}\big) = 0$ *with positive real part are simple, then*

1. *the equation* $\det\big(\mathbf{H}_\epsilon^\bullet(s) + s\mathbf{I} - \mathbf{\Lambda}\big) = 0$ *has exactly* N *non-negative solutions* $s_{\epsilon,1}^\bullet, \ldots, s_{\epsilon,N}^\bullet$, *with* $s_{\epsilon,1}^\bullet = 0$ *and* $s_{\epsilon,i}^\bullet = s_i - \epsilon\delta_i^\bullet + O(\epsilon^2)$ *for* $i = 2, \ldots, N$, *where*

$$\delta_i^\bullet := \delta^\bullet(s_i) = \frac{\sum_{j=1}^N \det\big(\mathbf{E}(s_i)_{\bullet 1}, \ldots, \mathbf{K}(s_i)_{\bullet j}, \ldots, \mathbf{E}(s_i)_{\bullet N}\big)}{\sum_{j=1}^N \det\big(\mathbf{E}(s_i)_{\bullet 1}, \ldots, \mathbf{E}^{(1)}(s_i)_{\bullet j}, \ldots, \mathbf{E}(s_i)_{\bullet N}\big)},$$

and $\mathbf{K}(s) = s\mu_p \widetilde{F}_p^e(s)\mathbf{Q}^{(2)} \circ \mathbf{P}\mathbf{\Lambda}$.

2. *We assume that the stability condition* $\boldsymbol{\pi}(\mathbf{\Lambda}^{-1} - \mathbf{M}_\epsilon^\bullet)\mathbf{e} > 0$ *is satisfied and we set* $\mathbf{A} = \big(\mathbf{\Lambda}^{-1}\mathbf{e}, \mathbf{a}_2, \ldots, \mathbf{a}_N\big)$ *(see Eq.* (4.14)*) and* $\mathbf{c} = \big(\boldsymbol{\pi}(\mathbf{\Lambda}^{-1} - \mathbf{M})\mathbf{e}, 0, \ldots, 0\big)$. *Then, the vector* $\mathbf{u}_\epsilon^\bullet$ *is the unique solution to the system of* N *linear equations*

$$\mathbf{u}_\epsilon^\bullet\big(\mathbf{A} - \epsilon\mathbf{B}^\bullet + O(\epsilon^2\mathbf{U})\big) = \mathbf{c} + \epsilon\mathbf{d}^\bullet,$$

where $\mathbf{B}^\bullet = \big(\mathbf{0}, \delta_2^\bullet \mathbf{a}_2{}^{(1)} - \mathbf{k}_2^\bullet, \ldots, \delta_N^\bullet \mathbf{a}_N{}^{(1)} - \mathbf{k}_N^\bullet\big)$ *and* $\mathbf{d}^\bullet = \big(\mu_p \boldsymbol{\pi}\mathbf{Q}^{(2)} \circ \mathbf{P}\mathbf{e}, 0, \ldots, 0\big)$, *with* $\mathbf{k}_i^\bullet$, $i = 2, \ldots, N$, *being a column vector with coordinates*

$$k_{i,j}^\bullet = (-1)^{m+j} \sum_{k=1}^{N-1} \det\Bigg(\Big(\mathbf{E}(s_i)_{\mathcal{N}\setminus\{m\}}^{\mathcal{N}\setminus\{j\}}\Big)_{\bullet 1}, \ldots, \Big(\mathbf{K}(s_i)_{\mathcal{N}\setminus\{m\}}^{\mathcal{N}\setminus\{j\}}\Big)_{\bullet k}, \ldots, \Big(\mathbf{E}(s_i)_{\mathcal{N}\setminus\{m\}}^{\mathcal{N}\setminus\{j\}}\Big)_{\bullet N-1}\Bigg), \quad j \in \mathcal{N},$$

and the choice of m *explained in Remark 4.6.*

Corollary 4.20. *The vector* $\mathbf{u}_\epsilon^\bullet$ *can be written in the form*

$$\mathbf{u}_\epsilon^\bullet = \mathbf{u} + \epsilon\mathbf{z}^\bullet + O(\epsilon^2\mathbf{e}),$$

where

$$\mathbf{z}^\bullet = \big(\mathbf{c}\mathbf{A}^{-1}\mathbf{B}^\bullet + \mathbf{d}^\bullet\big)\mathbf{A}^{-1}.$$

The next corollary gives us the denominator of $\widetilde{w}_\epsilon^\bullet(s)$.

Corollary 4.21. *If* $\det \mathbf{E}(s)$ *is evaluated according to Theorem 4.3 with* $\widetilde{G}(s) = \widetilde{F}_p(s)$, *then* $\det \mathbf{E}_\epsilon^\bullet(s)$ *can be written as perturbation of* $\det \mathbf{E}(s)$ *as follows*

$$\det \mathbf{E}_\epsilon^\bullet(s) = \det \mathbf{E}(s) + \epsilon s \mu_p \widetilde{F}_p^e(s) \sum_{k=1}^{N} k\big(\widetilde{F}_p(s)\big)^{k-1} \sum_{\substack{\Gamma\subset\mathcal{N} \\ |\Gamma|=k}} \sum_{\substack{S\subset\mathcal{N} \\ S\supset\Gamma}} \boldsymbol{\lambda}^S \zeta^{S^c}(s)$$
$$\times \det\Big(\big(\mathbf{Q}^{(1)} \circ \mathbf{P}\big)_S^{S\setminus\Gamma} \bowtie \big(\mathbf{Q}^{(2)} \circ \mathbf{P}\big)_S^{\Gamma}\Big) + O(\epsilon^2).$$

For the evaluation of the numerator of $\widetilde{w}_\epsilon^\bullet(s)$, we need the following result.

Corollary 4.22. *If $s\mathbf{u}\mathcal{E}(s)\mathbf{e}_i$ is evaluated according to Theorem 4.5 with $\widetilde{G}(s) = \widetilde{F}_p(s)$, then $s\mathbf{u}_\epsilon^\bullet\mathcal{E}_\epsilon^\bullet(s)\mathbf{e}_i$ can be written as perturbation of $s\mathbf{u}\mathcal{E}(s)\mathbf{e}_i$ as follows*

$$
\begin{aligned}
s\mathbf{u}_\epsilon^\bullet\mathcal{E}_\epsilon(s)\mathbf{e}_i =& s\mathbf{u}\mathcal{E}(s)\mathbf{e}_i \\
&+\epsilon s\Bigg[z_i^\bullet \sum_{k=1}^{N-1} \big(\widetilde{F}_p(s)\big)^k \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i\}\\ |\Gamma|=k;\\ S\subset\mathcal{N}\setminus\{i\}\\ S\supset\Gamma}} \boldsymbol{\lambda}^S\zeta^{S^c}(s) \det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^{S\setminus\Gamma} \bowtie \big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_S^{\Gamma}\Big) \\
&+z_i^\bullet \sum_{S\subset\mathcal{N}\setminus\{i\}} \boldsymbol{\lambda}^S\zeta^{S^c}(s) \det\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^S \\
&+\sum_{\substack{l=1\\ l\neq i}}^{N} z_l^\bullet(-1)^{l+i} \sum_{k=1}^{N-1} \big(\widetilde{F}_p(s)\big)^k \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\ |\Gamma|=k-1;\\ S\subset\mathcal{N}\setminus\{l,i\}\\ S\supset\Gamma;\\ R\subset S\cap T_{li}}} (-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s) \\
&\quad\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{S\setminus\Gamma} \bowtie \big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{\Gamma\cup\{i\}}\Big) \\
&+\sum_{\substack{l=1\\ l\neq i}}^{N} z_l^\bullet(-1)^{l+i} \sum_{k=0}^{N-2} \big(\widetilde{F}_p(s)\big)^k \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\ |\Gamma|=k}} \sum_{\substack{S\subset\mathcal{N}\setminus\{l,i\}\\ S\supset\Gamma;\\ R\subset S\cap T_{li}}} (-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s) \\
&\quad\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{(S\setminus\Gamma)\cup\{i\}} \bowtie \big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{\Gamma}\Big) \\
&+s\mu_p\widetilde{F}_p^e(s)\Bigg(u_i \sum_{k=1}^{N-1} k\big(\widetilde{F}_p(s)\big)^{k-1} \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{i\}\\ |\Gamma|=k}} \sum_{\substack{S\subset\mathcal{N}\setminus\{i\}\\ S\supset\Gamma}} \boldsymbol{\lambda}^S\zeta^{S^c}(s) \\
&\quad\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_S^{S\setminus\Gamma} \bowtie \big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_S^{\Gamma}\Big) \\
&+\sum_{\substack{l=1\\ l\neq i}}^{N} u_l(-1)^{l+i} \sum_{k=1}^{N-1} k\big(\widetilde{F}_p(s)\big)^{k-1} \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\ |\Gamma|=k-1}} \sum_{\substack{S\subset\mathcal{N}\setminus\{l,i\}\\ S\supset\Gamma;\\ R\subset S\cap T_{li}}} (-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s) \\
&\quad\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{S\setminus\Gamma} \bowtie \big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{\Gamma\cup\{i\}}\Big) \\
&+\sum_{\substack{l=1\\ l\neq i}}^{N} u_l(-1)^{l+i} \sum_{k=1}^{N-2} k\big(\widetilde{F}_p(s)\big)^{k-1} \sum_{\substack{\Gamma\subset\mathcal{N}\setminus\{l,i\}\\ |\Gamma|=k}} \sum_{\substack{S\subset\mathcal{N}\setminus\{l,i\}\\ S\supset\Gamma;\\ R\subset S\cap T_{li}}} (-1)^{|R|}\boldsymbol{\lambda}^{S\cup\{i\}}\zeta^{S^c}(s) \\
&\quad\times\det\Big(\big(\mathbf{Q}^{(1)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{(S\setminus\Gamma)\cup\{i\}} \bowtie \big(\mathbf{Q}^{(2)}\circ\mathbf{P}\big)_{S\cup\{l\}}^{\Gamma}\Big)\Bigg)\Bigg] + O(\epsilon^2),
\end{aligned}
$$

where $z_i^\bullet$, $i\in\mathcal{N}$, are the coordinates of the vector $\mathbf{z}^\bullet$ given in Corollary 4.20.

By combining Corollaries 4.21–4.22 and Proposition 4.13, we have the following proposition that connects the delay in the discard model $\widetilde{w}_\epsilon^\bullet(s)$ and the delay in the mixture model $\widetilde{w}_\epsilon(s)$.

Proposition 4.23. *If $\widetilde{w}_\epsilon^\bullet(s)$ is the Laplace transform of the queueing delay of the discard base model that is calculated as perturbation of $\widetilde{w}(s)$ (see Proposition 4.7), then there exist unique coefficients β, γ, α_k, β_k, γ_k, $k = 2, \dots, N$, and $\alpha''_{j,l}$, $\beta''_{j,l}$, $\gamma''_{j,l}$, $j = 1, \dots, \sigma$, $l = 1, \dots, r_j$, such that the Laplace transform $\widetilde{w}_\epsilon(s)$ of the queueing delay of the mixture model satisfies*

$$\begin{aligned}\widetilde{w}_\epsilon(s) =& \widetilde{w}_\epsilon^\bullet(s) + \epsilon \frac{1}{\mathbf{u}_\epsilon^\bullet \boldsymbol{\omega}} \widetilde{w}_\epsilon^\bullet(s) \Bigg[\left((\mathbf{z} - \mathbf{z}^\bullet)\boldsymbol{\omega} + \sum_{k=2}^{N} \frac{\alpha_k}{s - s_k} + \sum_{j=1}^{\sigma} \sum_{l=1}^{r_j} \frac{\alpha''_{j,l} \cdot (y_j)^{r_j - l + 1}}{(s + y_j)^{r_j - l + 1}} \right) \\ & - \mu_h \widetilde{F}_h^e(s) \left(\beta + \sum_{k=2}^{N} \frac{\beta_k}{s - s_k} + \sum_{j=1}^{\sigma} \sum_{l=1}^{r_j} \frac{\beta''_{j,l} \cdot (y_j)^{r_j - l + 1}}{(s + y_j)^{r_j - l + 1}} \right) \\ & + \mu_h \widetilde{F}_h^e(s) \widetilde{w}_\epsilon^\bullet(s) \left(\gamma + \sum_{k=2}^{N} \frac{\gamma_k}{s - s_k} + \sum_{j=1}^{\sigma} \sum_{l=1}^{r_j} \frac{\gamma''_{j,l} \cdot (y_j)^{r_j - l + 1}}{(s + y_j)^{r_j - l + 1}} \right) \Bigg] + O(\epsilon^2),\end{aligned}$$

where the vector $\mathbf{z}^\bullet$ given in Corollary 4.20.

Proof. The steps are exactly the same as in Proposition 4.13, but with different parameters that are in accordance to the discard base model. We first write the denominator and the numerator of $\widetilde{w}_\epsilon^\bullet(s)$ multiplied by $\big(p(s)\big)^r$ as perturbation of the respective quantities in the replace base model, and we have that

$$\big(p(s)\big)^r \det \mathbf{E}_\epsilon^\bullet(s) = \big(p(s)\big)^r \det \mathbf{E}(s) + \epsilon s \mu_p \widetilde{F}_p^e(s) \xi_{rM+N-1}(s) + O(\epsilon^2),$$

and,

$$\begin{aligned}\big(p(s)\big)^r s \mathbf{u}_\epsilon^\bullet \mathcal{E}_\epsilon^\bullet(s) \boldsymbol{\omega} =& \big(p(s)\big)^r s \mathbf{u} \mathcal{E}(s) \boldsymbol{\omega} + \epsilon s \Bigg[\sum_{i=1}^{N} \sum_{l=1}^{N} z_l \omega_i \xi'_{i,l,rM+N-1}(s) \\ & + s \mu_p \widetilde{F}_p^e(s) \sum_{i=1}^{N} \sum_{l=1}^{N} u_l \omega_i \xi_{i,l,rM+N-2}(s) \Bigg] + O(\epsilon^2),\end{aligned}$$

where the polynomials $\xi_{rM+N-1}(s)$, $\xi_{i,l,rM+N-2}(s)$, and $\xi'_{i,l,rM+N-1}(s)$ are defined according to the Eqs. (4.26), (4.30), and (4.31), respectively, and r is the maximum power of $p(s)$ that appears in the formulas. The $N-1$ common roots of the numerator and the denominator of $\widetilde{w}_\epsilon^\bullet(s)$ with positive real part are of the form $s_{\epsilon,k}^\bullet = s_k - \epsilon \delta_k^\bullet + O(\epsilon^2)$, $k = 2, \dots, N$, where the two equivalent definitions of $\delta_k^\bullet$ are as follows

$$\begin{aligned}\delta_k^\bullet =& \frac{\mu_p \widetilde{F}_p^e(s_k) \xi_{rM+N-1}(s_k) \widetilde{w}(s_k)}{\mathbf{u}\boldsymbol{\omega} \prod_{\substack{l=2 \\ l \neq k}}^{N} (s_k - s_l) \prod_{j=1}^{rM} (s_k + y_j)} \\ =& \frac{\mu_p \widetilde{F}_p^e(s_k) s_k \sum_{i=1}^{N} \sum_{l=1}^{N} u_l \omega_i \xi_{i,l,rM+N-2}(s_k)}{\mathbf{u}\boldsymbol{\omega} \prod_{\substack{l=2 \\ l \neq k}}^{N} (s_k - s_l) \prod_{j=1}^{rM} (s_k + y_j)} + \frac{\sum_{i=1}^{N} \sum_{l=1}^{N} z_l \omega_i \xi'_{i,l,rM+N-1}(s_k)}{\mathbf{u}\boldsymbol{\omega} \prod_{\substack{l=2 \\ l \neq k}}^{N} (s_k - s_l) \prod_{j=1}^{rM} (s_k + y_j)}.\end{aligned}$$

If we set now

$$d^\bullet(s) = \frac{\mu_p \widetilde{F}_p^e(s)\xi_{rM+N-1}(s)\widetilde{w}(s)}{\mathbf{u}\boldsymbol{\omega}\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+y_j)} - \sum_{k=2}^{N}\frac{\delta_k^\bullet}{s-s_k}$$

and

$$\begin{aligned} n^\bullet(s) =& \frac{\mu_p \widetilde{F}_p^e(s) s\sum_{i=1}^{N}\sum_{l=1}^{N} u_l\omega_i\xi_{i,l,rM+N-2}(s)}{\mathbf{u}\boldsymbol{\omega}\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+y_j)} \\ &+ \frac{\sum_{i=1}^{N}\sum_{l=1}^{N} z_l\omega_i\xi'_{i,l,rM+N-1}(s)}{\mathbf{u}\boldsymbol{\omega}\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+y_j)} - \sum_{k=2}^{N}\frac{\delta_k^\bullet}{s-s_k}, \end{aligned}$$

the denominator and the numerator of $\widetilde{w}_\epsilon^\bullet(s)$ multiplied by $\big(p(s)\big)^r$ can be written respectively as

$$\begin{aligned} &\big(p(s)\big)^r \det \mathbf{E}_\epsilon^\bullet(s) \\ &\quad = s\prod_{j=1}^{rM}(s+x_j)\prod_{k=2}^{N}(s-s_k+\epsilon\delta_k^\bullet+O(\epsilon^2))\big(1+\epsilon d^\bullet(s)+O(\epsilon^2)\big) \end{aligned} \tag{4.44}$$

and

$$\begin{aligned} &\big(p(s)\big)^r s\mathbf{u}_\epsilon^\bullet\mathcal{E}_\epsilon^\bullet(s)\boldsymbol{\omega} \\ &\quad = \mathbf{u}\boldsymbol{\omega}s\prod_{j=1}^{rM}(s+y_j)\prod_{k=2}^{N}\big(s-s_k+\epsilon\delta_k^\bullet+O(\epsilon^2)\big)\big(1+\epsilon n^\bullet(s)+O(\epsilon^2)\big). \end{aligned} \tag{4.45}$$

Note that both functions $d^\bullet(s)$ and $n^\bullet(s)$ are well-defined in the positive half-plane due to the definitions of $\delta_k^\bullet$ (page 114). By combining Eqs. (4.44) and (4.45), we obtain

$$\widetilde{w}_\epsilon^\bullet(s) = \widetilde{w}(s)\frac{1+\epsilon n^\bullet(s)+O(\epsilon^2)}{1+\epsilon d^\bullet(s)+O(\epsilon^2)} \quad\Rightarrow\quad \widetilde{w}(s) = \widetilde{w}_\epsilon^\bullet(s)\frac{1+\epsilon d^\bullet(s)+O(\epsilon^2)}{1+\epsilon n^\bullet(s)+O(\epsilon^2)}.$$

So,

$$\begin{aligned} \widetilde{w}_\epsilon(s) =& \widetilde{w}(s)\frac{1+\epsilon n(s)+O(\epsilon^2)}{1+\epsilon d(s)+O(\epsilon^2)} = \widetilde{w}_\epsilon^\bullet(s)\frac{1+\epsilon d^\bullet(s)+O(\epsilon^2)}{1+\epsilon n^\bullet(s)+O(\epsilon^2)}\cdot\frac{1+\epsilon n(s)+O(\epsilon^2)}{1+\epsilon d(s)+O(\epsilon^2)} \\ =& \widetilde{w}_\epsilon^\bullet(s)\Big(1+\epsilon\big((n(s)-n^\bullet(s))-(d(s)-d^\bullet(s))\big)+O(\epsilon^2)\Big) \\ =& \widetilde{w}_\epsilon^\bullet(s)+\epsilon\frac{1}{\mathbf{u}\boldsymbol{\omega}}\widetilde{w}_\epsilon^\bullet(s)\Bigg(\frac{\sum_{i=1}^{N}\sum_{l=1}^{N}(z_l-z_l^\bullet)\omega_i\xi'_{i,l,rM+N-1}(s)}{\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+y_j)} \\ &- \mu_h\widetilde{F}_h^e(s)\frac{s\sum_{i=1}^{N}\sum_{l=1}^{N}u_l\omega_i\xi_{i,l,rM+N-2}(s)}{\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+y_j)} \\ &+ \mu_h\widetilde{F}_h^e(s)\widetilde{w}(s)\frac{\xi_{rM+N-1}(s)}{\prod_{k=2}^{N}(s-s_k)\prod_{j=1}^{rM}(s+y_j)}\Bigg)+O(\epsilon^2) \end{aligned}$$

$$
\begin{aligned}
=&\widetilde{w}_\epsilon^\bullet(s)+\epsilon\frac{1}{\mathbf{u}\boldsymbol{\omega}}\widetilde{w}_\epsilon^\bullet(s)\Bigg[\left((z_i-z_l^\bullet)+\sum_{k=2}^{N}\frac{\alpha_k^\bullet}{s-s_k}+\sum_{j=1}^{rM}\frac{\alpha_j^{\bullet\prime}\cdot y_j}{s+y_j}\right)\\
&-\mu_h\widetilde{F}_h^e(s)\left(\beta+\sum_{k=2}^{N}\frac{\beta_k}{s-s_k}+\sum_{j=1}^{rM}\frac{\beta_j'\cdot y_j}{s+y_j}\right)\\
&+\mu_h\widetilde{F}_h^e(s)\widetilde{w}_\epsilon^\bullet(s)\left(\gamma+\sum_{k=2}^{N}\frac{\gamma_{i,k}}{s-s_k}+\sum_{j=1}^{rM}\frac{\gamma_j'\cdot y_j}{s+y_j}\right)\Bigg]+O(\epsilon^2).
\end{aligned}
$$

□

By using similar arguments as in the definition of the corrected replace approximations (see Section 4.3.4), we define the corrected discard approximations as follows.

Approximation 4.24. The corrected discard approximation of the survival function $\mathbb{P}(W_\epsilon>t)$ of the exact queueing delay is defined as

$$
\begin{aligned}
\widehat{\varphi}_{d,\epsilon}^\bullet(t):=&\mathbb{P}(W_\epsilon^\bullet>t)+\epsilon\frac{1}{\mathbf{u}_\epsilon^\bullet\boldsymbol{\omega}}\Big(\Theta_1^\bullet(t)+\Theta_2^\bullet(t)\Big),\ \text{where}\\
\Theta_1^\bullet(t)=&\left((\mathbf{z}-\mathbf{z}^\bullet)\boldsymbol{\omega}-\sum_{k=2}^{N}\frac{\alpha_k}{s_k}\right)\mathbb{P}(W_\epsilon^\bullet>t)-\left(\beta-\sum_{k=2}^{N}\frac{\beta_k}{s_k}\right)\mu_h\mathbb{P}(W_\epsilon^\bullet+C^e>t)\\
&+\left(\gamma-\sum_{k=2}^{N}\frac{\gamma_k}{s_k}\right)\mu_h\mathbb{P}(W_\epsilon^\bullet+W_\epsilon^{\bullet\prime}+C^e>t)\\
&+\sum_{j=1}^{\sigma}\sum_{l=1}^{r_j}\Bigg(\gamma''_{j,l}\mu_h\mathbb{P}\big(W_\epsilon^\bullet+W_\epsilon^{\bullet\prime}+C^e+E_{r_j-l+1}(y_j)>t\big)\\
&-\beta''_{j,l}\mu_h\mathbb{P}\big(W_\epsilon^\bullet+C^e+E_{r_j-l+1}(y_j)>t\big)+\alpha''_{j,l}\mathbb{P}\big(W_\epsilon^\bullet+E_{r_j-l+1}(y_j)>t\big)\Bigg),\\
\Theta_2^\bullet(t)=&\sum_{k=2}^{N}\frac{1}{s_k}\Bigg(\gamma_k\mu_h\mathbb{P}\big(t<W_\epsilon^\bullet+W_\epsilon^{\bullet\prime}+C^e<t+E(s_k)\big)\\
&-\beta_k\mu_h\mathbb{P}\big(t<W_\epsilon^\bullet+C^e<t+E(s_k)\big)+\alpha_k\mathbb{P}\big(t<W_\epsilon^\bullet<t+E(s_k)\big)\Bigg),
\end{aligned}
$$

$\mathbb{P}(W_\epsilon^\bullet>t)$ is the discard phase-type approximation of $\mathbb{P}(W_\epsilon>t)$, $W_\epsilon^{\bullet\prime}$ is independent and follows the same distribution of $W_\epsilon^\bullet$, and the coefficients β, γ, α_k, β_k, γ_k, $k=2,\dots,N$, and $\alpha''_{j,l}$, $\beta''_{j,l}$, $\gamma''_{j,l}$, $j=1,\dots,\sigma$, $l=1,\dots,r_j$, are calculated according to Proposition 4.13.

Approximation 4.24 can be made rigorous along the same lines as in Proposition 4.17. The simplified version of this approximation is found in the following lines.

Approximation 4.25. The simplified corrected discard approximation of the survival function $\mathbb{P}(W_\epsilon>t)$ of the exact queueing delay is defined as

$$
\widehat{\varphi}_{sd,\epsilon}^\bullet(t):=\mathbb{P}(W_\epsilon^\bullet>t)+\epsilon\frac{1}{\mathbf{u}_\epsilon^\bullet\boldsymbol{\omega}}\Theta_1^\bullet(t),
$$

where $\mathbb{P}(W_\epsilon^\bullet > t)$ is the replace phase-type approximation of $\mathbb{P}(W_\epsilon > t)$ and $\Theta_1^\bullet(t)$ is defined in Approximation 4.24.

In the next section, we perform numerical experiments to check the accuracy of the corrected phase-type and the simplified corrected phase-type approximations. In addition, we show that indeed the corrected approximations do not differ significantly from their simplified versions.

4.5 Numerical experiments

In Section 4.3.3, we pointed out that the first term of the corrected replace expansion is already a phase-type approximation of the queueing delay, a result that holds also for the discard expansion. In this section, we show that the addition of the correction term leads to improved approximations that are significantly more accurate than their phase-type counterparts. Therefore, we check the accuracy of the corrected phase-type approximations (see Approximations 4.16, 4.18, 4.24, and 4.25) by comparing them with the exact delay distribution and their corresponding phase-type approximations.

For the MArP arrival process of customers, we choose either a MMPP with two states or a MMPP with five states. What is left now is to fix values for the parameters of the mixture models and perform our numerical experiments. Thus, for the MMPP(2) arrival process we choose the parameters such that $\lambda_1 = 10$, $\lambda_1 = 1/2$, $p_{11} = 8/9$, and $p_{22} = 3/100$ (the rest of the parameters can be calculated by using Eqs. (4.2)–(4.3)). For the MMPP(5) model we choose:

$$\mathbf{P} = \begin{pmatrix} \frac{7}{27} & \frac{5}{27} & 0 & 0 & \frac{5}{9} \\ 0 & \frac{1}{29} & \frac{20}{29} & \frac{8}{29} & 0 \\ \frac{3}{25} & \frac{2}{5} & \frac{3}{10} & \frac{9}{50} & 0 \\ 0 & 0 & \frac{7}{36} & \frac{5}{18} & \frac{19}{36} \\ \frac{12}{47} & \frac{20}{47} & \frac{20}{47} & \frac{5}{47} & \frac{10}{47} \end{pmatrix},$$

and $\mathbf{\Lambda} = \mathrm{diag}\{11, 11, 13, 10, 8\}$.

Since it is more meaningful to compare approximations with exact results than with simulation outcomes, we choose the service time distribution such that we can find an exact formula for the queueing delay. As service time distribution we use a mixture of an exponential distribution with rate ν and a heavy-tailed one that belongs to a class of long-tailed distributions introduced in Abate and Whitt (1999b), i.e. the same mixture distribution we used in Section 3.4.1. For this combination of service time distributions, the survival queueing delay can be found explicitly, by following the same ideas as in Theorem 3.17. Although we do not have any restrictions for the parameters of the involved service time distributions, from a modelling point of view, it is counter-intuitive to fit a heavy-tailed service-time distribution with a mean smaller than the mean of the phase-type service-type distribution. For this reason, we select $\kappa = 2$ and $\nu = 3$.

Finally, note that we performed extensive numerical experiments for various values of the perturbation parameter ϵ in the interval $[0.001, 0.1]$. We chose to present only the case $\epsilon = 0.01$, since the qualitative conclusions for all other values of ϵ are similar to those presented in this section. For this choice of parameters, the load of the first system is equal to 0.909336 and of the second is 0.812845.

