Asymptotic analysis of stochastic systems

van Kreveld, L.

Publication date
2023
Document Version
Final published version

Asymptotic Analysis of Stochastic Systems

Lucas van Kreveld

Asymptotic Analysis of Stochastic Systems

Lucas van Kreveld

Printed by: Ipskamp Printing
Cover design: Kyra van der Velde
ISBN: 978-94-6473-086-9

Het hier beschreven onderzoek/de uitgave van dit proefschrift werd mede mogelijk gemaakt door steun van de Nederlandse organisatie voor Wetenschappelijk Onderzoek (NWO) door middel van het Zwaartekracht-project NETWORKS-024.002.003.

Asymptotic Analysis of Stochastic Systems

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 9 mei 2023, te 12.00 uur

door Lucas Renier van Kreveld
geboren te Houten

Promotiecommissie

Promotores:	prof. dr. M.R.H. Mandjes	Universiteit van Amsterdam
	prof. dr. ir. O.J. Boxma	Technische Universiteit Eindhoven
Copromotores:	dr. J.L. Dorsman	Universiteit van Amsterdam
Overige leden:	prof. dr. R. Nunez Queija	Universiteit van Amsterdam
	prof. dr. R.J.A. Laeven	Universiteit van Amsterdam
	prof. dr. P.J.C. Spreij	Universiteit van Amsterdam
	prof. dr. R.J. Boucherie	University of Twente
	prof. dr. ir. K.E.E.S. De Turck	Ghent University

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Contents

Part II: A scaling approach for queueing networks

Part III: Extremes of Markov additive processes

1 Introduction

One of the big philosophical questions in life is whether there is free will or everything is predetermined. In either perspective, randomness seems to be non-existent. The outcome of a coin toss is determined by the precise manner in which the coin is tossed. Whether this manner is freely chosen or the result of a series of chemical and physical reactions in your brain, there is no randomness involved. Even so, it is hard to predict the outcome of the coin toss simply because we are unable to take into account every little detail that has an effect on the outcome. One can imagine that the effects of processes underlying events more involved than a simple coin toss are nearly impossible to incorporate. In a world where no event can be predicted with perfect accuracy, the mathematical theory behind randomness, probability theory, serves as an incredibly useful alternative.

Simple random events (such as a coin toss) can be represented by *random variables*: variables that have multiple possible outcomes according to given probabilities (for a coin toss: heads and tails each with probability $\frac{1}{2}$). The set of all outcomes together with their probabilities is called the *distribution* of the random variable.

Rather than singular events, phenomena in practice often evolve over time (weather, traffic, stock values, download speed etc.). Where isolated events are modeled by random variables, these time-dependent phenomena can be modeled by *stochastic processes*. The idea is that at each point in time the process has a certain value, which is determined by the outcome of a random experiment. One could think of a gambler's profit after each spin of the roulette wheel, the number of cars in front of a given traffic light throughout the day, or the evolution of the temperature at a fixed location. Each of these examples comes with its own questions. What is the gambler's average profit or loss per spin? How likely is it that a car arriving at the traffic light finds more than five cars already present? What will be the temperature ten days from now? Mathematical analysis of stochastic processes is required to answer these questions.

Mathematical analysis is a very broad concept, as there exist numerous approaches to analyze a mathematical problem. On the one hand, we have exact approaches leading to provably correct mathematical formulas. Although their perfect accuracy makes them preferable to other approaches, exact solutions are often beyond reach and thus they are not always available for any given problem. On the other hand, we have approximate approaches including simulation and the exact study of simplified models. These approaches are more often available, but come at the cost of inaccuracy.

In this thesis, we focus on a set of approaches that can be both exact and approximate: *asymptotics*. While a proper description of this word's meaning is given in Section 1.2, a short definition could be that asymptotics are mathematical characterizations of limiting behavior. As an example, say that the value of a function $g(x)$ becomes arbitrarily close to the value of $f(x)$ when x is large enough, as in Figure 1.1, where $g(x) = \sqrt{x^2 - 2x + 4}$ and $f(x) = x - 1$.

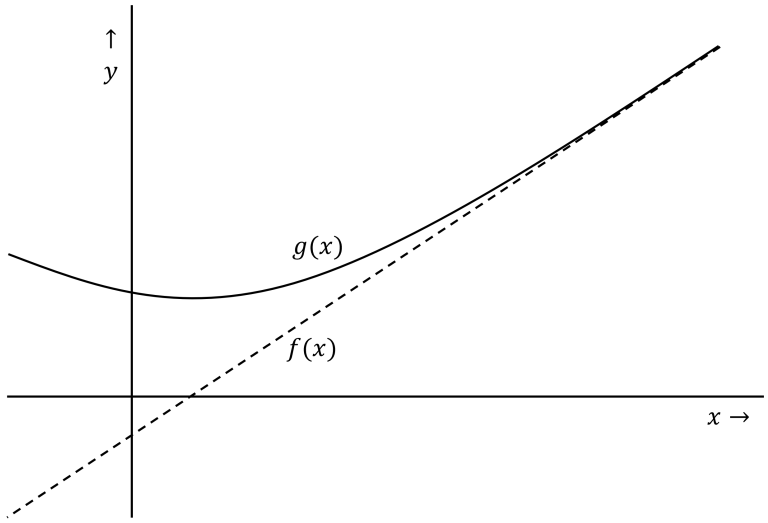

Figure 1.1: The asymptotic behavior of $g(x)$ as x grows large.

We say that $g(x)$ is asymptotically equal to $f(x)$ as x grows large.

This thesis describes a selection of problems and their solutions, where for each problem an asymptotic approach is used to obtain theoretical results on a certain stochastic process. Section 1.1 serves as a general introduction to stochastic processes, and describes two stochastic process "building blocks" that form the basis of the processes considered in this thesis. Then, an extensive introduction to asymptotics is given in Section 1.2, in which asymptotic approaches are divided into two types. Three unique combinations of asymptotic approach type and stochastic process building block separate the thesis into Parts I, II and III. A summary of each part is presented in Section 1.3, and Section 1.4 discusses some common notation.

1.1 Stochastic processes

Stochastic processes range from fairly simple to very involved. They can be subdivided into four classes, corresponding to the time component and value component being either discrete or continuous. We now describe the most fundamental examples of each of the classifications.

Named after the Russian mathematician Andrey Markov sr. in the early 20-th century, the discrete-time Markov chain [81] is the classic discrete-time discrete-valued stochastic process. In this model, time and value are indexed by the natural numbers (although any countable set suffices), and the value (or state) of the process at a given time step is assumed to be dependent only on the value at the previous time step. One can translate this model to continuous time by moving from unit time steps, to time steps that are exponentially distributed with rate depending on the current value. The resulting stochastic process is called a continuous-time Markov chain. We refer to [94] for extensive theory on both discrete-time and continuous-time Markov chains.

An important continuous-valued process in discrete time is the (continuous-valued) random walk. Such a process starts at some given initial value and at each time step adds an independent sample of the continuous random variable X to the current value of the process. This way, a

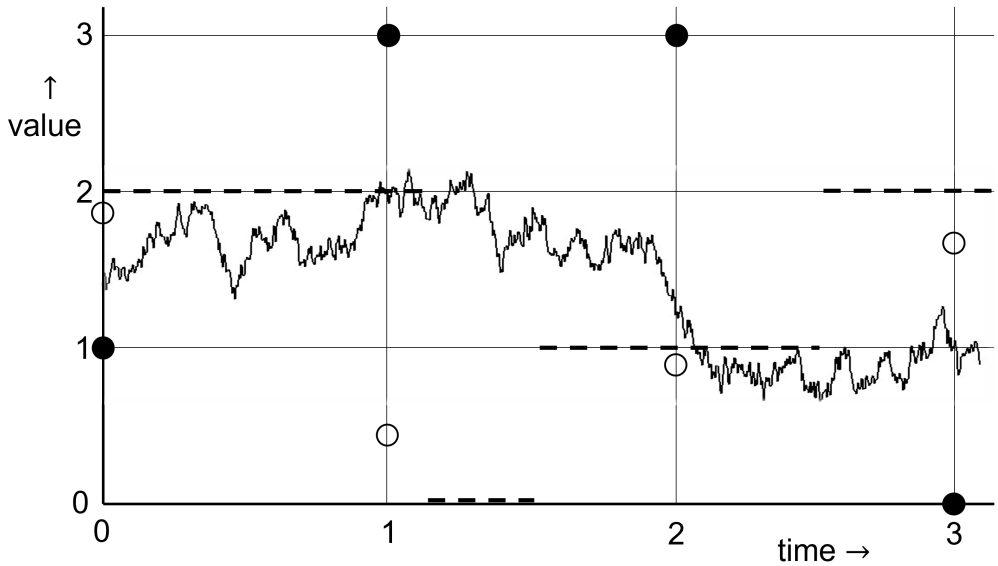

Figure 1.2: Examples of a discrete-time Markov chain (filled circles), continuous-time Markov chain (dashed line), random walk (empty circles) and Brownian motion (solid line).

discrete-time process with independent and identically distributed increments is formed. As an example of the final classification (continuous in both time and value) we have the Brownian motion (history and important results in [91]), which can be seen as a continuous-time analogue of a random walk. This process with continuous erratic movements up and down forms the basis of stock value models. The increments of a Brownian motion are still independent and identically distributed (normally distributed in fact), but the process is defined at any point in time, rather than only at time steps.

Figure 1.2 shows a sample path of all four stochastic processes described above. Note that discrete-time processes only have values on the vertical lines, and that discrete-valued processes only have values on the horizontal lines.

In the remainder of this section, we introduce two stochastic system "building blocks", *queues* and *Lévy processes*, in which stochastic processes are studied (in fact, the Lévy process is itself a stochastic process). We also point out how the more advanced models considered in this thesis are based on these building blocks.

1.1.1 First building block: queues

Being a stochastic research area with a very large variety of applications, queueing theory (an accessible introduction can be found in [57]) comprises the study of congestion caused by multiple entities requiring the same service. The words "entities" and "service" here are deliberately unspecific. Examples of their respective interpretations include customers and checking out at the cash register, cars and crossing an intersection, unfinished products and processing, data packets and downloading, and so on. Without thinking about it, we encounter dozens of queues every day.

With so many applications the need to study queues is clear. Whether we would like to obtain

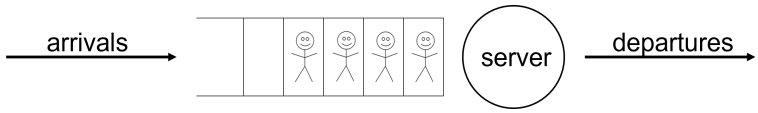

Figure 1.3: Graphical representation of a queue with one server.

information on waiting times or queue length, maximize the customer throughput, or reduce the cost caused by delay, mathematical results on a broad range of queueing models are valuable.

Queueing theory has its origin in a call center. In order to determine how many telephone operators were needed to handle incoming calls, the Danish mathematician Agner Erlang constructed a probabilistic framework to model call center systems [44]. Many mathematical objects are named after Erlang, including a class of probability distributions, various queueing models and a programming language.

The most standard model of a queue, however, is not Erlang's call center model, but the so-called $M/M/1$ FCFS queue. This system is defined by an arrival process, the service times and the service condition as follows:

- the times between consecutive arrivals are independent and exponentially distributed (first M),
- the service times of customers are exponentially distributed (second M), independent of each other and of the arrival process, and
- there is one server (1) that always serves customers in first-come, first-served (FCFS) order.

See Figure 1.3 for a graphical representation of the $M/M/1$ FCFS queue. Due to the fact that times between events (arrivals and service completions) are exponentially distributed, the number of customers in this queue can be modeled by a continuous-time Markov chain. Therefore, mathematical results on Markov chains can be used to reveal valuable information about the queue.

Much is known about the $M/M/1$ FCFS queue since the theory on continuous-time Markov chains is well developed. However, extensions of this basic model in various different directions can be considered. One could think for instance about a generalization of the times between arrivals and/or service times, a variable number of servers, or impatient customers.

Two other types of extensions are important in this thesis. In Part I, we consider different prioritization rules (scheduling policies) for the order in which to serve the available customers. Examples are policies where the priority goes to the customer that arrived last rather than first (last-come, first-served), or where the capacity of the server is shared equally among all customers (processor-sharing). Evaluating the performance of different scheduling policies is crucial, since the optimal policy strongly depends on the model and on the performance metric (average waiting time, making deadlines, etc.). We focus in Part I on an asymptotic performance metric, in that we study the probability that the waiting time exceeds t in the limit where t grows large.

Rather than considering a queue in isolation, we study in Part II a network of queues. The key feature of such models is that customers do not necessarily leave the system when their service is completed, but instead can be assigned a (possibly random) next queue to join [71]. Queueing networks frequently arise in computer systems and manufacturing, which motivated much research in this area. Particularly noteworthy is the pioneering work of R. Jackson [63] and J. Jackson [62], in which it was established that the lengths of the different queues behave essentially independently when the network has external arrivals and departures. If this condition does not hold, the dependence between the different queues of the network poses a challenge for mathematical analysis, especially if the network consists of many queues. In turns out in Part II that elegant results can be obtained by using an asymptotic alteration of the model in which the throughput of all queues grows large.

The core of Part III's topic lies (seemingly) outside of queueing theory. In the next subsection, we introduce a second stochastic system building block for that part of the thesis, and we briefly comment on its relation with queueing.

1.1.2 Second building block: Lévy processes

The Lévy process is among the most fundamental stochastic processes, named after the French mathematician Paul Lévy. Essentially, Lévy processes comprise all continuous-time stochastic processes that develop independently of time and of their past [77]. Because of this property, the Lévy process can be seen as the continuous-time counterpart of the random walk. See Figure 1.4 for an example.

Remark 1.1.1. Formally, a continuous-time and continuous-valued process is a Lévy process if

(i) the behavior of the process after any time t is independent of the values of the process before time t,

(ii) the increments of the process during equally sized time intervals have the same probability distribution, and

(iii) the process is continuous from the right and limits from the left exist.

This final condition is quite technical, and serves, informally stated, to exclude the possibility of infinitely many jumps of non-negligible size in any finite time interval. For a detailed account of Lévy processes we refer to [73]. ◇

A Lévy process can be seen as a generalization of the Brownian motion that we described at the start of Section 1.1. Because of its versatility, the Lévy process is frequently seen in, e.g., finance models, for instance as a representation of some stock value over time. In order to make profit, it is of key importance to learn about the underlying stochastic process. The ability to model stock values is one of the reasons why the study of the Lévy process is popular in mathematics.

Applications are not limited to finance, though. For instance, risk processes in insurance models are often represented by a Lévy process. Additionally, there exist intriguing relations

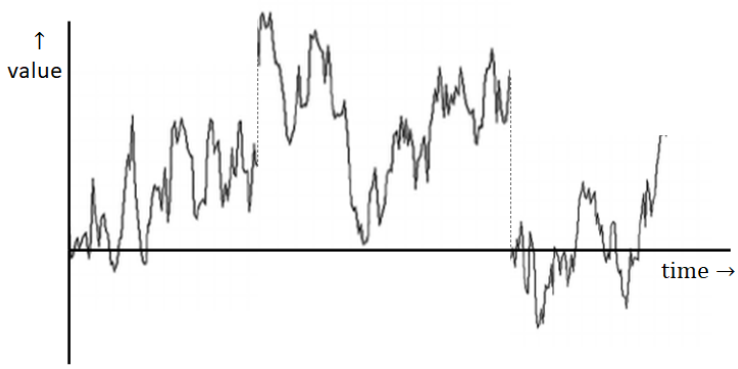

Figure 1.4: Sample path of a Lévy process.

between Lévy processes and queues, such as the maximum of a Lévy process having the same probability distribution as the workload of an associated queue.

In more advanced practical settings however, conditions (i) and (ii) are often not met. This is typically caused by the frequently occuring phenomenon that the dynamics of the Lévy process are influenced by external factors. As a consequence, Markov additive processes (MAPs) have received considerable attention [7, Ch. XI]. In a MAP, the external factors are modeled by a continuous-time Markov chain, and it behaves as a Lévy process with parameters depending on the current state of the Markov chain. In Part III of this thesis, we investigate the probability that the MAP exceeds a given threshold value. Among other things, we perform an asymptotic study of how this probability decreases when the threshold value grows large.

1.2 Asymptotics

The focus of this thesis lies on asymptotics of stochastic processes [126]. In this section, we introduce the term asymptotics through an example, explore advantages of asymptotic approaches and make a distinction between two types of asymptotics.

1.2.1 Motivating example

Suppose that we are interested in the value of a random walk at time step 5. Note that this value is equivalent to the sum of five independent samples of the same random variable. Computing the exact distribution of this sum is a tedious task because each possible outcome can be generated by many combinations of the individual random variables. With 50 random variables (time step 50 for a random walk), the possibilities are nearly countless and in most cases impossible to calculate even for a computer.

The better option is to make use of a famous mathematical result (to which Lévy made significant contributions): the central limit theorem [46]. It states that a sum of independent and identically distributed random variables behaves, after appropriate normalization, as a random variable with a normal distribution as the number of variables in the sum approaches infinity. This might seem a bit hypothetical, since in practice who ever considers an infinite

sum of random variables? The great property of the central limit theorem is that it gives remarkably accurate approximations for relatively small sums.

To demonstrate this property, consider an experiment where we toss a fair coin 50 times and count the number of heads. This number can be theoretically described as a sum of 50 independent random variables, each being either 0 or 1 with equal probability. When we use the central limit theorem for calculating, say, the probability of at least 30 heads, we get the approximate probability of 0.1015. For comparison, the actual probability (which can still be computed for this simple example) of 0.1013 is very close.

The central limit theorem is an example of an *asymptotic* result. Asymptotics are mathematical characterizations of limiting behavior, in all imaginable forms. One could think of one or more parameters approaching some boundary value. Or of analyzing how unlikely it becomes for a certain random variable to take larger and larger values. What kind of limit is taken varies, depending on the model and in what kind of results one is interested.

1.2.2 Advantages of asymptotic approaches

We now describe a number or reasons why asymptotics are highly relevant. First, the number of states of a stochastic system generally grows rapidly in its size, thus eliminating the possibility of exact numerical computations. If one can derive asymptotic results, these can be suitable as an approximation of the solution to the problem in practice. We already described this concept for sums of random variables: in case the number of variables in the sum is too large for exact computations, the central limit theorem can be used as an approximation.

Secondly, in order to avoid oversimplification of real-world problems, more complicated models generally involving many inter-dependencies must be considered. The dynamics of such models are hard to quantify in general, but they often exhibit surprisingly easy-to-interpret limiting behavior under the right asymptotic perspective. These types of asymptotic results can lead to a better understanding of models that seem extremely complicated. As an example, consider a production facility with a collection of machines is that subject to breakdown and one repairer, such that a queue is formed if multiple machines are broken at the same time. Moreover, items continuously arrive at the production facility requiring processing by the (working) machines. We thus have a layered queueing network [41], in which the machines are simultaneously customers and servers. Even though the additional layer significantly complicates exact analysis, the model shows interesting behavior when increasing the breakdown and repair rate. As these rates become arbitrarily large, the breakdown-repair layer vanishes (thus leaving a much simpler model) and is replaced by a reduced processing rate of the machines. The above-described asymptotic method, where events in one part of the model (breakdowns/repairs) follow up on each other much quicker than in the other part (items arriving/departing), is called separation of time scales.

Thirdly, asymptotics can help in situations where real-world data is insufficient. For instance, the long-term effect of an exotic plant on an ecosystem is hard to predict because it takes years before a new balance has settled. Another example is the difficulty in studying the frequency

of large-scale power blackouts, due to their rare occurrence. In these cases, data would be either of little help or too rare to obtain accurate predictions. With proper mathematical modeling however, concrete theoretical results can be realized by asymptotic analysis.

Finally, many key performance measures of stochastic systems describe some kind of limiting behavior, and are thus inherently asymptotic. In Markov chains the concept of stationarity plays an important role, which is essentially the state of the process as time approaches infinity. As a result, the effect of the initial state has been worn out. Another example is the so-called asymptotic drift of a Lévy process, which is the average increase (or decrease) of the process per time unit.

1.2.3 Types of asymptotics

Each of the stochastic systems in this thesis comes with its own challenge. In the scheduling setting of Part I, how do we avoid very large waiting times? In the queueing network setting of Part II, how do we keep track of the rapidly growing number of states? And in the MAPs of Part III, how do we estimate extremely unlikely events? All of these are questions in which asymptotics play a large role. In this subsection, we describe a number of asymptotic approaches sub-categorized in two types of asymptotics, and highlight their use in the stochastic systems we consider.

First type of asymptotics: threshold asymptotics

For every well-behaving random variable X, the probability that it exceeds some threshold value x approaches zero when x grows large. However, the speed at which this probability approaches zero is highly variable, and is of great importance when we are interested in events where X is much larger than its mean (average). Threshold asymptotics concern the precise way in which the tail probability $\mathbb{P}(X > x)$ (read: the probability that X is greater than x) converges to 0 as x goes to infinity.

Threshold asymptotics are particularly useful in the study of rare events. With rare we do not mean a probability of 10% or 1%, but rather of the order 0.001% or (way) smaller. Even though such probabilities would be negligible in most real-world scenarios, they become worthwhile considering when the corresponding event has large consequences. Think about stock market crashes, natural disasters, power blackouts, or epidemic breakouts.

The main challenge of rare events, as mentioned in Section 1.2.2, is the limited availability of historical data, thus making it difficult to properly estimate the probability. Even when a realistic model can be formulated, numerical estimations have large errors when the target event has low probability. To avoid these errors, probabilists have developed a technique where a new probability measure is defined under which the target event has a much larger probability. This change-of-measure approach [9, Chapter V] allows for exact results, but also enables a more efficient way for numerical estimation of the probability of the rare event. In Part III this technique plays a large role in understanding the probability that a MAP exceeds a given value.

In this thesis we also use an asymptotic technique related to rare events, namely large deviation theory [39]. This approach is focused on the tail probability of a sequence of probability distributions. To see why this can be challenging, consider that we are interested in the probability that a sample average of independent and identically distributed random variables is much larger than expected. The central limit theorem estimate is very accurate for values close to the mean, but as we go further along the tail, the error of this estimate typically increases. It is clear that a different approach is required to identify the tail probability, and it depends on the model which approach is most effective. In the queueing setting of Part I, we study the probability that a customer waits longer than t time units, as t grows large. We eventually characterize this probability in two steps. First, we work with a worst-case approach by arguing that the waiting time can not be larger than the length of a certain busy period. Then, bounds on such busy periods are derived.

Second type of asymptotics: scaling asymptotics

Rather than examining the tail behavior of a single random variable, sometimes it is useful to consider asymptotics of the stochastic system as a whole. This is typically modeled by letting multiple parameters depend on an artificial number n, and subsequently letting n approach infinity. We thus have a sequence of stochastic systems indexed by n, and we are interested in the behavior of the limiting model. The central question here is how to let each parameter depend on n in order to obtain meaningful results. Asymptotics of this kind are called scaling asymptotics, see also [126].

An example of a scaling asymptotic is the heavy-traffic regime in queueing systems [72]. In this limit the arrival rate of customers approaches the service rate of the system, resulting in a longer and longer average queue length. In fact, the average queue length grows arbitrarily large in the limit, so a normalization is applied to obtain non-trivial results.

A second model where scaling can be applied is the Erlang loss queue, often used to represent a call center (as described in Section 1.1.1). Here, calls (customers) arrive at an average rate of λ per time unit, and each has a mean service time of 1 time unit. The number of staff members (servers) is denoted by s, and when an arriving call finds all staff members occupied, the call is lost and removed from the system. Designers of call centers will try to balance the cost of lost calls and staff wages.

To understand the dynamics of large call centers, we scale the arrival rate to λn for a large number n, the question being how many staff members to hire. It turns out the cost is minimized if s equals $\lambda n + c\sqrt{n}$ (rounded up or down) for some $c > 0$, which is generally called the square root staffing rule. Elegant asymptotic results can be obtained when n approaches infinity [21]. A very similar scaling approach, generalized to queueing networks, is proposed in Part II to obtain the limiting distribution for the queue lengths.

	Building block	Type of asymptotic	Topic
Part I	Queues	Threshold	Scheduling in the $M/G/1$ queue
Part II	Queues	Scaling	Queue lengths in closed product-form queueing networks
Part III	Lévy processes	Threshold	Maximum of Markov additive processes

Table 1.1: An overview of the thesis's contents.

1.3 Outline of the thesis

This thesis consists of three parts, distinguished by stochastic system building block and type of asymptotic. For models based on queues we consider both threshold (Part I) and scaling (Part II) asymptotics, while for a model based on Lévy processes we consider only asymptotics of threshold type (Part III). A concise schematic overview of each part's contents is shown in Table 1.1. The topic of each part, summarized in the last column of Table 1.1, will be described in more detail below. We notify the reader that the text of this section is more in-depth than that of the first part of the introduction.

1.3.1 Part I

In the first part of the thesis, we consider the so-called $M/G/1$ queue. This model differs from the $M/M/1$ in that the service times are not necessarily exponentially distributed, but may have any distribution (G stands for General). An important (random) quantity of interest is the response time: the time that a customer spends in the system. Finding the best order in which to serve customers, as to minimize the response time, is a central problem in the $M/G/1$ queue.

A straightforward objective would be to minimize the mean response time, since this also minimizes the average total delay per time unit. If information on exact service times is available, it is known that this objective is reached by always serving the customer with the shortest remaining service time. In case only the distribution of the service times is available, the unique scheduling policy minimizing mean response time is the *Gittins policy* [53]. This policy gives each customer a rank based on the distribution of the remaining service time. A formal definition of the Gittins policy is given in Part I, but intuitively it prioritizes customers that have a large probability of a short remaining service time.

Besides the minimization of the mean response time, some applications require the avoidance of very large response times. To this end, we investigate in Part I the rate at which the response time tail probability decays to zero.

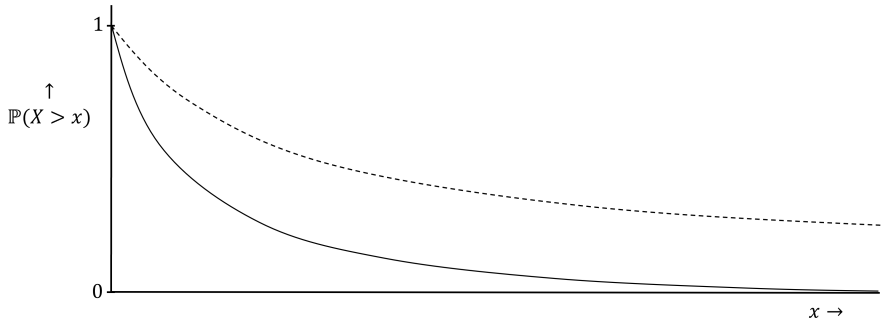

Figure 1.5: Tail probability of the random variable X in case $\mathbb{P}(X > x) = (1 + x/3)^{-2}$ (dashed line, heavy tail) or $\mathbb{P}(X > x) = e^{-x}$ (solid line, light tail).

In this setting, it is known that the performance of different scheduling policies depends on the rate at which the service time tail decays to zero. The terms *heavy-tailed* and *light-tailed* quantify this rate. A random variable is called heavy-tailed if its tail decays at a slow, say polynomial, rate, and light-tailed if its tail decays at a fast, say exponential, rate (precise formulations depend on the context; ours will be formally defined in Part I). The difference of the two categories is visualized in Figure 1.5: the polynomial tail decays to zero slower than the exponential tail.

In order to measure the performance of scheduling policies on the response time tail, a distinction is made between heavy-tailed and light-tailed service times because the queue is significantly more sensitive to congestion in the former case. Under heavy-tailed service times, we say that a policy is optimal when the response time tail is of the same polynomial order as the service time tail. For the light-tailed case, the optimality criterion entails that the exponential decay rate of the response time tail must be maximized. For a mathematical formulation of these optimality criteria see Sections 2.2 and 3.3.

Relevant literature

Initially, the focus of scheduling literature has been on light-tailed service times [120]. It became clear from several applications, however, that heavy tails [88] occur often in practice and therefore should not be neglected. In both scenarios, analysis of the response time tail has been carried out for most well-known policies, see the surveys [20, 29]. For example, it is known that under light-tailed service times, first-come first-served is an optimal policy, whereas in the heavy-tailed setting, processor-sharing (which divides the capacity of the server equally among all present customers) is optimal. At the same time, these two policies perform poorly in the opposite scenario. It was shown in fact, that when exact service times are not available, there exists no policy that is simultaneously optimal under heavy-tailed and light-tailed service times (unless one makes certain unconventional assumptions) [127].

This dichotomy has led to the asymptotic tail performance study of many individual policies. However, the literature on asymptotic tail performance has been missing out on the Gittins policy. In addition, very few works have considered large classes of policies. Among those, Núñez-Queija [95] formulated analytic conditions under which a policy is optimal in the heavy-tailed setting. In Chapter 2, a pivotal role is played by these conditions. A second study

considering a class of policies is the work of Nuyens et al. [97], in which it is assumed that the exact service times are known to the scheduler.

Our work in Part I forms a significant addition to the scarce literature described above. In the heavy-tailed setting, we consider a large class of policies without assuming that exact service times are known, while at the same time deriving optimality conditions that are easily verified. Particular attention is given to the Gittins policy, for which we characterize when it is optimal in the case of light-tailed or heavy-tailed service times.

Contributions

In Chapter 2 (based on [115]) we consider the recently introduced *SOAP* framework [113]: an analyzing tool for a class of policies in which customers are prioritized based on their *age*: the time they received service so far. Specifically, a SOAP policy prioritizes according to the value of a function of age. We use the framework to derive sufficient conditions on this function, that ensure the policy is optimal in the case of heavy-tailed service times. These conditions are general and lead to new results for important policies that have previously resisted analysis.

We focus on the Gittins policy in Chapter 3 (based on [116]). Characterizing Gittins's asymptotic tail behavior is important because if the Gittins policy has optimal tail asymptotics, then it simultaneously provides optimal mean response time *and* good tail performance. For heavy-tailed service times, we find that the Gittins policy always has asymptotically optimal tail performance. The situation is less clear-cut under light-tailed service times: Gittins's tail can be optimal, pessimal (i.e., minimal response time tail decay rate), or in between. To remedy this, we show that a modification of the Gittins policy avoids pessimal tail behavior while achieving near-optimal mean response time.

1.3.2 Part II

Motivated by applications in e.g. manufacturing, computer systems, communication networks and vehicle sharing, we consider in Part II a class of queueing networks with three features.

1. There are no external arrivals and departures (i.e. the network is closed), so the network has a fixed number of customers.

2. All queues of the network have either one server or an infinite number of servers. The single-server queue was described in Section 1.1.1. An infinite-server queue arises when there is always a server available, for example if there is self-service. Note that because of this there is no waiting in an infinite-server queue.

3. All single-server queues have at least one of the following characteristics:

   - exponential service times,

   - the processor-sharing scheduling policy, or

   - the preemptive resume last-come, first-served scheduling policy (in which services are interrupted by arriving customers and later resumed).

For a queueing network with these features, we are interested in the proportion of time that the customers are distributed over the queues in any way. We refer to the corresponding probability distribution as the joint stationary queue-length distribution. A powerful property of the type of network we consider is that it has a product form. This entails that the joint stationary queue-length distribution factorizes into one term for each queue (and a normalization constant).

From a probabilistic perspective, if a joint distribution allows such a factorization, the random variables involved are typically independent. There is, however, one additional condition that creates dependence: the fact that the sum of all queue lengths must equal the fixed total number of customers. This condition, which is specific to closed networks, significantly complicates exact analysis. Moreover, the potentially large number of ways in which the customers can be distributed over the queues forms a computational problem. To illustrate the complexity by an example: there are 2.5 billion ways in which customers can be distributed over a network of 50 customers and 10 queues. Many applications involve networks larger than this, and thus one cannot resort to exact computations in these cases.

Relevant literature

The first general classes of queueing networks for which product form was established are the class of Jackson networks [62], BCMP-networks [14] and Kelly networks [71]. A large account of work on (product-form) networks since then is presented in [25]. This book includes more general conditions under which product form holds as well as theoretical results for a variety of queueing network models.

One way to overcome the analytical and computational complications described above is to consider scaling asymptotics. Birman and Kogan [19] use integral representations for the generating function of the stationary distribution, leading to accurate approximations of the unknown constant in the stationary distribution. A popular scaling approach restricted to a single queue is the one of Halfin and Whitt [56], resulting in a relatively simple limiting distribution. This scaling approach forms the basis of important work on large call centers [21]. In Part II of the thesis, we obtain results by applying a similar scaling for queueing networks rather than for a single queue.

Contributions

In Chapter 4 (based on [123]), we analyze the queueing network with the three mentioned features in a Halfin–Whitt inspired scaling regime, where we jointly amplify the throughput of all queues and the number of customers in the network. In the limit, this leads to a simple stationary distribution for the scaled lengths of the queues. The simplicity of the limit result with respect to the original system provides intuition on the impact of the dependence between the queues on the network's behavior. We show how the scaling parameters can be chosen to obtain accurate approximations. In addition, we assess the practical applicability of our results through a series of numerical experiments, which illustrate the difference between pre-limit and limit behavior.

1.3.3 Part III

Lévy processes and the more general MAPs are the main tools to model capital in credit risk applications. Here, risk refers to the possibility of the capital of a firm dropping below zero at any point in time, given some initial capital reserve. To analyze this possibility, a MAP is defined such that the event of its maximum exceeding a given value directly corresponds to the capital dropping below zero (see Section 6.1 for the specific definition). For this reason, the maximum of a MAP is an important object of study. More specifically, we study the *distribution* of the maximum, since this process is random.

Relevant literature

Because of its general characterization, the Lévy process [16, 73] has been extensively studied. Particularly much attention has been given to risk models [8]. A prominent example is the Cramér-Lundberg model [28] for insurance companies, in which their capital increases at a constant rate through premium paid by customers, but decreases through occasional claims of variable size. The probability that the company ever goes bankrupt (ruin probability) is the most important object of study, which was asymptotically characterized by e.g. [17].

After the first considerations of MAPs [33, 92], researchers have worked on duplicating known results for Lévy processes to this more general model. The distribution of the MAP's maximum, which directly translates to the ruin probability, forms no exception. Although significant work has been done on specific MAPs such as the Markov-modulated compound Poisson process (see for instance [85]) or Brownian motion, much less is known about the maximum of the MAP in its full generality. Notable exceptions are [59] and [11], where results were obtained through respectively considering a first passage process and a martingale method. In Part III, we make contributions in two directions to the study of the MAP's maximum. In the first place, we lift the common assumption that the background Markov chain has a property called irreducibility. This property means that it is always possible to re-enter each state. Secondly, we characterize the tail probability of the maximum.

Contributions

In Chapter 5 (based on [124]) we consider a MAP that is spectrally positive (only upward jumps) or spectrally negative (only downward jumps), and allow for the finite-state background chain to be non-irreducible. In both the spectrally positive and negative cases, we derive a system of linear equations of which the solution characterizes the distribution of the maximum of the process. The general nature of our results allows for applicability in many specific models. Particularly, we develop a procedure for calculating the maximum of a spectrally positive or negative Lévy process over the class of phase-type distributed time intervals. This result for Lévy processes follows from letting the background chain of the MAP model the phase-type distribution.

We concentrate on the asymptotics of the ruin probability as the initial reserve grows large (Cramér-Lundberg asymptotics) in Chapter 6 (based on [125]). We work in the setting that the MAP is spectrally-positive and light-tailed, indicating that all jumps are positive and have

light-tailed distributions. By applying a change-of-measure argument, we express the exact asymptotics of the ruin probability in terms of the input parameters. The same argument also allows us to construct an algorithm that estimates the ruin probability through the alternative probability measure. Unlike direct simulation algorithms, for which the running times are inversely proportional to the (very small) ruin probability, the running time of our algorithm is independent of this probability. Numerical results show that our algorithm indeed leads to significantly lower running times, especially for small ruin probabilities.