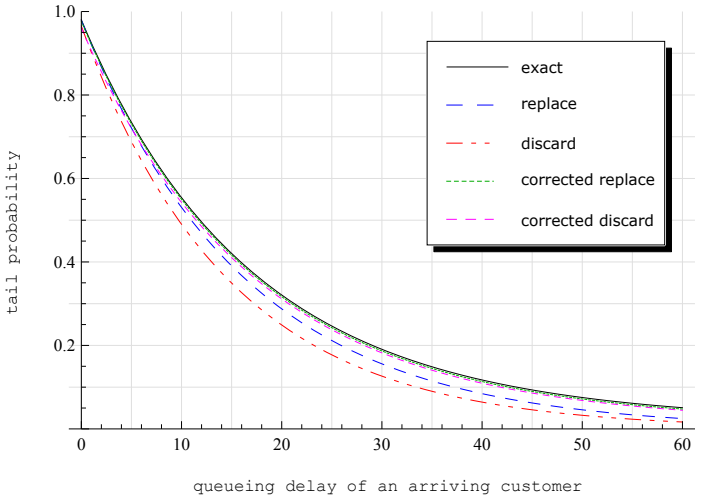


FIGURE 4.1: Exact queueing delay, phase-type and corrected phase-type approximations for perturbation parameter 0.01, MMPP(2) arrivals, and load of the system 0.908336.

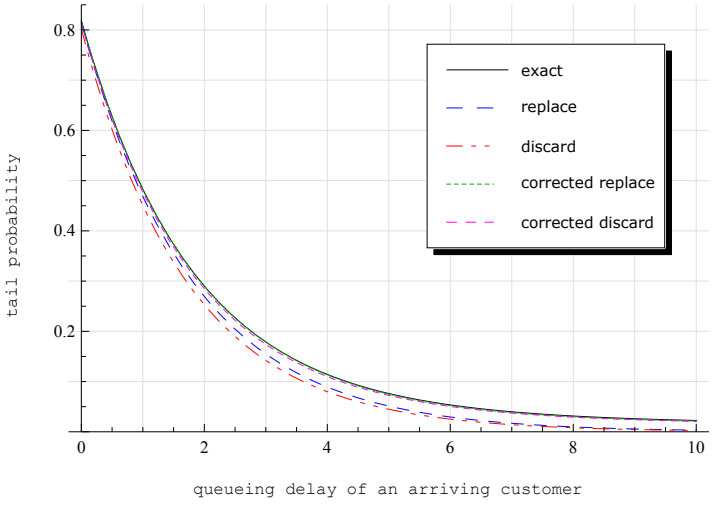


FIGURE 4.2: Exact queueing delay, phase-type and corrected phase-type approximations for perturbation parameter 0.01, MMPP(5) arrivals, and load of the system 0.812845.

As we observe from Figures 4.1–4.2, the phase-type approximations (replace and discard) give accurate estimates for small values of the queueing delay, while they are incapable of capturing the correct tail behaviour of the exact survival function of the queueing delay. On the contrary, both corrected phase-type approximations are highly accurate and give a small relative error at the tail. More precisely, we can observe the following:

- The corrected replace approximation gives better numerical estimates than the corrected discard approximation. Especially, in Figure 4.2 the corrected-replace approximation is hardly distinguishable from the exact distribution.
- The corrected discard approximation always underestimates the exact tail probability of the queueing delay. On the contrary, the corrected replace approximation may overestimate the exact survival function for small values, but always underestimates the tail of the exact queueing delay.
- The corrected phase-type approximations do not differ significantly from their simplified versions. The maximum observed absolute error between the two corrected replace approximations is smaller than 0.0011 for the MMPP(2) model and smaller than 0.00069 for the MMPP(5) model. The corresponding numbers for the corrected discard approximations are 0.0052 and 0.00167.
- Finally, we estimated the relative error at the tail for all four corrected phase-type approximations. We found that in the MMPP(2) model the relative error is smaller than 10% for all the approximations, while this number reduces to 7% in the MMPP(5) model.

4.6 Conclusions

To conclude, all corrected phase-type approximations are highly accurate and there is no significant difference between the Approximations 4.16 and 4.24 and their respective simplified versions 4.18 and 4.25. For this reason, the simplified versions of the approximations serve as excellent substitutes to their original corrected phase-type approximations for estimating the queueing delay. Finally, the corrected phase-type approximations give a small relative error at the tail, which can easily be verified to be $O(\epsilon)$.

CHAPTER 5

Truncated buffer approximations

5.1 Introduction

In Chapters 2–4, our main focus was on heavy-tailed models. More precisely, we truncated the heavy-tailed distributions in a way not only to construct accurate approximations for the performance measures under consideration, but also to derive error bounds. In this chapter, however, we no longer focus on heavy-tailed distributions. As we mentioned in Chapter 1, another example where truncations of the state space are involved is when considering networks of queues. Therefore, in this chapter, we aim at deriving error bounds for the queue lengths of a tandem queueing network when we truncate its state space.

Specifically, we consider the $M^X/M/1 \rightarrow \bullet/M/1$ tandem queueing network, where customers arrive in batches in the first queue. When the batch size distribution is degenerate at one – in other words, when customers arrive one at a time – this network is a special case of Jackson networks and admits a product form solution (Kelly, 1979; Latouche and Ramaswami, 1999; Lavenberg, 1983; Ramaswami and Taylor, 1996). This product form property may be interpreted as if the number of customers in the various queues are independent random variables in steady state. However, for general batch size distributions, a product form solution does not exist and neither has any other solution been identified.

In the absence of exact solutions, the queue lengths may be evaluated numerically. In Chapter 1, we briefly presented a number of techniques for the analysis of two-dimensional Markov processes. Here we focus on MAM and we consider the embedded Markov process at the moments upon which the state of the system changes. The embedded Markov chain giving the evolution of the queue lengths at each of the queues can be expressed as a QBD and a matrix-geometric solution can be found for the joint queue length distribution as long as either of the buffer sizes (waiting rooms in front of the queues) is finite (see Section 1.3.1). Thus, for the application of MAM, the finiteness of either one buffer is sufficient.

Besides the joint queue length distribution, it is also interesting to study the marginal queue lengths. For example, we could consider the case of truncating the buffer size of the second queue, so as the customer that finishes service in the first queue is lost if the truncated buffer of the second queue is full. Note that in this case, the number of customers in the first queue forms on its own a Markov chain (embedded on the same transition epochs as the two-dimensional Markov chain), which has a stationary distribution that can be found irrespectively of the truncation to the buffer size of the second queue (Kleinrock, 1976). On the other hand, if we only allow customers to enter the system until the number of customers in the first queue reaches a predetermined high level, the output process from the first queue is affected, and consequently this has an effect to the second queue. Therefore, it makes more sense to truncate the buffer size of the first queue to study the effect of truncation on the marginal queue length of the second queue.

Although the truncation of the state space leads to an approximate model that can be analysed numerically, it also introduces approximation errors. The goal of this chapter is to obtain a clear understanding of these types of errors. We explain now how to derive error bounds for the approximations of the queue lengths. Observe that even if we choose the truncation level relatively high, for the evaluation of the queue length distribution we exclude all sample paths where the number of customers in the first queue exceeds the truncation level. This means that these omitted sample paths are responsible for the observed approximation error. Therefore, with the aid of *extreme value theory* (EVT), which aims at providing statistical models for rare events (De Haan and Ferreira, 2007; Resnick, 2007a), we study the behaviour of the system under the assumption that the first queue reached a level greater than or equal to the truncation level and we derive an asymptotic upper bound for the approximations derived with MAM. Moreover, as we shall see in Section 5.6, the analysis forces us to distinguish three different cases in this model that relate to the arrival and service rates, which we study separately.

Outline

The rest of the chapter is organised as follows. In Section 5.2, we introduce the model under consideration and in Section 5.3, we truncate the state space of the first queue. We write the joint queue length probabilities as a sum of two terms that depend on the truncation level, where both terms are discussed in Section 5.3.1. We find that the first term relates to the steady state probabilities of the truncated model and we provide its connection with the exact probabilities. For the second term we prove that it is bounded by the mean cycle length during which the number of customers in the first queue exceeds the truncation level.

Afterwards, in Section 5.3.2, we combine the results of Section 5.3.1 to derive error bounds for the queue length distribution of the truncated system. In particular, the upper bound is an asymptotic bound expressed as a product of three factors. For the evaluation of these factors, we first perform in Section 5.4 an exponential change of measure. In addition, in Section 5.4.1, we relate the number of customers in the first queue with a random walk and we introduce the accompanying notation.

The first factor involved in the asymptotic upper bound is equal to the probability that the maximum number of the customers in the first queue exceeds the truncation

level, while the second is equal to the mean cycle length. An asymptotic result that links these two factors is derived in Section 5.5. The third factor, which is equal to the conditional mean cycle length given that number of customers in the first queue exceeds the truncation level, is discussed in Section 5.6. Since this factor requires a lot of different techniques for its evaluation, we first explain intuitively in Section 5.6.1 its asymptotic behaviour and later, in Section 5.6.2, we study it rigorously.

In Section 5.7, we perform numerical experiments to check the quality of the asymptotic error bound and we provide our conclusions in Section 5.8. Finally, in Appendix A.3, we provide useful results on random walks that we extend in the lattice case.

5.2 Presentation of the model

We consider an $\mathrm{M}^X/\mathrm{M}/1\rightarrow\bullet/\mathrm{M}/1$ tandem queueing network. Customers arrive in batches according to a Poisson stream with rate λ and join the first queue. The service times for each queue are exponential with rates μ_1 and μ_2, respectively. A customer that finishes service in the first queue moves to the second. The customer leaves the system after finishing his service in the second queue. We describe the system by a two-dimensional CTMC $\{(X_t, Y_t)\}_{t\geq 0}$, where X_t and Y_t are the queue lengths at time t of the first and the second queue, respectively, including customers in service in either queue. For this system, we are interested in evaluating its limiting distribution (X_∞, Y_∞); namely, the behaviour of the Markov process when $t \to \infty$.

We assume that an arriving batch is of size i with probability b_i, where $\sum_{i=1}^{\infty} b_i = 1$. We denote by B the generic r.v. of the batch sizes and we assume that its mean $\mathbb{E}B = \sum_{i=1}^{\infty} i b_i$ is finite. Moreover, we concentrate exclusively on the case where (X_t, Y_t) has a limit (X_∞, Y_∞) in distribution as $t \to \infty$. Thus, for stability reasons, we assume that $\lambda \mathbb{E}B/\mu_i < 1$, $i = 1, 2$.

Without loss of generality, we may take $\lambda + \mu_1 + \mu_2 = 1$ and assume that the servers always work (also when there is no job). However, service completions only lead to a departure if there is a customer in the corresponding queue. Otherwise, we assume that the customer in service is fictitious (and fictitious customers will be interrupted as soon as a real customer arrives). The artificial assumption of working on fictitious customers implies that in each state the outgoing transition rates add up to 1, and thus, the mean time spent in a state is 1 for all states. This trick of uniformisation allows us to convert the CTMC into a DTMC. The mean time between jump epochs is always 1 and, as a consequence, the embedded DTMC at jumps has the same equilibrium distribution as the original CTMC. Therefore, from now on, we only consider the DTMC (embedded at jump epochs) instead of the CTMC.

We introduce now some notation for our embedded Markov chain. We denote by (X_n, Y_n) the state of the Markov chain at the nth jump and we have that $(X_n, Y_n) \in (\Omega_1, \Omega_2)$, where $\Omega_1 = \Omega_2 = \mathbb{N}$. We denote the netput between the $(n-1)$st and the nth jump epoch in the first and second queue as Z_n and W_n, respectively, where

$$Z_n = \begin{cases} 0, & \text{with probability } \mu_2, \\ -1, & \text{with probability } \mu_1, \\ m, & \text{with probability } \lambda b_m, m = 1, 2, \ldots, \end{cases} \tag{5.1}$$

and

$$W_n = \begin{cases} -1, & \text{if } Z_n = 0, \\ 1, & \text{if } Z_n = -1 \text{ and } X_{n-1} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5.2}$$

Recall that due to uniformisation, $\lambda, \mu_1, \mu_2 < 1$ and the rates λ, μ_1, μ_2 can be seen as probabilities.

The number of customers X_n in the first queue satisfies the following Lindley recursion

$$X_0 = 0, \qquad X_{n+1} = \big(X_n + Z_{n+1}\big)^+, \quad n = 0, 1, \dots \tag{5.3}$$

Thus, $\{X_n\}_{n=0,1,\dots}$ evolves as a reflected at 0 discrete version of a random walk with increments $Z_1, Z_2, \dots$ Similarly, the number of customers Y_n in the second queue satisfies the recursion

$$Y_0 = 0, \qquad Y_{n+1} = \big(Y_n + W_{n+1}\big)^+, \quad n = 0, 1, \dots \tag{5.4}$$

The initial state of the system is $(X_0, Y_0) = (0, 0)$ and we define the first return time to the origin as $T_{(0,0)} = \inf\{n \geq 1 : X_n = Y_n = 0 \mid X_0 = Y_0 = 0\}$, which is also called *cycle length*. Therefore, since we have a two-dimensional positive recurrent irreducible Markov chain, it is known that

$$\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) = \frac{1}{\mathbb{E}T_{(0,0)}} \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}} \mathbb{1}\left(X_n \geq x, Y_n \geq y\right)\right].$$

From Eqs. (5.1) and (5.2), we can easily verify that the two-dimensional Markov chain (X_n, Y_n) is a QBD with an infinite state space. Therefore, MAM cannot be applied to evaluate the joint queue length distribution. Moreover, the model does not admit a product form solution (Latouche and Ramaswami, 1999, Theorem 15.1.1). Thus, we truncate the state space to find an approximation by using MAM. In the next section, we truncate the state space and we find error bounds for the approximation of the joint queue length distribution that stems from this truncation.

5.3 State space truncation and error bounds

As we mentioned in Section 5.1, truncation of the buffer in the second queue (considering lost customers) does not influence the Markov chain of the first queue. Thus, since we are interested in the effect of truncation to the marginal queue length distribution of the second queue, we truncate the space Ω_1 of the first queue at level N, which we call *truncation level*. In other words, $\{X_n\}_{n\geq 0}$ takes values in the space $\Omega_1^{(N)} = \mathbb{N}_N$.

The arriving customers are admitted in the system by applying the *Partial Batch Acceptance Strategy* (PBAS), i.e. if the batch size is larger than the number of available free positions in the buffer (which has capacity $N - 1$) then we accept only so many customers until they are in total N customers waiting in front of the first queue and we dismiss the remaining ones. Moreover, we denote by $\big(X_n^{(N)}, Y_n^{(N)}\big)$ the approximate Markov chain associated with the truncation level N and by $\big(Z_n^{(N)}, W_n^{(N)}\big)$

the corresponding netput process. Observe that definitions (5.1), (5.2), (5.3), and (5.4) are still valid (but with the notation adapted to the truncated system) for the processes $X_n^{(N)}$, $Y_n^{(N)}$, and $W_n^{(N)}$, respectively. However, the definition of $Z_{n+1}^{(N)}$ takes two alternative forms depending on the value of $X_n^{(N)}$. More precisely, if $X_n^{(N)} = N$, then

$$Z_{n+1}^{(N)} = \begin{cases} 0, & \text{with probability } \lambda + \mu_2, \\ -1, & \text{with probability } \mu_1, \end{cases} \tag{5.5}$$

while in case $X_n^{(N)} = N - m$, $m \in \{1, \ldots, N\}$

$$Z_{n+1}^{(N)} = \begin{cases} 0, & \text{with probability } \mu_2, \\ -1, & \text{with probability } \mu_1, \\ k, & \text{with probability } \lambda b_k \text{ for } k < m, \\ m, & \text{with probability } \lambda \sum_{i=m}^{\infty} b_i. \end{cases} \tag{5.6}$$

The steady state probability can then be split as follows

$$\begin{aligned} \mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) = & \frac{1}{\mathbb{E}T_{(0,0)}} \underbrace{\mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}} \mathbb{1}\left(X_n \geq x, Y_n \geq y\right) \cdot \mathbb{1}\left(\max_{1 \leq l \leq T_{(0,0)}} X_l < N\right)\right]}_{=\mathbb{I}} \\ & + \frac{1}{\mathbb{E}T_{(0,0)}} \underbrace{\mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}} \mathbb{1}\left(X_n \geq x, Y_n \geq y\right) \cdot \mathbb{1}\left(\max_{1 \leq l \leq T_{(0,0)}} X_l \geq N\right)\right]}_{=\mathbb{II}}. \end{aligned}$$

In Section 5.3.1, we discuss each of the terms $\mathbb{I}$ and $\mathbb{II}$ separately, while in Section 5.3.2, we explain how the results of Section 5.3.1 can be combined to generate error bounds.

5.3.1 Partition of the queue length probabilities

Term $\mathbb{I}$

We denote by $\boldsymbol{m} = (m_1, m_2)$ the two-dimensional states of the Markov chain (X_n, Y_n), where m_1 and m_2 are non-negative integers. If $\boldsymbol{P}$ is the transition probability matrix of the Markov chain and $\boldsymbol{P}^{(N)}$ its truncation, then we have that

$$\boldsymbol{P}^{(N)}(\boldsymbol{m}, \boldsymbol{n}) = \boldsymbol{P}(\boldsymbol{m}, \boldsymbol{n}), \qquad \forall \boldsymbol{m}, \boldsymbol{n} \text{ with } m_1, n_1 \in \mathbb{N}_{N-1}. \tag{5.7}$$

In other words, the entries in the two matrices $\boldsymbol{P}^{(N)}$ and $\boldsymbol{P}$ coincide as long as both two-dimensional Markov chains (original and truncated) live within the boundaries. Therefore, if we set $\nu = \inf\{n \geq 0 : X_n \geq N\}$ and $\nu^{(N)} = \inf\{n \geq 0 : X_n^{(N)} \geq N\}$, then it holds that $(X_n : n < \nu) \overset{\mathfrak{D}}{=} (X_n^{(N)} : n < \nu^{(N)})$. Finally, we define as $T_{(0,0)}^{(N)} = \inf\{n \geq 1 : X_n^{(N)} = Y_n^{(N)} = 0 \mid X_0^{(N)} = Y_0^{(N)} = 0\}$ the first return time to the origin for the truncated system. Observe that $T_{(0,0)} = T_{(0,0)}^{(N)}$ when

$\mathbb{1}\left(\max_{1\le l\le T_{(0,0)}} X_l < N\right) = 1$. Thus, since term $\mathbb{I}$ contains the sample paths of the truncated system, we obtain:

$$\begin{aligned}\mathbb{I} =& \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}^{(N)}} \mathbb{1}\left(X_n^{(N)} \ge x, Y_n^{(N)} \ge y\right) \cdot \mathbb{1}\left(\max_{1\le l\le T_{(0,0)}^{(N)}} X_l^{(N)} < N\right)\right] \\ \le& \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}^{(N)}} \mathbb{1}\left(X_n^{(N)} \ge x, Y_n^{(N)} \ge y\right)\right] = \mathbb{E}T_{(0,0)}^{(N)} \mathbb{P}\big(X_\infty^{(N)} \ge x, Y_\infty^{(N)} \ge y\big).\end{aligned}$$

Due to monotonicity, the mean return time to the origin of the truncated system is smaller than the corresponding mean return time of the original system; i.e. $\mathbb{E}T_{(0,0)} \ge \mathbb{E}T_{(0,0)}^{(N)}$. Moreover, the truncation is done in a way such that $\mathbb{P}\big(X_\infty^{(N)} \ge x, Y_\infty^{(N)} \ge y\big) \le \mathbb{P}\big(X_\infty \ge x, Y_\infty \ge y\big)$. We formulate these properties in the following theorem.

Theorem 5.1. *If N is the level at which we truncate the state space Ω_1, then the following inequalities hold:*

$$\mathbb{P}\big(X_\infty^{(N)} \ge x, Y_\infty^{(N)} \ge y\big) \le \mathbb{P}\big(X_\infty \ge x, Y_\infty \ge y\big), \quad \forall (x,y) \in (\Omega_1, \Omega_2), \text{ and} \tag{5.8}$$

$$\mathbb{E}T_{(0,0)}^{(N)} \le \mathbb{E}T_{(0,0)}. \tag{5.9}$$

Proof. Since the number of customers in the first queue of the truncated system cannot exceed N, it is immediately obvious that inequality (5.8) holds true for all $x \ge N+1$ and $y \in \Omega_2$. Therefore, we are only interested in the cases where $x \le N$. We prove the theorem, by using techniques from cost structure models (van Houtum et al., 1998).

For some appropriately chosen cost structure, most performance characteristics of Markovian systems can be represented by average costs. Thus, if $\boldsymbol{\pi}$ is the stationary distribution of the two-dimensional Markov chain (X_n, Y_n) and $c(\boldsymbol{m})$ represent the costs per period of time the system is in state $\boldsymbol{m}$, then the average costs g are given by

$$g = \sum_{\boldsymbol{m}} c(\boldsymbol{m})\boldsymbol{\pi}(\boldsymbol{m}). \tag{5.10}$$

If we define $g = \mathbb{P}\big(X_\infty \ge x, Y_\infty \ge y\big)$, then an appropriate cost structure is $c(\boldsymbol{m}) = 1$ when $\boldsymbol{m} \in D$, and $c(\boldsymbol{m}) = 0$ otherwise, for $D = \{(m_1, m_2) : m_1 \ge x \text{ and } m_2 \ge y\}$. In addition, if we set $\widetilde{g} = \mathbb{P}\big(X_\infty^{(N)} \ge x, Y_\infty^{(N)} \ge y\big)$, then our goal is to prove that $\widetilde{g}$ is a lower bound for g; namely $\widetilde{g} \le g$. To do so, we study the expected costs over a finite number of periods. We denote by $u_t(\boldsymbol{m})$ the expected costs in the first $t \ge 0$ periods when starting in state $\boldsymbol{m}$. Similar to $u_t(\boldsymbol{m})$, we define $\widetilde{u}_t(\boldsymbol{m})$ as the expected costs over the first t periods in the truncated model when starting in state $\boldsymbol{m}$. Defining $u_0 = \widetilde{u}_0 = 0$, we will prove by induction that for all $t = 0, 1, 2, \ldots$ and all recurrent states $\boldsymbol{m}$ in the truncated model

$$\widetilde{u}_t(\boldsymbol{m}) \le u_t(\boldsymbol{m}). \tag{5.11}$$

From this, it follows that

$$\widetilde{g} = \lim_{t\to\infty} \frac{1}{t}\widetilde{u}_t(\boldsymbol{m}) \le \lim_{t\to\infty} \frac{1}{t} u_t(\boldsymbol{m}) = g. \tag{5.12}$$

Thus, we first need to establish precedences between states in the original model. We say that state $\boldsymbol{m}$ has precedence over state $\boldsymbol{n}$, or is more attractive than state $\boldsymbol{n}$, if $\boldsymbol{m}$ and $\boldsymbol{n}$ satisfy the following precedence relation:

$$u_t(\boldsymbol{m}) \leq u_t(\boldsymbol{n}), \qquad \text{for all } t = 0, 1, 2, \ldots \tag{5.13}$$

In other words, starting in $\boldsymbol{m}$ yields lower total expected costs than starting in $\boldsymbol{n}$. Now, the first and crucial step is the determination of a set $\mathcal{P}$ of pairs $(\boldsymbol{m}, \boldsymbol{n})$ satisfying (5.13). These pairs are called *precedence pairs*. We prove (5.13) for the set of pairs $\mathcal{P} = \{(\boldsymbol{m}, \boldsymbol{n}) : m_1 \leq n_1 \ \& \ m_2 \leq n_2\}$ by induction over t. Taking $t = 1$ in (5.13) directly leads to

$$c(\boldsymbol{m}) \leq c(\boldsymbol{n}), \qquad \forall (\boldsymbol{m}, \boldsymbol{n}) \in \mathcal{P}. \tag{5.14}$$

Assume (5.13) holds for t. To prove inequality (5.13) for $t+1$ for all pairs in $\mathcal{P}$, it suffices to do so for the pairs in the smaller set $\mathcal{P}_- = \{(\boldsymbol{m}, \boldsymbol{n}) : \boldsymbol{n} = \boldsymbol{m} + \boldsymbol{e}_1 \text{ or } \boldsymbol{n} = \boldsymbol{m} + \boldsymbol{e}_2\}$, where $\boldsymbol{e}_i$ the ith unit vector. Clearly $\mathcal{P}_-$ is a subset of $\mathcal{P}$ and it is easily seen that the inequalities (5.13) for the pairs in $\mathcal{P}_-$ generate the ones for all pairs in $\mathcal{P}$, by using transitivity of the operation $\leq$. To establish (5.13) for $t+1$, we have to show for each $(\boldsymbol{m}, \boldsymbol{n}) \in P_-$ that

$$u_{t+1}(\boldsymbol{m}) = c(\boldsymbol{m}) + \sum_{\boldsymbol{i}} p(\boldsymbol{m}, \boldsymbol{i}) u_t(\boldsymbol{i}) \leq c(\boldsymbol{n}) + \sum_{\boldsymbol{j}} p(\boldsymbol{n}, \boldsymbol{j}) u_t(\boldsymbol{j}) = u_{t+1}(\boldsymbol{n}). \tag{5.15}$$

Because of (5.14), it suffices to show that

$$\sum_{\boldsymbol{i}} p(\boldsymbol{m}, \boldsymbol{i}) u_t(\boldsymbol{i}) \leq \sum_{\boldsymbol{j}} p(\boldsymbol{n}, \boldsymbol{j}) u_t(\boldsymbol{j}). \tag{5.16}$$

We prove (5.16) for the pairs $(\boldsymbol{m}, \boldsymbol{m} + \boldsymbol{e}_j)$, $j = 1, 2$, together. We consider first the case $\boldsymbol{m} = (m_1, m_2)$ with $m_1 \geq 1$ and $m_2 \geq 1$. Therefore, Eq. (5.16) takes the form

$$\begin{aligned} &\lambda \sum_{i=1}^{\infty} b_i u_t(\boldsymbol{m} + i\boldsymbol{e}_1) + \mu_1 u_t(\boldsymbol{m} - \boldsymbol{e}_1 + \boldsymbol{e}_2) + \mu_2 u_t(\boldsymbol{m} - \boldsymbol{e}_2) \\ &\qquad \leq \lambda \sum_{i=1}^{\infty} b_i u_t(\boldsymbol{m} + \boldsymbol{e}_j + i\boldsymbol{e}_1) + \mu_1 u_t(\boldsymbol{m} + \boldsymbol{e}_j - \boldsymbol{e}_1 + \boldsymbol{e}_2) \\ &\qquad\quad + \mu_2 u_t(\boldsymbol{m} + \boldsymbol{e}_j - \boldsymbol{e}_2), \qquad j = 1, 2. \end{aligned} \tag{5.17}$$

Now, we compare both sides of inequality (5.17). The first terms of each side, both corresponding to arrivals of customers, are ordered as desired by the induction hypothesis. The same holds for the second and third terms, which correspond to service completions at the first and the second queue, respectively. So, (5.17) holds. We also consider the cases $m_1 = 0$ and $m_2 = 0$ separately. If $m_1 = 0$, then the coefficient of μ_1 on the left hand side of (5.17) is identically equal to zero. Similarly, if $m_2 = 0$, then the coefficient of μ_2 on the left hand side of (5.17) is identically equal to zero. Therefore, (5.17) holds also for these two cases and the proof of (5.13) is complete.

The last step is to prove (5.11) by induction. For $t = 0$, inequality (5.11) trivially holds. Assuming (5.11) holds for t, we prove it for $t+1$. The expected costs over $t+1$

periods are equal to

$$u_{t+1}(\boldsymbol{m}) = c(\boldsymbol{m}) + \sum_{\boldsymbol{n}} p(\boldsymbol{m},\boldsymbol{n}) u_t(\boldsymbol{n}). \tag{5.18}$$

By using the PBAS, the transitions from $\boldsymbol{m}$ to $\boldsymbol{n}$ with $n_1 > N$ and $n_2 = m_2$ are redirected to the state $\widetilde{\boldsymbol{n}}$ with $n_1 = N$ and $n_2 = m_2$. The new transition probability to $\boldsymbol{n}$ is zero and to $\widetilde{\boldsymbol{n}}$ it is increased by $p(\boldsymbol{m},\boldsymbol{n})$. We denote the new transition probabilities by $\widetilde{p}(\boldsymbol{m},\boldsymbol{n})$. The costs per state are not altered. It follows that $u_{t+1}(\boldsymbol{m}) \geq \widetilde{u}_{t+1}(\boldsymbol{m})$ since we have constructed the truncated model by redirecting outgoing transitions to more attractive states; i.e. $(\widetilde{\boldsymbol{n}},\boldsymbol{n}) \in \mathcal{P}$. Namely, we have

$$u_{t+1}(\boldsymbol{m}) \geq c(\boldsymbol{m}) + \sum_{\boldsymbol{n}} \widetilde{p}(\boldsymbol{m},\boldsymbol{n}) u_t(\boldsymbol{n}) \geq c(\boldsymbol{m}) + \sum_{\boldsymbol{n}} \widetilde{p}(\boldsymbol{m},\boldsymbol{n}) \widetilde{u}_t(\boldsymbol{n}) = \widetilde{u}_{t+1}(\boldsymbol{m}),$$

where the second inequality follows from the induction hypothesis.