1.4 Notation

Throughout the manuscript, we often need to asymptotically compare two functions as their argument approaches infinity (or zero). Suppose $h(x) := f(x)/g(x)$. Assuming $x \to \infty$ (or $x \downarrow 0$), we denote

- $f(x) = o(g(x))$ if $h(x) \to 0$,
- $f(x) = O(g(x))$ if there exists $c \geq 0$ such that $h(x) \to c$,
- $f(x) = \Theta(g(x))$ if there exists $c > 0$ such that $h(x) \to c$,
- $f(x) = \Omega(g(x))$ if there exists $c \geq 0$ such that $1/h(x) \to c$,
- $f(x) = \omega(g(x))$ if $1/h(x) \to 0$.

Part I

The asymptotic response time tail in the $M/G/1$ queue

Introduction

The scheduling of jobs in single-server queues has been an important topic of study over the past decades. On one hand, much attention has been devoted to identifying scheduling policies that minimize the mean response time (commonly referred to as sojourn time) in a variety of settings. For example, in preemptive settings it is widely known that Shortest Remaining Processing Time (SRPT) minimizes the mean response time [106] regardless of the job size (service time) distribution when job sizes are known to the scheduler. When sizes are unknown to the scheduler but the job size distribution is known, the optimal scheduling policy is the Gittins policy [2], which, intuitively put, prioritizes the job that is most likely to finish soon (formal definition in Section 2.3.2 or Definition 3.1.1). If the job size distribution is also unknown, then the Randomized Multi-Level Feedback (RMLF) policy minimizes the competitive ratio for mean response time (largest possible ratio between RMLF and SRPT under any arrival and job size sequence) [69].

On the other hand, in many applications it is more important to avoid large response times rather than just minimizing the mean response time. Thus, a significant amount of research has been devoted to analyzing the distribution of response times under a large variety of scheduling policies, ranging from classical policies such as First-Come First-Served (FCFS) and SRPT, to newer ones such as Processor Sharing (PS) and its many generalizations [1, 6, 96]. In some simple settings it is possible to precisely characterize the response time distribution, but in general research focuses on characterizing the *tail* of the response time distribution. This characterization marks the asymptotic rate at which that tail decreases.

The task of characterizing the response time tail is more complex than that of optimizing the mean response time. Initially, response time tail asymptotics were studied in the case of light-tailed job size distributions, e.g., [29, 96, 120] and the references therein. Light-tailed job sizes are essentially those with exponential decay rate (for our definition of light-tailed see Section 3.3), and scheduling under these job size distributions aims to minimize the (exponential) decay rate of the response time tail. In this context, it has been shown that FCFS maintains the optimal (lightest) response time tail [120], whereas under SRPT it is the heaviest possible. This is a stark contrast to the optimality of SRPT for the mean response time.

While the focus of response time tail asymptotics was initially on light-tailed settings, a shift occurred in the late 1990s when it was observed that heavy-tailed distributions occur

frequently in computer and communications systems, e.g., in file sizes in the web [36], in I/O patterns [100], the length of network sessions [98], and more. These observations triggered much research into the impact of heavy-tailed phenomena on the design and performance of computer and communications systems. The resulting literature has demonstrated that heavy-tailed traffic characteristics have a dramatic effect on the waiting times and response times experienced by users and that scheduling and priority mechanisms need to be designed with heavy-tailed phenomena in mind.

A key observation from the research that followed is that scheduling policies that perform well under light-tailed settings may not perform well under heavy-tailed settings, and vice versa. A prime example is FCFS, which has the optimal response time tail under light-tailed job sizes, but has a response time tail as bad as possible under heavy-tailed job sizes. More precisely, assume that the job size X is regularly varying with index $-\alpha$ (that is, $\mathbb{P}(X > x) = L(x)x^{-\alpha}$, where the function $L(\cdot)$ is slowly varying, i.e., $L(ax)/L(x) \to 1$ for any $a > 0$) and denote this with $\mathrm{RV}(-\alpha)$. Then, the response time in a $GI/GI/1$ FCFS queue is known to be $\mathrm{RV}(1-\alpha)$ [35]. A worse index is not possible (as long as the server is always working when there is at least one job), since the residual busy period of such a queue is $\mathrm{RV}(1-\alpha)$. The response time in a $GI/GI/1$ SRPT queue, on the other hand, has the same index as the job size ($-\alpha$) in this setting. Since the response time of a job can never be smaller than its size, the response time index $-\alpha$ is optimal. So, SRPT is optimal in this heavy-tailed setting, whereas FCFS performs the worst in terms of response time tail index – the exact opposite of the light-tailed scenario.

Observations like the one described above have led to much research on the impact of the service discipline on delay asymptotics; cf. the surveys [20, 29]. In these works, the tail decay of several well-known scheduling policies is analyzed, under both light-tailed and heavy-tailed job sizes. The aim of Part I is to analyze the tail performance not only for individual policies, but rather for the broad class of SOAP [113] policies (to be introduced in the next section). A key concept is the notion of *tail optimality*, which is satisfied by a scheduling policy if it guarantees an asymptotically optimal response time tail (precise criteria for tail optimality to be specified in Definitions 2.2.2 and 3.3.3.

Model and contributions

We consider in Part I the $M/G/1$ queue: a single-server queue with a Poisson arrival process and general job size distribution. The specific range in which the (heavy-tailed and light-tailed) job size distribution falls will be specified in Definitions 2.2.1 and 3.3.2.

The scheduling policies we study in this part are *SOAP* (Schedule Ordered by Age-based Priority) policies, a broad class of policies introduced by Scully et al. [113]. At each point in time, a SOAP policy assigns to each job in the system a *rank* (priority). The rank of a job is based on a single variable called *age*: the amount of time the job has been served so far. Particularly, a SOAP policy is defined as follows:

Definition I.1. A *SOAP policy* is a policy π specified by a piecewise monotonic and piecewise differentiable *rank function* $r : \mathbb{R}_+ \to \mathbb{R}$, mapping a job's age to its rank. At every moment in

time, a SOAP policy *serves the job of minimum rank*, breaking ties in FCFS order.

Two observations are in place. Firstly, note that a rank function uniquely specifies a SOAP policy, but not vice versa. For example, any linear operation on a rank function yields the same policy. Secondly, since the ranks of the jobs may change, SOAP policies are typically preemptive. Throughout Part I we assume a preemptive-resume model with no preemption overhead. That is, preemption does not cost any time and preempted jobs do not lose their service progress.

We give two examples of rank functions corresponding to known service policies. First, consider the rank function $r(a) = a$, in which the rank of each job is equal to its age. Since we serve the job of minimum rank, the SOAP policy corresponding to this rank function serves the job that has received the least service. This is exactly the prioritization rule of the Foreground-Background policy. Secondly, with X denoting the job size, consider the SOAP policy with rank function $r(a) = \mathbb{E}(X - a | X > a)$. Note that the right hand side is the expected remaining processing time of a job with age a. Therefore, this rank function corresponds to the Shortest Expected Remaining Processing Time first policy.

The main benefit of the SOAP framework is the possibility to obtain results on scheduling policies by analyzing their rank function. Most notably, in their original paper [113], Scully et al. use two intuitive insights to identify how a tagged job is delayed by other jobs, given the rank function. Firstly, they simplify the analysis by arguing that the smallest non-increasing upper bound on the tagged job's rank function does not affect its response time. Secondly, it is shown that delay due to jobs already in the system upon arrival can be translated to the waiting time in a queueing system with server vacations. With these two insights, Scully et al. succeed in identifying the Laplace-Stieltjes transform (LST) of the response time under arbitrary SOAP policies.

Unfortunately, the expression for the above LST cannot be applied to characterize the asymptotic response time tail. The reason for this is that the accuracy of numerically inverted LSTs decreases further along the tail. However, instead of using transforms, we can still use the concept of analyzing rank functions to find asymptotic results for classes of SOAP policies. We do so in Chapters 2 and 3, where the asymptotic response time tail of SOAP policies is characterized based on the properties of the corresponding rank function.

More specifically, our contributions are as follows. In Chapter 2, we establish an easily verifiable condition for a SOAP policy to be tail optimal in case of heavy-tailed job sizes. We show that this condition is met by a few important scheduling policies for which tail optimality has not been shown before. In Chapter 3, we widen the established condition for heavy-tailed job sizes, and derive a second tail-optimality condition for light-tailed job sizes. This allows us in particular to characterize the tail behavior of the acclaimed Gittins policy.

2 Heavy-tailed job sizes

2.1 Introduction

Given the prominence of heavy-tailed phenomena in computer and communications systems, scheduling in the $M/G/1$ queue has often been directed towards heavy-tailed job sizes. A driving question in this context is to characterize which policies have the optimal response time tail asymptotics, i.e., which policies have a response time tail that is of the same order as the tail of the job size distribution under regularly varying job sizes. This notion of "tail equivalence" (which we refer to as tail optimality) has driven research for decades and there is a variety of common policies that have been shown to be tail-optimal, including Processor Sharing (PS) [130], Foreground-Background (FB) [95], and Preemptive Shortest Job First (PSJF) [95].

However, despite significant progress, there are still many important policies for which we do not know if they are tail-optimal or not. Examples are the Gittins policy and Randomized Multi-Level Feedback (RMLF). Further, no precise characterization of which properties a scheduling policy must have in order to be tail-optimal is known.

The first attempt to obtain a general set of conditions that ensure tail optimality was by Núñez-Queija [95], who provided analytic conditions that can be used to simplify the analysis of scheduling policies when studying the response time tail. It was these analytic conditions that enabled the first analysis of policies such as SRPT, PSJF, and FB. However, the conditions are not directly defined in terms of the prioritization rules of a policy, and so they do not provide insight into which policies are tail-optimal. For that, the most general result to this point is by Nuyens et al. [97], who introduce a set of properties based on job sizes that are sufficient conditions for tail optimality. These properties ensure that the scheduler always prioritizes jobs with small sizes and are satisfied by both SRPT and PSJF, but not by policies that do not make use of job sizes, such as Gittins, RMLF, FB, etc. Thus, there is a considerable gap between the sufficient conditions outlined by [97] and a general characterization of tail-optimal scheduling policies.

Contributions

In this chapter, we provide sufficient conditions that ensure optimality of the tail of the response time distribution for scheduling policies in M/G/1 queues with job size distributions that are intermediately regularly varying. Our results provide guidelines on how scheduling policies can perform prioritization in order to ensure tail optimality without having access to job sizes, and are thus complementary to the conditions in [97], which focus on prioritization based on job size. The conditions are general and are satisfied by important policies such as

the Shortest Expected Remaining Processing Time first (SERPT) and RMLF policies, for which no previous analysis of the response time tail is known. Additionally, the sufficient conditions are satisfied by policies that use limited preemption, for the first time highlighting the preemption frequency needed to achieve tail optimality.

The key building block underlying the sufficient conditions we develop is the SOAP framework, recently introduced in [113]. Using this framework, our sufficient conditions for tail optimality are defined in terms of the rank function of a policy. The formal conditions can be found in Section 2.3, but intuitively the conditions ensure that old jobs do not receive priority over other jobs for too long. Specifically, for a job J, part one of the condition bounds the consecutive time that other jobs outrank J, and part two bounds the first age at which jobs will never in the future outrank J.

In general, there are three typical approaches for proving tail optimality, see [20] for a survey. The first relies on a relationship between the tail behavior of a random variable Y and the behavior of its Laplace-Stieltjes transform (LST) $\mathbb{E}(e^{-sY})$ for $s \downarrow 0$ ([18], p. 333). An expression for the response time LST of the single-server queue under SOAP is indeed available (see [113]); however, it depends in such an intricate way on the rank function that this approach proved unsuitable for determining the tail behavior of the response time. Another common approach is to perform a sample path analysis of the policies, as was pursued by [97]. However, again, the form of dependence in the rank function makes this difficult. Hence, in proving our main result, Theorem 2.3.2, we have adapted the probabilistic method developed by Núñez-Queija [95], which exploits a Markov-type inequality. While the method of [95] does not apply directly off-the-shelf, we are able to extend it to apply to the analysis of our sufficient conditions. This extension requires technical effort and, in particular, relies on a new analysis of the fractional moments of busy periods that is of independent interest (see Theorem 2.5.1).

To conclude, we summarize the contributions of the chapter below:

- We provide a set of sufficient conditions for tail optimality of the response time tail, when job sizes are intermediately regularly varying, for policies that do not have access to job size information. These conditions highlight that tail optimality depends on imposing a bound on the amount of consecutive time that a job has priority over others.

- Our sufficient conditions provide a proof of tail optimality for a number of well-known scheduling policies, including the Gittins policy, RMLF, and the Shortest Expected Remaining Processing Time first policy. Tail optimality of these policies is a long-standing open question given their optimality among policies that do not use precise job size information.

- Our sufficient conditions provide the first insight into how much preemption is needed in order to maintain tail optimality. We specifically state which preemption frequencies guarantee tail optimality.

- Our proof of sufficiency includes an interesting foundational result for $M/G/1$ queues: a bound on the fractional moments of the $M/G/1$ busy period. Previously, only expressions for its integer moments were known.

2.2 Preliminaries

We consider an $M/G/1$ queue with arrival rate λ and job size X. We write $F(x) = \mathbb{P}(X \leqslant x)$ and $\overline{F}(x) = 1 - F(x)$ for the distribution function and tail of X, respectively. The system load is denoted by $\rho := \lambda\mathbb{E}(X) < 1$. We write T for response time and T_x for size-conditional response time, that is, the response time for jobs of size x. Our focus is on the case where X has a heavy tail. Roughly speaking, the heavy-tailed job size distributions we study are those which are asymptotically Pareto. The specific class we study, described below, is slightly more general in that it also includes distributions whose tails oscillate between Pareto tails of different shape parameters. In this context, the maximal and minimal shape parameters are respectively called the upper and lower Matuszewska indices [18, Section 2.1] of the job size distribution function.

Definition 2.2.1. *We say that a job size distribution is* HT *if* $x_{\max} = \infty$ *and both of the following hold:*

(i) The tail $\overline{F}(\cdot)$ *is of intermediate regular variation [34], meaning*

$$\liminf_{\varepsilon \downarrow 0} \liminf_{x \to \infty} \frac{\overline{F}((1+\varepsilon)x)}{\overline{F}(x)} = 1.$$

(ii) There exist $\beta > \alpha > 1$ *such that the upper and lower Matuszewska indices of* $\overline{F}(\cdot)$ *are in* $(-\beta, -\alpha)$.

The class of intermediately regularly varying functions contains the class of regularly varying functions [48], which in turn contains Pareto distributions and power law distributions, among others. Roughly speaking, one can think of Definition 2.2.1 as saying that $\overline{F}(\cdot)$ is bounded between two power law distributions, as the following Potter bound formalizes.

Lemma 2.2.1 ([18, Proposition 2.2.1])**.** *If the job size distribution is HT, then there exist constants* $C, x_0 > 0$ *such that for all* $x_2 \geqslant x_1 \geqslant x_0$,

$$\frac{1}{C}\left(\frac{x_2}{x_1}\right)^{-\beta} \leqslant \frac{\overline{F}(x_2)}{\overline{F}(x_1)} \leqslant C\left(\frac{x_2}{x_1}\right)^{-\alpha}.$$

We mentioned in the introduction that tail optimality holds if the response time tail is of the same order as the job size tail. The following criterion formalizes this.

Definition 2.2.2 (Tail Optimality)**.** *Consider an* $M/G/1$ *queue with an HT job size distribution. We call a scheduling policy* tail-optimal *among preemptive work-conserving policies if*

$$\lim_{x \to \infty} \frac{\mathbb{P}(T/(1-\rho) > x)}{\overline{F}(x)} = 1.$$

Here, the term work-conserving entails that the server is always serving a job as long as there is at least one. Definition 2.2.2 states that tail optimality holds if large jobs have a response time of approximately $1/(1-\rho)$ times their size, which is the best possible asymptotic tail decay in the heavy-tailed case [29, 127].

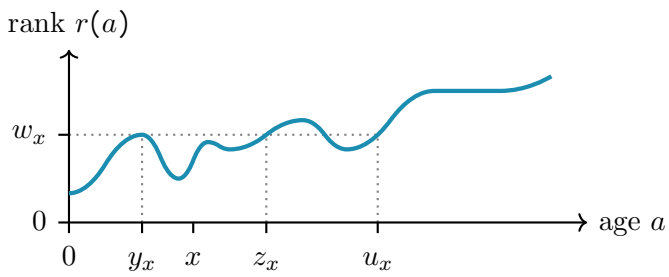

Figure 2.1: Illustration of w_x, y_x, z_x, and u_x.

2.3 Overview of results

Our main result, Theorem 2.3.2, gives sufficient conditions for tail optimality in terms of properties of the rank function. Thus, it characterizes properties of the prioritization of SOAP policies that guarantee optimal tail behavior. We first state a version of our main theorem, proven below Theorem 2.3.2, in which the condition for tail optimality is slightly simplified. It states that a policy is tail-optimal if its rank function is bounded between two power functions in a specific way.

Theorem 2.3.1 (Simplified Result)**.** *Consider an $M/G/1$ queue whose job size distribution is HT using a SOAP scheduling policy whose rank function obeys*

$$r(a) \in \Omega(a^{\gamma}) \cap O(a^{\delta})$$

for some $\delta > \gamma > 0$. If

$$\frac{\delta}{\gamma} - \frac{\gamma}{\delta} < \frac{\alpha - 1}{\beta},$$

then the policy is tail-optimal.

The condition of Theorem 2.3.1 is easy to interpret and is suitable for tail-optimality proofs for many of the policies presented in Section 2.3. However, the result holds under more general conditions. To state these conditions formally, we need some notation. Let

- w_x be the worst rank attained by a job of size x,
- $y_x \leqslant x$ be the earliest age with rank w_x,
- $z_x \geqslant x$ be the earliest age after x with rank $\geqslant w_x$, and
- $u_x \geqslant z_x$ be the latest age with rank $\leqslant w_x$.

Figure 2.1 illustrates these quantities and they are defined formally in Definitions 2.A.1, 2.A.7, and 2.A.9.

Our sufficient conditions on the rank function are defined in terms of two quantities: $z_x - y_x$ and u_x.

Assumption 2.3.1.

(i) There exists $\zeta \in [0, \infty]$ such that $z_x - y_x = O(x^{\zeta})$.

(ii) There exists $\eta \in [\max\{1, \zeta\}, \infty]$ *such that* $u_x = O(x^\eta)$.