Last, we need to prove the inequality $\mathbb{E}T^{(N)}_{(0,0)} \leq \mathbb{E}T_{(0,0)}$. Observe that $\mathbb{E}T_{(0,0)}$ and $\mathbb{E}T^{(N)}_{(0,0)}$ are by definition the expected first return times to the state $(0,0)$ in the original and the truncated system, respectively. By applying the *strong law of large numbers* for ergodic Markov chains (Kijima, 1997), we obtain that $\mathbb{E}T^{(N)}_{(0,0)} = 1/\mathbb{P}\big(X^{(N)}_\infty = 0, Y^{(N)}_\infty = 0\big)$ and $\mathbb{E}T_{(0,0)} = 1/\mathbb{P}\big(X_\infty = 0, Y_\infty = 0\big)$. Therefore, it is sufficient to show that the inequality $\mathbb{P}\big(X^{(N)}_\infty = 0, Y^{(N)}_\infty = 0\big) \geq \mathbb{P}\big(X_\infty = 0, Y_\infty = 0\big)$ holds. To prove this inequality, we use again a cost structure approach. More precisely, we set $c(\boldsymbol{m}) = 1$ when $\boldsymbol{m} = (0,0)$ and $c(\boldsymbol{m}) = 0$ otherwise. As a result, $g = \mathbb{P}\big(X_\infty = 0, Y_\infty = 0\big)$ and $\widetilde{g} = \mathbb{P}\big(X^{(N)}_\infty = 0, Y^{(N)}_\infty = 0\big)$. By repeating the steps we followed to prove the inequality $\mathbb{P}\big(X^{(N)}_\infty \geq x, Y^{(N)}_\infty \geq y\big) \leq \mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big)$, we can now prove that $\widetilde{g}$ is an upper bound for g. This can easily be seen, since with these costs we redirect transitions to less attractive states. The precedence set for this cost model is $\acute{\mathcal{P}} = \{(\boldsymbol{m},\boldsymbol{n}) : m_1 \geq n_1 \;\&\; m_2 \geq n_2\}$. □

We now turn our attention to the second term.

Term $\mathbb{III}$

We set $M^{T_{(0,0)}} = \max_{1 \leq n \leq T_{(0,0)}} X_n$ for the maximum queue length of the first queue before the first return time to the state $(0,0)$. Thus, we have

$$\begin{aligned}
\mathbb{III} &= \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}} \mathbb{1}\left(X_n \geq x, Y_n \geq y\right) \cdot \mathbb{1}\left(\max_{1 \leq l \leq T_{(0,0)}} X_l \geq N\right)\right] \\
&\leq \mathbb{E}\left[T_{(0,0)} \cdot \mathbb{1}\left(\max_{1 \leq l \leq T_{(0,0)}} X_l \geq N\right)\right] = \mathbb{E}\big[T_{(0,0)} \cdot \mathbb{1}\left(M^{T_{(0,0)}} \geq N\right)\big] \\
&= \mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big]\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big),
\end{aligned}$$

which shows that term $\mathbb{III}$ evolves in some sense like $M^{T_{(0,0)}}$.

In the next section, we combine the results we have derived so far in order to derive error bounds for the truncated steady state probability $\mathbb{P}\big(X^{(N)}_\infty \geq x, Y^{(N)}_\infty \geq y\big)$.

5.3.2 Error bounds

In this section, we provide error bounds for the steady state probability of the truncated system, i.e. the probability $\mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$. From Theorem 5.1, and more precisely from Eq. (5.8), we obtain that the truncated probability is always underestimating the exact probability $\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big)$. To construct an upper bound for the difference $\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) - \mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$, we combine the results for the upper bounds of terms $\mathbb{I}$ and $\mathbb{III}$ (see pages 126 and 128, respectively) with Eq. (5.9) and we get

$$0 \leq \mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) - \mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big) \leq \mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big] \frac{\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)}{\mathbb{E}T_{(0,0)}}. \tag{5.19}$$

The error between the exact steady state probability and its truncated approximation is upper bounded by $\mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big]\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)/\mathbb{E}T_{(0,0)}$ according to Eq. (5.19). All factors involved in the upper bound are hard to evaluate exactly. Instead, we derive an asymptotic upper bound by examining the behaviour of the factors $\mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big]$ and $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ as $N \to \infty$.

To study the asymptotic behaviour of the factors $\mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big]$ and $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$, an exponential change of measure is first required. Therefore, in Section 5.4, we perform such an exponential change of measure. Afterwords, in Section 5.5, we provide asymptotic results for the probability $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$, where the latter is treated in conjunction with the factor $\mathbb{E}T_{(0,0)}$. Asymptotic results for the conditional expectation $\mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big]$ are derived in Section 5.6. The expression for the asymptotic upper bound is then formulated in Eq. (5.44).

Remark 5.2. The probabilities $\mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$ can be calculated numerically with the aid of MAM by interpreting the truncated Markov chain as a QBD. As we explained in Chapter 1, algorithms to solve for the stationary distribution of a QBD exist if the phase space is finite. Therefore, for the numerical implementation of existing algorithms (Ramaswami and Latouche, 1986; Latouche and Ramaswami, 1999), we conveniently interchange the coordinates of the state space. In other words, the states now have the form (m_2, m_1), where m_2 is the number of customers in the second queue and m_1 is the number of customers in the fist queue. In this case, the phase coordinate takes values $m_1 \in \{0, 1, \ldots, N\}$, and consequently the $(N+1)$-dimensional levels $l(m_2)$ take the form $l(m_2) = \{(m_2, 0), (m_2, 1), \ldots, (m_2, N)\}$, $m_2 = 0, 1, \ldots$

5.4 Exponential change of measure

Recall that the conditional expectation $\mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big]$ and the probability $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ involved in the upper bound, which is given in Eq. (5.19), relate to the event that the number of customers in the first queue exceeds the truncation level during a cycle. According to *Large Deviations Theory* (LDT) (Shwartz and Weiss, 1995), the number of customers in the first queue reaches a very high level due to a 'conspiracy', where the arrival and service rates change so as to make the probability

of crossing level N a more likely event. Thus, in order to estimate these two terms, we need to perform first an exponential change of measure.

To perform an exponential change of measure, we define as $\kappa(\alpha)$ the *cumulant generating function* (c.g.f.) of the r.v.'s $Z_1, Z_2, \ldots$ Then, the c.g.f. $\kappa(\alpha)$ takes the form

$$\kappa(\alpha) = \ln \mathbb{E}e^{\alpha Z_1} = \ln\left(\mu_2 + \mu_1 e^{-\alpha} + \lambda \mathbb{E}e^{\alpha B}\right) = \ln\left(\mu_2 + \mu_1 e^{-\alpha} + \lambda M_B(\alpha)\right),$$

where $M_B(\alpha)$ is the *moment generating function* (m.g.f.) of the batch sizes. By differentiating with respect to α, it is ease to verify that $\kappa'(0) = \lambda \mathbb{E}B - \mu_1 < 0$. We assume now that there exists a solution $\gamma > 0$ to the *Lundberg equation* $\kappa(\gamma) = 0$. The parameter γ is called the *adjustment coefficient* and conditions for its existence can be found in Asmussen and Albrecher (2010, page 91).

If F is the distribution of the $Z \overset{\mathfrak{D}}{=} Z_n$, we define $\breve{F}$ to be the probability distribution with density $e^{\gamma x}$ w.r.t. F, i.e. $\breve{F}(dx) = e^{\gamma x} F(dx)$ (obvious notations like $\breve{\kappa}(\alpha)$, $\breve{\mathbb{P}}$, $\breve{\mathbb{E}}$, etc, are used for quantities under the exponential change of measure). In the next theorem, we show that, under this exponential change of measure, the r.v. Z has a positive mean, thus making the event of crossing the level N in the first queue more probable.

Theorem 5.3. *Under the probability measure $\breve{\mathbb{P}}$ the arrival rate of the batches is equal to $\breve{\lambda} = \lambda + (1 - e^{-\gamma})\mu_1$, the batch size distribution is equal to*

$$\breve{\mathbb{P}}(B = n) = \frac{e^{\gamma n}}{M_B(\gamma)} \mathbb{P}(B = n), \quad n = 1, 2, \ldots, \tag{5.20}$$

and the customers are served with rates $\breve{\mu}_1 = e^{-\gamma}\mu_1$ and $\breve{\mu}_2 = \mu_2$ in each server, respectively. In addition, it holds that $\breve{\mathbb{E}}Z = \breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1 > 0$.

Proof. According to Asmussen (2003, Proposition XIII.1.1) the c.g.f. of $\breve{F}$ is equal to

$$\begin{aligned}
\breve{\kappa}(\alpha) &= \kappa(\alpha + \gamma) - \kappa(\gamma) = \ln\left(\mu_2 + \mu_1 e^{-\alpha-\gamma} + \lambda M_B(\alpha + \gamma)\right) \\
&= \ln\left(\mu_2 + e^{-\gamma}\mu_1 e^{-\alpha} + \lambda M_B(\gamma)\frac{M_B(\alpha+\gamma)}{M_B(\gamma)}\right) \\
&= \ln\left(\mu_2 + e^{-\gamma}\mu_1 e^{-\alpha} + \lambda M_B(\gamma)\breve{M}_B(\alpha)\right).
\end{aligned}$$

By substituting in the above relation $\alpha = 0$, we obtain that

$$\breve{\kappa}(0) = 0 \quad \Rightarrow \quad \mu_2 + e^{-\gamma}\mu_1 + \lambda M_B(\gamma) = 1 \quad \Rightarrow \quad \lambda M_B(\gamma) = \lambda + (1 - e^{-\gamma})\mu_1.$$

Therefore,

$$\breve{\kappa}(\alpha) = \ln\left(\mu_2 + e^{-\gamma}\mu_1 e^{-\alpha} + \left(\lambda + (1 - e^{-\gamma})\mu_1\right)\breve{M}_B(\alpha)\right),$$

which shows that the arrival rate of the batches under $\breve{\mathbb{P}}$ is equal to $\breve{\lambda} = \lambda + (1 - e^{-\gamma})\mu_1$, the customers are served with rates $\breve{\mu}_1 = e^{-\gamma}\mu_1$ and $\breve{\mu}_2 = \mu_2$ in each queue, respectively, and the density function of the batch sizes is given by Eq. (5.20).

As a last step, we need to show that $\breve{\mathbb{E}}Z = \breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1 > 0$. First observe that since $\kappa(\alpha)$ is a strictly convex function and its slope at 0 is negative, this means that

$\kappa'(\gamma) > 0$. Now, according to Asmussen (2003, Proposition XIII.1.1), it holds that $\breve{\mathbb{E}}Z = \kappa'(\gamma) = \breve{\kappa}'(0)$. We calculate,

$$\breve{\kappa}'(\alpha) = \frac{-\breve{\mu}_1 e^{-\alpha} + \breve{\lambda}\frac{d}{d\alpha}\breve{M}_B(\alpha)}{\breve{\mu}_2 + \breve{\mu}_1 e^{-\alpha} + \breve{\lambda}\breve{M}_B(\alpha)},$$

where the substitution of $\alpha = 0$ in the above formula gives $\breve{\kappa}'(0) = \breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1$, and the result is immediate. □

Remark 5.4. In the special case that the batches are of size 1, the equation $\kappa(\gamma) = 0$ simplifies to $e^{-\gamma} = \lambda/\mu_1$. Thus, we find $\breve{\lambda} = \mu_1$ and $\breve{\mu}_1 = \lambda$, which means that the service and arrival rates are interchanged.

With the aid of the above change of measure, our goal is to provide estimates for $\mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big]$ and $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$. To achieve our objective, we define a random walk with increments Z_n. As we shall see in Section 5.5, the event $\{M^{T_{(0,0)}} \geq N\}$ connects with the first passage time of the random walk above the truncation level N. We proceed with introducing the notation for the random walk in the next section.

5.4.1 Random walk notation

We define the random walk $U_n = Z_1 + \cdots + Z_n$, with $U_0 = 0$. It is easy to see that the Lindley process $\{X_n\}$ has the same transition mechanism as the random walk $\{U_n\}$ except when the random walk crosses from positive to negative values (the Lindley process then stays at 0). With the aid of the random walk $\{U_n\}$ we derive our results. Inspired by Asmussen (2003, page 221), we use the following notation:

τ_+ the first (strict) *ascending ladder epoch* or the *entrance time* to $(0, \infty)$; namely, $\tau_+ = \inf\{n \geq 1 : U_n > 0\}$. The distribution of τ_+ may be defective, i.e. $\mathbb{P}(\tau_+ = \infty) = \mathbb{P}(U_n \leq 0 \text{ for all } n \geq 1) > 0$.

U_{τ_+} the first (strict) *ascending ladder height* (defined on $\{\tau_+ < \infty\}$ only).

H_+ the (strict) *ascending ladder height distribution* $H_+(x) = \mathbb{P}(U_{\tau_+} \leq x)$. Here H_+ is concentrated on $(0, \infty)$ and may be defective, i.e. $\|H_+\| = \mathbb{P}(\tau_+ < \infty) < 1$.

τ_- the first (weak) *descending ladder epoch* or the entrance time to $(\infty, 0]$; namely, $\tau_- = \inf\{n \geq 1 : U_n \leq 0\}$.

U_{τ_-} the first (weak) *descending ladder height* (defined on $\{\tau_- < \infty\}$ only).

H_- the (weak) *descending ladder height distribution* $H_-(x) = \mathbb{P}(U_{\tau_-} \leq x)$. Here H_- is concentrated on $(-\infty, 0]$ and may be defective, i.e. $\|H_-\| = \mathbb{P}(\tau_- < \infty) < 1$.

$\tau(N)$ the time $\inf\{n \geq 1 : U_n > N\}$ of *first passage* to level $N \geq 0$ or the entrance to (N, ∞). The distribution of $\tau(N)$ may be defective. Clearly, $\tau(0) = \tau_+$.

$B(N)$ the *overshoot* $U_{\tau(N)} - N$. Clearly, $B(0)$ is the ascending ladder height U_{τ_+}. Moreover, since $U_{\tau(N)}$ is strictly greater than N, the overshoot $B(N)$ can only take positive values.

$B(\infty)$ a r.v. having the limiting distribution (if it exists) of $B(N)$.

We also define as $\tau_+(n)$ the ladder epoch at which the nth record $U_{\tau_+(n)}$ is achieved. Formally,

$$\tau_+(n+1) = \inf\{k > \tau_+(n) : U_k > U_{\tau_+(n)}\}. \tag{5.21}$$

In this definition, we use the convention that $\tau_+(0) = 0$ and consequently we have $U_{\tau_+(0)} = 0$. Finally, if we let $\mathfrak{F}_n = \sigma(Z_1, \ldots, Z_n)$ be the σ-algebra generated by $Z_1, \ldots, Z_n$, then $\mathfrak{F}_{\tau(N)}$ follows the usual definition of a stopping time σ-algebra.

Since the results of Sections 5.5 and 5.6 are based on the connection between the probability space $\mathbb{P}$ with $\breve{\mathbb{P}}$, we provide in Appendix A.3 some background theory on the latter connection.

5.5 Asymptotic approximation for the maximum

In this section, we derive an asymptotic approximation for $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$, with the aid of extreme value theory. Observe that the number of customers in the first queue $\{X_n\}_{n=0,1,\ldots}$ forms a one-dimensional Markov chain on its own. Therefore, we denote as $T_0 = \inf\{n \geq 1 : X_n = 0 \mid X_0 = 0\}$ the return time to the origin of the first queue only and we define $M^{T_0} = \max_{1 \leq n \leq T_0} X_n$. We show that the probability $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ exhibits a similar tail behaviour with the probability $\mathbb{P}\big(M^{T_0} \geq N\big)$. Thus, we first discuss the behaviour of $\mathbb{P}\big(M^{T_0} \geq N\big)$ as $N \to \infty$.

We define $\tau_1 = \inf\{n : X_n \geq N\}$. Observe that if the maximum M^{T_0} before the first return time to the origin is greater than or equal to N, this means that τ_1 is smaller than T_0. In other words, $\mathbb{P}\big(M^{T_0} \geq N\big) = \mathbb{P}(\tau_1 < T_0)$. Moreover, the Lindley process X_n has the same transition mechanism as the random walk U_n until T_0 because X_n does not hit zero before T_0. Thus, it also holds that $\{\tau_1 < T_0\} = \{\tau(N-1) < \tau_-\}$, and consequently $\mathbb{P}\big(M^{T_0} \geq N\big) = \mathbb{P}(\tau(N-1) < \tau_-)$. For the latter probability, a variant of the *Cramér-Lundberg approximation* is already known by Asmussen (2003, Corrolary XIII.5.9). Therefore, we provide the following lemma without proof.

Lemma 5.5. *If $B(N)$ converges in $\breve{\mathbb{P}}$ as $N \to \infty$, say to $B(\infty)$, then*

$$e^{\gamma(N-1)}\mathbb{P}\big(M^{T_0} \geq N\big) = \breve{\mathbb{E}}e^{-\gamma B(N-1)}\mathbb{1}\left(\tau_1 < T_0\right) \to C_1,$$

where $C_1 = \breve{\mathbb{P}}(\tau_- = \infty)C_0$ and $C_0 = \breve{\mathbb{E}}e^{-\gamma B(\infty)}$.

We continue with showing that the tail behaviour of $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ is similar to the tail behaviour of $\mathbb{P}\big(M^{T_0} \geq N\big)$. For this purpose, note that both $T_{(0,0)}$ and T_0 are regeneration cycles for the Markov chain X_n. Thus, if we denote $M_i^{T_0} \overset{\mathfrak{D}}{=} M^{T_0}$ as the maximum of X_n in the ith cycle T_0, where M^{T_0} is the generic cycle maximum, and similarly $M_i^{T_{(0,0)}} \overset{\mathfrak{D}}{=} M^{T_{(0,0)}}$ as the maximum of X_n in the ith cycle $T_{(0,0)}$, we have that (Iglehart, 1972; Rootzén, 1988; Asmussen, 1998)

$$\max_{i=1,\ldots,\frac{n}{\mathbb{E}T_{(0,0)}}} M_i^{T_{(0,0)}} \approx \max_{i=1,\ldots,n} X_i \approx \max_{i=1,\ldots,\frac{n}{\mathbb{E}T_0}} M_i^{T_0}. \tag{5.22}$$

From Lemma 5.5, we know the tail behaviour of M^{T_0}. Therefore, we can derive asymptotics for the maximum $\max_{i=1,\dots,n} X_i$. As such, Eq. (5.22) indicates that in order to study the asymptotic behaviour of $M^{T_{(0,0)}}$, we first need to study the asymptotics of $\max_{i=1,\dots,n} X_i$, as $n \to \infty$.

Classically, extreme value theory focuses on finding constants a_n, b_n, such that

$$\frac{\max_{i=1,\dots,n} X_i - a_n}{b_n} \xrightarrow{\mathfrak{D}} H, \tag{5.23}$$

where H is some non-degenerate r.v. and $\xrightarrow{\mathfrak{D}}$ denotes convergence in distribution. This is equivalent to showing that the probability $\mathbb{P}\big(\max_{i=1,\dots,n} X_i \le a_n x + b_n\big)$ has a limit, for any x. In our case, we prove that given the tail behaviour of M^{T_0} from Lemma 5.5, there exist constants a_n, b_n, such that (5.23) holds with H following the *Gumbel* function $\Lambda(x) = e^{-e^{-x}}$, $x \in \mathbb{R}$ (Gümbel, 1958).

The asymptotic behaviour of the probability $\mathbb{P}\big(M^{T_{(0,0)}} \ge N\big)$ is given in the following theorem. To establish this asymptotic result, we use Eq. (5.22) to first derive the asymptotics of $\max_{i=1,\dots,n} X_i$, as $n \to \infty$, and later connect these asymptotics with the probability $\mathbb{P}\big(M^{T_{(0,0)}} \ge N\big)$.

Theorem 5.6. *It holds that*

$$\mathbb{P}\big(M^{T_{(0,0)}} \ge N\big) \sim \frac{\mathbb{E}T_{(0,0)}}{\mathbb{E}T_0} C_1 e^{-\gamma(N-1)}, \qquad N \to \infty,$$

where C_1 is defined in Lemma 5.5.

Proof. We first find the asymptotics for $\max_{i=1,\dots,n} X_i$. We set $b_n = 1/\gamma$ and $a_n = \big(\ln(n/\mathbb{E}T_0) + \ln C_1\big)/\gamma$. Thus, from Asmussen (1998, Lemma 1.1), we obtain that as $n \to \infty$

$$\begin{aligned}
\mathbb{P}\Big(\max_{i=1,\dots,n} X_i \le a_n + b_n x\Big) &= \mathbb{P}\big(M^{T_0} \le a_n + b_n x\big)^{\frac{n}{\mathbb{E}T_0}} \\
&= \exp\left\{\frac{n}{\mathbb{E}T_0} \ln \mathbb{P}\big(M^{T_0} \le a_n + b_n x\big)\right\} \\
&= \exp\left\{\frac{n}{\mathbb{E}T_0} \ln\Big(1 - \mathbb{P}\big(M^{T_0} > a_n + b_n x\big)\Big)\right\} \\
&\approx \exp\left\{-\frac{n}{\mathbb{E}T_0} C_1 e^{-\gamma(a_n + b_n x)}\big(1 + o(1)\big)\right\} \\
&\to e^{-e^{-x}}, \quad n \to \infty,
\end{aligned}$$

where at the third step we used $\mathbb{P}\big(M^{T_0} \ge N\big) = C_1 e^{-\gamma(N-1)}\big(1+o(1)\big)$ and $\ln(1-x) \approx x$. From Asmussen (1998, Lemma 1.1), we also have that as $n \to \infty$

$$\begin{aligned}
&\mathbb{P}\big(M^{T_{(0,0)}} \le a_n + b_n x\big)^{\frac{n}{\mathbb{E}T_{(0,0)}}} = \mathbb{P}\Big(\max_{i=1,\dots,n} X_i \le a_n + b_n x\Big) \\
\Rightarrow\ &\mathbb{P}\big(M^{T_{(0,0)}} \le a_n + b_n x\big)^{\frac{n}{\mathbb{E}T_{(0,0)}}} = e^{-e^{-x}} \\
\Rightarrow\ &\frac{n}{\mathbb{E}T_{(0,0)}} \ln \mathbb{P}\big(M^{T_{(0,0)}} \le a_n + b_n x\big) = -e^{-x}
\end{aligned}$$

$$\Rightarrow \frac{n}{\mathbb{E}T_{(0,0)}} \ln\Big(1-\mathbb{P}\big(M^{T_{(0,0)}} > a_n + b_n x\big)\Big) = -e^{-x}$$
$$\Rightarrow \frac{n}{\mathbb{E}T_{(0,0)}} \mathbb{P}\big(M^{T_{(0,0)}} > a_n + b_n x\big) = e^{-x}\big(1+o(1)\big)$$
$$\Rightarrow \mathbb{P}\big(M^{T_{(0,0)}} > a_n + b_n x\big) = \frac{\mathbb{E}T_{(0,0)}}{n} e^{-x}\big(1+o(1)\big).$$

By setting now $y = a_n + b_n x$, we write

$$\mathbb{P}\big(M^{T_{(0,0)}} > y\big) = \frac{\mathbb{E}T_{(0,0)}}{n} e^{\frac{-y+a_n}{b_n}}\big(1+o(1)\big) = \frac{\mathbb{E}T_{(0,0)}}{n} e^{-\gamma y} e^{\ln n \ln \frac{C_1}{\mathbb{E}T_0}}\big(1+o(1)\big)$$
$$= \frac{\mathbb{E}T_{(0,0)} C_1}{\mathbb{E}T_0} e^{-\gamma y}\big(1+o(1)\big), \quad y \to \infty,$$

which completes the proof. □

Remark 5.7. Practically, Theorem 5.6 states that, as $N \to \infty$,

$$\frac{\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)}{\mathbb{E}T_{(0,0)}} \sim \frac{\mathbb{P}\big(M^{T_0} \geq N\big)}{\mathbb{E}T_0}.$$

Therefore, if we replace $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)/\mathbb{E}T_{(0,0)}$ with $\mathbb{P}\big(M^{T_0} \geq N\big)/\mathbb{E}T_0$ in Eq. (5.19), then we can derive an asymptotic bound, where the constant term $\mathbb{E}T_0$ is computable.

At this point, observe that the limiting distribution of the overshoot $B(\infty)$ is required for the evaluation of the constants C_0 and C_1, which were defined in Lemma 5.5. Since, the distribution $B(\infty)$ is found in Lemma A.17, we can find explicit expressions for these constants, by using properties of lattice random walks. Thus, we conclude this section by providing explicit expressions for the constants C_0 and C_1. We also calculate the mean $\mathbb{E}T_0$.

Evaluation of the constants C_0 and C_1

For the evaluation of the constants C_0 and C_1 we need the limiting distribution of the overshoot $B(\infty)$. According to Lemma A.17, the distribution of $B(\infty)$ is found through the ladder height distribution under the probability measure $\breve{\mathbb{P}}$. For this reason, let $\breve{H}_+$ be the distribution function of the ascending ladder height with respect to $\breve{\mathbb{P}}$ and $\breve{l}_+$ be its corresponding mean. We have the following result.

Proposition 5.8. *For a discrete-time lattice random walk, $B(\infty)$ exists with respect to $\breve{\mathbb{P}}$. In this case, C_0 is given in terms of the ladder height distributions by*

$$C_0 = \breve{\mathbb{E}} e^{-\gamma B(\infty)} = \frac{\big(1-\|H_+\|\big)\big(1-\big\|\breve{H}_-\big\|\big)}{(e^{\gamma}-1)\kappa'(\gamma)}.$$

Proof. By using Proposition 5.8 and Lemma A.17, we obtain that

$$C_0 = \breve{\mathbb{E}} e^{-\gamma B(\infty)} = \sum_{n=1}^{+\infty} e^{-\gamma n} \breve{\mathbb{P}}(B(\infty)=n) = \frac{1}{\breve{l}_+} \sum_{n=1}^{+\infty} e^{-\gamma n}\big(1-\breve{H}_+(n-1)\big)$$

$$= \frac{1}{\breve{l}_+} \sum_{n=1}^{+\infty} e^{-\gamma n} \sum_{k=n}^{+\infty} \breve{h}_+(k) = \frac{1}{\breve{l}_+} \sum_{k=1}^{+\infty} \breve{h}_+(k) \sum_{n=1}^{k} e^{-\gamma n}$$

$$= \frac{1}{\breve{l}_+(e^{\gamma}-1)} \sum_{k=1}^{+\infty} \breve{h}_+(k)(1-e^{-\gamma k}) = \frac{1}{\breve{l}_+(e^{\gamma}-1)} \Big(\big\| \breve{H}_+ \big\| - \sum_{k=1}^{+\infty} e^{-\gamma k} \breve{h}_+(k) \Big)$$

$$= \frac{1}{\breve{l}_+(e^{\gamma}-1)} \Big(1 - \sum_{k=1}^{+\infty} h_+(k) \Big) = \frac{1}{\breve{l}_+(e^{\gamma}-1)} (1 - \|H_+\|),$$

where we interchanged the infinite summations due to *Tonelli's theorem* (Tonelli, 1909), and the last equality holds because of Eq. (A.10). Finally, we need to calculate $\breve{l}_+$. From *Wald's equation* (Wald, 1944) we have that

$$\breve{l}_+ = \breve{\mathbb{E}}\tau_+ \breve{\mathbb{E}} Z.$$

But, from Asmussen (2003, Theorem VIII.2.4(ii)), we know $\breve{\mathbb{E}}\tau_+ = \big(1 - \big\| \breve{H}_- \big\|\big)^{-1}$. Also, recall from Theorem 5.3 that $\breve{\mathbb{E}} Z = \kappa'(\gamma)$. Combining all these, the result is immediate. □

Lemma 5.9. *For a downward skip-free (or left-continuous) random walk, the constant C_1 in Lemma 5.5 is equal to*

$$C_1 = -\frac{\mathbb{E}Z}{\breve{\mathbb{E}} Z}(1-e^{-\gamma})e^{-\gamma}\mu_1 = -\frac{\kappa'(0)}{\kappa'(\gamma)}(1-e^{-\gamma})e^{-\gamma}\mu_1.$$

Proof. From Proposition 5.8, it is evident that we need to find exact values for the terms $1 - \|H_+\|$ and $1 - \big\| \breve{H}_- \big\|$. Observe that the random walk U_n is downward skip-free.

We start with the evaluation of the term $1 - \|H_+\|$. We set $f_n = \mathbb{P}(Z = n)$. Under the probability measure $\mathbb{P}$, it holds that $\mathbb{E}Z = \kappa'(0) < 0$. Therefore, according to Asmussen (2003, Corollary VIII.5.6), $\|H_+\| = 1 + \mathbb{E}Z/f_{-1}$, where from Eq. (5.2) we know that $f_{-1} = \mathbb{P}(Z = -1) = \mu_1$.

By the definition of the descending ladder height distribution, we have that

$$1 - \big\| \breve{H}_- \big\| = \breve{\mathbb{P}}(\tau_- = \infty) = \breve{\mathbb{P}}(U_n \geq 1 \text{ for all } n \geq 1).$$

We set now $\breve{f}_n = \breve{\mathbb{P}}(Z = n)$ and $T_1 = \inf\{n : U_n = -1\}$. Since U_n is a downward skip-free random walk with an upward drift under the probability measure $\breve{\mathbb{P}}$, it holds from Brown et al. (2010, Proposition 11) that

$$1 - \big\| \breve{H}_- \big\| = \breve{f}_{-1} \cdot \frac{1 - \breve{\mathbb{P}}(T_1 < \infty)}{\breve{\mathbb{P}}(T_1 < \infty)}.$$

Thus, it is left to find the probability $\breve{\mathbb{P}}(T_1 < \infty)$, which according to Brown et al. (2010, Lemma 2) is equal to the unique value $s \in (0,1)$ that satisfies the equation $\breve{\mathbb{E}} s^Z = 1$. From Theorem 5.3, we know that $\breve{\mathbb{E}} e^{\alpha Z_1} = e^{\kappa(\alpha+\gamma)}$. Therefore, $\breve{\mathbb{E}} e^{-\gamma Z} = e^{\kappa(0)} = 1$, and consequently $s = e^{-\gamma} \in (0,1)$ is the unique solution to the equation $\breve{\mathbb{E}} s^Z = 1$. As a result, $\breve{\mathbb{P}}(T_1 < \infty) = e^{-\gamma}$. We also find $\breve{f}_{-1} = \breve{\mathbb{P}}(Z = -1) = e^{-\gamma}\mu_1$. Combining all the above and Lemma 5.5, the result is immediate. □

The mean return time $\mathbb{E}T_0$

We turn now our attention to the evaluation of the mean return time $\mathbb{E}T_0$. Observe that the Markov chain X_n is ergodic, because $\lambda\mathbb{E}B < \mu_1$. Therefore, as we already saw in the proof of Theorem 5.1, the ergodicity of X_n results in $\mathbb{E}T_0 = 1/\mathbb{P}\big(X_\infty = 0\big)$. Since X_n corresponds to an $\mathrm{M}^X/\mathrm{M}/1$ queue, $\mathbb{P}\big(X_\infty = 0\big)$ is the probability that the queue will be empty in the long-run. By applying *Little's formula* (Little, 1961), we find that the fraction of time the server is busy (and consequently the queue is not empty) is equal to $\rho_1 = \lambda\mathbb{E}B/\mu_1$, with $\lambda\mathbb{E}B$ being the average number of customers entering the system per unit time. Consequently, $\mathbb{P}\big(X_\infty = 0\big) = 1 - \rho_1 = 1 - \lambda\mathbb{E}B/\mu_1$. Thus, we proved the following lemma.

Lemma 5.10. *The mean return time $\mathbb{E}T_0$ is calculated by the formula*

$$\mathbb{E}T_0 = \frac{1}{1 - \lambda\mathbb{E}B/\mu_1}.$$

Remark 5.11. In this section, we provided an asymptotic approximation for the probability $\mathbb{P}\big(M^{T_0} \geq N\big)$. Alternatively, the same probability could be found numerically by solving a system of linear equations for first passage probabilities. More precisely, if we define $q_i = \mathbb{P}\big(X_n \text{ hits } N \text{ before it hits } 0 \;\big|\; X_0 = i\big)$, then we get by first step analysis a system of linear equations where q_0 is the desired probability.

5.6 Asymptotics for the conditional mean return time

The goal of this section is to study the asymptotic behaviour of the conditional mean $\mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big]$, i.e. to study the limit $\lim_{N\to\infty} \frac{1}{N}\mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big]$. As we pointed out in Section 5.1, we recognise three different scenarios for this conditional expectation. The analysis involved in the estimation of its asymptotic behaviour for all three cases requires a variety of techniques and is thus complicated. Therefore, in order to increase the readability of this section, we split it in two parts: the *intuitive* and the *rigorous* part.

In the intuitive part, first we explain how the three different scenarios for the asymptotic behaviour appear. We use graphs as a visual aid to explain our arguments and we provide estimates for the asymptotic behaviour. The intuitive part is covered in Section 5.6.1. Finally, in Section 5.6.2, the rigorous part provides the mathematical proofs of the findings in Section 5.6.1.

Before continuing with our analysis, we introduce the following notation:

τ_1: for the time at which the first queue reaches or exceeds level N. Recall that is was defined in Section 5.5 as $\tau_1 = \inf\{n : X_n \geq N\}$.

τ_2: for the return time to 0 in the first queue after reaching its maximum value.

τ_3: for the first time the second queue empties after the first queue reaches its maximum value. The time τ_3 can either coincide with or happen before $T_{(0,0)}$.

5.6.1 Intuitive illustrations

In this section, our goal is to describe the behaviour of both queues, given that the number of customers in the first queue has reached a very high level before the first return time $T_{(0,0)}$ to the empty state $(0,0)$. Our description is based on intuition and common sense. Therefore, neither the graphs nor the notation we use in this section are very precise. We explain these conventions in the following lines.

With respect to the graphs, we want to point out that they depict only qualitative behaviours between the slopes but not quantitative ones. For example, in Figure 5.4, the difference in the slopes $\breve{\mu}_1 - \mu_2$ and $\mu_1 - \mu_2$ indicates that the former is smaller than the latter, but it does not indicate how much smaller it is. With respect to the notation, we write $a \approx b$ to denote that a is approximately equal to b, without explicitly determining the degree of accuracy. Finally, we denote as $\#Q_1$ and $\#Q_2$ the number of customers in the first and the second queue, respectively.

The behaviour of the first queue

Observe that the behaviour of the fist queue is not affected by what happens in the second queue. Therefore, we describe in Figure 5.1 the behaviour of the first queue until time $T_{(0,0)}$, given that $\#Q_1$ reached or exceeded level N.

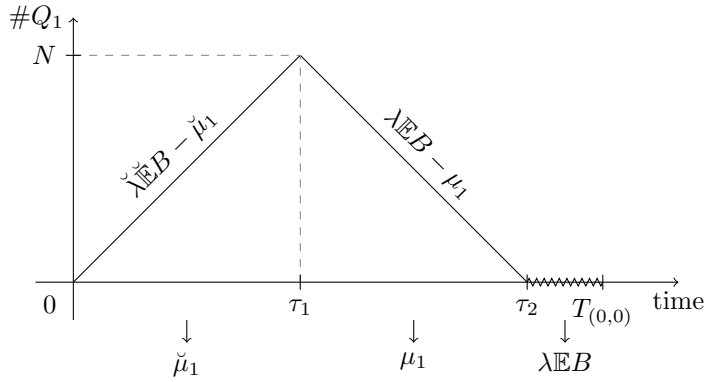


FIGURE 5.1: The asymptotic behaviour of Q_1, given that the number of customers before the first return time $T_{(0,0)}$ has reached (and possibly exceeded) the truncation level N.

We have the following observations. Since $N \to \infty$, the time its takes the queue from τ_1 to reach its maximum value (something above N) before $T_{(0,0)}$ is negligible (compared to τ_1). Moreover, until τ_1, the departure rate of the customers is asymptotically equal to $\breve{\mu}_1$ because the system is overloaded ($\breve{\lambda}\breve{\mathbb{E}}B > \breve{\mu}_1$). On the other hand, after τ_1 all the rates are back to normal. As we have already mentioned, τ_2 is the point at which the first queue reaches 0 after reaching its maximum value within cycle $T_{(0,0)}$. Since during the time interval $[\tau_1, \tau_2]$ the first queue is always full, the departure rate of customers is equal to μ_1.

Next, we describe the behaviour of the second queue before time $T_{(0,0)}$. To do so, we recognise three different cases that arise from the relation between the rates μ_1, μ_2, and $\breve{\mu}_1$.

Case 1: $\mu_1 < \mu_2$

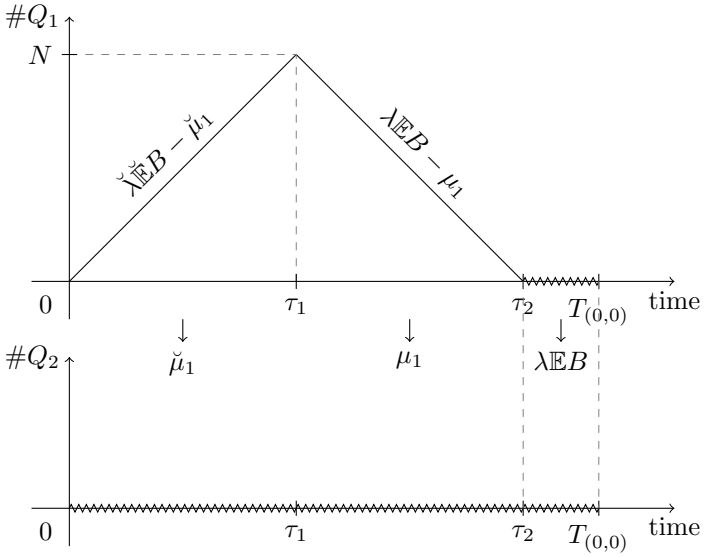


FIGURE 5.2: The asymptotic behaviours of Q_1 and Q_2, given that the number of customers in Q_1 before the first return time $T_{(0,0)}$ has reached (and possibly exceeded) the truncation level N, when $\mu_1 < \mu_2$.

From Theorem 5.3, it always holds that $\breve{\mu}_1 < \mu_1$. Therefore, in this case, the second queue behaves asymptotically as a stable M/M/1 queue in all time intervals (but with different arrival rates of customers). Thus, the number of customers in the second queue is bounded by the number of customers in a stable M/M/1 queue until $T_{(0,0)}$; see Figure 5.2. As a consequence, the time interval $[\tau_2, T_{(0,0)}]$ is negligible compared to $[0, \tau_2]$ and we expect that $\mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big] \approx \mathbb{E}\big[\tau_2 \;\big|\; M^{T_{(0,0)}} \geq N\big]$, where from Euclidean geometry we can easily verify that

$$\tau_1 \approx \frac{N}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1}, \qquad \tau_2 - \tau_1 \approx \frac{N}{\mu_1 - \lambda\mathbb{E}B}. \tag{5.24}$$

Case 2: $\breve{\mu}_1 < \mu_2 < \mu_1$

Since $\breve{\mu}_1 < \mu_2$, the second queue behaves asymptotically as a stable M/M/1 queue with arrival rate $\breve{\mu}_1$ and service rate μ_2 until time τ_1; see Figure 5.3. This means that the number of customers in the second queue at time τ_1 is bounded by the number of customers in the latter M/M/1 queue. From τ_1 onwards, the arrival rate of customers in the second queue is equal to μ_1, which is greater than the service rate μ_2. Therefore, the number of customers in the second queue grows linearly with rate $\mu_1 - \mu_2$ up until τ_2. After τ_2, the output rate from the first queue is equal to $\lambda\mathbb{E}B$ and the customers in the second queue reduce linearly with rate $\lambda\mathbb{E}B - \mu_2$ until the queue empties at

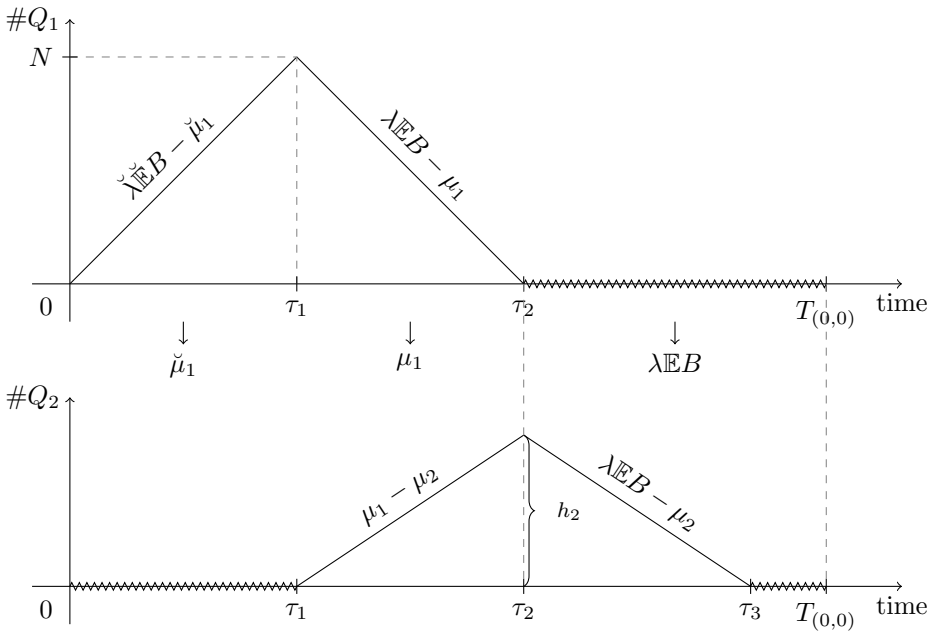


FIGURE 5.3: The asymptotic behaviours of Q_1 and Q_2, given that the number of customers in Q_1 before the first return time $T_{(0,0)}$ has reached (and possibly exceeded) the truncation level N, when $\breve{\mu}_1 < \mu_2 < \mu_1$.

time τ_3. We calculate

$$\begin{aligned} h_2 &\approx (\mu_1 - \mu_2)\frac{N}{\mu_1 - \lambda\mathbb{E}B}, \\ \tau_3 - \tau_2 &\approx \frac{h_2}{\mu_2 - \lambda\mathbb{E}B} = \frac{\mu_1 - \mu_2}{\mu_2 - \lambda\mathbb{E}B} \cdot \frac{N}{\mu_1 - \lambda\mathbb{E}B}. \end{aligned} \tag{5.25}$$

Obviously, in this case $\mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big] \approx \mathbb{E}\big[\tau_3 \;\big|\; M^{T_{(0,0)}} \geq N\big]$, because the time interval $[\tau_3, T_{(0,0)}]$ is negligible compared to $[0, \tau_3]$.

Case 3: $\mu_2 < \breve{\mu}_1 < \mu_1$

Since $\breve{\mu}_1 > \mu_2$, the number of customers in the second queue grows linearly with rate $\breve{\mu}_1 - \mu_2$ up until time τ_1; see Figure 5.4. For the remaining time intervals, the second queue behaves in a similar manner as in Case 2. Therefore, $\mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big] \approx \mathbb{E}\big[\tau_3 \;\big|\; M^{T_{(0,0)}} \geq N\big]$, where

$$\begin{aligned} h_1 &\approx (\breve{\mu}_1 - \mu_2)\frac{N}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1}, \\ h_2 &\approx h_1 + (\mu_1 - \mu_2)\frac{N}{\mu_1 - \lambda\mathbb{E}B} = N \cdot \left(\frac{\breve{\mu}_1 - \mu_2}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1} + \frac{\mu_1 - \mu_2}{\mu_1 - \lambda\mathbb{E}B}\right), \end{aligned} \tag{5.26}$$

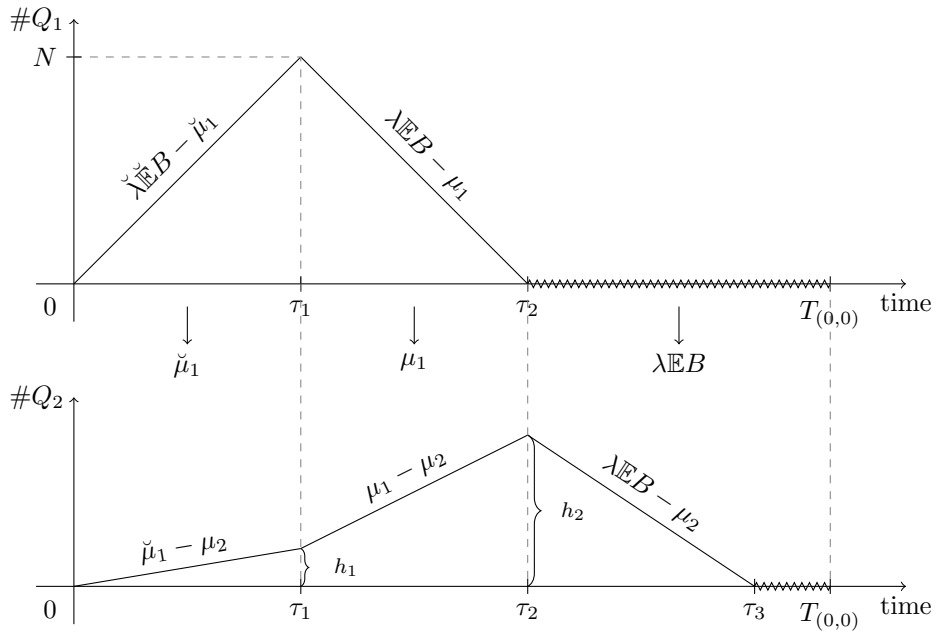


FIGURE 5.4: The asymptotic behaviours of Q_1 and Q_2, given that the number of customers in Q_1 before the first return time $T_{(0,0)}$ has reached (and possibly exceeded) the truncation level N, when $\mu_2 < \breve{\mu}_1 < \mu_1$.

$$\tau_3 - \tau_2 \approx \frac{h_2}{\mu_2 - \lambda \mathbb{E}B} = \frac{N}{\mu_2 - \lambda \mathbb{E}B} \cdot \left(\frac{\breve{\mu}_1 - \mu_2}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1} + \frac{\mu_1 - \mu_2}{\mu_1 - \lambda \mathbb{E}B} \right).$$

We now proceed with making the above results rigorous in the next section.

5.6.2 Rigorous proofs

In this section, we prove rigorously the results presented earlier in an intuitive way. From Section 5.6.1, we understand that we need to find estimates for the conditional expectations $\mathbb{E}\big[\tau_1 \;\big|\; M^{T_{(0,0)}} \geq N\big]$, $\mathbb{E}\big[\tau_2 - \tau_1 \;\big|\; M^{T_{(0,0)}} \geq N\big]$, $\mathbb{E}\big[\tau_3 - \tau_2 \;\big|\; M^{T_{(0,0)}} \geq N\big]$, and $\mathbb{E}\big[T_{(0,0)} - \tau_3 \;\big|\; M^{T_{(0,0)}} \geq N\big]$. Thus, we split the time interval $[0, T_{(0,0)}]$ in the following four sub-intervals: $[0, \tau_1]$, $[\tau_1, \tau_2]$, $[\tau_2, \tau_3]$, and $[\tau_3, T_{(0,0)}]$. In the following, we consider each of these intervals separately. At the end, we sum all these conditional expectations to find an expression for $\mathbb{E}\big[T_{(0,0)} \;\big|\; M^{T_{(0,0)}} \geq N\big]$.

The sub-interval $[0, \tau_1]$

To find an approximation for $\mathbb{E}\big[\tau_1 \;\big|\; M^{T_{(0,0)}} \geq N\big]$, we need to show that the number of customers in the first queue grows linearly with rate $1/\big(\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1\big)$. We prove this in the next theorem by using Renewal Theory arguments and the relation between the probability measures $\mathbb{P}$ and $\breve{\mathbb{P}}$.

Theorem 5.12. *If τ_1 is the first time the number of customers in the first queue reaches or exceeds level N, then it holds that*

$$\mathbb{E}\big[\tau_1 \,\big|\, M^{T_{(0,0)}} \geq N\big] = \frac{1}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1}\big(N + o(N)\big).$$

To prove the above theorem, we need the following two lemmas.

Lemma 5.13. *Let $\{F(N)\}_{N\geq 0}$ be any family of events with $F(N) \in \mathfrak{F}_{\tau(N-1)}$ that satisfies $\breve{\mathbb{P}}\big(F(N)\big) \to 1$, as $N \to \infty$. It then holds that $\mathbb{P}\big(F(N) \,\big|\, \tau_1 < T_{(0,0)}\big) \to 1$.*

Proof. We follow the idea of Asmussen and Albrecher (2010, Theorem V.7.1). Thus, the goal is to prove that the complement $\overline{F(N)}$ conditioned on the event $\{\tau_1 < T_{(0,0)}\}$ converges to zero. First observe that $\{\tau_1 < T_{(0,0)}\} \cap \{\tau(N-1) < \infty\} = \{\tau_1 < T_{(0,0)}\}$, because $\{\tau_1 < T_{(0,0)}\} \subset \{\tau(N-1) < T_{(0,0)}\} \subset \{\tau(N-1) < \infty\}$. To this end, we define $G_{N-1} = \overline{F(N)} \cap \{\tau_1 < T_{(0,0)}\}$ and we apply Lemma A.14 to get

$$\begin{aligned}
\mathbb{P}\big(\overline{F(N)}; \tau_1 < T_{(0,0)}\big) =& \mathbb{P}\big(\overline{F(N)}; \tau_1 < T_{(0,0)}; \tau(N-1) < \infty\big) \\
=& e^{-\gamma(N-1)}\breve{\mathbb{E}}e^{-\gamma B(N-1)}\mathbb{1}\left(G_{N-1}\right) \\
\leq& e^{-\gamma(N-1)}\breve{\mathbb{E}}\mathbb{1}\left(G_{N-1}\right) = e^{-\gamma(N-1)}\breve{\mathbb{P}}\big(\overline{F(N)}; \tau_1 < T_{(0,0)}\big) \\
\leq& e^{-\gamma(N-1)}\breve{\mathbb{P}}\big(\overline{F(N)}\big).
\end{aligned}$$

Conditioning now on the event $\{\tau_1 < T_{(0,0)}\}$, we obtain

$$\begin{aligned}
\mathbb{P}\big(\overline{F(N)} \,\big|\, \tau_1 < T_{(0,0)}\big) =& \frac{\mathbb{P}\big(\overline{F(N)}; \tau_1 < T_{(0,0)}\big)}{\mathbb{P}(\tau_1 < T_{(0,0)})} \leq \frac{e^{-\gamma(N-1)}\breve{\mathbb{P}}(\overline{F(N)})}{\mathbb{P}(\tau_1 < T_{(0,0)})} \\
\sim& \frac{e^{-\gamma(N-1)}\breve{\mathbb{P}}(\overline{F(N)})}{\frac{\mathbb{E}T_{(0,0)}}{\mathbb{E}T_0}C_1 e^{-\gamma(N-1)}} \to 0,
\end{aligned}$$

where in the last step we used Theorem 5.6. □

Lemma 5.14. *For $z > 1/\big(\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1\big)$, it holds that*

(a) $\displaystyle\lim_{N\to\infty}\int_0^z \breve{\mathbb{P}}\big(\tau_1 > yN\big)dy = \frac{1}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1}$, *and*

(b) $\displaystyle\lim_{N\to\infty}\int_0^z \mathbb{P}\Big(\tau_1 > yN \,\Big|\, \tau_1 < T_{(0,0)}\Big)dy = \frac{1}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1}$.

Proof. To prove the lemma, we first need to show that

$$\frac{\tau_1}{N} \xrightarrow{\breve{\mathbb{P}}} \frac{1}{\breve{\mathbb{E}}Z}, \quad a.s. \quad N \to \infty. \tag{5.27}$$

For this reason, we define the *fluid scaled process* (Whitt, 2002)

$$\overline{X}_N(t) = \frac{X_{\lfloor Nt \rfloor}}{N}.$$

We know from Doney et al. (2009, Theorem 2.1) that $\lim_{N\infty} \breve{\mathbb{E}}\tau_1/N = 1/\breve{\mathbb{E}}Z$. Thus, it holds that $\lim_{N\to\infty} \overline{X}_N(t) = \breve{\mathbb{E}}Zt$. Since the infimum is a continuous function, from the *Continuous Mapping Theorem* (Mann and Wald, 1943), we obtain

$$\frac{\tau_1}{N} = \inf\{t : \overline{X}_N(t) \geq 1\} \to \inf\{t : \breve{\mathbb{E}}Zt \geq 1\} = \frac{1}{\breve{\mathbb{E}}Z}.$$

We now prove (a). If we set $\theta = 1/\big(\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1\big)$, then for all $\epsilon > 0$, we know from Eq. (5.27) that

$$\lim_{N\to\infty} \breve{\mathbb{P}}\Big(\tau_1 \in \big((\theta-\epsilon)N, (\theta+\epsilon)N\big)\Big) = 1.$$

However,

$$\breve{\mathbb{P}}\Big(\tau_1 \in \big((\theta-\epsilon)N, (\theta+\epsilon)N\big)\Big) \leq \breve{\mathbb{P}}\big(\tau_1 > (\theta-\epsilon)N\big) \leq 1,$$

which results in

$$\lim_{N\to\infty} \breve{\mathbb{P}}\big(\tau_1 > yN\big) = 1, \quad \forall y < \theta.$$

In other words, we obtain

$$\lim_{N\to\infty} \breve{\mathbb{P}}\big(\tau_1 > yN\big) = \mathbb{1}\left(y < \theta\right). \tag{5.28}$$

By the *Bounded Convergence Theorem* (BCT) (Wade, 1974) and the fact that $z > \theta$, we have

$$\lim_{N\to\infty} \int_0^z \breve{\mathbb{P}}\big(\tau_1 > yN\big)dy = \int_0^z \mathbb{1}\left(y < \theta\right)dy = \theta,$$

which completes the proof. For (b), we combine Eq. (5.28) and Lemma 5.13, and we get

$$\lim_{N\to\infty} \mathbb{P}\Big(\tau_1 > yN \;\Big|\; \tau_1 < T_{(0,0)}\Big) = \mathbb{1}\left(y < \theta\right).$$

Thus, by applying here again the BCT and taking into account the relation $z > \theta$, the result is immediate. □

We proceed now with the proof of Theorem 5.12.