Intuitively, ζ and η have the following interpretations. The smaller ζ is, the more quickly the system can preempt jobs. The smaller η is, the more each job is shielded from getting stuck behind larger jobs. Note that any rank function trivially satisfies Assumption 2.3.1 with $\zeta = \eta = \infty$, but, as suggested by the intuitive interpretations, we would like ζ and η to be small. Our main result states just how small ζ and η need to be to ensure tail optimality.

Theorem 2.3.2 (Main Theorem)**.** *Consider an* $M/G/1$ *queue whose job size distribution is HT and a SOAP scheduling policy whose rank function obeys Assumption 2.3.1. If*

$$\zeta - \frac{1}{\eta} < \frac{\alpha - 1}{\beta}, \tag{2.1}$$

then the policy is tail-optimal.

As we prove now, Theorem 2.3.2 immediately implies its simplified version, Theorem 2.3.1.

Proof of Theorem 2.3.1. Precomposing any strictly increasing function with the rank function r yields an equivalent rank function that encodes the same SOAP policy, so we may assume without loss of generality that

$$r(a) \in \Omega(a^{\gamma/\delta}) \cap O(a).$$

This implies $w_x = O(x)$ and thus $u_x = \Omega(x^{\delta/\gamma})$. Therefore, Assumption 2.3.1 holds with $\zeta = \eta = \delta/\gamma$, so tail optimality follows from Theorem 2.3.2. □

The proof of Theorem 2.3.2 makes up the bulk of the remainder of the chapter. However, a key component of our proof that we would like to highlight here is an analysis of the fractional moments of a busy period. The bounds we obtain are potentially of interest beyond the study of the tail of response time. In particular, let B_U be the length of a busy period with initial work U. Thus, a standard busy period would be B_X. We develop the following representation of the nth moment of a busy period for $n \in \mathbb{Z}_+$:

$$\mathbb{E}(B_U^n) = \sum_{i=1}^{I} d_i \frac{\mathbb{E}(U^{b_i})}{(1-\rho)^{a_i}} \prod_{j=1}^{J_i} \lambda \mathbb{E}(X^{c_{ij}}),$$

where $I, J_i, a_i, b_i, c_{ij}, d_i \in \mathbb{Z}_+$ are constants that depend on n (see Lemma 2.5.2 and Corollary 2.5.1). Moreover, we show that this representation extends in a natural way to fractional moments of order $p = n - q > 0$, where $q \in (0, 1)$. Instead of equality, we obtain an upper bound in the case of fractional moments. We defer the full statement, which requires heavy notation, to Theorem 2.5.1.

Applications of Theorems 2.3.1 and 2.3.2

To illustrate the generality of the sufficient conditions in Theorems 2.3.1 and 2.3.2 it is interesting to consider how they can be applied to understand the response time tail of common policies. In this section, we illustrate the application of the theorems to understand tail optimality of policies for which no analysis is known. We focus on four examples: FB with limited preemption, Gittins, SERPT, and RMLF.

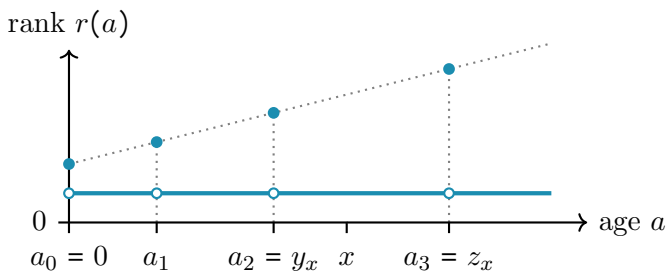

Figure 2.2: Rank function (as in (2.2)) of FB with limited preemption.

2.3.1 FB with limited preemption

Our first example focuses on a policy that is known to be tail-optimal (FB) but limits the amount of preemption it may use. We consider FB here, but the same analysis can be performed for other policies that satisfy the conditions of Theorem 2.3.2. FB is particularly interesting because it is the optimal policy for job size distributions with a decreasing failure rate when no job size information is known. FB works by always serving the job of least age, sharing the processor equally in the case of ties. That is, FB is the SOAP policy with rank function $r(a) = a$. As a result, it preempts jobs frequently and rarely works on a single job without interruption. In situations where there is a cost to preemption this is a signficant drawback. Thus, it is important to understand the performance of FB when preemption is limited.

To this end, we study a variation of FB with limited preemption (FB-LP) where preempting a job is only allowed when its age is one of a limited set of *checkpoints* $A \subseteq \mathbb{R}_+$. Specifically, FB-LP is the SOAP policy with rank function

$$r(a) = \begin{cases} a + 2 & \text{if } a \in A, \\ 1 & \text{otherwise.} \end{cases} \tag{2.2}$$

Figure 2.2 illustrates an example of FB-LP where A is a sequence $a_0 = 0, a_1, a_2, \ldots$.

The design of the FB-LP policy amounts to choosing the set of checkpoints A. In the extreme where $A = \mathbb{R}_+$, FB-LP is the same as using ordinary FB, which is tail-optimal but has frequent preemption and processor sharing. In the other extreme, setting $A = \varnothing$ is the same as using FCFS, which never preempts jobs but has the worst possible response time tail behavior [35]. We therefore ask:

How frequently must checkpoints occur in order to ensure tail optimality?

We can answer this question using Theorem 2.3.2.

Consider a sequence of checkpoints $A = \{0, a_1, a_2, \ldots\}$. When $x \in (a_i, a_{i+1}]$, we have $y_x = a_i$ and $z_x = a_{i+1}$, as shown in Figure 2.2. This means if $a_{i+1} - a_i = O(a_i^{\zeta})$, then Assumption 2.3.1 holds with the same value of ζ and $\eta = \infty$. By Theorem 2.3.2, tail optimality holds if $\zeta < (\alpha - 1)/\beta$, implying the following result.

Corollary 2.3.1. *Consider an $M/G/1$ queue whose job size distribution is HT. The FB-LP policy with checkpoints $a_0 = 0, a_1, a_2, \ldots$ is tail-optimal if*

$$a_{i+1} - a_i = O\big(a_i^{\zeta}\big)$$

for some $\zeta < (\alpha - 1)/\beta$.

2.3.2 The Gittins policy

Our next example application of Theorem 2.3.2 is the Gittins policy, which is the policy that minimizes mean response time of the $M/G/1$ queue when the job size distribution is known but individual job sizes are unknown. Gittins can be viewed as a SOAP policy whose rank function depends on the job size distribution [113, Example 3.6]. Its rank function is given by

$$r(a) = \inf_{b>a} \frac{\int_a^b \overline{F}(t)\,\mathrm{d}t}{\overline{F}(a) - \overline{F}(b)}. \tag{2.3}$$

While Gittins is known to be optimal for the mean response time, the response time tail behavior of Gittins has resisted analysis. In this section we show that, under an HT job size distribution and an additional technical condition (the latter is removed in Chapter 3, see Remark 2.3.1), the Gittins policy is tail-optimal.

Given Theorem 2.3.1, it suffices to bound the Gittins rank function. Because $b = \infty$ is a possibility for the infimum in (2.3), by Lemma 2.2.1,

$$r(a) \leqslant \int_a^\infty \frac{\overline{F}(t)}{\overline{F}(a)}\,\mathrm{d}t \leqslant O(1) \int_a^\infty \left(\frac{t}{a}\right)^{-\alpha} \mathrm{d}t = O(a).$$

By Theorem 2.3.1, Gittins is tail-optimal if its rank function satisfies $r(a) = \Omega\big(a^{\gamma}\big)$, where $\frac{1}{\gamma} - \gamma < \frac{\alpha-1}{\beta}$. However, this is not the case for all HT job size distributions. For example, if the job size distribution has positive mass at some value x, then Gittins has $r(x-) = 0$.

Fortunately, under a mild additional condition, we can prove a lower bound on the Gittins rank function. Suppose that for sufficiently large x, the job size distribution has a well defined density $f(x) = -\frac{\mathrm{d}}{\mathrm{d}x}\overline{F}(x)$ and hazard rate $h(x) = f(x)/\overline{F}(x)$. Then for sufficiently large ages a,

$$\begin{aligned} r(a) &= \inf_{b>a} \frac{\int_a^b \overline{F}(t)\,\mathrm{d}t}{\overline{F}(a) - \overline{F}(b)} = \inf_{b>a} \frac{\int_a^b \overline{F}(t)\,\mathrm{d}t}{\int_a^b h(t)\overline{F}(t)\,\mathrm{d}t} \\ &\geqslant \inf_{b>a} \frac{\int_a^b \overline{F}(t)\,\mathrm{d}t}{\big(\sup_{c\in(a,b)} h(c)\big)\int_a^b \overline{F}(t)\,\mathrm{d}t} = \inf_{b>a} \frac{1}{h(b)}. \end{aligned}$$

This means that if $h(a) = O(a^{-\gamma})$ for $\gamma > 0$, then $r(a) = \Omega(a^{\gamma})$, so Theorem 2.3.1 yields the following result.

Corollary 2.3.2. *Consider an $M/G/1$ queue whose job size distribution is HT. The Gittins policy is tail-optimal if the job size distribution has hazard rate $h(x) = O(x^{-\gamma})$ for some $\gamma > 0$ satisfying*

$$\frac{1}{\gamma} - \gamma < \frac{\alpha - 1}{\beta}.$$

In particular, $h(x) = O(x^{-\min\{1,\beta/(\alpha-1)\}})$ *is sufficient for the Gittins policy to be tail-optimal.*

Remark 2.3.1. The above condition on the hazard rate turns out not to be necessary. Specifically, in Chapter 3 we weaken Assumption 2.3.1 and show that tail-optimality still holds under the new assumption. We also show that the Gittins policy always satisfies the new assumption, and is thus tail-optimal without any additional condition. However, the argumentation in this subsection is still valuable as it will be used to determine the tail performance of Gittins when the job sizes are light-tailed. ◇

2.3.3 Shortest Expected Remaining Processing Time

Shortest Expected Remaining Processing Time (SERPT) is a variation of SRPT for settings when the precise remaining sizes of jobs are not known, but the expected remaining size can be computed given knowledge of the job size distribution. As the name implies, SERPT always serves whichever job has the least expected remaining size. Like Gittins, SERPT is a SOAP policy whose rank function depends on the job size distribution [113, Example 3.5]:

$$r(a) = \mathbb{E}(X - a \mid X > a) = \frac{\int_a^\infty \overline{F}(t)\,\mathrm{d}t}{\overline{F}(a)}.$$

We show that SERPT is always tail-optimal. By Lemma 2.2.1 and with the notation of Section 1.4, the rank function is bounded by

$$\Omega(a) = O(1)\int_a^\infty \left(\frac{t}{a}\right)^{-\beta} \mathrm{d}t \leqslant r(a) \leqslant O(1)\int_a^\infty \left(\frac{t}{a}\right)^{-\alpha} \mathrm{d}t = O(a),$$

so Theorem 2.3.1 implies tail optimality.

The *Monotonic SERPT* (M-SERPT) policy is a variant of SERPT introduced by Scully et al. [114]. Its rank function is the increasing envelope of SERPT's:

$$r(a) = \max_{0 \leqslant b \leqslant a} \mathbb{E}(X - b \mid X > b).$$

As with SERPT, Lemma 2.2.1 implies $r(a) \in \Omega(a) \cap O(a)$ for M-SERPT, so M-SERPT is also tail-optimal.

Corollary 2.3.3. *In an* $M/G/1$ *queue whose job size distribution obeys is HT, SERPT and M-SERPT are both tail-optimal.*

The tail optimality of M-SERPT is particularly significant because M-SERPT has mean response time within a factor of 5 of Gittins's [114]. Thus, for all HT distributions, M-SERPT is within a constant factor of optimal for both the mean response time and the tail of the response time.

2.3.4 Randomized Multi-Level Feedback

The Randomized Multi-Level Feedback (RMLF) policy, to be introduced below, is designed to have low mean response time when neither individual job sizes nor the job size distribution

is known. Originally introduced in the worst-case scheduling literature [15, 69], RMLF was studied in the stochastic $GI/GI/1$ setting by Bansal et al. [13], who showed that RMLF is $O\big(\log \frac{1}{1-\rho}\big)$-competitive with SRPT for mean response time. However, no previous results exist for the tail of the response time under RMLF.

Here, we seek to apply our sufficient condition for tail optimality to RMLF. Unfortunately, RMLF does not fit into the SOAP framework as stated so far because not every job follows the same rank function: each job chooses a random parameter $v \in [0,1]$ and then follows rank function

$$r_v(a) = \min\{2^n \mid n \in \mathbb{N}, 2^{n+v} > a\}.$$

Nevertheless, we still have $a/2 \leqslant r_v(a) \leqslant 2a$ for all ages $a \geqslant 1$ and parameters $v \in [0,1]$, so it seems that some adaptation of Theorem 2.3.1 ought to imply tail optimality of RMLF. This is indeed the case, but stating the adaptation requires some new terminology.

While RMLF is not a SOAP policy, it is what [110] calls a *SOAP Bubble* policy. The SOAP Bubble class of policies is a superset of the SOAP class. Much like SOAP, under a SOAP Bubble policy, every job's rank is a function of its age, and the system always serves the job of minimal rank, but *different jobs can have different rank functions*. Specifically, a SOAP Bubble policy is characterized by *lower and upper rank functions* $r^-, r^+ : \mathbb{R}_+ \to \mathbb{R}$, and the rank function r_j of each job j can be any function obeying $r^-(a) \leqslant r_j(a) \leqslant r^+(a)$. Therefore, RMLF is a SOAP Bubble policy with lower and upper rank functions

$$r^-(a) = \min\{2^n \mid n \in \mathbb{N}, 2^{n+1} > a\},$$
$$r^+(a) = \min\{2^n \mid n \in \mathbb{N}, 2^n > a\}.$$

In Appendix 2.E, we formulate adaptations of Theorems 2.3.1 and 2.3.2 that apply to SOAP Bubble policies. For example, Theorem 2.E.2 is the same as Theorem 2.3.1, except its precondition is $r^-(a) = \Omega(a^\gamma)$ and $r^+(a) = O(a^\delta)$. Applying Theorem 2.E.2 to RMLF with $\gamma = \delta = 1$ implies that RMLF is tail-optimal.

Corollary 2.3.4. *In an $M/G/1$ queue whose job size distribution is HT, RMLF is tail-optimal.*

2.4 Tail optimality of SOAP policies

In the remainder of the chapter we present a proof of our main result Theorem 2.3.2, namely that a SOAP policy is tail-optimal under certain conditions on the rank function. The foundation of the proof is an adaptation of a result by Núñez-Queija [95], which gives sufficient conditions for tail optimality. Hence, proving Theorem 2.3.2 amounts to verifying these conditions when Assumption (2.1) holds. The six steps of which this verification consists are presented in this section. Some of these steps rely on technical lemmas that require more background information. These lemmas are extensively introduced in Section 2.5, Section 2.6 and Appendix 2.A, and proven in Appendices 2.B–2.D.

We now present a slight reformulation of the conditions in Núñez-Queija [95], relating to the conditional response time of a policy.

Condition 2.4.1. *T_x is stochastically increasing in x (that is, $\mathbb{P}(T_x > t) \leqslant \mathbb{P}(T_y > t)$ for all $t \geqslant 0$ and all $0 \leqslant x \leqslant y$).*

Condition 2.4.2. *We have $\lim_{x\to\infty} \mathbb{E}(T_x)/x = 1/(1-\rho)$.*

Condition 2.4.3.

(i) For all $\varepsilon > 0$,

$$\lim_{x\to\infty} \mathbb{P}\left(T_X < \frac{(1-\varepsilon)x}{1-\rho} \;\middle|\; X > x\right) = 0.$$

(ii) For all $\varepsilon > 0$,

$$\lim_{x\to\infty} \frac{1}{\overline{F}(x)} \mathbb{P}\left(T_X > \frac{(1+\varepsilon)x}{1-\rho} \;\middle|\; X \leqslant x\right) = 0.$$

Based on these conditions, Núñez-Queija [95] deduces the following tail-optimality result.

Proposition 2.4.1. *If the job size distribution is HT and Conditions 2.4.1–2.4.3 hold, then*

$$\lim_{x\to\infty} \frac{1}{\overline{F}(x)} \mathbb{P}\left(T > \frac{x}{1-\rho}\right) = 1.$$

Remark 2.4.1. Proposition 2.4.1 differs from Núñez-Queija [95, Theorem 2.3] in that, instead of assuming Condition 2.4.3 directly, Núñez-Queija [95, Lemmas 2.1 and 2.2] proves it starting from a stronger condition. This adapted version is more appropriate for our analysis. ◇

Since Proposition 2.4.1 immediately implies Theorem 2.3.2, what remains is to prove that Conditions 2.4.1–2.4.3 hold if the rank function parameters ζ, η satisfy $\zeta - \frac{1}{\eta} < \frac{\alpha-1}{\beta}$. We break down this proof in the following six steps.

Step 1: Express the tails of Condition 2.4.3 in terms of moments of T_x.
Step 2: Bound the moments obtained in Step 1. These bounds are used in Steps 4-6.
Step 3: Verify Condition 2.4.1.
Step 4: Verify Condition 2.4.2.
Step 5: Verify Condition 2.4.3(i).
Step 6: Verify Condition 2.4.3(ii).

In the remainder of this section we go through each step individually and refer to later sections for more technical details.

Step 1: *From tails to moments.*
To relate the tails of Condition 2.4.3 to moments of T_x, we need the following lemma, which does not rely on any specifics of the $M/G/1$ model or SOAP.

Let

$$g_x^p(t) := t^p - \mathbb{E}(T_x)^p - p\mathbb{E}(T_x)^{p-1}(t - \mathbb{E}(T_x)). \tag{2.4}$$

We can think of $g_x^p(t)$ as t^p minus the first two terms of the Taylor series of t^p about $t = \mathbb{E}(T_x)$.

Lemma 2.4.1. *For all $p, t > 0$,*

$$\mathbb{P}(T_x > t) \leqslant \frac{\mathbb{E}(T_x^p) - \mathbb{E}(T_x)^p}{g_x^p(t)} \quad \text{if } t > \mathbb{E}(T_x),$$
$$\mathbb{P}(T_x < t) \leqslant \frac{\mathbb{E}(T_x^p) - \mathbb{E}(T_x)^p}{g_x^p(t)} \quad \text{if } t < \mathbb{E}(T_x).$$

Proof. Note that $g_x^p(t)$ is decreasing in t for $t < \mathbb{E}(T_x)$ and increasing for $t > \mathbb{E}(T_x)$. Therefore, if $t < \mathbb{E}(T_x)$, then

$$\mathbb{P}(T_x < t) = \mathbb{P}(T_x < t \text{ and } g_x^p(T_x) > g_x^p(t)) \leqslant \mathbb{P}(g_x^p(T_x) > g_x^p(t)),$$

and if $t > \mathbb{E}(T_x)$, then

$$\mathbb{P}(T_x > t) = \mathbb{P}(T_x > t \text{ and } g_x^p(T_x) > g_x^p(t)) \leqslant \mathbb{P}(g_x^p(T_x) > g_x^p(t)).$$

In both cases, Markov's inequality implies the desired bound. □

Step 2: *Moment bounds.*
The bounds presented in this step will be used to verify Conditions 2.4.2 and 2.4.3 (Steps 4-6). First, we split a job's response time T_x into two independent non-negative components, *waiting time* $Q[w_x]$ and *residence time* R_x [113]:

$$T_x =_{\mathrm{d}} Q[w_x] + R_x, \tag{2.5}$$

where $=_{\mathrm{d}}$ denotes equality in distribution. We bound $\mathbb{E}(Q[w_x])$ and $\mathbb{E}(R_x)$ in Lemmas 2.4.2 and 2.4.3 below, subject to the precondition of Theorem 2.3.2. For more details we refer to Section 2.6.

In the sequel we use the following notation. We write $f(x) = \check{o}(g(x))$ if there exists $\delta > 0$ such that $f(x) = o(x^{-\delta} g(x))$.

Lemma 2.4.2. *If* (2.1) *holds, then there exists $\beta' > \beta$ such that for all $p \in (0, \beta')$,*

$$\mathbb{E}(Q[w_x]^p) \leqslant \check{o}(x^p),$$
$$\mathbb{E}(R_x^p) \leqslant \left(\frac{x}{1-\rho}\right)^p + \check{o}(x^p).$$

Lemma 2.4.3. *If* (2.1) *holds, then*

$$\mathbb{E}(R_x) \geqslant \frac{x}{1-\rho} - \check{o}(x).$$

The proofs of Lemmas 2.4.2 and 2.4.3, presented in Section 2.6, require additional analysis of $M/G/1$ busy periods and SOAP policies, which is the purpose of Section 2.5 and Appendix 2.A. Specifically, Section 2.5 derives a general bound for fractional busy period moments, and Appendix 2.A describes how to express waiting and residence times in terms of busy periods.

We now use the moment bounds of Lemmas 2.4.2 and 2.4.3 for $Q[w_x]$ and R_x to obtain moment bounds for T_x.

Lemma 2.4.4. *If* (2.1) *holds, then*

$$\mathbb{E}(T_x^p) \leqslant \left(\frac{x}{1-\rho}\right)^p + \check{o}(x^p) \qquad \text{for all } p \in [1, \beta'),$$
$$\mathbb{E}(T_x)^p \geqslant \mathbb{E}(R_x)^p \geqslant \left(\frac{x}{1-\rho}\right)^p - \check{o}(x^p) \quad \text{for all } p > 0.$$

Proof. Note that

$$(x \pm \check{o}(x))^p = x^p \pm \check{o}(x^p). \tag{2.6}$$

The lower bound on $\mathbb{E}(T_x)^p$ follows directly from (2.5), (2.6), and Lemma 2.4.3. For the upper bound on $\mathbb{E}(T_x^p)$, we use Minkowski's inequality to compute

$$\begin{aligned}
\mathbb{E}(T_x^p) &\leqslant \left(\mathbb{E}(Q[w_x]^p)^{1/p} + \mathbb{E}(R^p)^{1/p}\right)^p && \text{[by (2.5), Minkowski]}\\
&\leqslant \left(\check{o}(x) + \left(\left(\frac{x}{1-\rho}\right)^p + \check{o}(x^p)\right)^{1/p}\right)^p && \text{[by Lemma 2.4.2]}\\
&= \left(\frac{x}{1-\rho}\right)^p + \check{o}(x^p). && \text{[by (2.6)]}
\end{aligned}$$

□

Recall that the response time moments of Lemma 2.4.1 are of the form $\mathbb{E}(T_x^p) - \mathbb{E}(T_x)^p$. We provide an upper bound for this quantity in the following lemma, which is a direct consequence of Lemma 2.4.4.

Lemma 2.4.5. *If* (2.1) *holds, then there exists* $p > \beta$ *such that* $\mathbb{E}(T_x^p) - \mathbb{E}(T_x)^p = \check{o}(x^p)$ *in the* $x \to \infty$ *limit.*

Proof. Choose $p \in (\beta, \beta')$ in Lemma 2.4.4. □

Remark 2.4.2. Núñez-Queija [95] uses a slightly different version of Lemma 2.4.5, showing that Condition 2.4.3 holds if there exists $p > \beta$ such that $\mathbb{E}(|T_x - \mathbb{E}(T_x)|^p) = \check{o}(x^p)$. Unfortunately, working with the absolute central moment is difficult unless p is an even integer, which suffices for the simple policies considered by Núñez-Queija [95] but not for the broad class of SOAP policies we consider. Our Lemma 2.4.5 is easier to work with for odd and fractional p and, as shown below, still allows us to verify Condition 2.4.3. ◇

Step 3: *Verification of Condition 2.4.1.*
Condition 2.4.1 is immediate for all SOAP policies (see Lemma 2.A.3).

Step 4: *Verification of Condition 2.4.2.*
Condition 2.4.2 follows from choosing $p = 1$ in Lemma 2.4.4.

Step 5: *Verification of Condition 2.4.3(i).*
Let $p > \beta$ be as in Lemma 2.4.5. We compute

$$
\begin{aligned}
&\lim_{x\to\infty} \frac{1}{x^p} g_x^p\left(\frac{(1\pm\varepsilon)x}{1-\rho}\right) \\
&\quad = \lim_{x\to\infty} \frac{1}{x^p}\left(\left(\frac{(1\pm\varepsilon)x}{1-\rho}\right)^p - \mathbb{E}(T_x)^p - p\mathbb{E}(T_x)^{p-1}\left(\frac{(1\pm\varepsilon)x}{1-\rho} - \mathbb{E}(T_x)\right)\right) && \text{[by (2.4)]} \\
&\quad = \frac{(1\pm\varepsilon)^p - 1 - p(1\pm\varepsilon-1)}{(1-\rho)^p} && \text{[by Condition 2.4.2]} \\
&\quad = \frac{(1\pm\varepsilon)^p - (1\pm\varepsilon p)}{(1-\rho)^p} > 0,
\end{aligned}
$$

and therefore

$$
g_x^p\left(\frac{(1\pm\varepsilon)x}{1-\rho}\right) = \Omega(x^p). \tag{2.7}
$$

Combining this with Condition 2.4.1 and Lemma 2.4.1 implies Condition 2.4.3(i):

$$
\begin{aligned}
\lim_{x\to\infty} \mathbb{P}\left(T_X < \frac{(1-\varepsilon)x}{1-\rho} \;\middle|\; X > x\right) &\leqslant \lim_{x\to\infty} \mathbb{P}\left(T_x < \frac{(1-\varepsilon)x}{1-\rho}\right) && \text{[by Condition 2.4.1]} \\
&\leqslant \lim_{x\to\infty} \frac{\mathbb{E}(T_x^\beta) - \mathbb{E}(T_x)^\beta}{g_{T_x}^\beta\left(\frac{(1-\varepsilon)x}{1-\rho}\right)} && \text{[by Lemma 2.4.1]} \\
&= \lim_{x\to\infty} \frac{\check{o}(x^\beta)}{\Omega(x^\beta)} && \text{[by (2.7) and Lemma 2.4.5]} \\
&= 0.
\end{aligned}
$$

Step 6: *Verification of Condition 2.4.3(ii).*
We begin by applying Lemma 2.4.1:

$$
\begin{aligned}
\mathbb{P}\left(T_X > \frac{(1+\varepsilon)x}{1-\rho} \text{ and } X \leqslant x\right) &= \int_0^x \mathbb{P}\left(T_t > \frac{(1+\varepsilon)x}{1-\rho}\right) \mathrm{d}F(t) \\
&\leqslant \int_0^x \frac{\mathbb{E}(T_t^p) - \mathbb{E}(T_t)^p}{g_t^p\left(\frac{(1+\varepsilon)x}{1-\rho}\right)} \mathrm{d}F(t).
\end{aligned} \tag{2.8}
$$

We would like to apply (2.7) to the denominator, but the variables in the subscript and function argument do not match. To make them match, observe in (2.4) that $g_x^p(t)$ is decreasing in $\mathbb{E}(T_x)$ as long as $\mathbb{E}(T_x) < t$. By Conditions 2.4.1 and 2.4.2, for all sufficiently large x and all $t \in (0, x)$,

$$
\mathbb{E}(T_t) \leqslant \mathbb{E}(T_x) < \frac{(1+\varepsilon)x}{1-\rho},
$$

so for sufficiently large x, we may replace t with x in the subscript in the denominator in (2.8). Using this along with Lemma 2.4.5 and (2.7) gives us

$$
\begin{aligned}
\mathbb{P}\left(T_X > \frac{(1+\varepsilon)x}{1-\rho} \text{ and } X \leqslant x\right) &\leqslant \int_0^x \frac{\mathbb{E}(T_t^p) - \mathbb{E}(T_t)^p}{g_x^p\left(\frac{(1+\varepsilon)x}{1-\rho}\right)} \mathrm{d}F(t) \\
&\leqslant \frac{1}{\Omega(x^p)} \int_0^x \left(\left(\frac{t}{1-\rho}\right)^p + \check{o}(t^p) - \left(\frac{t}{1-\rho} - \check{o}(t)\right)^p\right) \mathrm{d}F(t) \\
&\leqslant O(x^{-p}) \int_0^x \check{o}(t^p)\, \mathrm{d}F(t).
\end{aligned}
$$

We continue the computation by integrating by parts. We also apply Lemma 2.2.1, in which the constants $x_0, C > 0$ are associated to X:

$$\begin{aligned}
\mathbb{P}\left(T_X > \frac{(1+\varepsilon)x}{1-\rho} \text{ and } X \leqslant x\right) &\leqslant O(x^{-p}) - O(x^{-p}) \int_{x_0}^{x} \check{o}(t^p)\,\mathrm{d}\overline{F}(t) \\
&= O(x^{-p}) - O(x^{-p})\left(\check{o}(x^p)\overline{F}(x) - O(1) - \int_{x_0}^{x} \check{o}(t^{p-1})\overline{F}(t)\,\mathrm{d}t\right) \\
&\leqslant O(x^{-p}) + O(x^{-p}) \int_{x_0}^{x} \check{o}(t^{p-1}) \cdot C\overline{F}(x)\left(\frac{t}{x}\right)^{-\beta} \mathrm{d}t \\
&= O(x^{-p}) + O(x^{-(p-\beta)})\overline{F}(x) \int_{x_0}^{x} \check{o}(t^{(p-\beta)-1})\,\mathrm{d}t \\
&= O(x^{-p}) + \check{o}(1)\overline{F}(x).
\end{aligned}$$

Lemma 2.2.1 and the fact that $p > \beta > \alpha$ imply that this is $\check{o}(\overline{F}(x))$, so Condition 2.4.3(ii) holds.

Finally, the proof of our main result combines all of the pieces outlined in this section.

Proof of Theorem 2.3.2. Conditions 2.4.1–2.4.3 have been proven above and the theorem now follows from Proposition 2.4.1. □

The bulk of the remainder of the chapter is devoted to proving Lemmas 2.4.2 and 2.4.3, which give bounds on moments of size-conditional waiting and residence times. Our proofs of these lemmas rely on detailed analysis of fractional moments of busy periods (Section 2.5) and on new general results about SOAP policies (Appendix 2.A).

2.5 Fractional moments of busy periods

A key component of our proof of Theorem 2.3.2 is an analysis of the fractional moments of an $M/G/1$ queue. We write B for the length of a busy period and B_U for the length of a busy period with initial work U.

We denote the LST of a random variable V by

$$\widetilde{V}(s) := \mathbb{E}(\exp(-sV)).$$

We shall also encounter the *excess* $\mathcal{E}V$ of a random variable V. It has distribution

$$\mathbb{P}(\mathcal{E}V > x) = \int_0^x \frac{\mathbb{P}(V > t)}{\mathbb{E}(V)}\,\mathrm{d}t$$

and has LST

$$\widetilde{\mathcal{E}V}(s) = \frac{1 - \widetilde{V}(s)}{s\mathbb{E}(V)}. \tag{2.9}$$

Letting

$$\sigma(s) := s + \lambda(1 - \widetilde{B}(s)), \tag{2.10}$$

we can write the LSTs of B and B_U as

$$\begin{aligned} \widetilde{B}(s) &= \widetilde{X}(\sigma(s)), \\ \widetilde{B_U}(s) &= \widetilde{U}(\sigma(s))). \end{aligned} \tag{2.11}$$

Although the expression for $\widetilde{B}(s)$ is recursive, it suffices for extracting moments.

Let $\mathcal{D}$ be the derivative operator.

Lemma 2.5.1. *The derivative of $\sigma(s)$ satisfies*

$$\mathcal{D}\sigma(s) = \frac{1}{1 - \lambda(-\mathcal{D})\widetilde{X}(\sigma(s))}. \tag{2.12}$$

Proof. Differentiating (2.10) yields

$$\mathcal{D}\sigma(s) = 1 - \lambda\mathcal{D}\sigma(s) \cdot \mathcal{D}\widetilde{X}(\sigma(s)),$$

which rearranges to the desired equation. □

Lemma 2.5.2. *For all $n \in \mathbb{Z}_+$,*

$$(-\mathcal{D})^n \widetilde{B_U}(s) = \sum_{i=1}^{I} d_i (\mathcal{D}\sigma(s))^{a_i} \cdot (-\mathcal{D})^{b_i} \widetilde{U}(\sigma(s)) \prod_{j=1}^{J_i} \lambda(-\mathcal{D})^{c_{ij}} \widetilde{X}(\sigma(s)),$$

where $I, J_i, a_i, b_i, c_{ij}, d_i \in \mathbb{Z}_+$ are constants, independent of the system parameters λ and X, satisfying

$$\begin{aligned} a_i, b_i &\geqslant 1 && \text{for all } i, \\ c_{ij} &\geqslant 2 && \text{for all } i, j, \\ b_i + \sum_{j=1}^{J_i} (c_{ij} - 1) &= n && \text{for all } i, \\ b_1 &> \ldots > b_n, \\ a_1 &= b_1 = n, \\ d_1 &= 1, \\ J_1 &= 0. \end{aligned}$$

Proof. See Appendix 2.B.

As an immediate consequence, we have the following.

Corollary 2.5.1. *For all $n \in \mathbb{Z}_+$,*

$$\mathbb{E}(B_U^n) = \sum_{i=1}^{I} d_i \frac{\mathbb{E}(U^{b_i})}{(1-\rho)^{a_i}} \prod_{j=1}^{J_i} \lambda \mathbb{E}(X^{c_{ij}}),$$

where $I, J_i, a_i, b_i, c_{ij}, d_i \in \mathbb{Z}_+$ are as in Lemma 2.5.2.

The main result of this subsection is that nearly the same formula works for fractional moments, though it gives an upper bound instead of an exact result. To bound $\mathbb{E}(B_U^p)$ for $p = n - q$, $n \in \mathbb{Z}_+$, we start with the formula for $\mathbb{E}(B_U^n)$, then decrease some of the exponents by q. Specifically, for each i, we decrease a_i and one more exponent of our choice, either b_i or one of the c_{ij}.

Theorem 2.5.1. *Let $p = n - q > 0$ for $n \in \mathbb{Z}_+$ and $q \in (0,1)$. Then for all choices of $\chi_{ij} \in \{0,1\}$ such that $\sum_{j=0}^{J_i} \chi_{ij} = 1$ for all i, we have*

$$\mathbb{E}(B_U^p) \leqslant \sum_{i=1}^{I} d_i \frac{\mathbb{E}(U^{b_i - q\chi_{i0}})}{(1-\rho)^{a_i - q}} \prod_{j=1}^{J_i} \lambda \mathbb{E}(X^{c_{ij} - q\chi_{ij}}),$$

where $I, J_i, a_i, b_i, c_{ij}, d_i \in \mathbb{Z}_+$ are as in Lemma 2.5.2.

Proof. See Appendix 2.B.

Remark 2.5.1. Remerova et al. [104, Lemma 3] discuss the finiteness of $\mathbb{E}(f(B))$ for the $M/G/1$ busy period B for a quite general class of functions $f(\cdot)$. They obtain bounds on moments of B that are more general but less sharp than Theorem 2.5.1. Bansal et al. [13] formulate a bound on fractional moments of $GI/GI/1$ busy periods, but their bound only characterizes the growth rate in the $\rho \to 1$ limit. Focusing on the $M/G/1$ setting, in which a recursive LST is known for busy periods, enables us to obtain a much sharper bound in Theorem 2.5.1 that characterizes the coefficient of each $1/(1-\rho)^b$ term. ◇

2.6 Proving tail optimality

Recall from Section 2.4 that the proof of our main result, Theorem 2.3.2, is complete once we prove Lemmas 2.4.2 and 2.4.3. This section is devoted to proving these last two lemmas. In doing so, we require some definitions and preliminary results related to SOAP from Appendix 2.A. It is therefore recommended to look through that appendix section before thoroughly reading this section.

Many of the lemma statements in this section use similar preconditions on a parameter p. For convenience, we name these conditions $\Phi(p)$ and $\Psi(p)$:

$$\begin{aligned} \Phi(p) &\iff \zeta - \frac{1}{\eta} < \frac{\alpha - 1}{p} \text{ or } p \leqslant 0, \\ \Psi(p) &\iff 1 - \frac{1}{\zeta} < \frac{\alpha - 1}{p} \text{ or } p \leqslant 0. \end{aligned}$$

For all $p \geqslant q$, we have

$$\begin{aligned} \Phi(p) &\Rightarrow \Psi(p), \\ \Phi(p) &\Rightarrow \Phi(q), \\ \Psi(p) &\Rightarrow \Psi(q). \end{aligned} \tag{2.13}$$

The latter two implications are straightforward, and $\Phi(p) \Rightarrow \Psi(p)$ follows from $\zeta \leqslant \eta$ and the fact that $\Psi(p)$ is vacuously true for $\zeta \leqslant 1$.