Proof of Theorem 5.12. Observe that it is sufficient to show

$$\mathbb{E}\left[\frac{\tau_1}{N} \;\middle|\; \tau_1 < T_{(0,0)}\right] = \frac{1}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1}\big(1 + o(1)\big).$$

Let z be a value such that $z > 1/\big(\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1\big)$; namely z is greater than the value we want to show that τ_1/N converges to. The expectation $\mathbb{E}\left[\frac{\tau_1}{N} \;\middle|\; \tau_1 < T_{(0,0)}\right]$ can be split in two terms as follows

$$\begin{aligned}\mathbb{E}\left[\frac{\tau_1}{N}\middle|\tau_1 < T_{(0,0)}\right] &= \int_0^z \mathbb{P}\Big(\tau_1 > yN \;\Big|\; \tau_1 < T_{(0,0)}\Big)dy \\ &\quad + \int_z^\infty \mathbb{P}\Big(\tau_1 > yN \;\Big|\; \tau_1 < T_{(0,0)}\Big)dy. \end{aligned} \tag{5.29}$$

From Lemma 5.14, we know that the first term of Eq. (5.29) gives the desired convergence. Thus, we only need to prove that the second term of Eq. (5.29) vanishes as $N \to \infty$. We write

$$\begin{aligned}
&\int_z^\infty \mathbb{P}\big(\tau_1 > yN \,\big|\, \tau_1 < T_{(0,0)}\big)dy \\
&\qquad = \int_z^\infty \mathbb{P}\big(\tau_1 > yN; \tau_1 < T_{(0,0)}\big)dy \,\Big/\, \mathbb{P}\big(\tau_1 < T_{(0,0)}\big) \\
&\qquad = \int_z^\infty \breve{\mathbb{E}}\big[e^{-\gamma U_{\tau(N-1)}}; \tau_1 > yN; \tau_1 < T_{(0,0)}\big]dy \,\Big/\, \mathbb{P}\big(\tau_1 < T_{(0,0)}\big) \\
&\qquad = \int_z^\infty \breve{\mathbb{E}}\big[e^{-\gamma\left(B(N-1)+N-1\right)}; \tau_1 > yN; \tau_1 < T_{(0,0)}\big]dy \,\Big/\, \mathbb{P}\big(\tau_1 < T_{(0,0)}\big) \\
&\qquad \le e^{-\gamma(N-1)} \int_z^\infty \breve{\mathbb{P}}\big(\tau_1 > yN; \tau_1 < T_{(0,0)}\big)dy \,\Big/\, \mathbb{P}\big(\tau_1 < T_{(0,0)}\big) \\
&\qquad \le e^{-\gamma(N-1)} \int_z^\infty \breve{\mathbb{P}}\big(\tau_1 > yN\big)dy \,\Big/\, \mathbb{P}\big(\tau_1 < T_{(0,0)}\big) \\
&\qquad = e^{-\gamma(N-1)} \left(\breve{\mathbb{E}}\left[\frac{\tau_1}{N}\right] - \int_0^z \breve{\mathbb{P}}\big(\tau_1 > yN\big)dy\right) \Big/ \mathbb{P}\big(\tau_1 < T_{(0,0)}\big),
\end{aligned}$$

where the second equality holds because of the relation (A.6). From Doney et al. (2009, Theorem 2.1), we have that $\lim_{N\to\infty} \breve{\mathbb{E}}\left[\frac{\tau_1}{N}\right] = \frac{1}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1}$, while from Lemma 5.14, we also know that for $N \to \infty$ the integral $\int_0^z \breve{\mathbb{P}}\big(\tau_1 > yN\big)dy$ converges to the same number. Combining all the above, we obtain

$$\begin{aligned}
&\int_z^\infty \mathbb{P}\Big(\tau_1 > yN \,\Big|\, \tau_1 < T_{(0,0)}\Big)dy \\
&\qquad \le \frac{e^{-\gamma(N-1)}\left(\breve{\mathbb{E}}\left[\frac{\tau_1}{N}\right] - \int_0^z \breve{\mathbb{P}}\big(\tau_1 > yN\big)dy\right)}{\mathbb{P}\big(\tau_1 < T_{(0,0)}\big)} \\
&\qquad \sim \frac{e^{-\gamma(N-1)}\left(\breve{\mathbb{E}}\left[\frac{\tau_1}{N}\right] - \int_0^z \breve{\mathbb{P}}\big(\tau_1 > yN\big)dy\right)}{\frac{\mathbb{E}T_{(0,0)}}{\mathbb{E}T_0} C_1 e^{-\gamma(N-1)}} \to 0, \qquad N \to \infty,
\end{aligned}$$

where at the last step we used Theorem 5.6. □

In the next theorem, we find the mean number of customer in the second queue at time τ_1.

Theorem 5.15. *If τ_1 is the first time the number of customers in the first queue reaches of exceeds level N, then the number of customers in the second queue at time τ_1 satisfies*

$$\mathbb{E}\big[Y_{\tau_1} \,\big|\, M^{T_{(0,0)}} \ge N\big] = \frac{(\breve{\mu}_1 - \mu_2)^+}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1}\big(N + o(N)\big).$$

Proof. We follow the same idea as in the proof of Theorem 5.12. Thus, we let z be a value such that $z > 1/\big(\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1\big)$ and we write

$$\mathbb{E}\left[\frac{Y_{\tau_1}}{N} \;\middle|\; M^{T_{(0,0)}} \geq N\right] = \int_0^z \mathbb{P}\big(Y_{\tau_1} > yN \mid \tau_1 < T_{(0,0)}\big)dy + \int_z^\infty \mathbb{P}\big(Y_{\tau_1} > yN \mid \tau_1 < T_{(0,0)}\big)dy. \tag{5.30}$$

The goal is to show that the first term in Eq. (5.30) gives the desired convergence, while the second term vanishes as $N \to \infty$.

We start our discussion with the first term in Eq. (5.30). To prove this part, we construct a process stochastically larger than Y_n. Recall from Eq. (5.2) that $W_n = 1$ if $Z_n = -1$ and $X_{n-1} > 0$. This means that customers arrive in the second queue only if there is a real departure of a customer from the first queue. This dependence of W_n on X_{n-1} only makes the analysis harder. Therefore, we eliminate this dependence by constructing a stochastically larger process that allows the fictitious departures from the first queue to generate real arrivals in the second queue. This auxiliary process Y'_n satisfies the Lindley recursion $Y'_{n+1} = (Y'_n + W'_{n+1})^+$, where the netput process W'_n is constructed such that

$$W'_n = \begin{cases} -1, & \text{if } Z_n = 0, \\ K, & \text{if } Z_n = -1, \\ 0, & \text{otherwise,} \end{cases}$$

with $K \geq 1$. More precisely, when $\breve{\mu}_1 > \mu_2$, which happens in Case 3, K is equal to one, giving $\breve{\mathbb{E}}W'_n = \breve{\mu}_1 - \mu_2$. However, in Cases 1 and 2, where $\breve{\mu}_1 < \mu_2$, K is greater than one so as to have $\breve{\mathbb{E}}W'_n = \epsilon > 0$, for some $\epsilon > 0$. This trick allows us to focus only on the case $\breve{\mathbb{E}}W'_n > 0$ and thus have a uniformised treatment for all different cases.

Let now $V'_n = \sum_{i=1}^n W'_n$ be the random walk defined by the increments W'_n. To prove the desired convergence, we first find a connection between V'_n and Y'_n, and then between the processes Y'_n and Y_n. To this end, observe that the random walk V'_n is skip-free in both directions. Consequently, we can easily verify for every sample path that $\big|V'_{k_1} - V'_{k_2}\big| \leq |k_1 - k_2|$ and we write

$$\frac{V'_{\tau_1}}{N} = \frac{V'_{\tau(N)}}{N} + \frac{V'_{\tau_1} - V'_{\tau(N)}}{N} \geq \frac{V'_{\tau(N)}}{N} + \frac{\tau_1 - \tau(N)}{N}. \tag{5.31}$$

We have from Gut (2009, Theorem 4.2.1.(iii)) that

$$\frac{V'_{\tau(N)}}{N} \xrightarrow{\breve{\mathbb{P}}} \frac{\breve{\mathbb{E}}W'}{\breve{\mathbb{E}}Z}, \quad a.s. \quad N \to \infty, \tag{5.32}$$

and also from Ross (1996, Proposition 3.3.1) we obtain

$$\frac{\tau(N)}{N} \xrightarrow{\breve{\mathbb{P}}} \frac{1}{\breve{\mathbb{E}}Z}, \quad a.s. \quad N \to \infty. \tag{5.33}$$

Combining now Eqs. (5.27), (5.31), (5.32), and (5.33), we write for $\theta_2 = \breve{\mathbb{E}}W'/\breve{\mathbb{E}}Z$

$$\breve{\mathbb{P}}\Big(V'_{\tau(N)} + \tau_1 - \tau(N) \in \big((\theta_2 - \epsilon)N, (\theta_2 + \epsilon)N\big)\Big) \leq \breve{\mathbb{P}}\big(V'_{\tau_1} > (\theta_2 - \epsilon)N\big) \leq 1,$$

which results in

$$\lim_{N\to\infty} \breve{\mathbb{P}}\big(V'_{\tau_1} > yN\big) = \mathbb{1}\left(y < \theta_2\right). \tag{5.34}$$

Moreover, it holds that $Y'_{\tau_1} \geq V'_{\tau_1}$, because the Lindley process Y'_n dominates the random walk V'_n. Thus, from the inequality

$$\breve{\mathbb{P}}\big(V'_{\tau_1} > (\theta_2 - \epsilon)N\big) \leq \breve{\mathbb{P}}\big(Y'_{\tau_1} > (\theta_2 - \epsilon)N\big) \leq 1$$

and Eq. (5.34), we obtain

$$\lim_{N\to\infty} \breve{\mathbb{P}}\big(Y'_{\tau_1} > yN\big) = \mathbb{1}\left(y < \theta_2\right). \tag{5.35}$$

Similarly, between the processes Y_n and Y'_n, the inequality $Y_n \leq Y'_n$ holds. Consequently, for the first term of Eq. (5.30), by applying the BCT we have

$$\begin{aligned}\int_0^z \mathbb{P}\big(Y_{\tau_1} > yN \;\big|\; \tau_1 < T_{(0,0)}\big)dy \leq & \int_0^z \mathbb{P}\big(Y'_{\tau_1} > yN \;\big|\; \tau_1 < T_{(0,0)}\big)dy \\ & \to \int_0^z \mathbb{1}\left(y < \theta_2\right)dy = \theta_2, \quad N \to \infty\end{aligned}$$

where $\lim_{N\to\infty} \mathbb{P}\big(Y'_{\tau_1} > yN \;\big|\; \tau_1 < T_{(0,0)}\big) = \mathbb{1}\left(y < \theta_2\right)$ because of Lemma 5.13 and Eq. (5.35).

For the second term of Eq. (5.30), we observe that $Y_{\tau_1} \leq \tau_1$, because Y_n is a skip-free process in both directions. As a result,

$$\int_z^\infty \mathbb{P}\big(Y_{\tau_1} > yN \;\big|\; \tau_1 < T_{(0,0)}\big)dy \leq \int_z^\infty \mathbb{P}\big(Y_{\tau_1} > yN \;\big|\; \tau_1 < T_{(0,0)}\big)dy.$$

However, in Theorem 5.12, we already proved that the right hand side of the above inequality converges to zero as $N \to \infty$, which completes the proof. □

The sub-interval $[\tau_1, \tau_2]$

So far, we studied the behaviour of both queues until time τ_1. We proceed with studying their behaviour in the next rime interval.

Theorem 5.16. *If τ_1 is the first time the number of customers in the first queue reaches of exceeds level N and τ_2 is the first time the first queue empties after τ_1, then it holds that*

$$\mathbb{E}\big[\tau_2 - \tau_1 \;\big|\; M^{T_{(0,0)}} \geq N\big] = \frac{1}{\mu_1 - \lambda\mathbb{E}B}\big(N + o(N)\big).$$

Proof. To study the asymptotic behaviour of the expectation $\mathbb{E}\big[\tau_2 - \tau_1 \;\big|\; M^{T_{(0,0)}} \geq N\big]$, we condition on the number of customers in front of the first queue at time τ_1 and we define the function

$$g(k) = \mathbb{E}\big[\tau_2 - \tau_1 \;\big|\; M^{T_{(0,0)}} \geq N; X_{\tau_1} = k\big].$$

By conditioning on the first jump after τ_1, it is apparent that the function $g(k)$ satisfies the equation

$$g(k) = 1 + \sum_{i=-1}^{\infty} \mathbb{P}(Z_1 = i)g(k+i), \quad k \geq 1, \tag{5.36}$$

where $g(0) = 0$. We can rewrite Eq. (5.36) as follows:

$$\sum_{i=-1}^{\infty} \mathbb{P}(Z_1 = i)\big(g(k) - g(k+i)\big) = 1 \quad \Rightarrow \quad \mathbb{E}h(Z) = 1,$$

where $h(i) = g(k) - g(k+i)$. Therefore, it is immediately obvious that Eq. (5.36) has the following solution

$$g(k) = -\frac{k}{\mathbb{E}Z_1} = \frac{k}{\mu_1 - \lambda \mathbb{E}B}.$$

Thus, since the number of customers in front of the first queue at time τ_1 is equal to $X_{\tau_1} = N - 1 + B(N-1)$ and the system is saturated until τ_1, we can write

$$\begin{aligned}
\mathbb{E}\big[\tau_2 - \tau_1 \;\big|\; M^{T_{(0,0)}} \geq N\big] &= \sum_{k=1}^{\infty} \mathbb{E}\big[\tau_2 - \tau_1; B(N-1) = k \;\big|\; M^{T_{(0,0)}} \geq N\big] \\
&= \sum_{k=1}^{\infty} \mathbb{E}\big[\tau_2 - \tau_1 \;\big|\; M^{T_{(0,0)}} \geq N; X_{\tau_1} = N - 1 + k\big] \cdot \breve{\mathbb{P}}\big(B(N-1) = k\big) \\
&= \sum_{k=1}^{\infty} \frac{N-1+k}{\mu_1 - \lambda \mathbb{E}B} \cdot \breve{\mathbb{P}}\big(B(N-1) = k\big) = \frac{N-1}{\mu_1 - \lambda \mathbb{E}B} + \frac{\breve{\mathbb{E}}\big(B(N-1)\big)}{\mu_1 - \lambda \mathbb{E}B}.
\end{aligned}$$

However, from Lemma A.17, we obtain that $\breve{\mathbb{E}}\big(B(N-1)\big) \to \breve{\mathbb{E}}\big(B(\infty)\big)$. Consequently, we have that

$$\lim_{N\to\infty} \frac{1}{N} \mathbb{E}\big[\tau_2 - \tau_1 \;\big|\; M^{T_{(0,0)}} \geq N\big] = \frac{1}{\mu_1 - \lambda \mathbb{E}B},$$

which completes the proof. □

Theorem 5.17. *If τ_1 is the first time the number of customers in the first queue reaches of exceeds level N and τ_2 is the first time the first queue empties after τ_1, then the number of customers in the second queue at time τ_2 satisfies*

$$\mathbb{E}\big[Y_{\tau_2} \;\big|\; M^{T_{(0,0)}} \geq N\big] = \left(\frac{(\breve{\mu}_1 - \mu_2)^+}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1} + \frac{(\breve{\mu}_1 - \mu_2)^+}{\mu_1 - \lambda \mathbb{E}B} \right) \big(N + o(N)\big).$$

Proof. There are always customers in the first queue during the time interval $[\tau_1 + 1, \tau_2]$, so every departure from the first queue is an arrival to the second one. Therefore, the dynamics of the second queue in the time interval $[\tau_1 + 1, \tau_2]$ are expressed by the sum $\sum_{n=\tau_1+1}^{\tau_2} W'_n$, where the process W'_n was introduced in Theorem 5.15. Observe that the sum $\sum_{n=\tau_1+1}^{\tau_2} W'_n$ can become smaller than $-Y_{\tau_1}$. Thus, to compensate for the negative 'positions' of the random walk V'_n and avoid the situation $Y_{\tau_2} < 0$, we write

$$Y_{\tau_2} = Y_{\tau_1} + \sum_{n=\tau_1+1}^{\tau_2} W'_n + \sum_{n=\tau_1+1}^{\tau_2} \mathbb{1}\left(Y'_n = 0\right).$$

Conditioning on $\{M^{T_{(0,0)}} \geq N\}$ and taking expectations yields

$$\mathbb{E}\big[Y_{\tau_2} \;\big|\; M^{T_{(0,0)}} \geq N\big] = \mathbb{E}\big[Y_{\tau_1} \;\big|\; M^{T_{(0,0)}} \geq N\big] + \underbrace{\mathbb{E}\left[\sum_{n=\tau_1+1}^{\tau_2} W'_n \;\middle|\; M^{T_{(0,0)}} \geq N\right]}_{\text{Wald's equation}}$$

$$
\begin{aligned}
&+\mathbb{E}\left[\sum_{n=\tau_1+1}^{\tau_2}\mathbb{1}\left(Y_n'=0\right)\,\middle|\,M^{T_{(0,0)}}\geq N\right]\\
&=\mathbb{E}\big[Y_{\tau_1}\,\big|\,M^{T_{(0,0)}}\geq N\big]+\mathbb{E}\big[\tau_2-\tau_1\,\big|\,M^{T_{(0,0)}}\geq N\big]\cdot\mathbb{E}W'\\
&\quad+\mathbb{E}\left[\sum_{n=\tau_1+1}^{\tau_2}\mathbb{1}\left(Y_n'=0\right)\,\middle|\,M^{T_{(0,0)}}\geq N\right]\\
&\leq\mathbb{E}\big[Y_{\tau_1}\,\big|\,M^{T_{(0,0)}}\geq N\big]+\mathbb{E}\big[\tau_2-\tau_1\,\big|\,M^{T_{(0,0)}}\geq N\big]\cdot\mathbb{E}W'\\
&\quad+\mathbb{E}\left[\sum_{n=\tau_1+1}^{\infty}\mathbb{1}\left(Y_n'=0\right)\,\middle|\,M^{T_{(0,0)}}\geq N\right],
\end{aligned}
$$

where $\sum_{n=\tau_1+1}^{\infty}\mathbb{1}\left(Y_n'=0\right)$ follows a geometric distribution and is independent of N. By dividing with N, we get that

$$
\begin{aligned}
\mathbb{E}\left[\frac{Y_{\tau_2}}{N}\,\middle|\,M^{T_{(0,0)}}\geq N\right]&\leq\mathbb{E}\left[\frac{Y_{\tau_1}}{N}\,\middle|\,M^{T_{(0,0)}}\geq N\right]+\mathbb{E}\left[\frac{\tau_2-\tau_1}{N}\,\middle|\,M^{T_{(0,0)}}\geq N\right]\cdot\mathbb{E}W'\\
&\quad+\frac{1}{N}\mathbb{E}\left[\sum_{n=\tau_1+1}^{\infty}\mathbb{1}\left(Y_n'=0\right)\,\middle|\,M^{T_{(0,0)}}\geq N\right]\\
&\to\frac{(\breve{\mu}_1-\mu_2)^+}{\breve{\lambda}\breve{\mathbb{E}}B-\breve{\mu}_1}+\frac{\mathbb{E}W'}{\mu_1-\lambda\mathbb{E}B}=\frac{(\breve{\mu}_1-\mu_2)^+}{\breve{\lambda}\breve{\mathbb{E}}B-\breve{\mu}_1}+\frac{(\mu_1-\mu_2)^+}{\mu_1-\lambda\mathbb{E}B},
\end{aligned}
$$

as $N\to\infty$, and the proof is complete. □

The sub-interval $[\tau_2,\tau_3]$

Recall that τ_2 is the time the fist queue hit zero after the number of its customers reached or exceeded the truncation level. In the next theorem, we find how long it takes for the second queue to empty starting at τ_2.

Theorem 5.18. *If τ_1 is the first time the number of customers in the first queue reaches of exceeds level N, τ_2 is the first time the first queue empties after τ_1, and τ_3 is the time the second queue empties after τ_2, then it holds that*

$$
\mathbb{E}\big[\tau_3-\tau_2\,\big|\,M^{T_{(0,0)}}\geq N\big]=\frac{1}{\mu_2-\lambda\mathbb{E}B}\cdot\left(\frac{(\breve{\mu}_1-\mu_2)^+}{\breve{\lambda}\breve{\mathbb{E}}B-\breve{\mu}_1}+\frac{(\mu_1-\mu_2)^+}{\mu_1-\lambda\mathbb{E}B}\right)\big(N+o(N)\big).
$$

Proof. Note that at time τ_2 the first queue is empty and from Theorem 5.17, we know that there exists some $y_0>0$ such that $Y_{\tau_2}=y_0N$. Moreover, time τ_3 is defined as $\tau_3=\inf\{n>\tau_2:Y_n=0\}$. The idea is to see our two-dimensional Markov chain as a *Markov Additive Process* (MAP) (Çinlar, 1972) during the time interval $[\tau_2,\tau_3]$. For notational convenience, we shift the time index and we assume that the process starts at 0 instead of τ_2. Therefore, we observe the system in the time interval $[0,\tau_Y(0)]$, with $X_0=0$, $Y_0=y_0N$, and $\tau_Y(0)=\inf\{n>0:Y_n=0\}$.

Observe that the second queue never hits zero until time $\tau_Y(0)$. This means that $Y_n=(Y_n)^+$ for $n\in[0,\tau_Y(0)]$, which shows that the increments $W_{n+1}=Y_{n+1}-Y_n$

are conditionally independent given $(Z_i)_{i\geq 0}$. Thus, the process $(X_n, S_n)_{n\geq 0}$, with $S_n = -\sum_{i=1}^{n} W_i$ and $X_0 = S_0 = 0$, defines a MAP (or *Markov Random Walk* (MRW)) that satisfies

$$\mathbb{P}\big(X_{n+1} \in A, S_{n+1} - S_n \in B \;\big|\; X_n, W_n\big) = \mathbb{P}\big(X_n, A \times B\big),$$

for all $n \geq 0$ and $A \in \mathfrak{B}(\mathbb{N})$, $B \in \mathfrak{B}(\mathbb{Z})$, with $\mathfrak{B}$ denoting the *Borel σ-algebra* on a given state space. For this auxiliary process, $\tau_Y(0) = \inf\{n > 0 : S_n = y_0 N\}$. We are interested in finding $\mathbb{E}[\tau_Y(0)]$.

Since $\{X_n\}_{n\geq 0}$ is an ergodic Markov chain with some stationary distribution π_i, $i = 0, 1, \ldots$, the *Markov Renewal Theorem* as formulated in Alsmeyer (1994, Theorem 2.1) takes the form (with $d = 1$, *shift function* $\gamma(\cdot) = 0$, ξ the unique stationary distribution of X_n, and $\mu(x) = \mathbb{E}[S_1 | X_0 = x]$)

$$\lim_{k\to\infty} \mathbb{E}\Big(\sum_{n\geq 0} g(X_n, k - S_n)\Big) = \frac{1}{\mu_2 - \lambda\mathbb{E}B} \sum_{m=0}^{\infty} \sum_{n\in\mathbb{Z}} \pi_m g(m, n), \tag{5.37}$$

for every measurable function $g : \mathbb{N} \times \mathbb{Z} \to \mathbb{R}$ that satisfies

$$\sum_{m=0}^{\infty} \sum_{n\in\mathbb{Z}} \pi_m \, |g(m, n)| < \infty. \tag{5.38}$$

Letting now $g(m, n) = \mathbb{1}\,(n = 0)$ and $k = y_0 N$ in Eq. (5.37) gives

$$\mathbb{E}\Big(\sum_{n\geq 0} \mathbb{1}\,(S_n = y_0 N)\Big) \to \frac{1}{\mu_2 - \lambda\mathbb{E}B}, \tag{5.39}$$

which shows that the mean number of renewals at $y_0 N$ converges at $1/(\mu_2 - \lambda\mathbb{E}B)$. We use now the following elementary result: if $a_k = \mathbb{E}[\text{number of renewals at } k]$ with $\lim_{k\to\infty} a_k = 1/(\mu_2 - \lambda\mathbb{E}B)$ then it holds that

$$\frac{1}{y_0 N} \sum_{k=1}^{y_0 N - 1} a_k \to \frac{1}{\mu_2 - \lambda\mathbb{E}B}, \qquad N \to \infty.$$

In other words,

$$\frac{\mathbb{E}[\tau_Y(0)]}{y_0 N} \to \frac{1}{\mu_2 - \lambda\mathbb{E}B}, \qquad N \to \infty. \tag{5.40}$$

which completes the proof. □

The following theorem shows that asymptotically the mean number of customers in the first queue is finite when the second queue empties.

Theorem 5.19. *If τ_1 is the first time the number of customers in the first queue reaches of exceeds level N, τ_2 is the first time the first queue empties after τ_1, and τ_3 is the time the second queue empties after τ_2, then it holds that*

$$\mathbb{E}\big[X_{\tau_3} \;\big|\; M^{T_{(0,0)}} \geq N\big] = \frac{\lambda\mathbb{E}B + \mu_2}{2(\mu_2 - \lambda\mathbb{E}B)^2}\big(1 + o(1)\big).$$

Proof. To prove the theorem, the idea is to consider the total number of customers in the system without taking into account the internal transitions between the two queues. As long as the second queue does not empty, the total number of customers in the system can be described in a simple way. Following the proof of Theorem 5.18, we also assume here for notational convenience that the process starts at 0 instead of τ_2 with $X_0 = 0$, $Y_0 = y_0 N$, and $\tau_Y(0) = \inf\{n > 0 : Y_n = 0\}$. Therefore, for $n \leq \tau_Y(0)$, we write

$$X_n + Y_n = X_0 + Y_0 + \sum_{i=1}^{n} Z_i \mathbb{1}(Z_i > 0) - \sum_{i=1}^{n} \mathbb{1}(Z_i = 0). \tag{5.41}$$

We define now the process

$$A_n = \sum_{i=1}^{n} Z_i \mathbb{1}(Z_i > 0) - \sum_{i=1}^{n} \mathbb{1}(Z_i = 0) - (\lambda \mathbb{E}B - \mu_2)n,$$

for $n \leq \tau_Y(0)$. Observe that $\mathbb{E}\big[Z_1 \mathbb{1}(Z_1 > 0) - \mathbb{1}(Z_1 = 0) - (\lambda \mathbb{E}B - \mu_2)\big] = 0$ and $\mathbb{E}\big[Z_{n+1}\mathbb{1}(Z_{n+1} > 0) - \mathbb{1}(Z_{n+1} = 0) - (\lambda \mathbb{E}B - \mu_2) \mid Z_1, \ldots, Z_n\big] = 0$. As a result, $\{Z_n \mathbb{1}(Z_n > 0) - \mathbb{1}(Z_n = 0)\big] - (\lambda \mathbb{E}B - \mu_2)\}_{n \geq 1}$ is a sequence of *absolutely fair* random variables and A_n is a martingale with respect to the filtration generated by $(Z_n)_{n \geq 0}$ (Feller, 1971, page 210). Consequently, $\mathbb{E}|A_n| < \infty$. In addition, since $\tau_Y(0)$ is a stopping time with finite expectation for fixed N according to Eq. (5.40), we apply *Doob's optional sampling theorem* to get $\mathbb{E}A_{\tau_Y(0)} = \mathbb{E}A_0 = 0$. Moreover, we have that $X_0 = Y_{\tau_Y(0)} = 0$. Thus, setting $n = \tau_Y(0)$ in Eq. (5.41) and taking expectation yields

$$\begin{aligned} \mathbb{E}X_{\tau_Y(0)} &= \mathbb{E}X_0 + \mathbb{E}A_{\tau_Y(0)} + (\lambda \mathbb{E}B - \mu_2)\mathbb{E}[\tau_Y(0)] \\ &= y_0 N - (\mu_2 - \lambda \mathbb{E}B)\mathbb{E}[\tau_Y(0)]. \end{aligned}$$

As a last step, we want to show that the right hand side of the above equation converges to a constant. To this end, we follow the notation introduced for the MAP defined in Theorem 5.18 and we write

$$S_{\tau_Y(0)} = y_0 N + R_{\tau_Y(0)}, \tag{5.42}$$

where $R_{\tau_Y(0)}$ is the overshoot of S_n at time $\tau_Y(0)$. From *Wald's equation for Markov random walks* (Fuh and Lai, 1998; Moustakides, 1999), by taking expectations in Eq. (5.42), we obtain

$$S_{\tau_Y(0)} = (\mu_2 - \lambda \mathbb{E}B)\mathbb{E}[\tau_Y(0)] \quad \Rightarrow \quad \mathbb{E}[\tau_Y(0)] = \frac{y_0 N}{\mu_2 - \lambda \mathbb{E}B} + \frac{\mathbb{E}R_{\tau_Y(0)}}{\mu_2 - \lambda \mathbb{E}B}.$$

To find now $\mathbb{E}R_{\tau_Y(0)}$, we use Eq. (5.37) with $g(m,n) = \mathbb{E}\big[S_1 - n; S_1 > n \mid X_0 = m\big]$, for $n \geq 0$, and $g(m,n) = 0$ otherwise. Therefore, we have

$$\begin{aligned} \mathbb{E}R_{\tau_Y(0)} \to & \frac{1}{\mu_2 - \lambda \mathbb{E}B} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \pi_m \mathbb{E}\big[S_1 - n; S_1 > n \mid X_0 = m\big] \\ = & \frac{1}{\mu_2 - \lambda \mathbb{E}B} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \sum_{k=n}^{\infty} \pi_m \mathbb{P}\big(S_1 > k \mid X_0 = m\big) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\mu_2 - \lambda \mathbb{E}B} \sum_{m=0}^{\infty} \pi_m \sum_{k=0}^{\infty} \sum_{n=0}^{k-1} \mathbb{P}\big(S_1 > k \mid X_0 = m\big) \\
&= \frac{1}{\mu_2 - \lambda \mathbb{E}B} \sum_{m=0}^{\infty} \pi_m \sum_{k=0}^{\infty} k \mathbb{P}\big(S_1 > k \mid X_0 = m\big) \\
&= \frac{1}{\mu_2 - \lambda \mathbb{E}B} \sum_{m=0}^{\infty} \pi_m \frac{\mathbb{E}\big[S_1^2 \mid X_0 = m\big]}{2} = \frac{\lambda \mathbb{E}B + \mu_2}{2(\mu_2 - \lambda \mathbb{E}B)},
\end{aligned}
$$

where π_i, $i = 0, 1, \ldots$, is the stationary distribution of the Markov chain $\{X_n\}_{n \geq 0}$. Moreover, the function g satisfies the condition (5.38) since

$$
\sum_{m=0}^{\infty} \sum_{n \in \mathbb{Z}} \pi_m \, |g(m,n)| = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \pi_m \mathbb{E}\big[S_1 - n; S_1 > n \mid X_0 = m\big] = \frac{\mathbb{E}\big[S_1^2\big]}{2} < \infty.
$$

Combining all the above, we have

$$
\mathbb{E}[\tau_Y(0)] - \frac{y_0 N}{\mu_2 - \lambda \mathbb{E}B} \to \frac{\lambda \mathbb{E}B + \mu_2}{2(\mu_2 - \lambda \mathbb{E}B)^2},
$$

and the proof is complete. □

The sub-interval $[\tau_3, T_{(0,0)}]$

At time τ_3, the system is in state $(X_{\tau_3}, 0)$, where X_{τ_3} is a finite r.v. with finite expectation according to Theorem 5.19. Therefore, since $(0,0)$ is a recurrent state for our ergodic Markov chain (X_n, Y_n), it holds that the hitting time of $(0,0)$ is finite. We formulate this result in the following proposition.