We prove Lemmas 2.4.2 and 2.4.3 by way of the following more general statements.

Lemma 2.6.1. *For all $p > 0$ satisfying $\Phi(p)$, in the $x \to \infty$ limit,*

$$\mathbb{E}(Q[w_x]^p) \leqslant \check{o}(x^p).$$

Lemma 2.6.2. *For all $p > 0$ satisfying $\Psi(p-1)$, in the $x \to \infty$ limit,*

$$\mathbb{E}(R_x^p) \leqslant \left(\frac{x}{1-\rho}\right)^p + \check{o}(x^p).$$

Lemma 2.6.3. *If $\zeta < 1$ or $\eta < \infty$, then in the $x \to \infty$ limit,*

$$\mathbb{E}(R_x) \geqslant \frac{x}{1-\rho} - \check{o}(x).$$

Lemmas 2.6.1 and 2.6.2 immediately imply Lemma 2.4.2, and Lemma 2.6.3 immediately implies Lemma 2.4.3, so it remains only to prove Lemmas 2.6.1–2.6.3. In the remainder of this section we prove Lemma 2.6.2. We also give the main ideas of the proofs of Lemmas 2.6.1 and 2.6.3, deferring their full proofs to Appendix 2.D.

To prove Lemma 2.6.2, we use Lemma 2.A.1 to bound residence times using a busy period, namely $R_x \leqslant_{\text{st}} B_x[w_x]$. We can apply Theorem 2.5.1 to bound moments of the busy period $B_x[w_x]$ in terms of moments of its initial work, which is simply x, and its job size, which is $X_0[w_x]$. Thus, to bound moments of R_x, it suffices to bound moments of $X_0[w_x]$, which is the purpose of the following lemma.

Lemma 2.6.4. *For all $p > 0$ satisfying $\Psi(p)$, in the $x \to \infty$ limit,*

$$\mathbb{E}(X_0[w_x]^{p+1}) = \check{o}(x^p).$$

Proof. By (2.15) and Lemma 2.A.5,

$$\mathbb{E}(X_0[w_x]^{p+1}) = \mathbb{E}(X\langle z_x\rangle^{p+1}) = \int_0^{z_x} (p+1)t^p\overline{F}(t)\,\mathrm{d}t.$$

Hence for $x \to \infty$, Assumption 2.3.1 implies

$$\begin{aligned}\mathbb{E}(X_0[w_x]^{p+1}) &= \int_0^{O(x^{\max\{1,\zeta\}})} O(t^{p-\alpha})\,\mathrm{d}t \\ &= O(x^{\max\{0,(p-(\alpha-1))\max\{1,\zeta\}\}}).\end{aligned} \tag{2.14}$$

If $\zeta \leqslant 1$, then (2.14) is $O(x^{\max\{0,p-(\alpha-1)\}}) = \check{o}(x^p)$. If instead $\zeta > 1$, then $\Psi(p)$ implies (2.14) is $\check{o}(x^p)$. □

Armed with bounds on moments of $X_0[w_x]$, we are now ready to prove Lemma 2.6.2.

Proof of Lemma 2.6.2. Let $p = n - q$ for $n \in \mathbb{Z}_+$ and $q \in (0,1)$. We again apply Theorem 2.5.1, choosing $\chi_{i0} = 1$ for all i. Using that and Lemma 2.A.1, we obtain

$$\begin{aligned}\mathbb{E}(R_x^p) &\leqslant \mathbb{E}(B_x^p[w_x]) && \text{[by Lemma 2.A.1]}\\ &\leqslant \left(\frac{x}{1-\rho_0[w_x]}\right)^p + \sum_{i=2}^{I} d_i \frac{x^{b_i-q}}{(1-\rho_0[w_x])^{a_i-q}} \prod_{j=1}^{J_i} \lambda\mathbb{E}(X_0[w_x]^{c_{ij}}). && \text{[by Theorem 2.5.1]}\end{aligned}$$

Recall from Lemma 2.5.2 that

$$b_i - q + \sum_{j=1}^{J_i}(c_{ij} - 1) = n - q = p.$$

This means for all i and j, we have $c_{ij} - 1 \leqslant p - b_i \leqslant p - 1$, so $\Psi(c_{ij} - 1)$ holds by (2.13). We can therefore apply Lemma 2.6.4, which yields

$$\begin{aligned}\sum_{i=2}^{I} d_i \frac{x^{b_i - q}}{(1 - \rho_0[w_x])^{a_i - q}} \prod_{j=1}^{J_i} \lambda \mathbb{E}(X_0[w_x]^{c_{ij}}) &= \sum_{i=2}^{I} O(x^{b_i - q}) \prod_{j=1}^{J_i} \check{o}(x^{c_{ij}-1}) \\ &= \sum_{i=2}^{I} \check{o}\big(x^{b_i - q + \sum_{j=1}^{J_i}(c_{ij}-1)}\big) \\ &= \check{o}(x^p). \qquad \square\end{aligned}$$

It remains only to prove Lemma 2.6.1 and Lemma 2.6.3. The proof of Lemma 2.6.1, an upper bound on moments of waiting time, follows essentially the same outline as the proof of Lemma 2.6.2: we use Lemma 2.A.2 to bound waiting time in terms of busy periods, use Theorem 2.5.1 to bound the moments of those busy periods in terms of moments of $X_i[w_x]$, then use Lemma 2.6.5 below to bound those moments. Finally, we prove Lemma 2.6.3 by combining several results from Appendix 2.A.

Lemma 2.6.5. *For all $p > 0$ satisfying $\Phi(p)$, in the $x \to \infty$ limit,*

$$\sum_{k=1}^{K[w_x]} \mathbb{E}(X_k[w_x]^{p+1}) = \check{o}(x^p).$$

Proofs of Lemmas 2.6.1, 2.6.3, and 2.6.5. See Appendix 2.D.

2.7 Discussion

Over the past decades, much effort has been given to the task of designing policies that maintain the optimal response time tail, i.e., a response time tail that is equally heavy as the tail of the job size distribution. While the analysis of individual policies has been successful in many cases, e.g., SRPT and FB, there are many important policies that have resisted analysis and, further, little is known about which scheduling mechanisms provably lead to tail optimality. In this chapter, we provide general sufficient conditions on the type of prioritization that ensures tail optimality in policies that do not have access to job sizes. Our sufficient conditions enable the first results on tail optimality for Gittins, RMLF, SERPT, and FB with limited preemption.

Although our sufficient conditions define a broad class of tail-optimal policies, it must be stressed that they are not necessary. For instance, Processor Sharing (PS), which is known to be tail-optimal, does not use job sizes and does not satisfy our sufficient conditions since it is not a SOAP policy. Thus, it is important to continue to develop both necessary and sufficient

conditions for tail optimality. An interesting open question is to identify sufficient conditions that unify the results in [97] for size-based policies with the results in this chapter on policies that do not have access to job size information. Additionally, the only necessary condition known for tail optimality is given by [127], which proves that all tail-optimal policies must "remain stable when faced with the arrival of a job with infinite size." It is not known if this condition is also sufficient. To this end, Guillemin et al. [55] have developed an interesting probabilistic method to prove tail optimality of the response time and service time for a large class of $M/G/1$ processor-sharing queues (with and without impatience, and with finite or infinite capacity). It would be interesting to investigate whether their approach is applicable to a class of SOAP policies. See also Sections 2.4 and 3 of [22], where the approaches of [55] and [95] are compared and unified.

Another interesting research topic is to weaken the goal and, instead of trying to characterize policies that are tail-optimal, characterize classes of policies with response time tails that are neither optimal (of the same order as the job size tail) nor pessimal (one order higher). It is known that the orders of the job size and response time tails can differ by any number $\gamma \in (0, 1]$ [26, 87], and so a natural question is: what forms of prioritization achieve these intermediate response time tails?

It also remains to be seen what tail-optimality results extend to more complicated queueing models. This includes single-server systems with variable service rate, such as systems using computational sprinting [86, 101], as well as systems with multiple servers and systems with a non-Poisson arrival process.

Finally, it is worth considering tail optimality among light-tailed job size distributions. Are there sufficient conditions on prioritization that ensure tail optimality in the light-tailed setting? The results of [127] highlight that if a policy is tail-optimal under heavy-tailed job sizes it cannot be tail-optimal under light-tailed job sizes, and thus it is clear that the sufficient conditions must change. In Chapter 3, we characterize the tail performance of SOAP policies when the job size distribution is light-tailed.

Appendix 2.A SOAP background

Recall from Definition I.1 that a SOAP policy is a policy defined by a rank function $r : \mathbb{R}_+ \to \mathbb{R}$ mapping each job's age, or attained service, to its rank. The scheduler always serves the job of minimum rank, so lower rank means higher priority.

In this section we give background on how to analyze the mean response time of SOAP policies. Appendices 2.A.1 and 2.A.2 review the response time analysis in [113], adapting the notation slightly to suit our needs. These expressions are hard to work with directly, and the complexity grows when considering higher moments. As such, we introduce new concepts and results in Appendices 2.A.3 and 2.A.4 which help simplify the analysis.

2.A.1 Core SOAP concepts

All of the definitions in the remainder of this section are given in terms of a generic SOAP policy with rank function r. The way [113] analyzes response time of SOAP policies is with the "tagged job" approach, following the journey of a specific job from arrival to departure. Suppose the tagged job has size x. One of their key insights is that to determine the tagged job's response time, its current rank is less important than the worst rank it will attain in its remaining time in the system.

Definition 2.A.1. *The* worst future rank *of a job of size x at age a, written $w_x(a)$, is*

$$w_x(a) = \sup_{a \leqslant b < x} r(b).$$

The worst ever rank *of a job of size x is $w_x = w_x(0)$.*

When the tagged job initially enters the system, there may be a number of other jobs already present. Any other job with rank w_x or less is "relevant" to the tagged job, meaning it will receive some amount of service during the tagged job's time in the system.

Definition 2.A.2. *The amount of w-*relevant work *a job has is the amount of service it needs to either finish or attain rank greater than w. Similarly, the amount of w-relevant work in a system is the total amount of w-relevant work of all jobs in the system.*

To find the response time of the tagged job, we need to know the amount of w_x-relevant work it encounters upon arrival. Because the arrival process is Poisson, this means finding the steady-state distribution of the amount of w_x-relevant work in the system, for which we need the following definition.

Definition 2.A.3.

*(i) The kth w-*relevant age interval *is $(b_k[w], c_k[w])$, where*

$$\begin{aligned}
b_0[w] &= 0 \\
c_0[w] &= \inf\{a \geqslant 0 \mid r(a) > w\} \\
b_k[w] &= \inf\{a > c_{k-1}[w] \mid r(a) \leqslant w\} \quad \text{for all } k \geqslant 1 \\
c_k[w] &= \inf\{a > b_k[w] \mid r(a) > w\} \quad \text{for all } k \geqslant 1.
\end{aligned}$$

Additionally, let $K[w]$ be the number of w-relevant age intervals, namely the maximum k such that $b_k[w] < \infty$. It may be that $K[w] = \infty$.

*(ii) The kth w-*relevant job segment *is*

$$X_k[w] =_{\mathrm{d}} \max\{0, \min\{X, c_k[w]\} - b_k[w]\}.$$

For convenience, we define $X_k[w] = 0$ for $k > K[w]$.

(iii) The kth w-relevant load *is*

$$\rho_k[w] = \lambda \mathbb{E}(X_k[w]).$$

For convenience, we also define

$$\rho_\Sigma[w] = \sum_{k=0}^{K[w]} \rho_k[w].$$

The tagged job can also be delayed by jobs arriving after it. The following definition helps us quantify this delay.

Definition 2.A.4. *The* w-relevant busy period*, written* $B[w]$*, is the length of an* $M/G/1$ *busy period with arrival rate* λ *and job size* $X_0[w]$*. Similarly, the* w-relevant busy period with initial work U*, written* $B_U[w]$*, is the length of such a busy period with initial work* U*.*

2.A.2 SOAP response time analysis

To study the response time of SOAP policies, we introduce the following random variables.

Definition 2.A.5. *The* residence time *of a job of size* x*, written* R_x*, is a random variable with transform*

$$\widetilde{R_x}(s) = \exp\left(-\int_0^x (s + \lambda(1 - B[w_x(a)-]))\,\mathrm{d}a\right).$$

We can write the residence time as the limit of a sum of increasingly many busy periods with increasingly little initial work:

$$R_x =_{\mathrm{d}} \lim_{n\to\infty} \sum_{k=1}^{n} B_{x/n}[w_x(kx/n)-].$$

Definition 2.A.6. *The* rank-w waiting time*, written* $Q[w]$*, is a random variable that has the same distribution as a particular busy period:*

$$Q[w] =_{\mathrm{d}} B_{U[w]}[w-],$$

where $U[w]$ *is the steady-state amount of* w*-relevant work, which is a random variable with LST [113, Lemma 5.2]*

$$\widetilde{U}[w](s) = \frac{1 - \rho_\Sigma[w] + \sum_{i=1}^{K[w]} \rho_i[w]\widetilde{\mathcal{E}X_i}[w](s)}{1 - \rho_0[w]\widetilde{\mathcal{E}X_0}[w](s)}.$$

Scully et al. [113, Theorem 5.4] show that for any SOAP policy, T_x is the independent sum of waiting and residence times, namely $T_x =_{\mathrm{d}} Q[w_x] + R_x$, which implies the following formula for mean response time.

Corollary 2.A.1 ([113, Theorem 5.5])**.** *Under any SOAP policy,*

$$\mathbb{E}(Q[w]) = \frac{\frac{\lambda}{2}\sum_{i=0}^{K[w]} \mathbb{E}(X_i[w]^2)}{(1 - \rho_0[w])(1 - \rho_0[w-])},$$

$$\mathbb{E}(R_x) = \int_0^x \frac{1}{1 - \rho_0[w_x(a)-]}\,\mathrm{d}a,$$

$$\mathbb{E}(T_x) = \mathbb{E}(Q[w_x]) + \mathbb{E}(R_x).$$

2.A.3 Stochastic response time bounds

The next two lemmas, proven in Appendix 2.C, bound the residence and waiting time in terms of busy periods. The main concept in their proofs is the observation that jobs with rank larger than w will not be served before the w-relevant busy period ends.

Recall for two random variables A and B that $A \leqslant_{\text{st}} B$ if $\mathbb{P}(A > t) \leqslant \mathbb{P}(B > t)$ for all t (we say that A is stochastically bounded by B).

Lemma 2.A.1. *For any SOAP policy, the residence time R_x of a job of size x is stochastically bounded by*

$$R_x \leqslant_{\text{st}} B_x[w_x].$$

Lemma 2.A.2. *For any SOAP policy, the rank-w waiting time $Q[w]$ is stochastically bounded by*

$$Q[w] \leqslant_{\text{st}} \begin{cases} \mathcal{E}B_{X_0[w]}[w] & \text{w.p. } \pi_0[w], \\ \mathcal{E}B_{X_1[w]}[w] & \text{w.p. } \pi_1[w], \\ \vdots & \\ 0 & \text{w.p. } 1 - \rho_\Sigma[w], \end{cases}$$

where

$$\pi_0[w] := \frac{\rho_0[w](1 - \rho_\Sigma[w])}{1 - \rho_0[w]},$$
$$\pi_k[w] := \frac{\rho_k[w]}{1 - \rho_0[w]} \qquad \text{for all } k \geqslant 1.$$

These lemmas allow us to express response time moments in terms of busy period moments, which can be further analyzed using Theorem 2.5.1.

2.A.4 Additional SOAP bounds

The purpose of the next few definitions is to formalize Assumption 2.3.1. All of them relate to the worst rank for a job of size x.

Definition 2.A.7. *The* maximum relevant age *of a job of size x is the latest age at which another job can possibly outrank it:*

$$u_x = c_{K[w_x]}[w_x] = \sup\{a > 0 \mid r(a) \leqslant w_x\}.$$

The next two definitions are due to Scully et al. [114].

Definition 2.A.8. *A* hill age *is an age b such that $r(a) < r(b)$ for all ages $a < b$. An age that is not a hill age is called a* valley age.

Definition 2.A.9. *The* previous and next hill ages *of x are, respectively,*

$$y_x = c_0[w_x-],$$
$$z_x = c_0[w_x].$$

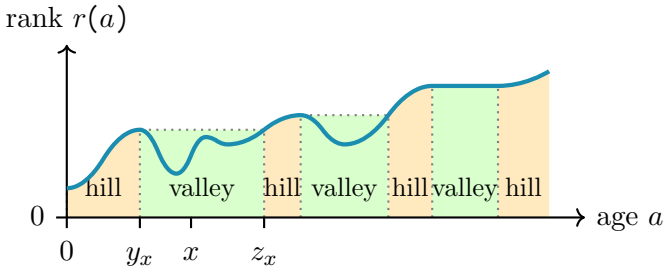

Figure 2.3: Hills and valleys.

Note that it may be that $y_x = x = z_x$. This occurs when the rank function is strictly increasing at x and x is a "running maximum", meaning $r(a) < r(x)$ for all ages $a \in [0, x)$. For any x such that $y_x < z_x$, we call the interval (y_x, z_x) a *valley*, and any interval that does not overlap with a valley is called a *hill*. Figure 2.3 illustrates hills and valleys, including previous and next hill ages.

The next definition is not specific to SOAP but, as we will soon see, can be helpful when analyzing the moments of a SOAP policy's response time.

Definition 2.A.10.

(i) The a-cutoff job segment *is*

$$X\langle a\rangle =_{\mathrm{d}} \min\{X, a\}.$$

(ii) The a-cutoff load *is*

$$\rho\langle a\rangle = \lambda \mathbb{E}(X\langle a\rangle).$$

The y_x- and z_x-cutoff job segments give us another way to write $X_0[w_x-]$ and $X_0[w_x]$:

$$\begin{aligned} X_0[w_x-] &=_{\mathrm{d}} X\langle y_x\rangle, \\ X_0[w_x] &=_{\mathrm{d}} X\langle z_x\rangle. \end{aligned} \tag{2.15}$$

The following lemma follows immediately from Definitions 2.A.5 and 2.A.6, observing that w-relevant busy periods are stochastically increasing in w.

Lemma 2.A.3.

(i) $Q[w]$ *is stochastically increasing in* w.

(ii) R_x *is stochastically increasing in* x.

(iii) T_x *is stochastically increasing in* x.

Note that Lemma 2.A.3 completes Step 3 of the proof described in Section 2.4. In fact, the only part of the proof outlined in Section 2.4 that remains to prove is Step 2, specifically Lemmas 2.4.2 and 2.4.3. We prove these lemmas in Section 2.6 with the help of the useful results given in the remainder of this appendix.

The next lemmas follow from integration by parts.

Lemma 2.A.4. *For any $p > 0$,*

$$\mathbb{E}(X_k[w]^p) = \int_{b_k[w]}^{c_k[w]} p(t - b_k[w])^{p-1}\overline{F}(t)\,\mathrm{d}t.$$

Lemma 2.A.5. *For any $p > 0$,*

$$\mathbb{E}(X\langle a\rangle^p) = \int_0^a pt^{p-1}\overline{F}(t)\,\mathrm{d}t.$$

Previous and next hill ages are also useful for bounding moments of $X_k[w_x]$ for $k \geqslant 1$, specifically by combining Lemma 2.A.4 with the following lemma.

Lemma 2.A.6. *For all ranks w and $k \geqslant 1$, if $x \in (b_k[w], c_k[w])$, then*

$$y_x \leqslant b_k[w] < x < c_k[w] \leqslant z_x.$$

Proof. By Definition 2.A.9, we have $x \in (y_x, z_x)$, where y_x is the first age at which the rank function reaches rank w_x, and z_x is the first age at which the rank function strictly exceeds w_x. Because $k \geqslant 1$, there must be some age $a \leqslant b_k[w]$ at which $r(a) > b_k[w]$, so $w_x > w$. But by Definition 2.A.3, a job's rank is at most w during $(b_k[w], c_k[w])$, so $y_x, z_x \notin (b_k[w], c_k[w])$. We therefore must have $y_x \leqslant b_k[w]$ and $z_x \geqslant c_k[w]$. □

Appendix 2.B Proofs for Section 2.5

Given a random variable V, it is well known how to obtain positive integer moments of V from its LST: for all $n \in \mathbb{Z}_+$,

$$(-\mathcal{D})^n\widetilde{V}(s) = \mathbb{E}(V^n \exp(-sV)), \tag{2.16}$$

so in particular $\mathbb{E}(V^n) = (-\mathcal{D})^n\widetilde{V}(0)$. Obtaining fractional moments of V from its LST is trickier but also possible: for all $p \geqslant 0$, letting $p = n - q > 0$ for $n \in \mathbb{Z}_+$ and $q \in (0, 1)$, we have

$$\begin{aligned}
\mathbb{E}(V^p) &= \int_0^\infty t^{n-q}\,\mathrm{d}\mathbb{P}(V \leqslant t) \\
&= \int_{t=0}^\infty \frac{t^{n-q}}{\Gamma(q)} \int_{s=0}^\infty (st)^{q-1}\exp(-st)\cdot t\,\mathrm{d}s\,\mathrm{d}\mathbb{P}(V \leqslant t) \\
&= \int_{s=0}^\infty \frac{1}{s^{1-q}\Gamma(q)} \int_{t=0}^\infty t^n \exp(-st)\,\mathrm{d}\mathbb{P}(V \leqslant t)\,\mathrm{d}s \\
&= \int_0^\infty \frac{1}{s^{1-q}\Gamma(q)}(-\mathcal{D})^n\widetilde{V}(s)\,\mathrm{d}s.
\end{aligned} \tag{2.17}$$

Lemma 2.5.2. *For all $n \in \mathbb{Z}_+$,*

$$(-\mathcal{D})^n\widetilde{B_U}(s) = \sum_{i=1}^I d_i(\mathcal{D}\sigma(s))^{a_i}\cdot(-\mathcal{D})^{b_i}\widetilde{U}(\sigma(s))\prod_{j=1}^{J_i}\lambda(-\mathcal{D})^{c_{ij}}\widetilde{X}(\sigma(s)),$$

where $I, J_i, a_i, b_i, c_{ij}, d_i \in \mathbb{Z}_+$ are constants, independent of the system parameters λ and X, satisfying

$$\begin{aligned} a_i, b_i &\geqslant 1 && \text{for all } i, \\ c_{ij} &\geqslant 2 && \text{for all } i, j, \\ b_i + \sum_{j=1}^{J_i} (c_{ij} - 1) &= n && \text{for all } i, \\ b_1 &> \ldots > b_n, \\ a_1 &= b_1 = n, \\ d_1 &= 1, \\ J_1 &= 0. \end{aligned}$$

Proof. We proceed by induction on n. The base case of $n = 0$ is immediate by (2.11), so we turn to the inductive step. By relabeling, we can have $a_1 > \ldots > a_n$ without loss of generality. We address the constraint on the $i = 1$ constants at the end of the proof.

For $a, b, c_j \in \mathbb{Z}_+$, let

$$\tau_{a,b,\langle c_1,\ldots,c_J\rangle}(s) = (\mathcal{D}\sigma(s))^a \cdot (-\mathcal{D})^b \widetilde{U}(\sigma(s)) \prod_{j=1}^{J} \lambda(-\mathcal{D})^{c_j} \widetilde{X}(\sigma(s)).$$

We abbreviate $c = \langle c_1, \ldots, c_J\rangle$. Call $b + \sum_{j=1}^{J}(c_j - 1)$ the *degree* of $\tau_{a,b,c}(s)$. For the inductive step, it suffices to show that the derivative of a term with degree n is a sum of terms with degree $n + 1$. Using Lemma 2.5.1, we compute

$$\begin{aligned} -\mathcal{D}\tau_{a,b,c}(s) = \tau_{a+1,b+1,c}(s) &+ a\tau_{a+2,b,\langle c_1,\ldots,c_J,2\rangle}(s) \\ &+ \sum_{j=1}^{J} \tau_{a+1,b,\langle c_1,\ldots,c_j+1,\ldots,c_J\rangle}(s), \end{aligned} \tag{2.18}$$

where in the second term on the right-hand side, we append c with an extra 2, and in the third term, we increase the jth element of c by 1. Note that each term has degree $n + 1$, as desired.

We now address the constraint on the $i = 1$ term, again by induction on n. The base case of $n = 0$ is immediate by (2.11), and the inductive step follows from plugging $a = b = n$ into (2.18). □

Lemma 2.B.1. *Let $p = m - q > 0$ for $m \in \mathbb{Z}_+$ and $q \in (0, 1)$. Then for any nonnegative random variable V, we have*

$$\int_0^\infty \frac{1}{s^{1-q}\Gamma(q)} (-\mathcal{D})^m \widetilde{V}(\sigma(s)) \cdot \mathcal{D}\sigma(s)\, \mathrm{d}s \leqslant \frac{\mathbb{E}(V^p)}{(1-\rho)^{1-q}}.$$

Proof. We first show that for all $s > 0$,

$$\frac{\sigma(s)}{s} \leqslant \frac{1}{1-\rho}. \tag{2.19}$$

By (2.10),

$$\begin{aligned}\sigma(s) &= s + \lambda(1 - \tilde{X}(\sigma(s)))\\ &= s + \lambda \int_0^\infty (1 - \exp(-x\sigma(s)))\,\mathrm{d}\mathbb{P}(X \leqslant x)\\ &\leqslant s + \lambda\mathbb{E}(X)\sigma(s),\end{aligned}$$

which implies (2.19). Making a change of variable $\sigma = \sigma(s)$, we compute

$$\begin{aligned}&\int_0^\infty \frac{1}{s^{1-q}\Gamma(q)}(-\mathcal{D})^m\widetilde{V}(\sigma(s)) \cdot \mathcal{D}\sigma(s)\,\mathrm{d}s\\ &\quad\leqslant \frac{1}{(1-\rho)^{1-q}}\int_0^\infty \frac{1}{(\sigma(s))^{1-q}\Gamma(q)}(-\mathcal{D})^m\widetilde{V}(\sigma(s)) \cdot \mathcal{D}\sigma(s)\,\mathrm{d}s \qquad \text{[by (2.19)]}\\ &\quad= \frac{1}{(1-\rho)^{1-q}}\int_0^\infty \frac{1}{\sigma^{1-q}\Gamma(q)}(-\mathcal{D})^m\widetilde{V}(\sigma)\,\mathrm{d}\sigma\\ &\quad= \frac{\mathbb{E}(V^p)}{(1-\rho)^{1-q}}. \qquad \text{[by (2.17)]}\end{aligned}$$

□

Proof of Theorem 2.5.1. Let V be a nonnegative random variable. By (2.16), for all $m \in \mathbb{Z}_+$ and $s > 0$,

$$(-\mathcal{D})^m\widetilde{V}(\sigma(s)) \leqslant (-\mathcal{D})^m\widetilde{V}(0) = \mathbb{E}(V^m), \tag{2.20}$$

which when applied to (2.12) implies

$$\mathcal{D}\sigma(s) \leqslant \frac{1}{1-\rho}. \tag{2.21}$$

Combining (2.21) and Lemma 2.5.2 yields

$$\begin{aligned}\mathbb{E}(B_U^p) &= \int_0^\infty \frac{1}{s^{1-q}\Gamma(q)}(-\mathcal{D})^n\widetilde{B_U}(s)\,\mathrm{d}s\\ &= \int_0^\infty \frac{1}{s^{1-q}\Gamma(q)}\Bigg(\sum_{i=1}^I d_i(\mathcal{D}\sigma(s))^{a_i} \cdot (-\mathcal{D})^{b_i}\widetilde{U}(\sigma(s))\prod_{j=1}^{J_i}\lambda(-\mathcal{D})^{c_{ij}}\widetilde{X}(\sigma(s))\Bigg)\mathrm{d}s\\ &\leqslant \sum_{i=1}^I \frac{d_i}{(1-\rho)^{a_i-1}}\int_0^\infty \frac{1}{s^{1-q}\Gamma(q)}\Bigg((-\mathcal{D})^{b_i}\widetilde{U}(\sigma(s))\prod_{j=1}^{J_i}\lambda(-\mathcal{D})^{c_{ij}}\widetilde{X}(\sigma(s))\Bigg) \cdot \mathcal{D}\sigma(s)\,\mathrm{d}s.\end{aligned} \tag{2.22}$$

It remains only to bound the integral in (2.22), which we do separately for each value of i. If $\chi_{i0} = 1$, then applying Lemma 2.B.1 with $V = U$ and $m = b_i$ along with (2.20) yields

$$\begin{aligned}&\int_0^\infty \frac{1}{s^{1-q}\Gamma(q)}\Bigg((-\mathcal{D})^{b_i}\widetilde{U}(\sigma(s))\prod_{j=1}^{J_i}\lambda(-\mathcal{D})^{c_{ij}}\widetilde{X}(\sigma(s))\Bigg) \cdot \mathcal{D}\sigma(s)\,\mathrm{d}s\\ &\quad\leqslant \int_0^\infty \frac{1}{s^{1-q}\Gamma(q)}\Bigg((-\mathcal{D})^{b_i}\widetilde{U}(\sigma(s)) \cdot \mathcal{D}\sigma(s)\,\mathrm{d}s\Bigg)\prod_{j=1}^{J_i}\lambda\mathbb{E}(X^{c_{ij}}) \qquad \text{[by (2.20)]}\\ &\quad\leqslant \frac{\mathbb{E}(U^{b_i-q})}{(1-\rho)^{1-q}}\prod_{j=1}^{J_i}\lambda\mathbb{E}(X^{c_{ij}}), \qquad \text{[by Lemma 2.B.1]}\end{aligned}$$

which gives the desired bound for the ith summand in (2.22). The case where $\chi_{ij} = 1$ for some $j \geqslant 1$ is very similar, except we apply Lemma 2.B.1 with $V = X$ and $m = c_{ij}$. □

Remark 2.B.1. Note that one might hope to get a simpler expression for $(-1)^n \frac{\mathrm{d}^n}{\mathrm{d}s^n}\widetilde{U}(\sigma(s))$ by applying the following compact form of Faà di Bruno's formula for derivatives of composite functions [68]:

$$\frac{\mathrm{d}^n\widetilde{U}(\sigma(s))}{\mathrm{d}s^n} = \sum_{k=1}^{n} \frac{\mathrm{d}^k\widetilde{U}(y)}{\mathrm{d}y^k}\bigg|_{y=\sigma(s)} B_{n,k}(\sigma'(s), \sigma^{(2)}(s), \ldots, \sigma^{(n-k+1)}(s)), \tag{2.23}$$

where the partial or incomplete exponential Bell polynomials $B_{n,k}$ are given by

$$B_{n,k}(x_1, \ldots, x_{n-k+1}) = \sum \frac{n!}{j_1! \ldots j_{n-k+1}!}\left(\frac{x_1}{1!}\right)^{j_1} \cdots \left(\frac{x_{n-k+1}}{(n-k+1)!}\right)^{j_{n-k+1}},$$

where the sum is taken over all nonnegative integers $(j_1, \ldots, j_{n-k+1})$ with $j_1 + \ldots + j_{n-k+1} = k$ and $j_1 + 2j_2 + \ldots + (n-k+1)j_{n-k+1} = n$. In particular, $B_{n,n}(x) = x^n$, so that the leading $(k = n)$ term of $(-\mathcal{D})^n\widetilde{B_U}(s)$ is bounded by $\mathbb{E}(U^n)(\sigma'(s))^n$. In the proof of Lemma 2.6.2, where we have a busy period with initial work x and job size $X_0[w_x]$, that would very quickly lead to the leading term $\left(\frac{x}{1-\rho_0[w_x]}\right)^p$. However, to show that the remaining $n-1$ terms in (2.23) are $\check{o}(x^p)$ requires a detailed study of the higher derivatives of $\sigma(s)$. ◇

Appendix 2.C Proofs for Appendix 2.A

Lemma 2.A.1. *For any SOAP policy, the residence time R_x of a job of size x is stochastically bounded by*

$$R_x \leqslant_{\mathrm{st}} B_x[w_x].$$

Proof. By Definition 2.A.5,

$$R_x = \lim_{n\to\infty} \sum_{k=1}^{n} B_{x/n}[w_x(kx/n)-].$$

It is clear from Definition 2.A.4 that $B_U[w]$ is stochastically increasing in w for any U. Definition 2.A.1 implies $w_x \geqslant w_x(a)$ for all $a \geqslant 0$, so

$$R_x \leqslant_{\mathrm{st}} \lim_{n\to\infty} \sum_{k=1}^{n} B_{x/n}[w_x] =_{\mathrm{d}} B_x[w_x].$$ □

Lemma 2.A.2. *For any SOAP policy, the rank-w waiting time $Q[w]$ is stochastically bounded by*

$$Q[w] \leqslant_{\mathrm{st}} \begin{cases} \mathcal{E}B_{X_0[w]}[w] & \text{w.p. } \pi_0[w], \\ \mathcal{E}B_{X_1[w]}[w] & \text{w.p. } \pi_1[w], \\ \vdots & \\ 0 & \text{w.p. } 1-\rho_\Sigma[w], \end{cases}$$

where

$$\pi_0[w] := \frac{\rho_0[w](1-\rho_\Sigma[w])}{1-\rho_0[w]},$$

$$\pi_k[w] := \frac{\rho_k[w]}{1-\rho_0[w]} \qquad \text{for all } k \geqslant 1.$$

Proof. Following the approach of [113, Section 5], one can think of $Q[w]$ as defined by the following process. A job J with initial rank r arrives at a random time. Because the system uses FCFS tiebreaking between jobs of the same rank, job J is first served when

- all jobs that arrived *before* J either complete or have rank strictly greater than r, and
- all jobs that arrived *after* J either complete or have rank greater than or equal to r.

Then $Q[w]$ is the amount of time from J's arrival to its first service.

Define $Q'[w]$ in the same way as $Q[w]$ but in a system that breaks rank ties by prioritizing all other jobs over J. Clearly, $Q[w] \leqslant_{\mathrm{st}} Q'[w]$. But we can succinctly describe $Q'[w]$: it is either 0 or the excess of a w-relevant busy period with some amount of initial work. Specifically, the initial work is a kth w-relevant job segment for some $k \geqslant 0$. Thus, letting $\pi_k[w]$ be the steady-state probability that the system is in a w-relevant busy period started by a kth w-relevant segment, we have

$$Q'[w] =_{\mathrm{d}} \begin{cases} \mathcal{E}B_{X_0[w]}[w] & \text{w.p. } \pi_0[w], \\ \mathcal{E}B_{X_1[w]}[w] & \text{w.p. } \pi_1[w], \\ \vdots & \\ 0 & \text{w.p. } 1 - \rho_\Sigma[w]. \end{cases}$$

All that remains is to compute the probabilities $\pi_k[w]$. For $k \geqslant 1$, each job's kth w-relevant segment starts a w-relevant busy period with expected length $\mathbb{E}(X_k[w])/(1 - \rho_0[w])$. Note that the possibility of a job completing before reaching its kth w-relevant segment is not a problem: this corresponds to the outcome $X_k[w] = 0$, in which case we think of the segment as starting a w-relevant busy period of length 0. Since jobs arrive at rate λ, we have for $k \geqslant 1$ that

$$\pi_k = \frac{\rho_k[w]}{1 - \rho_0[w]}.$$

The $k = 0$ case is similar, except that a job's 0th w-relevant segment only starts a w-relevant busy period if the system has no w-relevant work. Thus, the arrival rate of jobs whose 0th w-relevant segment starts a w-relevant busy period is $\lambda(1 - \rho_\Sigma[w])$, so

$$\pi_0 = \frac{\rho_0[w](1 - \rho_\Sigma[w])}{1 - \rho_0[w]}.$$

□

Appendix 2.D Proofs for Section 2.6

Lemma 2.6.1. *For all $p > 0$ satisfying $\Phi(p)$, in the $x \to \infty$ limit,*

$$\mathbb{E}(Q[w_x]^p) \leqslant \check{o}(x^p).$$

Proof. By Lemma 2.A.2,

$$\mathbb{E}(Q[w_x]^p) \leqslant \sum_{k=0}^{K[w_x]} \pi_k[w_x] \cdot \mathbb{E}(\mathcal{E}B_{X_k[w_x]}[w_x]^p), \tag{2.24}$$

where

$$\pi_0[w_x] = \frac{\rho_0[w_x](1-\rho_\Sigma[w_x])}{1-\rho_0[w_x]} = O(1)\cdot\rho_0[w_x],$$
$$\pi_k[w_x] = \frac{\rho_k[w_x]}{1-\rho_0[w_x]} = O(1)\cdot\rho_k[w_x] \quad \text{for all } k \geqslant 1.$$

We start by bounding each term of the sum in (2.24). Observe first that for any random variable V and any $p \geqslant 0$,

$$\mathbb{E}(\mathcal{E}V^p) = \frac{\mathbb{E}(V^{p+1})}{(p+1)\mathbb{E}(V)}.$$

Then for all $k \geqslant 0$, we compute

$$\begin{aligned}\pi_k[w_x]\cdot\mathbb{E}(\mathcal{E}B_{X_k[w_x]}[w_x]^p) &= O(1)\cdot\rho_k[w_x]\cdot\mathbb{E}(\mathcal{E}B_{X_k[w_x]}[w_x]^p)\\ &= O(1)\cdot\rho_k[w_x]\cdot\frac{\mathbb{E}(B_{X_k[w_x]}[w_x]^{p+1})}{\mathbb{E}(X_k[w_x])}\\ &= O(1)\cdot\mathbb{E}(B_{X_k[w_x]}[w_x]^{p+1}).\end{aligned} \tag{2.25}$$

Bounding the right-hand side of (2.25) requires bounding fractional busy period moments. We therefore apply Theorem 2.5.1 to the $(p+1)$th moment above, letting $p+1 = n-q$ for $n \in \mathbb{Z}_+$ and $q \in (0,1)$. We choose $\chi_{i0} = 1$ for all i such that $b_i \geqslant 2$ and $\chi_{i1} = 1$ for all other i. This choice requires checking that $J_i \geqslant 1$ for all i such that $b_i = 1$, which holds by Lemma 2.5.2 and the fact that $n \geqslant 2$. The choice ensures that

$$\begin{aligned} b_i - q\chi_{i0} &\geqslant 1\\ c_{ij} - q\chi_{ij} &> 1,\end{aligned}$$

which will allow the use of Lemmas 2.6.4 and 2.6.5 later in the proof.