Proposition 5.20. *If τ_1 is the first time the number of customers in the first queue reaches of exceeds level N, τ_2 is the first time the first queue empties after τ_1, and τ_3 is the time the second queue empties after τ_2, then it holds that*

$$
\mathbb{E}\big[T_{(0,0)} - \tau_3 \mid M^{T_{(0,0)}} \geq N\big] = O(1).
$$

Adding the results of Theorems 5.12, 5.16, and 5.18 and Proposition 5.20 yields

$$
\begin{aligned}
\mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big] = \Bigg(&\frac{1}{\mu_2 - \lambda \mathbb{E}B} \cdot \Big(\frac{(\breve{\mu}_1 - \mu_2)^+}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1} + \frac{(\mu_1 - \mu_2)^+}{\mu_1 - \lambda \mathbb{E}B} \Big) \\
&+ \frac{1}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1} + \frac{1}{\mu_1 - \lambda \mathbb{E}B} \Bigg) \big(N + o(N)\big).
\end{aligned} \tag{5.43}
$$

To sum up, we have studied the asymptotic behaviour of all the factors involved in the upper bound (5.19). Thus, by combining the results of Theorem 5.6, Lemma 5.10, and Eq. (5.43), we derive the following expression for the asymptotic upper bound.

Asymptotic upper bound. As $N \to \infty$,

$$\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) - \mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$$
$$\lesssim$$
$$\left(\frac{1}{\mu_2 - \lambda\mathbb{E}B} \cdot \left(\frac{(\breve{\mu}_1 - \mu_2)^+}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1} + \frac{(\mu_1 - \mu_2)^+}{\mu_1 - \lambda\mathbb{E}B}\right) + \frac{1}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1} + \frac{1}{\mu_1 - \lambda\mathbb{E}B}\right)$$
$$\times N C_1 e^{-\gamma(N-1)}\left(1 - \frac{\lambda\mathbb{E}B}{\mu_1}\right), \tag{5.44}$$

where C_1 is calculated according to Lemma 5.9. ▲

5.7 Numerical experiments

In this section, we perform numerical experiments to check the quality of our asymptotic upper bound (5.44). As we explained in Section 5.1, the queue lengths have a product form solution only for single arrivals. Since it is more meaningful to compare approximations with exact results than with simulation outcomes, we choose single arrivals for the batch size distribution. By using Eq. (5.44), we first derive the exact expression for the asymptotic upper bound of this particular model in Section 5.7.1. Afterwards, in Section 5.7.2, we fix values for the parameters to perform our numerical experiments.

5.7.1 Special case: single arrivals

When we assume single arrivals, our tandem network reduces to the M/M/1→ •/M/1 queue. Thus, the exact joint queue length distribution is found by the formula

$$\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) = \rho_1^x \rho_2^y, \qquad x, y \geq 0,$$

where $\rho_i = \lambda/\mu_i$. Let now N be the truncation level for the number of customers in the first queue. To find the asymptotic upper bound (5.44), we perform an exponential change of measure according to Section 5.4. Therefore, from Remark 5.4, we have that the adjustment coefficient is equal to $\gamma = \ln\big(\mu_1/\lambda\big)$ and the rates under the new measure $\breve{\mathbb{P}}$ take the form $\breve{\lambda} = \mu_1$ and $\breve{\mu}_1 = \lambda$. Using Lemma 5.9, we also find that $C_1 = \lambda(1 - \rho_1)$.

Note that for single arrivals it always holds that $\breve{\mu}_1 = \lambda < \mu_2$, because of stability. Thus, Case 3 does not appear in the case of single arrivals. From Eq. (5.43), we get

$$\mathbb{E}\big[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\big] \sim \frac{N}{\mu_1 - \lambda}\left(2 + \frac{(\mu_1 - \mu_2)^+}{\mu_2 - \lambda}\right).$$

From all the above, we conclude that the *asymptotic upper error bound* (*a.u.e.b.*) in Eq. (5.44) takes the form

$$a.u.e.b. := N\left(2 + \frac{(\mu_1 - \mu_2)^+}{\mu_2 - \lambda}\right)\rho_1^N(1 - \rho_1)$$

$$= N\left(2 + \frac{\left(\frac{\rho_2}{\rho_1} - 1\right)^+}{1-\rho_2}\right)\rho_1^N(1-\rho_1). \tag{5.45}$$

Moreover, we call here *constant coefficient bound* (*c.c.b*) the part of the *a.u.e.b* that does not depend on the truncation level N. Thus,

$$c.c.b. = \left(2 + \frac{\left(\frac{\rho_2}{\rho_1} - 1\right)^+}{1-\rho_2}\right)\rho_1(1-\rho_1). \tag{5.46}$$

The usefulness of this term will become clear in the next section. We proceed now with our numerical experiments.

5.7.2 Numerical results

For our numerical experiments, we focus on the marginal distribution of the second queue. As we explained in Section 5.7.1, only Cases 1 and 2 appear for this choice of the batch size distribution. Therefore, we fix the parameters λ, μ_1, and μ_2, to have examples of both cases. Note that for a given combination of $\{\rho_1, \rho_2\}$ there exists a unique combination of $\{\lambda, \mu_1, \mu_2\}$, because due to uniformisation, the rates are connected through the equation $\lambda + \mu_1 + \mu_2 = 1$. Thus, instead of deciding on 3 parameters (arrival and service rates), we decide on combinations of $\{\rho_1, \rho_2\}$.

For each combination of the loads, we choose a number of truncation levels N. Since we assumed throughout the chapter that $\rho_1 < 1$, we know that $\lim_{N\to\infty} \rho_1^N = 0$ and we choose N such that it is a multiple of $(1-\rho_1)^{-1}$. Therefore, we take $N = \lceil n(1-\rho_1)^{-1} \rceil$, $n = 1, \ldots, 7$, and we calculate for each N the truncated approximation $\mathbb{P}\big(Y_\infty^{(N)} \geq y\big)$ of $\mathbb{P}\big(Y_\infty \geq y\big) = \rho_2^y$, $y \geq 0$, according to Remark 5.2. In addition, note that we performed extensive numerical experiments for various combinations $\{\rho_1, \rho_2\}$ of the loads. We chose to present here the combinations $\{\rho_1 = 0.7, \rho_2 = 0.6\}$ (Case 1) and $\{\rho_1 = 0.7, \rho_2 = 0.8\}$ (Case 2), since the qualitative results were similar among the various combinations we tested.

To check the quality of our asymptotic upper error bound, we calculate the differences between the exact and the truncated approximation of the marginal queue length distribution and we compare them with the *a.u.e.b.* (5.45). We summarise our findings in Tables 5.1 and 5.2. The last line of these tables represents the *a.u.e.b.* From the tables, we observe that the truncated approximations become more accurate as N increases, which is in accordance with our expectations. The same also holds for the asymptotic bound. However, the bound is at least 5 times greater than the observed error, which makes it rather pessimistic.

In an attempt to understand why the bound is pessimistic, we perform another numerical experiment. More precisely, we calculate the standardised differences $e^{\gamma(N-1)}\big(\mathbb{P}(Y_\infty \geq y) - \mathbb{P}(Y_\infty^{(N)} \geq y)\big)/N$ and we compare them with the *c.c.b.*, which is the factor of the bound that does not depend on N; see Eq. (5.46). The results from this experiment are displayed in Tables 5.3 and 5.4. We observe that for each value of y, the standardised differences seem to converge to a constant, which is different for every y. Moreover, the *c.c.b.* is far from realistic and at least 5 times greater than the exact standardised differences. This is an indication that most probably the inaccuracy of the bound is due to an overestimation of this constant factor *c.c.b.* Moreover, even

though the *c.c.b.* is smaller when $\rho_1 = 0.7$ and $\rho_2 = 0.6$, as expected from Eq. (5.46), it does not mean that it captures better the behaviour of the standardised differences.

y	$N=4$	$N=7$	$N=10$	$N=14$	$N=17$	$N=20$	$N=24$
5	0.043164	0.016910	0.006206	0.001578	0.000557	0.000195	0.000047
10	0.005249	0.003332	0.001684	0.000542	0.000212	0.000080	0.000020
15	0.000451	0.000359	0.000239	0.000108	0.000051	0.000022	6.5×10^{-6}
20	0.000036	0.000032	0.000024	0.000014	8.3×10^{-6}	4.3×10^{-6}	1.5×10^{-6}
25	2.8×10^{-6}	2.6×10^{-6}	2.2×10^{-6}	1.5×10^{-6}	1.1×10^{-6}	6.2×10^{-7}	2.7×10^{-7}
30	2.2×10^{-7}	2.1×10^{-7}	1.9×10^{-7}	1.4×10^{-7}	1.1×10^{-7}	7.4×10^{-8}	3.8×10^{-8}
a.u.e.b.	0.432180	0.296475	0.152537	0.052901	0.022332	0.009096	0.002643

TABLE 5.1: Observed errors between the original marginal distribution of the second queue and its QBD approximation for $\rho_1 = 0.7$ and $\rho_2 = 0.6$. The last line corresponds to the asymptotic upper error bound for each truncation level.

y	$N=4$	$N=7$	$N=10$	$N=14$	$N=17$	$N=20$	$N=24$
5	0.153260	0.055785	0.019824	0.004957	0.001843	0.000669	0.000169
10	0.081620	0.040864	0.017412	0.004932	0.001725	0.000600	0.000146
15	0.031387	0.019249	0.009657	0.003277	0.001343	0.000525	0.000143
20	0.010969	0.007716	0.004334	0.001695	0.000769	0.000329	0.000099
25	0.003695	0.002865	0.001752	0.000759	0.000373	0.000173	0.000058
30	0.001225	0.001019	0.000667	0.000312	0.000163	0.000081	0.000029
a.u.e.b.	0.782040	0.469420	0.230016	0.077317	0.032202	0.012994	0.003744

TABLE 5.2: Observed errors between the original marginal distribution of the second queue and its QBD approximation for $\rho_1 = 0.7$ and $\rho_2 = 0.8$. The last line corresponds to the asymptotic upper error bound for each truncation level.

y	$N=4$	$N=7$	$N=10$	$N=14$	$N=17$	$N=20$	$N=24$
5	0.031460	0.020534	0.015380	0.011637	0.009867	0.008570	0.007292
10	0.003825	0.004046	0.004175	0.004002	0.003764	0.003509	0.003186
15	0.000329	0.000436	0.000592	0.000796	0.000903	0.000965	0.000996
20	0.000026	0.000038	0.000061	0.000106	0.000147	0.000188	0.000234
25	2.1×10^{-6}	3.2×10^{-6}	5.5×10^{-6}	0.000011	0.000018	0.000027	0.000041
30	1.6×10^{-7}	2.5×10^{-7}	4.7×10^{-7}	1.1×10^{-6}	1.9×10^{-6}	3.2×10^{-6}	5.9×10^{-6}
c.c.b.	0.42	0.42	0.42	0.42	0.42	0.42	0.42

TABLE 5.3: Standardised observed errors between the original marginal distribution of the second queue and its QBD approximation for $\rho_1 = 0.7$ and $\rho_2 = 0.6$. The last line corresponds to the constant coefficient bound, which is independent of the truncation level.

Last, we calculate the differences $e^{\gamma(N-1)}\big(\mathbb{P}(Y_\infty \geq y) - \mathbb{P}(Y_\infty^{(N)} \geq y)\big)$. From Tables 5.5 and 5.6, we view that these differences grow as N grows, but it is not clear if they grow linearly in N.

5.8 Conclusions

In this chapter, we addressed the problem of how to derive error bounds for the queue length distribution of a tandem network of two queues when we truncate the

y	$N=4$	$N=7$	$N=10$	$N=14$	$N=17$	$N=20$	$N=24$
5	0.111705	0.067738	0.049126	0.036544	0.032626	0.029381	0.025861
10	0.059490	0.049620	0.043148	0.036365	0.030545	0.026358	0.022293
15	0.022877	0.023374	0.023932	0.024159	0.023783	0.023055	0.021799
20	0.007995	0.009369	0.010740	0.012498	0.013621	0.014453	0.015081
25	0.002693	0.003479	0.004342	0.005599	0.006618	0.007631	0.008843
30	0.000893	0.001238	0.001654	0.002305	0.002895	0.003566	0.004535
c.c.b.	0.57	0.57	0.57	0.57	0.57	0.57	0.57

TABLE 5.4: Standardised observed errors between the original marginal distribution of the second queue and its QBD approximation for $\rho_1 = 0.7$ and $\rho_2 = 0.8$. The last line corresponds to the constant coefficient bound, which is independent of the truncation level.

y	$N=4$	$N=7$	$N=10$	$N=14$	$N=17$	$N=20$	$N=24$
5	0.125843	0.143740	0.153805	0.162918	0.167754	0.171416	0.175029
10	0.015303	0.028326	0.041754	0.056035	0.064001	0.070198	0.076478
15	0.001317	0.003057	0.005926	0.011150	0.015352	0.019302	0.023904
20	0.000105	0.000272	0.000615	0.001496	0.002515	0.003775	0.005639
25	8.2×10^{-6}	0.000022	0.000055	0.000159	0.000312	0.000550	0.001007
30	6.4×10^{-7}	1.8×10^{-6}	4.7×10^{-6}	0.000015	0.000032	0.000065	0.000142

TABLE 5.5: Observed errors between the original marginal distribution of the second queue and its QBD approximation multiplied by $e^{\gamma(N-1)}$ for $\rho_1 = 0.7$ and $\rho_2 = 0.6$.

background state space. In doing so, we truncated the buffer size of the first queue and, with the aid of extreme value analysis, we derived an upper bound for the exact queue length probabilities (see Section 5.3.2). We studied further the asymptotic behaviour of the factors involved in this bound and we derived in Eq. (5.44) an asymptotic upper error bound.

The conclusions we can draw for the asymptotic upper bound are summarised as follows:

- The bound depends only on the truncation level and the parameters of the model, i.e. it is uniform in the values x and y of $\mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$.

- The bound is rather pessimistic. Moreover, the bound becomes more pessimistic as the truncation level increases.

y	$N=4$	$N=7$	$N=10$	$N=14$	$N=17$	$N=20$	$N=24$
5	0.446821	0.474166	0.491262	0.509115	0.519269	0.527160	0.535052
10	0.237961	0.347341	0.431488	0.511621	0.554648	0.587623	0.620665
15	0.091507	0.163618	0.239323	0.338237	0.404312	0.461115	0.523188
20	0.031980	0.065585	0.107405	0.174972	0.231559	0.289079	0.361945
25	0.010773	0.024356	0.043420	0.078387	0.112517	0.152636	0.212239
30	0.003573	0.008666	0.016546	0.032278	0.049222	0.071330	0.108849

TABLE 5.6: Observed errors between the original marginal distribution of the second queue and its QBD approximation multiplied by $e^{\gamma(N-1)}$ for $\rho_1 = 0.7$ and $\rho_2 = 0.8$.

- From the performed numerical experiments, we discovered that the undesired behaviour of the bound is most probably attributed to the constant term of the bound, i.e. the factor *c.c.b.* that is independent of the truncation level N.

The above observations indicate that further modifications are important to improve the accuracy of the asymptotic upper bound. One possible direction is to make the bound dependent on the values x and y.

Appendix

A.1 Subexponential distributions

In Chapters 2–3, to study the tail behaviour of the ruin probability we consider that the claim sizes belong to the the class of subexponential distributions $\mathcal{S}$. Following Teugels (1975), we give the following definition for the class $\mathcal{S}$.

Definition A.1. A distribution F concentrated on $[0,\infty)$ belongs to the class of subexponential distributions $\mathcal{S}$ if and only if

$$\lim_{u\to\infty}\frac{1-F^{*n}(u)}{1-F(u)}=n, \qquad n=1,2,\dots$$

When a distribution F belongs to $\mathcal{S}$, it is known that $\overline{F}$ decays slower than any exponential rate (Asmussen and Albrecher, 2010). Two very useful known properties of subexponentiality are the following, which are given without proof; see Asmussen and Albrecher (2010).

Property A.2. The class $\mathcal{S}$ is closed under tail-equivalence. That is, if $\overline{A}(u)\sim a\overline{F}(u)$ for some $F\in\mathcal{S}$ and some constant $a>0$, then $A\in\mathcal{S}$.

Property A.3. Let $F\in\mathcal{S}$ and let A be any distribution with a lighter tail, i.e. $\overline{A}(u)=o\left(\overline{F}(u)\right)$. Then for the convolution $A*F$ of A and F we have $A*F\in\mathcal{S}$ and $\overline{(A*F)}(u)\sim\overline{F}(u)$.

A.2 Results on perturbation theory

In this section, we provide some preliminary results on linear algebra, matrix functions, and perturbation theory that are needed in the analysis of Chapter 4. We introduce an $N\times N$ matrix function $\mathbf{E}(s)$ with a single parameter $s>0$ and we set $\mathcal{N}=\{1,\dots,N\}$. We say that the matrix function $\mathbf{E}(s)$ is *regular* if $\det\mathbf{E}(s)$ is not identically zero as a function of s. In addition, if $\mathbf{E}(s)$ is regular (we denote it as $\det\mathbf{E}(s)\not\equiv 0$), then the eigenvalues of $\mathbf{E}(s)$ are the solutions of the equations $\det\mathbf{E}(s)=0$ (De Terán, 2011). Throughout our analysis, we assume that matrix $\mathbf{E}(s)$ is regular and that r is a simple eigenvalue of it. In addition, we assume that $\mathbf{E}(s)$ is analytic in the neighbourhood of

r. We use the notation $\mathbf{E}^{(n)}(s)$ for the nth derivative with respect to s of the matrix function $\mathbf{E}(s)$. Thus, $\mathbf{E}(s)$ can be written as a Taylor series in the following form:

$$\mathbf{E}(s) = \mathbf{E}^{(0)}(r) + (s-r)\mathbf{E}^{(1)}(r) + \cdots = \sum_{n=0}^{\infty} \frac{(s-r)^n}{n!}\mathbf{E}^{(n)}(r). \tag{A.1}$$

To avoid redundant notation, in the forthcoming analysis we use the conventions that $\mathbf{E} = \mathbf{E}^{(0)}(r) = \mathbf{E}(r)$ and $\mathbf{E}^{(n)} = \mathbf{E}^{(n)}(r)$.

As a consequence of the fact that the multiplicity of the eigenvalue r is one, the dimension of the nullspace of $\mathbf{E}$ is equal to one. Our first goal is to find the form of the eigenvectors of the nullspace of matrix $\mathbf{E}$. We prove the following theorem, which gives us exactly the form of these eigenvectors.

Theorem A.4. *If $\mathbf{C}$ is an $N \times N$ matrix with determinant equal to zero, i.e. $\det \mathbf{C} = 0$, and nullspace of dimension one, then a right $N \times 1$ eigenvector that corresponds to the simple eigenvalue zero is $\mathbf{t}$ with coordinates $t_j = (-1)^{1+j} \det \mathbf{C}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{1\}}$, $j \in \mathcal{N}$.*

Proof. We need to prove that the inner product of every row of $\mathbf{C}$ with $\mathbf{t}$ is equal to zero. More precisely, if $\mathbf{c}_i$ denotes the ith row of matrix $\mathbf{C}$, we need to show that

$$\mathbf{c}_i\mathbf{t} = 0, \qquad i \in \mathcal{N}.$$

If c_{ij} is the (i,j) element of matrix $\mathbf{C}$, for the first row we have

$$\mathbf{c}_1\mathbf{t} = \sum_{j=1}^{N} c_{1j}(-1)^{1+j} \det \mathbf{C}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{1\}} \overset{\text{def.}}{=} \det \mathbf{C} = 0.$$

For an arbitrary row $i = 2, \ldots, N$, we have

$$\mathbf{c}_i\mathbf{t} = \sum_{j=1}^{N} c_{ij}(-1)^{1+j} \det \mathbf{C}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{1\}}.$$

We expand the determinant of each matrix $\mathbf{C}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{1\}}$, $j \in \mathcal{N}$, in minors of the ith row of matrix $\mathbf{C}$. Observe that the ith row of the initial matrix is indexed by $i-1$ in every matrix $\mathbf{C}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{1\}}$ due to the removal of the first row of $\mathbf{C}$. Note also that every column k placed to the right of the jth column of matrix $\mathbf{C}$, after the removal of the jth column is shifted one position to the left, therefore it is indexed by $k-1$. After the above observations, we have

$$\mathbf{c}_i\mathbf{t} = \sum_{j=1}^{N} c_{ij}(-1)^{1+j} \det \mathbf{C}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{1\}} = \sum_{j=1}^{N} c_{ij}(-1)^{1+j} \sum_{k\neq j} c_{ik}(-1)^{i-1+k-\mathbb{1}_{\{k>j\}}} \det \mathbf{C}^{\mathcal{N}\setminus\{j,k\}}_{\mathcal{N}\setminus\{1,i\}}$$

$$= (-1)^i \sum_{j=1}^{N}\sum_{k\neq j} c_{ij}c_{ik}(-1)^{j+k-\mathbb{1}_{\{k>j\}}} \det \mathbf{C}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{1\}} = 0,$$

because for any two arbitrary columns m and l, with $m > l$, only the summands

$$c_{il}c_{im}(-1)^{l+m-1} \det \mathbf{C}^{\mathcal{N}\setminus\{l,m\}}_{\mathcal{N}\setminus\{1,i\}} \quad \text{and} \quad c_{im}c_{il}(-1)^{m+l} \det \mathbf{C}^{\mathcal{N}\setminus\{l,m\}}_{\mathcal{N}\setminus\{1,i\}}$$

appear in the expression of $\mathbf{c}_i\mathbf{t}$ and they cancel out with one another. Since all summands of the above double sum are coupled and cancelled out, the double sum is equal to zero. Thus, we have proven that the inner product of any column of $\mathbf{C}$ with $\mathbf{t}$ is equal to zero. Consequently, $\mathbf{t}$ is an eigenvector of matrix $\mathbf{C}$ that corresponds to its eigenvalue zero. □

Remark A.5. If the null space of an $N \times N$ matrix $\mathbf{C}$ has dimension one, then $\text{rank}\mathbf{C} = N - 1$. Therefore, there exists at least one sub-matrix of $\mathbf{C}$ such that its determinant is not equal to zero. More precisely, there exists at least one combination of row-column (m, n) with $\det \mathbf{C}^{\mathcal{N}\setminus\{n\}}_{\mathcal{N}\setminus\{m\}} \neq 0$. Thus, if all determinants $\det \mathbf{C}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{1\}}$, $j \in \mathcal{N}$, are equal to zero, we can choose the coordinates of the right eigenvector $\mathbf{t}$, which corresponds to the eigenvalue zero, as $t_j = (-1)^{m+j} \det \mathbf{C}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}$, $j \in \mathcal{N}$.

Remark A.6. If $\mathbf{t}$ is an arbitrary eigenvector that belongs to the null space of $\mathbf{C}$, then any other eigenvector $\mathbf{z}$ that belongs to the same null space is proportional to $\mathbf{t}$. Namely, there exists $\sigma \in \mathbb{R}$ such that $\mathbf{z} = \sigma\mathbf{t}$.

From Theorem A.4 and Remark A.5, we have as consequence the following corollary for the right eigenvectors of the matrix $\mathbf{E}$.

Corollary A.7. *If $m \in \mathcal{N}$ is such that $\det \mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}} \neq 0$ for at least one $j \in \mathcal{N}$, a right eigenvector $\mathbf{t}$ of the null space of $\mathbf{E}$ has coordinates*

$$t_j = (-1)^{m+j} \det \mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}, \qquad j \in \mathcal{N}.$$

We now perturb the matrix function $\mathbf{E}(s)$ by $\epsilon\mathbf{K}(s)$. Namely, we consider the matrix $\mathbf{E}(s) + \epsilon\mathbf{K}(s)$, where we assume that the matrix $\mathbf{K}(s)$ is analytic in the neighbourhood of r. If $\mathbf{K}^{(n)}$ is the nth derivative of the matrix function $\mathbf{K}(s)$ at $s = r$, the Taylor series of matrix $\mathbf{K}(s)$ around r is:

$$\mathbf{K}(s) = \mathbf{K} + (s - r)\mathbf{K}^{(1)} + \cdots = \sum_{n=0}^{\infty} \frac{(s - r)^n}{n!} \mathbf{K}^{(n)}, \tag{A.2}$$

where $\mathbf{K}^{(n)} = \mathbf{K}^{(n)}(r)$ and $\mathbf{K} = \mathbf{K}^{(0)}$. Our goal is to find the form of the eigenvectors of the null space of $\mathbf{E}(s) + \epsilon\mathbf{K}(s)$. Thus, as a first step we find the roots of the equation

$$\det\big(\mathbf{E}(s) + \epsilon\mathbf{K}(s)\big) = 0. \tag{A.3}$$

At this point, we need the following result from perturbation theory, which gives us the root of a function $f(s)$ when it is perturbed by a small amount. We include its proof for completeness.

Theorem A.8. *Let r be a simple root of an analytic function $f(s)$. For some function $h(s, \epsilon)$ and for all small real values ϵ, we define the perturbed function*

$$F(s, \epsilon) = f(s) + h(s, \epsilon). \tag{A.4}$$

If $h(s, \epsilon)$ is analytic in s and ϵ near $(r, 0)$, then $F(s, \epsilon)$ has a unique simple root $(x(\epsilon), \epsilon)$ near $(r, 0)$ for all small values of ϵ. Moreover, $x(\epsilon)$ is an analytic function in ϵ, and if $\frac{\partial}{\partial s^n} h(s, 0) = 0$, $n = 0, 1, \ldots$, then it holds

$$x(\epsilon) = r - \epsilon \frac{\frac{\partial}{\partial \epsilon} h(r, 0)}{f^{(1)}(r)} + O(\epsilon^2). \tag{A.5}$$

Proof. From the *Implicit function theorem* (Dini, 1907), we know that there exist a unique function x with $x(0) = r$, such that for all small values of ϵ, it holds that $F\big(x(\epsilon), \epsilon\big) = 0$ close to $(r, 0)$. Moreover, the function x is analytic in ϵ. To find the linear Taylor polynomial approximation of $x(\epsilon)$, which is defined as

$$x(\epsilon) = x(0) + \epsilon x^{(1)}(0) + O(\epsilon^2),$$

we differentiate the function $F\big(x(\epsilon), \epsilon\big) = 0$ as a function of ϵ. By using the chain rule we obtain

$$\frac{\partial}{\partial x(\epsilon)} F\big(x(\epsilon), \epsilon\big) x^{(1)}(\epsilon) + \frac{\partial}{\partial \epsilon} F\big(x(\epsilon), \epsilon\big) = 0$$

$$\Rightarrow$$

$$\big(f^{(1)}(x(\epsilon)) + \frac{\partial}{\partial x(\epsilon)} h(x(\epsilon), \epsilon)\big) x^{(1)}(\epsilon) + \frac{\partial}{\partial \epsilon} h(x(\epsilon), \epsilon) = 0.$$

In the latter equation, we substitute $\epsilon = 0$ and we solve it with respect to $x^{(1)}(0)$. Since r is a simple root of the function f, it holds that $f^{(1)}(r) \neq 0$ (Krantz, 1999). Thus, we have

$$f^{(1)}(r) x^{(1)}(0) + \frac{\partial}{\partial \epsilon} h(r, 0) = 0 \;\Rightarrow\; x^{(1)}(0) = -\frac{\frac{\partial}{\partial \epsilon} h(r, 0)}{f^{(1)}(r)},$$

which completes the proof. □

From Theorem A.8, we have the following lemma, which we give without proof.

Lemma A.9. *If the functions $f(s)$ and $h(s, \epsilon)$ satisfy the assumptions of Theorem A.8, and $g(s)$ is an analytic function with $g(r) \neq 0$, then the perturbed function*

$$G(s, \epsilon) = f(s)g(s) + h(s, \epsilon)g(s)$$

has the same unique simple root $(x(\epsilon), \epsilon)$ near $(r, 0)$, for all small values of ϵ, with the perturbed function $F(s, \epsilon) = f(s) + h(s, \epsilon)$. Namely $x(\epsilon) = r - \epsilon \frac{\partial}{\partial \epsilon} h(r, 0)/f^{(1)}(r) + O(\epsilon^2)$.

We also need the following property for the determinant of a square matrix.

Proposition A.10. *If $\mathbf{C}$ and $\mathbf{D}$ are $N \times N$ matrices with columns $\mathbf{C}_{\bullet i}$ and $\mathbf{D}_{\bullet i}$, $i \in \mathcal{N}$, respectively, then*

$$\det(\mathbf{C}_{\bullet 1} + \epsilon \mathbf{D}_{\bullet 1}, \ldots, \mathbf{C}_{\bullet N} + \epsilon \mathbf{D}_{\bullet N}) = \underbrace{\det(\mathbf{C}_{\bullet 1}, \ldots, \mathbf{C}_{\bullet N})}_{\det(\mathbf{C})} + \epsilon \sum_{i=1}^{N} \det(\mathbf{C}_{\bullet 1}, \ldots, \mathbf{D}_{\bullet i}, \ldots, \mathbf{C}_{\bullet N}) + O(\epsilon^2).$$

Proof. The result is an immediate consequence of the additive property of determinants. □

By combining the results of Theorem A.8 and Proposition A.10, we show in the following corollary how we can find the roots of the equation $\det\big(\mathbf{E}(s) + \epsilon \mathbf{K}(s)\big) = 0$.

Corollary A.11. *The number* $r_\epsilon = r - \epsilon\delta + O(\epsilon^2)$*, where*

$$\delta = \frac{\sum_{j=1}^{N} \det\left(\mathbf{E}_{\bullet 1}, \ldots, \mathbf{K}_{\bullet j}, \ldots, \mathbf{E}_{\bullet N}\right)}{\sum_{j=1}^{N} \det\left(\mathbf{E}_{\bullet 1}, \ldots, \mathbf{E}_{\bullet j}{}^{(1)}, \ldots, \mathbf{E}_{\bullet N}\right)},$$

is a simple root of the determinant $\det\big(\mathbf{E}(s) + \epsilon\mathbf{K}(s)\big) = 0$.

Proof. According to Proposition A.10,

$$\det\big(\mathbf{E}(s) + \epsilon\mathbf{K}(s)\big) = \det \mathbf{E}(s) + \epsilon \sum_{j=1}^{N} \det\left(\mathbf{E}_{\bullet 1}, \ldots, \mathbf{K}_{\bullet j}(s), \ldots, \mathbf{E}_{\bullet N}\right) + O(\epsilon^2).$$

Note that $\det \mathbf{E}(s)$ is an analytic function in r and its derivative is defined as

$$\frac{d}{ds} \det \mathbf{E}(s) = \sum_{j=1}^{N} \det\left(\mathbf{E}_{\bullet 1}, \ldots, \mathbf{E}^{(1)}_{\bullet j}(s), \ldots, \mathbf{E}_{\bullet N}(s)\right).$$

Since r is a simple eigenvalue of $\mathbf{E}(s)$, by the definition of the multiplicity of a root of an analytic function, it holds that $\frac{d}{ds} \det \mathbf{E}(s) \,\big|_{s=r} \neq 0$ (see Krantz (1999)). In addition, the function $\sum_{j=1}^{N} \det\left(\mathbf{E}_{\bullet 1}, \ldots, \mathbf{K}_{\bullet j}(s), \ldots, \mathbf{E}_{\bullet N}\right)$ is also analytic in the neighbourhood of r. The result is then immediate from Theorem A.8. □

According to Corollary A.11, the eigenvalue r_ϵ of the matrix $\mathbf{E}(s) + \epsilon\mathbf{K}(s)$ is simple. Consequently, the dimension of the null space of the matrix $\mathbf{E}(r_\epsilon) + \epsilon\mathbf{K}(r_\epsilon)$ is equal to one. We apply Theorem A.4 to find the eigenvectors of the matrix $\mathbf{E}(s) + \epsilon\mathbf{K}(s)$ that correspond to its eigenvalue r_ϵ. Before that though, we do the following simplification. From Eqs. (A.1)–(A.2), we have the Taylor expansion

$$\mathbf{E}(s) + \epsilon\mathbf{K}(s) = \sum_{n=0}^{\infty} \frac{(s-r)^n}{n!} \Big(\mathbf{E}^{(n)} + \epsilon\mathbf{K}^{(n)}\Big).$$

Evaluating this at the point $r_\epsilon = r - \epsilon\delta + O(\epsilon^2)$, we obtain

$$\mathbf{E}(r_\epsilon) + \epsilon\mathbf{K}(r_\epsilon) = \mathbf{E} - \epsilon\delta\mathbf{E}^{(1)} + \epsilon\mathbf{K} + O(\epsilon^2\mathbf{U}) = \mathbf{E} + \epsilon\big(\mathbf{K} - \delta\mathbf{E}^{(1)}\big) + O(\epsilon^2\mathbf{U}).$$

In the next theorem, we find the form of the right eigenvectors of a perturbed matrix.

Theorem A.12. *A right eigenvector of matrix* $\mathbf{E} + \epsilon\big(\mathbf{K} - \delta\mathbf{E}^{(1)}\big)$ *that corresponds to the eigenvalue* r_ϵ *is*

$$\mathbf{w} = \mathbf{t} - \epsilon\delta\mathbf{t}^{(1)} + \epsilon\mathbf{k} + O(\epsilon^2\mathbf{e}),$$

where $\mathbf{t}$ *is a right eigenvector of* $\mathbf{E}$ *defined as in Corollary A.7 and* $\mathbf{t}^{(1)}$ *is its derivative at* r_ϵ*. Moreover,* $\mathbf{k}$ *is an* $N \times 1$ *vector with coordinates*

$$k_j = (-1)^{m+j} \sum_{k=1}^{N-1} \det\left(\left(\mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet 1}, \ldots, \left(\mathbf{K}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet k}, \ldots, \left(\mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet N-1}\right), j \in \mathcal{N},$$

where the choice of $m \in \mathcal{N}$ *is explained in Corollary A.7.*

Proof. According to Remark A.5 and Corollary A.7, there exists an $m \in \mathcal{N}$ such that the vector $\mathbf{t}$ with coordinates

$$t_j = (-1)^{m+j} \det \mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}, \qquad j \in \mathcal{N},$$

is a right eigenvector of matrix $\mathbf{E}$. We prove that a right eigenvector that corresponds to the matrix $\mathbf{E} + \epsilon\big(\mathbf{K} - \delta\mathbf{E}^{(1)}\big)$ is $\mathbf{w}$ with coordinates

$$w_j = (-1)^{m+j} \det \Big(\mathbf{E} + \epsilon\big(\mathbf{K} - \delta\mathbf{E}^{(1)}\big)\Big)^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}, \qquad j \in \mathcal{N}.$$

By using Proposition A.10, the above equation simplifies to

$$\begin{aligned}
w_j &= (-1)^{m+j} \det \Big(\mathbf{E} + \epsilon\big(\mathbf{K} - \delta\mathbf{E}^{(1)}\big)\Big)^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}} \\
&= (-1)^{m+j} \det \mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}} + \epsilon(-1)^{m+j} \\
&\quad \times \sum_{k=1}^{N-1} \det \left(\left(\mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet 1}, \ldots, \left((\mathbf{K} - \delta\mathbf{E}^{(1)})^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet k}, \ldots, \left(\mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet N-1} \right) \\
&= (-1)^{m+j} \mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}} - \epsilon(-1)^{m+j}\delta \sum_{k=1}^{N-1} \det \left(\left(\mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet 1}, \ldots, \left(\mathbf{E}^{(1)\,\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet k}, \right. \\
&\quad \left. \ldots, \left(\mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet N-1} \right) \\
&\quad + \epsilon(-1)^{m+j} \sum_{k=1}^{N-1} \det \left(\left(\mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet 1}, \ldots, \left(\mathbf{K}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet k}, \ldots, \left(\mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet N-1} \right) \\
&= t_j - \epsilon\delta t_j^{(1)} + \epsilon k_j,
\end{aligned}$$

where $t_j^{(1)} = \dfrac{d}{ds} t_j(s)\Big|_{s=r}$ and

$k_j = (-1)^{m+j} \sum_{k=1}^{N-1} \det \left(\left(\mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet 1}, \ldots, \left(\mathbf{K}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet k}, \ldots, \left(\mathbf{E}^{\mathcal{N}\setminus\{j\}}_{\mathcal{N}\setminus\{m\}}\right)_{\bullet N-1} \right).$

Observe that $\mathbf{t}$ is not identically equal to zero, because it is an eigenvector of $\mathbf{E}$. Thus, the vector $\mathbf{w}$ is also not identically equal to zero. Therefore, according to Remark A.5, $\mathbf{w}$ is an eigenvector of the matrix $\mathbf{E} + \epsilon\big(\mathbf{K} - \delta\mathbf{E}^{(n)}\big)$, which completes the proof. □

A.3 Random walks and related results

In this section, we follow the notation introduced in Chapter 5 (and more precisely Section 5.4.1) and we provide some results that connect the probability measure $\mathbb{P}$ with $\breve{\mathbb{P}}$. We also discuss about the ladder height distribution and the overshoot, which play a crucial role in the Cramér-Lundberg approximation derived in Section 5.5.

The results of this section are widely known in the non-lattice case. Here, we present their lattice equivalents, where in most cases the extension is simple. The following lemma gives the connection between the probability measures $\mathbb{P}$ and $\breve{\mathbb{P}}$.

Lemma A.13. *For any Borel-measurable function g on n variables such that the expectations exist, the following relations hold:*

$$\mathbb{E}g(Z_1,\dots,Z_n) = \breve{\mathbb{E}}e^{-\gamma U_n} g(Z_1,\dots,Z_n), \tag{A.6}$$

$$\breve{\mathbb{E}}g(Z_1,\dots,Z_n) = \mathbb{E}e^{\gamma U_n} g(Z_1,\dots,Z_n). \tag{A.7}$$

Proof. This proof is the lattice equivalent of Asmussen (1982, Lemma 2.1). We write

$$\begin{aligned}\mathbb{E}g(Z_1,\dots,Z_n) &= \sum_{z_1\in\mathbb{Z}}\dots\sum_{z_n\in\mathbb{Z}} g(z_1,\dots,z_n)\mathbb{P}(Z=z_1)\cdots\mathbb{P}(Z=z_n)\\ &= \sum_{z_1\in\mathbb{Z}}\dots\sum_{z_n\in\mathbb{Z}} g(z_1,\dots,z_n)e^{-\gamma z_1}\breve{\mathbb{P}}(Z=z_1)\cdots e^{-\gamma z_n}\breve{\mathbb{P}}(Z=z_n)\\ &= \breve{\mathbb{E}}e^{-\gamma U_n} g(Z_1,\dots,Z_n).\end{aligned}$$

The remaining relation is proven analogously. □

Lemma A.14. *For any event $G_N \in \mathfrak{F}_{\tau(N)}$,*

$$\mathbb{P}(G_N;\tau(N)<\infty) = e^{-\gamma N}\breve{\mathbb{E}}e^{-\gamma B(N)}\mathbb{1}\left(G_N\right). \tag{A.8}$$

Proof. This proof is the lattice equivalent of Asmussen (1982, Lemma 2.2). We include it here for completeness. The definitions of $\mathfrak{F}_n$ and $\mathfrak{F}_{\tau(N)}$ imply the existence of n-dimensional Borel sets $A_{N,n}$ such that $G_N \cap \{\tau(N)<\infty\} = \{(Z_1,\dots,Z_n)\in A_{N,n}\}$. But, $U_n = N + B(N)$ on $\{\tau(N)=n\}$ so that by (A.6)

$$\begin{aligned}\mathbb{P}(G_N;\tau(N)<\infty) &= \sum_{n=1}^{\infty}\mathbb{P}\big((Z_1,\dots,Z_n)\in A_{N,n}\big)\\ &= \sum_{n=1}^{\infty}\breve{\mathbb{E}}e^{-\gamma U_n}\mathbb{1}\left((Z_1,\dots,Z_n)\in A_{N,n}\right)\\ &= e^{-\gamma N}\breve{\mathbb{E}}e^{-\gamma B(N)}\sum_{n=1}^{\infty}\mathbb{1}\left((Z_1,\dots,Z_n)\in A_{N,n}\right)\\ &= e^{-\gamma N}\breve{\mathbb{E}}e^{-\gamma B(N)}\mathbb{1}\left(G_N\right).\end{aligned}$$

□

For $N=0$, the above lemma takes the form:

Corollary A.15. *For any event $G\in\mathfrak{F}_{\tau_+}$, it holds that $\mathbb{P}(G) = \breve{\mathbb{E}}[e^{-\gamma U_{\tau_+}};G]$, since $G\subset\{\tau_+<\infty\}$.*

In the following lemma, we study the properties of the ladder height distribution $\breve{H}_+$. We also use the notation $\breve{h}_+$ for its probability mass function.

Lemma A.16. *The ladder height distribution $\breve{H}_+$ and the the renewal measure $\breve{\mathcal{U}}(dx) = \sum_{n=0}^{\infty}\breve{H}_+^{*n}(dx)$ are lattice and aperiodic ($d=1$). In addition, the ladder height distribution $\breve{H}_+$ is proper – i.e. $\left\|\breve{H}_+\right\| = 1$ – and its probability mass function is defined as $\breve{h}_+(n) = e^{\gamma n}h_+(n)$, $n\geq 1$.*

Proof. It is immediately obvious that the ladder height distribution is lattice since the increments Z_n take values on $\{-1, 0, 1, 2, \dots\}$. However, due to the definition of τ_+, the ascending ladder height U_{τ_+} takes values on $\{1, 2, \dots\}$. Now, observe that the nth ladder height can be written as:

$$U_{\tau_+(n)} = U_{\tau_+(1)} + (U_{\tau_+(2)} - U_{\tau_+(1)}) + \cdots + (U_{\tau_+(n)} - U_{\tau_+(n-1)}). \tag{A.9}$$

If we set $S_n = U_{\tau_+(n)} - U_{\tau_+(n-1)}$, $n = 1, \dots$, then by the definition (5.21) of $\tau_+(n)$ we deduct that S_n can only take positive integer values. Therefore, the epochs $U_{\tau_+(0)} = 0$, $U_{\tau_+(n)} = S_1 + \cdots + S_n$ constitute a renewal process. The associated renewal sequence $\breve{u}_0, \breve{u}_1, \dots$ is defined by $\breve{u}_k = \breve{\mathbb{P}}(U_{\tau_+(n)} = k, \text{for some } n \geq 1)$, i.e. $\breve{u}_k$ is the probability of a renewal at k. Note also that $\breve{u}_0$ =1, because $U_{\tau_+(0)} = 0$. Thus, according to Asmussen (2003, Lemma I.2.1), the renewal measure $\breve{\mathcal{U}}$ is supported on $\{0, 1, \dots\}$ and it is aperiodic.

From Theorem 5.3, we know that $\breve{\mathbb{E}}Z > 0$. Therefore, according to Asmussen (2003, Theorem VIII.2.4), the ladder height distribution $\breve{H}_+$ is proper, i.e. $\left\|\breve{H}_+\right\| = 1$. As a last step to find a connection between the ladder height distributions H_+ and $\breve{H}_+$, we use Corollary A.15. If we take $G = \{U_{\tau_+} = n\}$, where $n \geq 1$, we have that

$$\begin{aligned} h_+(n) = \mathbb{P}(U_{\tau_+} = n) &= \breve{\mathbb{E}}\left[e^{-\gamma U_{\tau_+}} \mathbb{1}\left(U_{\tau_+} = n\right)\right] = e^{-\gamma n}\mathbb{P}(U_{\tau_+} = n) \\ &= e^{-\gamma n}\breve{h}_+(n), \end{aligned} \tag{A.10}$$

which completes the proof. □

In Lemma A.17, we prove a weak convergence result of the overshoot $B(N)$ and we find the distribution of its limit $B(\infty)$. We denote weak convergence as $\overset{\breve{d}}{\Rightarrow}$. Thus, if ξ_n and ξ are random elements of a metric space $\mathfrak{T}$, $\xi_n \overset{\breve{d}}{\Rightarrow} \xi$ means that for any $f \in C_b(\mathfrak{T})$ (the bounded continuous function on $\mathfrak{T}$), it holds that $\breve{\mathbb{E}}f(\xi_n) \to \breve{\mathbb{E}}f(\xi_n)$ (Asmussen, 1982, page 147).

Lemma A.17. *$B(N) \overset{\breve{d}}{\Rightarrow} B(\infty)$, where $\breve{\mathbb{P}}(B(\infty) = n) = \dfrac{1 - \breve{H}_+(n-1)}{\breve{l}_+}$, $n \geq 1$.*

Proof. Our random walk has positive mean under the tilted probability measure. Thus, for all $N > 0$, at least one n exists with certainty so as $U_n > N$. Since $\tau(N)$ is the smallest index for which this is true, $U_{\tau(N)}$ is called the point of first entry into (N, ∞). The variable $U_{\tau(N)} - N$ is the amount by which the level N is overshot at the first entry and we want to find its distribution; namely, we want to find the probabilities $\breve{\mathbb{P}}(B(N) = n) = \breve{\mathbb{P}}(U_{\tau(N)} = N + n)$, for $n \in \mathbb{Z}^+$. In other words, $\breve{\mathbb{P}}(B(N) = n)$ corresponds to the probability that the level N is overshot exactly by an amount n.

We define now U'_1 as the point of the first entry into $(0, \infty)$ and by induction U'_{n+1} as the point of the first entry into (U'_n, ∞). The sequence $U'_1, U'_2, \dots$ coincides with the ladder heights $U_{\tau_+(1)}, U_{\tau_+(2)}, \dots$ and forms a renewal process: the differences $U'_{n+1} - U'_n$ are evidently mutually independent and have the same distribution as U_{τ_+}. Consequently, the event that the level N is overshot by an amount n, with $n \geq 1$ occurs if some renewal epoch U'_n equals $m \leq N$ and the following inter-arrival time $U'_{n+1} - U'_n$ is equal to $N - m + n$.

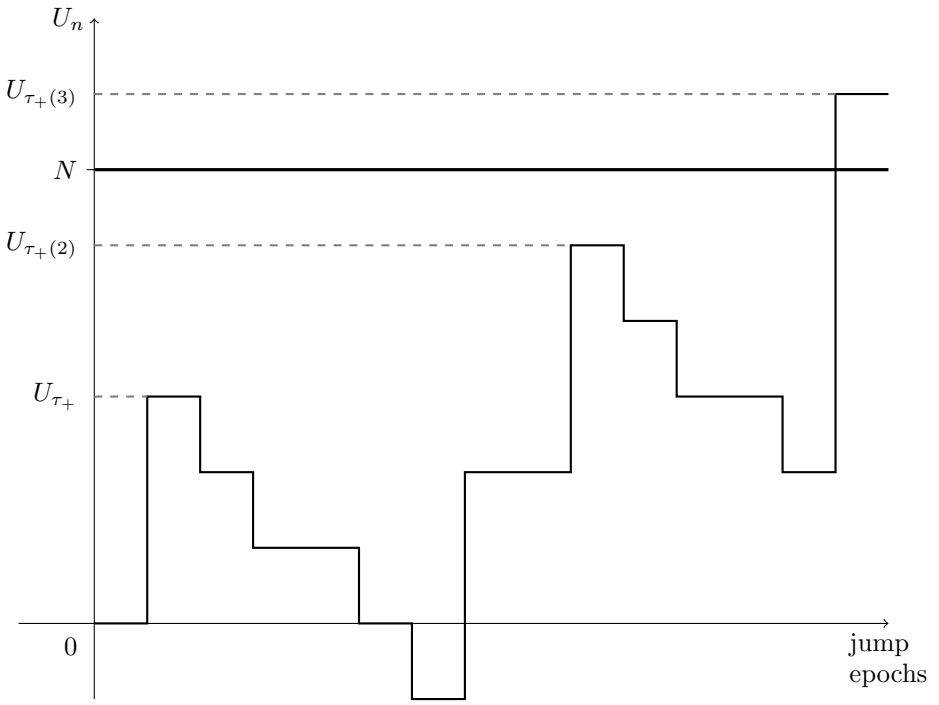


FIGURE A.1: Ladder height sample path.

Combining all the above, we have

$$\breve{\mathbb{P}}(B(N)=n)=\breve{\mathcal{U}}*\breve{h}_+(N+n)=\sum_{m=0}^{N}\breve{u}_m\breve{h}_+(N+n-m)$$
$$=\sum_{m=0}^{N}\breve{u}_{N-m}\breve{h}_+(n+m)\to\frac{1}{\breve{l}_+}\sum_{j=0}^{\infty}\breve{h}_+(n+j),\qquad N\to\infty,$$

where in the last step we applied the lattice version of the *Renewal Theorem* (see Feller (1971, page 363) and Asmussen (2003, page 157)). Now, since $\sum_{j=0}^{\infty}\breve{h}_+(n+j)=1-\breve{H}_+(n-1)$, the result is immediate. □

Summary

Error analysis of structured Markov chains

This dissertation is concerned with the error analysis of structured Markov chains. Performance measures for various classes of structured Markov processes can be found algorithmically with the aid of Matrix-Analytic Methods (MAM), where the latter techniques combine probability and matrix theory.

Although there exist cases in which performance measures for structured models can be found explicitly, often in practice, a system is described by a stochastic model which can only be analysed numerically if the background state space is truncated. From an application point of view, truncation of the background state space can be interpreted as approximating general distributions with phase-type distributions, and/or infinite waiting rooms with finite waiting rooms. On the one hand the approximation of several input parameters leads to an approximate model that can be analysed exactly or numerically, but on the other hand it introduces approximation errors.

The goal of this dissertation is to obtain a rigorous understanding of these approximation errors for a number of practically relevant classes of stochastic systems arising in risk and queueing theory. More precisely, since MAM are very successful in the numerical analysis of structured Markov processes, we combine MAM with other techniques, such as Laplace transforms, perturbation analysis, and extreme value theory, in order to derive algorithms that yield provably accurate estimates of performance measures for a wide class of systems. In addition, we focus on relating the incurred errors to the truncation levels.

In Chapters 2–4, we consider heavy-tailed models. To preserve the heavy-tailed property within the context of structured Markov processes, we should allow for a (doubly) infinite state space, which makes the numerical evaluation of performance measures cumbersome if not impossible. Specifically, in Chapters 2–3, we consider the classical compound Poisson risk model and we find approximations, with their accompanying error bounds, for the ruin probability. In broad terms, a risk reserve process is a model for the time evolution of the reserve of an insurance company, where the initial reserve is non-negative. Claims for money arrive according to a

Poisson process, the claim sizes are i.i.d. and independent of the aforementioned Poisson process, and premiums flow in at a rate one per unit time. Furthermore, the probability of ultimate ruin, is defined as the probability that the reserve ever drops below zero.

In Chapter 2, we assume that the claim sizes are heavy-tailed and we show how to approximate the heavy-tailed claim size distribution with a hyperexponential one in order to meet a predetermined accuracy for the ruin probability. In addition, we perform an extensive numerical study to compare our approximations with well-established approximations for heavy-tailed risk models. Motivated by statistical theory, we describe in Chapter 3 how the claim sizes can be written as a mixture of a phase-type and a heavy-tailed distribution. From this representation of the claim size distribution, we derive, with the aid of perturbation analysis, a series expansion for the ruin probability. Our proposed approximations consist of the first two terms of this series expansion and we refer to them collectively as corrected phase-type approximations. Finally, we prove that the corrected phase-type approximations provide small absolute and relative errors and we check their accuracy through numerical experiments.

In Chapter 4, we extend the applicability of the corrected phase-type approximations to a more involved queueing model. In particular, we consider a single server queue with FIFO discipline where customers arrive according to a Markovian Arrival Process (MArP) and their service times follow the same distribution as the claim sizes in Chapter 3, i.e. a mixture of a phase-type and a heavy-tailed distribution. For this model, we focus on the evaluation of the queuing delay, where significant correlations between arrivals of load-generating events make the numerical evaluation of such a performance measure a challenging problem. We show that the developed approximations capture the exact tail behaviour and provide bounded relative errors. We exhibit their performance with numerical examples.

Finally, in Chapter 5, we no longer focus on heavy-tailed models. However, we consider a tandem queue with batch arrivals. Customers arrive in batches according to a Poisson stream and join the first queue, while the service times in each queue are exponential. A customer leaves the system after completing service in both queues. For this model, the joint queue length distribution can be represented by a doubly-infinite Quashi-Birth-Death (QBD) process and we can apply MAM to find the steady-state distribution only if the number of customers in front of either queue is finite. To find approximations for the queue lengths we exploit the latter property, and we truncate the number of customers of the first queue. We connect our two dimensional queuing process with a two dimensional random walk and with the aid of large deviations theory we find an asymptotic upper bound for our approximations. We recognise three possible cases for the bound, study its qualitative characteristics, and test its accuracy through numerical experiments.

Bibliography

J. Abate and W. Whitt. Computing Laplace transforms for numerical inversion via continued fractions. *INFORMS Journal on Computing*, 11(4):394–405, 1999a. 23, 50

J. Abate and W. Whitt. Explicit $M/G/1$ waiting-time distributions for a class of long-tail service-time distributions. *Operations Research Letters*, 25(1):25–31, 1999b. 25, 33, 62, 117

I. Adan. *A compensation approach for queueing problems*. PhD thesis, Centrum voor Wiskunde en Informatica, 1994. 17

I. J. B. F. Adan and V. G. Kulkarni. Single-server queue with Markov-dependent inter-arrival and service times. *Queueing Systems. Theory and Applications*, 45(2): 113–134, 2003. 74, 75, 77, 78

I. J. B. F. Adan, J. Wessels, and W. H. M. Zijm. Analysing multiprogramming queues by generating functions. *SIAM Journal on Applied Mathematics*, 53(4):1123–1131, 1993. 17

S. Ahn, J. H. T. Kim, and V. Ramaswami. A new class of models for heavy tailed distributions in finance and insurance risk. *Insurance: Mathematics & Economics*, 51(1):43–52, 2012. 15

G. Alsmeyer. On the Markov renewal theorem. *Stochastic processes and their applications*, 50(1):37–56, 1994. 148

E. Altman, K. E. Avrachenkov, and R. Núñez-Queija. Perturbation analysis for denumerable Markov chains with application to queueing models. *Advances in Applied Probability*, 36(3):839–853, 2004. 17

S. Asmussen. Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the $GI/G/1$ queue. *Advances in Applied Probability*, 14(1):143–170, 1982. 163, 164

S. Asmussen. Ladder heights and the Markov-modulated $M/G/1$ queue. *Stochastic Processes and their Applications*, 37(2):313–326, 1991. 17

S. Asmussen. Phase-type representations in random walk and queueing problems. *Annals of Probability*, 20(2):772–789, 1992a. 11, 23

S. Asmussen. Light traffic equivalence in single-server queues. *The Annals of Applied Probability*, 2(3):555–574, 1992b. 24

S. Asmussen. Extreme value theory for queues via cycle maxima. *Extremes. Statistical Theory and Applications in Science, Engineering and Economics*, 1(2):137–168, 1998. 132, 133

S. Asmussen. Matrix-analytic models and their analysis. *Scandinavian Journal of Statistics. Theory and Applications*, 27(2):193–226, 2000. 11, 73

S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, New York, 2003. 6, 10, 15, 17, 18, 25, 48, 73, 74, 130, 131, 132, 135, 164, 165

S. Asmussen and H. Albrecher. *Ruin Probabilities*. Advanced Series on Statistical Science & Applied Probability, 14. World Scientific, Second edition, 2010. 17, 18, 25, 26, 37, 57, 58, 74, 130, 141, 157

S. Asmussen and M. Bladt. Renewal theory and queueing algorithms for matrix-exponential distributions. In *Matrix-analytic methods in stochastic models*, volume 183 of *Lecture Notes in Pure and Applied Mathematics*, pages 313–341. Marcel Dekker, Inc., New York, 1997. 11

S. Asmussen and G. Koole. Marked point processes as limits of Markovian arrival streams. *Journal of Applied Probability*, 30(2):365–372, 1993. 14

S. Asmussen and C. A. O'Cinneide. *Matrix-Exponential Distributions*. John Wiley & Sons, Inc., 2004. 11

S. Asmussen and D. Perry. On cycle maxima, first passage problems and extreme value theory for queues. *Communications in Statistics. Stochastic Models*, 8(3): 421–458, 1992. 17

S. Asmussen and M. Pihlsgård. Performance analysis with truncated heavy-tailed distributions. *Methodology and Computing in Applied Probability*, 7(4):439–457, 2005. 15

S. Asmussen and V. Ramaswami. Probabilistic interpretations of some duality results for the matrix paradigms in queueing theory. *Communications in Statistics. Stochastic Models*, 6(4):715–733, 1990. 10

S. Asmussen, O. Nerman, and M. Olson. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23(4):419–441, 1996. 12, 48

S. Asmussen, F. Avram, and M. R. Pistorius. Russian and American put options under exponential phase-type Lévy models. *Stochastic Processes and their Applications*, 109(1):79–111, 2004. 3

K. E. Avrachenkov and J. B. Lasserre. The fundamental matrix of singularly perturbed Markov chains. *Advances in Applied Probability*, 31(3):679–697, 1999. 18

A. L. Badescu, E. C. Cheung, and D. Landriault. Dependent risk models with bivariate phase-type distributions. *Journal of Applied Probability*, 46(1):113–131, 2009. 3

N. G. Bean, M. Fackrell, and P. Taylor. Characterization of matrix-exponential distributions. *Stochastic Models*, 24(3):339–363, 2008. 11

J. A. Beekman. A ruin function approximation. *Transactions of the Society of Actuaries*, 21:41–48, 1969. 24

D. Bini, B. Meini, S. Steffé, and B. Van Houdt. Structured Markov chains solver: software tools. In *Proceeding from the 2006 workshop on Tools for solving structured Markov chains*, page 14. ACM, 2006. 8

D. A. Bini, B. Meini, and V. Ramaswami. Analyzing M/G/1 paradigms through QBDs: the role of the block structure in computing the matrix G. In *Advances in Matrix Analytic Methods for Stochastic Models*, pages 73—86. Notable Publications Neshanic Station, NJ, 2000. 9

D. A. Bini, G. Latouche, and B. Meini. *Numerical Methods for Structured Markov Chains*. Numerical Mathematics and Scientific Computation. Oxford University Press, 2005. 2, 8

M. Bladt and M. F. Neuts. Matrix-exponential distributions: Calculus and interpretations via flows. *Stochastic Models*, 19(1):113–124, 2003. 11

M. Bladt, B. F. Nielsen, and G. Samorodnitsky. Calculation of ruin probabilities for a dense class of heavy tailed distributions. *Scandinavian Actuarial Journal*, 2014. Accepted. 15

J. P. C. Blanc. On a numerical method for calculating state probabilities for queueing systems with more than one waiting line. In *Proceedings of the 2nd international conference on computational and applied mathematics (Leuven, 1986)*, volume 20, pages 119–125, 1987. 18

J. P. C. Blanc. The power-series algorithm applied to cyclic polling systems. *Communications in Statistics. Stochastic Models*, 7(4):527–545, 1991. 18

J. P. C. Blanc. The power-series algorithm applied to the shortest-queue model. *Operations Research*, 40(1):157–167, 1992. 18

J. Blanchet and B. Zwart. Asymptotic expansions of defective renewal equations with applications to perturbed risk models and processor sharing queues. *Mathematical Methods of Operations Research*, 72(2):311–326, 2010. 18

P. Bloomfield and D. R. Cox. A low traffic approximation for queues. *Journal of Applied Probability*, 9(4):832–840, 1972. 24

V. A. Bolotin. Modeling call holding time distributions for CCS network design and performance analysis. *Selected Areas in Communications, IEEE Journal on*, 12(3): 433–438, 1994. 15

A. A. Borovkov and S. Foss. Stochastically recursive sequences. *Siberian Advances in Mathematics*, 2(1):16–81, 1992. 15, 32

O. J. Boxma and D. Perry. A queueing model with dependence between service and interarrival times. *European Journal of Operational Research*, 128(3):611–624, 2001. 74

L. Breuer. Parameter estimation for a class of BMAPs. In *Advances in Matrix-Analytic Methods for Stochastic Models*, pages 87–97. Notable Publications Neshanic Station, NJ, 2000. 14

L. Breuer. An EM algorithm for Batch Markovian Arrival Processes and its comparison to a simpler estimation procedure. *Annals of Operations Research*, 112(1-4):123–138, 2002. 14

L. Breuer and D. Baum. *An Introduction to Queueing Theory: and Matrix-Analytic Methods*. Springer, 2005. 3

M. Brown. Error bounds for exponential approximations of geometric convolutions. *The Annals of Probability*, 18(3):1388–1402, 1990. 25, 32, 37, 44

M. Brown, E. A. Peköz, and S. M. Ross. Some results for skip-free random walk. *Probability in the Engineering and Informational Sciences*, 24(4):491–507, 2010. 135

E. Çinlar. Markov additive processes. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 24(2):85–93, 1972. 147

J. W. Cohen. *The Single Server Queue*. North-Holland Publishing Co., 1982. 17, 30

J. W. Cohen and O. J. Boxma. *Boundary Value Problems in Queueing System Analysis*. North-Holland, 1983. 17

D. R. Cox. A use of complex probabilities in the theory of stochastic processes. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 313–319. Cambridge University Press, 1955a. 11

D. R. Cox. The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 433–441. Cambridge University Press, 1955b. 11

S. Cox, Y. Lin, R. Tian, and L. Zuluaga. Bounds for ruin probabilities and value at risk. Technical report, Actuarial Research Clearing House Newsletters 1, 2008. 48

D. J. Daley and T. Rolski. A light traffic approximation for a single-server queue. *Mathematics of Operations Research*, 9(4):624–628, 1984. 24

D. J. Daley and T. Rolski. Light traffic approximations in queues. *Mathematics of Operations Research*, 16(1):57–71, 1991. 24

R. Davis and S. Resnick. Tail estimates motivated by extreme value theory. *The Annals of Statistics*, 12(4):1467–1487, 1984. 47

L. De Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer-Verlag, 2007. 122

F. De Terán. On the perturbation of singular analytic matrix functions: a generalization of Langer and Najman's results. *Operators and Matrices*, 5(4):553–564, 2011. 92, 157

F. De Vylder. A practical solution to the problem of ultimate ruin probability. *Scandinavian Actuarial Journal*, 1978(2):114–119, 1978. 24

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. 12

U. Dini. *Lezioni di Analisi Infinitesimale*, volume 1. Fratelli Nistri, Pisa, 1907. 160

R. Doney, R. Maller, and M. Savov. Renewal theorems and stability for the reflected process. *Stochastic Processes and their Applications*, 119(4):1270 – 1297, 2009. 142, 143

D. E. Duffy, A. A. McIntosh, M. Rosenstein, and W. Willinger. Statistical analysis of CCSN/SS7 traffic data from working CCS subnetworks. *IEEE Journal on Selected Areas in Communications*, 12(3):544–551, 1994. 15

P. Embrechts and G. Samorodnitsky. Ruin theory revisited: stochastic models for operational risk. In *Risk Management for Central Bank Foreign Reserve*, pages 243–261. ECB, 2004. 48

P. Embrechts and N. Veraverbeke. Estimates for the probability of ruin with special emphasis on the possibility of large claims. *Insurance: Mathematics & Economics*, 1(1):55–72, 1982. 15, 25, 32

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events: for Insurance and Finance*, volume 33 of *Applications of Mathematics*. Springer-Verlag, 1997. 15, 48, 54, 59, 74

M. Fackrell. An alternative characterization for matrix exponential distributions. *Advances in Applied Probability*, 41(4):1005–1022, 2009. 11

M. W. Fackrell. *Characterization of Matrix-Exponential distributions*. PhD thesis, University of Adelaide, School of Applied Mathematics, 2003. 10, 11

G. Fayolle and R. Iasnogorodski. Two coupled processors: the reduction to a Riemann-Hilbert problem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(3):325–351, 1979. 17

A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31(3–4):245–279, 1998. 3, 12, 15, 28

W. Feller. *An Introduction to Probability Theory and its Applications, Vol. II*. John Wiley & Sons Inc., Second edition, 1971. 28, 110, 149, 165

W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18(2):149–171, 1993. 13

L. Flatto and H. P. McKean. Two queues in parallel. *Communications on Pure and Applied Mathematics*, 30(2):255–263, 1977. 17

S. Foss, Z. Palmowski, and S. Zachary. The probability of exceeding a high boundary on a random time interval for a heavy-tailed random walk. *The Annals of Applied Probability*, 15(3):1936–1957, 2005. 15

C. D. Fuh and T. L. Lai. Wald's equations, first passage times and moments of ladder variables in markov random walks. *Journal of applied probability*, 35(3):566–580, 1998. 149

H. R. Gail, S. L. Hantler, and B. A. Taylor. Spectral analysis of $M/G/1$ and $G/M/1$ type Markov chains. *Advances in Applied Probability*, 28(1):114–165, 1996. 2

W. K. Grassmann and D. A. Stanford. Matrix analytic methods. In *Computational Probability*, volume 24 of *International Series in Operations Research & Management Science*, pages 153–203. Springer US, 2000. 7

E. J. Gümbel. *Statistics of extremes*. Columbia University Press, 1958. 133

R. Gusella. Characterizing the variability of arrival processes with indexes of dispersion. *IEEE Journal on Selected Areas in Communications*, 9(2):203–211, 1991. 14

A. Gut. *Stopped random walks: limit theorems and applications*. Springer, Second edition, 2009. 144

E. Haeusler and J. L. Teugels. On asymptotic normality of Hill's estimator for the exponent of regular variation. *The Annals of Statistics*, 13(2):743–756, 1985. 47

C. M. Harris and W. G. Marchal. Distribution estimation using Laplace transforms. *INFORMS Journal on Computing*, 10(4):448–458, 1998. 24

C. M. Harris, P. H. Brill, and M. J. Fischer. Internet-type queues with power-tailed interarrival times and computational methods for their analysis. *INFORMS Journal on Computing*, 12(4):257–260, 2000. 24

M. Haviv and Y. Ritov. On series expansions and stochastic matrices. *SIAM Journal on Matrix Analysis and Applications*, 14(3):670–676, 1993. 17

Q. M. He. *Fundamentals of Matrix-Analytic Methods*. Springer, 2014. 3, 10

B. Heidergott, A. Hordijk, and H. Leahu. Strong bounds on perturbations. *Mathematical Methods of Operations Research*, 70(1):99–127, 2009. 18

A. Heindl, Q. Zhang, and E. Smirni. ETAQA truncation models for the MAP/MAP/1 departure process. In *Proceedings First International Conference on the Quantitative Evaluation of Systems*, pages 100–109, 2004. 3

B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975. 47

M. Hofri. A generating-function analysis of multiprogramming queues. *International Journal of Computer and Information Sciences*, 7(2):121–155, 1978. 17

G. Hooghiemstra, M. Keane, and S. van de Ree. Power series for stationary distributions of coupled processor models. *SIAM Journal on Applied Mathematics*, 48(5):1159–1166, 1988. 18

R. A. Horn and C. R. Johnson, editors. *Matrix Analysis*. Cambridge University Press, 1986. 6

A. Horváth and M. Telek. Approximating heavy tailed behavior with phase type distributions. In *Advances in Algorithmic Methods for Stochastic Models, Proceedings of the Third International Conference on Matrix Analytic Methods*, pages 191–214, 2000. 3, 12, 15

A. Horváth and M. Telek. PhFit: A general phase-type fitting tool. In *Computer Performance Evaluation: Modelling Techniques and Tools*, volume 2324 of *Lecture Notes in Computer Science*, pages 82–91. Springer Berlin Heidelberg, 2002. 12

A. Horváth and M. Telek. Matching more than three moments with acyclic phase type distributions. *Stochastic Models*, 23(2):167–194, 2007. 12

G. Horváth and H. Okamura. A fast EM algorithm for fitting marked Markovian arrival processes with a new special structure. In *Computer Performance Engineering*, volume 8168 of *Lecture Notes in Computer Science*, pages 119–133. Springer Berlin Heidelberg, 2013. 14

D. L. Iglehart. Extreme values in the GI/G/1 queue. *The Annals of Mathematical Statistics*, 43(2):627–635, 1972. 132

J. R. Jackson. Jobshop-like queueing systems. *Management Science*, 10(1):131–142, 1963. 16

N. P. Jewell. Mixtures of exponential distributions. *The Annals of Statistics*, 10(2): 479–484, 1982. 34

M. A. Johnson and M. R. Taaffe. Matching moments to phase distributions: Mixtures of Erlang distributions of common order. *Communications in Statistics. Stochastic Models*, 5(4):711–743, 1989. 12

V. Kalashnikov. *Geometric sums: bounds for rare events with applications: risk analysis, reliability, queueing*, volume 413 of *Mathematics and its Applications*. Springer, 1997. 23, 32

V. Kalashnikov. Stability bounds for queueing models in terms of weighted metrics. In *Analytic Methods in Applied Probability*, volume 207 of *American Mathematical Society Translations Ser. 2*, pages 77–90. American Mathematical Society, 2002. 15

V. Kalashnikov and R. Norberg. Power tailed ruin probabilities in the presence of risky investments. *Stochastic Processes and their Applications*, 98(2):211–228, 2002. 15

V. Kalashnikov and G. Tsitsiashvili. Tails of waiting times and their bounds. *Queueing Systems. Theory and Applications*, 32(1-3):257–283, 1999. 24

F. P. Kelly. *Reversibility and Stochastic Networks*. Wiley, 1979. 121

D. G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, 24(3):338–354, 1953. 4

M. Kijima. *Markov processes for stochastic modeling*. Springer, 1997. 128

J. F. C. Kingman. Two similar queues in parallel. *Annals of Mathematical Statistics*, 32(4):1314–1323, 1961. 17

J. F. C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society. Series B. Methodological*, 24(2):383–392, 1962. 23, 32

L. Kleinrock. *Queueing Systems, Vol. 1: Theory*. Wiley-Interscience, 1976. 6, 122

S. A. Klugman, H. H. Panjer, and G. E. Willmot. *Loss models: from data to decisions*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Third edition, 2008. 48

D. Korshunov. How to measure the accuracy of the subexponential approximation for the stationary single server queue. *Queueing Systems. Theory and Applications*, 68 (3-4):261–266, 2011. 15

S. G. Krantz. *Handbook of complex variables*. Birkhäuser Boston Inc., 1999. 160, 161

D. P. Kroese, W. R. W. Scheinhardt, and P. G. Taylor. Spectral properties of the tandem Jackson network, seen as a Quasi-Birth-and-Death process. *The Annals of Applied Probability*, 14(4):2057–2089, 2004. 16

P. Lancaster, A. S. Markus, and F. Zhou. Perturbation theory for analytic matrix functions: the semisimple case. *SIAM Journal on Matrix Analysis and Applications*, 25(3):606–626, 2003. 92

G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. SIAM, 1999. 8, 9, 10, 16, 121, 124, 129

S. Lavenberg. *Computer performance modeling handbook*. Elsevier, 1983. 121

A. M. Law, W. D. Kelton, and W. D. Kelton. *Simulation modeling and analysis*. McGraw-Hill New York, 1991. 12

L. Lipsky. *Queueing Theory – a Linear Algebraic Approach*. Springer-Verlag, Second edition, 2009. 11

J. D. C. Little. A proof for the queuing formula: $L = \lambda W$. *Operations Research*, 9(3): 383–387, 1961. 136

D. M. Lucantoni. New results on the single-server queue with a batch Markovian arrival process. *Stochastic Models*, 7(1):1–46, 1991. 13, 17, 74

D. M. Lucantoni. The BMAP/G/1 queue: A tutorial. In *Models and Techniques for Performance Evaluation of Computer and Communications Systems*, volume 729 of *Lecture Notes in Computer Science*, pages 330–358. Springer Verlag, 1993. 13, 14, 74

D. M. Lucantoni, K. S. Meier-Hellstern, and M. F. Neuts. A single-server queue with server vacations and a class of nonrenewal arrival processes. *Advances in Applied Probability*, 22(3):676–705, 1990. 13, 17

D. M. Lucantoni, G. L. Choudhury, and W. Whitt. The transient BMAP/G/1 queue. *Communications in Statistics. Stochastic Models*, 10(1):145–182, 1994. 74

H. B. Mann and A. Wald. On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14(3):217–226, 1943. 142

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tool*. Princeton University Press, 2005. 48

K. S. Meier-Hellstern. A fitting algorithm for markov-modulated poisson processes having two arrival rates. *European Journal of Operational Research*, 29(3):370–377, 1987. 14

B. Meini. Solving M/G/1 type markov chains: recent advances and applications. *Communications in Statistics. Stochastic Models*, 14(1–2):479–496, 1998. 9

G. V. Moustakides. Extension of Wald's first lemma to Markov processes. *Journal of Applied Probability*, 36(1):48–59, 1999. 149

M. F. Neuts. Probability distributions of phase type. In *Liber Amicorum Prof. Emeritus H. Florin*, pages 173–206. Department of Mathematics, University of Louvain, Belgium, 1975. 10

M. F. Neuts. Renewal processes of phase type. *Naval Research Logistics Quarterly*, 25 (3):445–454, 1978. 13

M. F. Neuts. A versatile Markovian point process. *Journal of Applied Probability*, 16 (4):764–779, 1979. 12, 13

M. F. Neuts. *Structured Stochastic Matrices of $M/G/1$ Type and their Applications*, volume 5 of *Probability: Pure and Applied*. Marcel Dekker Inc., 1989. 8, 9, 13, 73

M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models*. Dover Publications Inc., 1994. Corrected reprint of the 1981 original. 8, 9, 11, 49

B. F. Nielsen. Modelling long-range dependent and heavy-tailed phenomena by matrix-analytic methods. In *Advances in Algorithmic Methods for Stochastic Models*, pages 265–278. Notable Publications Inc., 2000. 3

C. A. O'Cinneide. Phase-type distributions: open problems and a few properties. *Communications in Statistics. Stochastic Models*, 15(4):731–757, 1999. 12

M. Olvera-Cravioto, J. Blanchet, and P. Glynn. On the transition from heavy traffic to heavy tails for the $M/G/1$ queue: the regularly varying case. *The Annals of Applied Probability*, 21(2):645–668, 2011. 15, 25, 32

A. Ost. *Performance of communication systems: a model-based approach with matrix-geometric methods*. Springer, 2001. 3

A. G. Pakes. On the tails of waiting-time distributions. *Journal of Applied Probability*, 12(3):555–564, 1975. 15, 32

J. F. Pérez, J. Van Velthoven, and B. Van Houdt. Q-MAM: A tool for solving infinite queues using matrix-analytic methods. In *Proceedings of the 3rd International Conference on Performance Evaluation Methodologies and Tools*, ValueTools '08, pages 16:1–16:9. ICST, 2008. 8

N. U. Prabhu. On the ruin problem of collective risk theory. *Annals of Mathematical Statistics*, 32(3):757–764, 1961. 18

V. Ramaswami. The $N/G/1$ queue and its detailed analysis. *Advances in Applied Probability*, 12(1):222–261, 1980. 13, 17, 73

V. Ramaswami. A stable recursion for the steady state vector in Markov chains of M/G/1 type. *Communications in Statistics. Stochastic Models*, 4(1):183–188, 1988. 9

V. Ramaswami. A duality theorem for the matrix paradigms in queueing theory. *Communications in Statistics. Stochastic Models*, 6(1):151–161, 1990. 10, 23, 73

V. Ramaswami and G. Latouche. A general class of Markov processes with explicit matrix-geometric solutions. *Operations Research Spektrum*, 8(4):209–218, 1986. 9, 129

V. Ramaswami and P. G. Taylor. Some properties of the rate operators in level dependent Quasi-Birth-and-Death processes with countable number of phases. *Stochastic Models*, 12(1):143–164, 1996. 121

G. J. K. Regterschot and J. H. A. de Smit. The queue $M/G/1$ with Markov modulated arrivals and services. *Mathematics of Operations Research*, 11(3):465–483, 1986. 17

P. Reinecke, T. Krauß, and K. Wolter. Phase-type fitting using HyperStar. In *Computer Performance Engineering*, volume 8168 of *Lecture Notes in Computer Science*, pages 164–175. Springer Berlin Heidelberg, 2013. 12

S. I. Resnick. *Extreme values, regular variation, and point processes*. Springer, 2007a. 122

S. I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering. Springer, 2007b. 47

A. Riska and E. Smirni. MAMSolver: A matrix analytic methods tool. In *TOOLS '02: Proceedings of the 12th International Conference on Computer Performance Evaluation, Modelling Techniques and Tools*, pages 205–211. Springer-Verlag, 2002a. 8

A. Riska and E. Smirni. M/G/1-type markov processes: A tutorial. In *Performance Evaluation of Complex Systems: Techniques and Tools*, volume 2459 of *Lecture Notes in Computer Science*, pages 36–63. Springer Berlin Heidelberg, 2002b. 7, 9

A. Riska and E. Smirni. ETAQA Solutions for Infinite Markov Processes with Repetitive Structure. *INFORMS Journal on Computing*, 19(2):215–228, 2007. 8

A. Riska, V. Diev, and E. Smirni. An EM-based technique for approximating long-tailed data sets with PH distributions. *Performance Evaluation*, 55(1–2):147–164, 2004. 12

T. Rolski, H. Schmidli, V. Schmidt, and J. Teugels. *Stochastic Processes for Insurance and Finance*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., 1999. 15, 49

H. Rootzén. Maxima and exceedances of stationary Markov chains. *Advances in Applied Probability*, 20(2):371–390, 1988. 132

S. M. Ross. *Stochastic Processes*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., Second edition, 1996. 144

T. Rydén. An EM algorithm for estimation in Markov-modulated Poisson processes. *Computational Statistics & Data Analysis*, 21(4):431–447, 1996. 14

Y. Sasaki, H. Imai, M. Tsunoyame, and I. Ishii. Approximation of probability distribution functions by Coxian distribution to evaluate multimedia systems. *Systems and Computers in Japan*, 35(2):16–24, 2004. 15

R. Schassberger. *Warteschlangen*. Springer-Verlag, 1973. 11

E. Seneta. *Nonnegative Matrices and Markov Chains*. Springer Series in Statistics. Springer-Verlag, Second edition, 1981. 14

B. Sengupta. The semi-Markovian queue: theory and applications. *Communications in Statistics. Stochastic Models*, 6(3):383–413, 1990. 73

J. F. Shortle, P. H. Brill, M. J. Fischer, D. Gross, and D. M. B. Masi. An algorithm to compute the waiting time distribution for the $M/G/1$ queue. *INFORMS Journal on Computing*, 16(2):152–161, 2004. 24

A. Shwartz and A. Weiss. *Large deviations for performance analysis: queues, communication and computing*. Chapman & Hall, 1995. 129

K. Sigman. Light traffic for workload in queues. *Queueing Systems. Theory and Applications*, 11(4):429–442, 1992. 24

K. Sigman. Appendix: A primer on heavy-tailed distributions. *Queueing Systems. Theory and Applications*, 33(1–3), 1999. 15

D. S. Silvestrov. *Limit theorems for randomly stopped stochastic processes*. Probability and its Applications (New York). Springer-Verlag London Ltd., 2004. 18

S. R. Smits, M. Wagner, and A. G. de Kok. Determination of an order-up-to policy in the stochastic economic lot scheduling model. *International Journal of Production Economics*, 90(3):377–389, 2004. 74

D. Starobinski and M. Sidi. Modeling and analysis of power-tail distributions via classical teletraffic methods. *Queueing Systems. Theory and Applications*, 36(1-3): 243–267, 2000. 15, 24

Y. Takahashi. Asymptotic exponentiality of the tail of the waiting-time distribution in a $PH/PH/c$ queue. *Advances in Applied Probability*, 13(3):619–630, 1981. 11

T. Takine and T. Hasegawa. The workload in the MAP/G/1 queue with state-dependent services: its application to a queue with preemptive resume priority. *Communications in Statistics. Stochastic Models*, 10(1):183–204, 1994. 13, 77, 78

J. L. Teugels. The class of subexponential distributions. *The Annals of Probability*, 3 (6):1000–1011, 1975. 15, 32, 157

A. Thuümmler, P. Buchholz, and M. Telek. A novel approach for phase-type fitting with the EM algorithm. *IEEE Transactions on Dependable and Secure Computing*, 3(3):245–258, 2006. 12

L. Tonelli. Sull'integrazione per parti. *Atti della Accademia Nazionale dei Lincei (5)*, 18(2):246–253, 1909. 135

G. J. van Houtum, W. H. M. Zijm, I. J. B. F. Adan, and J. Wessels. Bounds for performance characteristics: A systematic approach via cost structures. *Communications in Statistics. Stochastic Models*, 14(1-2):205–224, 1998. Special issue in honor of Marcel F. Neuts. 126

J. Van Velthoven, B. Van Houdt, and C. Blondia. Simultaneous transient analysis of QBD Markov chains for all initial configurations using a level based recursion. In *Proceedings of the Fourth International Conference on Quantitative Evaluation of Systems*, QEST '07, pages 79–90. IEEE, 2007. 8

M. van Vuuren. *Performance analysis of manufacturing systems: queueing approximations and algorithms*. PhD thesis, Technische Universiteit Eindhoven, 2007. 16

E. Vatamidou, I. Adan, M. Vlasiou, and B. Zwart. Corrected phase-type approximations for the workload of the MAP/G/1 queue with heavy-tailed service times. *SIGMETRICS Performance Evaluation Review*, 41(2):53–55, 2013a. 19

E. Vatamidou, I. J. B. F. Adan, M. Vlasiou, and B. Zwart. Corrected phase-type approximations of heavy-tailed risk models using perturbation analysis. *Insurance: Mathematics and Economics*, 53(2):366–378, 2013b. 19

E. Vatamidou, I. Adan, M. Vlasiou, and B. Zwart. On the accuracy of phase-type approximations of heavy-tailed risk models. *Scandinavian Actuarial Journal*, 2014 (6):510–534, 2014a. 19

E. Vatamidou, I. J. B. F. Adan, M. Vlasiou, and B. Zwart. Corrected phase-type approximations of heavy-tailed queueing models in a Markovian environment. *Stochastic Models*, 30(4):598–638, 2014b. 19

B. von Bahr. Asymptotic ruin probabilities when exponential moments do not exist. *Scandinavian Actuarial Journal*, 1975(1):6–10, 1975. 15, 32

W. R. Wade. The bounded convergence theorem. *The American Mathematical Monthly*, 81(4):387–389, 1974. 142

A. Wald. On cumulative sums of random variables. *Annals of Mathematical Statistics*, 15(3):283–296, 1944. 135

D. Wallace. Asymptotic approximations to distributions. *Annals of Mathematical Statistics*, 29(3):635–654, 1958. 18

W. Whitt. *Stochastic-Process Limits: An introduction to stochastic-process limits and their application to queues*. Springer Series in Operations Research. Springer-Verlag, 2002. 141

B. Zwart. *Queueing Systems with Heavy Tails*. PhD thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2001. 15

Curriculum Vitae

Eleni Vatamidou was born in Thessaloniki, Greece, on February 2, 1986. She studied Mathematics at the Aristotle University of Thessaloniki, where she graduated with honours from the Faculty of Exact Sciences in July 2007. In the following year, she was involved with teaching mathematics and statistics to high school and university students. In September 2008, she was admitted to the master program Statistics and Modelling at the Department of Mathematics of the Aristotle University of Thessaloniki. She obtained her Master's degree in July 2010.

On November 15, 2010, she started her PhD research project at the Department of Mathematics and Computer Science of the Technical University of Eindhoven, under the supervision of Ivo Adan, Maria Vlasiou, and Bert Zwart. Her research programme was supported by NWO and a scholarship for post-graduate studies abroad from the Legacy of K. Katseas awarded by the Aristotle University of Thessaloniki. Eleni's research resulted in a number of publications and conference talks. For one of these publications, she received the Marcel Neuts Student Paper Award at the 8th International Conference on Matrix Analytic Methods in Stochastic Models (MAM) in January 2014. In addition, she received a Best Paper Award for her presentation at the 8th Samos Conference in Actuarial Science and Finance in June 2014.

During her stay in the Netherlands, Eleni was also involved in teaching, served as a referee for scientific publications, and elected as the PhD representative of the LNMB program (Landelijk Netwerk Mathematische Besliskunde – Dutch Network on the Mathematics of Operations Research) for the calendar year 2013. The PhD project ends with the realisation of this thesis, which Eleni defends at TU/e on June 22, 2015.

Cover design by Eleni Vatamidou with Canva
Printed by Gildeprint.