Applying Theorem 2.5.1 to (2.25) yields, for $x \to \infty$,

$$\begin{aligned}\pi_k[w_x]\cdot\mathbb{E}(\mathcal{E}B_{X_k[w_x]}[w_x]^p) &\leqslant O(1)\cdot\sum_{i=1}^{I} d_i\frac{\mathbb{E}(X_k[w_x]^{b_i-q\chi_{i0}})}{(1-\rho_0[w_x])^{a_i-q}}\prod_{j=1}^{J_i}\lambda\mathbb{E}(X_0[w_x]^{c_{ij}-q\chi_{ij}})\\ &= O(1)\cdot\sum_{i=1}^{I}\mathbb{E}(X_k[w_x]^{b_i-q\chi_{i0}})\prod_{j=1}^{J_i}\mathbb{E}(X_0[w_x]^{c_{ij}-q\chi_{ij}}).\end{aligned} \tag{2.26}$$

Recall from Lemma 2.5.2 that

$$b_i - q\chi_{i0} + \sum_{j=1}^{J_i}(c_{ij} - q\chi_{ij} - 1) = n - q = p+1 \tag{2.27}$$

(note that we are applying Theorem 2.5.1 to a $(p+1)$th moment, not a pth moment). This means for all i and j, we have $c_{ij} - q\chi_{ij} - 1 \leqslant p$, so $\Psi(c_{ij} - q\chi_{ij} - 1)$ holds by (2.13). Returning

to (2.26), applying Lemma 2.6.4 and (2.27) gives us

$$\begin{aligned}
&\pi_k[w_x] \cdot \mathbb{E}(\mathcal{E}B_{X_k[w_x]}[w_x]^p) \\
&\quad \leqslant \sum_{i=1}^{I} \mathbb{E}(X_k[w_x]^{b_i - q\chi_{i0}}) \prod_{j=1}^{J_i} \check{o}(x^{c_{ij} - q\chi_{ij} - 1}) && \text{[by Lemma 2.6.4]} \\
&\quad = \sum_{i=1}^{I} \mathbb{E}(X_k[w_x]^{b_i - q\chi_{i0}}) \max\{O(1), \check{o}(x^{\sum_{j=1}^{J_i}(c_{ij} - q\chi_{ij} - 1)})\} \\
&\quad = \sum_{i=1}^{I} \mathbb{E}(X_k[w_x]^{b_i - q\chi_{i0}}) \max\{O(1), \check{o}(x^{p+1-(b_i - q\chi_{i0})})\}, && \text{[by (2.27)]} \qquad (2.28)
\end{aligned}$$

where the $O(1)$ covers the $J_i = 0$ case, in which the product is empty.

We now return to bounding the right-hand side of (2.24), substituting in (2.28) and interchanging the order of summation:

$$\mathbb{E}(Q[w_x]^p) \leqslant \sum_{i=1}^{I} \max\{O(1), \check{o}(x^{p+1-(b_i - q\chi_{i0})})\} \sum_{k=0}^{K[w_x]} \mathbb{E}(X_k[w_x]^{b_i - q\chi_{i0}}).$$

It suffices to show that each term of the outer sum is $\check{o}(x^p)$. We would like to use Lemmas 2.6.4 and 2.6.5. We know $\Phi(b_i - q\chi_{i0} - 1)$ holds by (2.13) and (2.27). However, the lemmas also require $b_i - q\chi_{i0} > 1$, yet it may be the case that $b_i - q\chi_{i0} = 1$. To handle this case, we use the fact that by Definition 2.A.3,

$$\sum_{k=0}^{K[w_x]} \mathbb{E}(X_k[w_x]) = \mathbb{E}\left(\sum_{k=0}^{K[w_x]} X_k[w_x]\right) \leqslant \mathbb{E}(X) = O(1).$$

Combining this with Lemmas 2.6.4 and 2.6.5 gives us

$$\begin{aligned}
&\max\{O(1), \check{o}(x^{p+1-(b_i - q\chi_{i0})})\} \sum_{k=0}^{K[w_x]} \mathbb{E}(X_k[w_x]^{b_i - q\chi_{i0}}) \\
&\quad \leqslant \max\{O(1), \check{o}(x^{p+1-(b_i - q\chi_{i0})})\} \cdot \max\{O(1), \check{o}(x^{b_i - q\chi_{i0} - 1})\} \\
&\quad = \check{o}(x^p).
\end{aligned}$$

□

Lemma 2.6.3. *If $\zeta < 1$ or $\eta < \infty$, then in the $x \to \infty$ limit,*

$$\mathbb{E}(R_x) \geqslant \frac{x}{1-\rho} - \check{o}(x).$$

Proof. We consider the $\zeta < 1$ and $\eta < \infty$ cases separately.

Case 1: $\zeta < 1$. Definitions 2.A.3 and 2.A.9 imply (see also Figure 2.1)

$$w_x(a) = w_x \quad \text{for all } a \in [0, y_x). \qquad (2.29)$$

From this we compute

$$\begin{aligned}
\mathbb{E}(R_x) &= \int_0^x \frac{1}{1-\rho_0[w_x(a)-]}\,\mathrm{d}a && \text{[by Corollary 2.A.1]}\\
&\geqslant \int_0^{y_x} \frac{1}{1-\rho_0[w_x(a)-]}\,\mathrm{d}a\\
&= \frac{y_x}{1-\rho_0[w_x-]} && \text{[by (2.29)]}\\
&= \frac{y_x}{1-\rho\langle y_x\rangle} && \text{[by (2.15)]}\\
&\geqslant \frac{x-O(x^\zeta)}{1-\rho\langle x-O(x^\zeta)\rangle} && \text{[by Assumption 2.3.1 and Lemma 2.A.6]}\\
&= \frac{x}{1-\rho\langle\Omega(x)\rangle} - \check{o}(x).
\end{aligned}$$

For any $\rho' \in (0,\rho)$, we have

$$\frac{1}{1-\rho'} = \frac{1}{1-\rho}\cdot\frac{1}{1+\frac{\rho-\rho'}{1-\rho}} \geqslant \frac{1}{1-\rho} - \frac{\rho-\rho'}{(1-\rho)^2}. \tag{2.30}$$

By (2.30) with $\rho' = \rho\langle\Omega(x)\rangle$, it suffices to show that $\rho - \rho\langle\Omega(x)\rangle = \check{o}(1)$. This indeed holds by Lemma 2.A.5 and Definition 2.2.1:

$$\begin{aligned}
\rho - \rho\langle\Omega(x)\rangle &= \lambda\int_0^\infty \overline{F}(t)\,\mathrm{d}t - \lambda\int_0^{\Omega(x)} \overline{F}(t)\,\mathrm{d}t && \text{[by Lemma 2.A.5]}\\
&= \int_{\Omega(x)}^\infty O(t^{-\alpha})\,\mathrm{d}t && \text{[by Definition 2.2.1]}\\
&= O(x^{-(\alpha-1)})\\
&= \check{o}(1).
\end{aligned}$$

Case 2: $\eta < \infty$. A job's worst future rank $w_x(a)$ is decreasing in a by Definition 2.A.1, so for all $a \in [0,x)$,

$$w_x(a) \geqslant w_x(x-) = r(x-).$$

Applying this to Corollary 2.A.1 yields

$$\mathbb{E}(R_x) = \int_0^x \frac{1}{1-\rho_0[w_x(a)-]}\,\mathrm{d}a \geqslant \frac{x}{1-\rho_0[r(x-)]}.$$

By (2.30) with $\rho' = \rho_0[r(x-)]$, it suffices to show $\rho - \rho_0[r(x-)] = \check{o}(1)$.

Let $f(\cdot)$ be a strictly increasing function such that for sufficiently large t,

$$u_{t+} \leqslant f(t) \leqslant 2u_{t+}. \tag{2.31}$$

Definition 2.A.3 tells us that for all ages $a > f(t)$, we have $r(a) > w_{t+}$. But by Definitions 2.A.1 and 2.A.3, we have $r(x-) \leqslant r(c_0[r(x-)]) = w_{c_0[r(x-)]+}$, so it must be that $x \leqslant f(c_0[r(x-)])$. Because $f(\cdot)$ is strictly increasing, it is invertible, so Assumption 2.3.1 and (2.31) imply

$$c_0[r(x-)] \geqslant f^{-1}(x) = \Omega(x^{1/\eta}).$$

Combining this with (2.15) and Lemma 2.A.5, we compute, similarly to the previous case,

$$\begin{aligned}
\rho - \rho_0[r(x-)] &= \rho - \rho\langle \Omega(x^{1/\eta})\rangle && \text{[by (2.15)]}\\
&= \int_{\Omega(x^{1/\eta})}^{\infty} O(t^{-\alpha})\,\mathrm{d}t && \text{[by Lemma 2.A.5]}\\
&= O(x^{-(\alpha-1)/\eta})\\
&= \check{o}(1).
\end{aligned}$$

□

Lemma 2.6.5. *For all $p > 0$ satisfying $\Phi(p)$, in the $x \to \infty$ limit,*

$$\sum_{k=1}^{K[w_x]} \mathbb{E}(X_k[w_x]^{p+1}) = \check{o}(x^p).$$

Proof. We compute

$$\begin{aligned}
\sum_{k=1}^{K[w_x]} \mathbb{E}(X_k[w_x]^{p+1}) &= \sum_{k=1}^{K[w_x]} \int_{b_k[w_x]}^{c_k[w_x]} (p+1)(t - b_k[w_x])^p \overline{F}(t)\,\mathrm{d}t && \text{[by Lemma 2.A.4]}\\
&\leqslant \sum_{k=1}^{K[w_x]} \int_{b_k[w_x]}^{c_k[w_x]} (p+1)(z_x - y_x)^p \overline{F}(t)\,\mathrm{d}t && \text{[by Lemma 2.A.6]}\\
&\leqslant \sum_{k=1}^{K[w_x]} \int_{c_{k-1}[w_x]}^{c_k[w_x]} (p+1)(z_x - y_x)^p \overline{F}(t)\,\mathrm{d}t && \text{[by Definition 2.A.3]}\\
&\leqslant \int_0^{u_x} (p+1)(z_x - y_x)^p \overline{F}(t)\,\mathrm{d}t. && \text{[by Definition 2.A.7]}
\end{aligned}$$

Hence for $x \to \infty$, Assumption 2.3.1 implies

$$\begin{aligned}
\sum_{k=1}^{K[w_x]} \mathbb{E}(X_k[w_x]^{p+1}) &\leqslant \int_0^{O(x^\eta)} O(x^{\zeta p} t^{-\alpha})\,\mathrm{d}t\\
&= O(x^{\max\{0,\zeta p - \eta(\alpha-1)\}}),
\end{aligned}$$

which $\Phi(p)$ implies is $\check{o}(x^p)$. □

Appendix 2.E Generalization to SOAP Bubble policies

In this section we generalize our main results, Theorems 2.3.1 and 2.3.2, to SOAP Bubble policies, which are a superset of SOAP policies that is introduced in [110]. To review (see Section 2.3.4), a SOAP Bubble policy has *lower and upper rank functions*

$$r^-, r^+ : \mathbb{R}_+ \to \mathbb{R}.$$

A SOAP Bubble policy works like a SOAP policy, except each job j can have a different rank function r_j. Each job's rank function may be set arbitrarily, provided it remains within the "bubble" between the lower and upper rank functions, meaning for all jobs j and ages a,

$$r^-(a) \leqslant r_j(a) \leqslant r^+(a).$$

One can view ordinary SOAP policies as the special case with $r^-(a) = r(a) = r^+(a)$.

To upper bound the response time of a SOAP bubble policy, one can essentially replicate the analysis of SOAP policies, but replacing each use of r with either r^- or r^+ as appropriate. The intuition is that a tagged job has maximal response time if it follows r^+ while every other job follows r^-. Specifically,

- when defining worst future rank (Definition 2.A.1), replace r with r^+; and
- when defining w-relevant work, intervals, segments, and load (Definitions 2.A.2 and 2.A.3), replace r with r^-.

For details, see [110].

To generalize Theorems 2.3.1 and 2.3.2, we begin by defining some new notation. Recalling that $c_i[w]$ is defined using r^-, let

$$\begin{aligned}
w_x^- &= \sup_{0 \leq b < x} r^-(b), \\
w_x^+ &= \sup_{0 \leq b < x} r^+(b), \\
y_x^- &= c_0[w_x^- -], \\
z_x^- &= c_0[w_x^-], \\
y_x^+ &= c_0[w_x^+ -], \\
z_x^+ &= c_0[w_x^+], \\
u_x^+ &= c_{K[w_x^+]}[w_x^+].
\end{aligned}$$

Throughout our proofs, we can simply replace u_x with u_x^+, but y_x and z_x are more subtle.

- The main use of y_x and z_x is through Lemma 2.A.6, which is used in Lemma 2.6.5. The lemma statement now holds with y_x^- and z_x^-.
- There is one more use of y_x in Lemma 2.6.3, and this one needs to be replaced with y_x^+.
- There is one more use of z_x in Lemma 2.6.4, and this one needs to be replaced with z_x^+.

This implies the following generalizations of Assumption 2.3.1 and Theorem 2.3.2. The only substantial change is that we need two versions of ζ because there are two version of y_x and z_x. This ends up breaking the $\Phi(p) \Rightarrow \Psi(p)$ implication in (2.13), so we add some extra preconditions to our result.

Assumption 2.E.1.

(i) There exists $\zeta^- \in [0, \infty]$ such that $z_x^- - y_x^- = O(x^{\zeta^-})$.

(ii) There exists $\zeta^+ \in [\zeta^-, \infty]$ such that $z_x^+ - y_x^+ = O(x^{\zeta^+})$.

(iii) There exists $\eta^+ \in [\max\{1, \zeta^+\}, \infty]$ such that $u_x = O(x^{\eta^+})$.

Theorem 2.E.1. *Consider an $M/G/1$ queue whose job size distribution is HT and a SOAP*

Bubble scheduling policy whose lower and upper rank functions obey Assumption 2.E.1. If

$$\zeta^- - \frac{1}{\eta^+} < \frac{\alpha - 1}{\beta},$$
$$1 - \frac{1}{\zeta^+} < \frac{\alpha - 1}{\beta},$$
$$\text{and either } \zeta^+ < 1 \text{ or } \eta^+ < \infty,$$

then the policy is tail-optimal, i.e., $\lim_{x\to\infty} \frac{1}{\overline{F}(x)} \mathbb{P}\big(T > \frac{x}{1-\rho}\big) = 1$.

One can obtain the following simplified condition in much the same way as done in Theorem 2.3.1.

Theorem 2.E.2. *Consider an $M/G/1$ queue whose job size distribution is HT using a SOAP Bubble scheduling policy whose lower and upper rank functions obey*

$$r^-(a) = \Omega(a^\gamma),$$
$$r^+(a) = O(a^\delta)$$

for some $\delta > \gamma > 0$. If

$$\frac{\delta}{\gamma} - \frac{\gamma}{\delta} < \frac{\alpha - 1}{\beta},$$

then the policy is tail-optimal, i.e., $\lim_{x\to\infty} \frac{1}{\overline{F}(x)} \mathbb{P}\big(T > \frac{x}{1-\rho}\big) = 1$.

3 Extension for heavy-tailed job sizes, light-tailed job sizes, and the Gittins policy

3.1 Introduction

In this chapter we stay within the context of scheduling in the $M/G/1$ queue and its effect on the response time T. In general, a queueing system will have a response time *distribution*, and there are a variety of metrics one might hope to minimize. There is significant work on minimizing *mean response time* $\mathbb{E}(T)$, which is the average response time of all jobs in a long arrival sequence [2, 52, 53, 105].

Much less is known about minimizing the *tail of response time* $\mathbb{P}(T > t)$, which is the probability a job has response time greater than a parameter $t \geqslant 0$. In light of the difficulty of studying the tail directly, theorists have studied the *asymptotic tail of response time*, which is the asymptotic decay of $\mathbb{P}(T > t)$ in the $t \to \infty$ limit [20, 29, 95, 120]. The content of Chapter 2 also falls within this category, particularly focusing on heavy-tailed job sizes. In the current chapter, we consider both heavy and light tails, and give special attention to the Gittins policy (Definition 3.1.1).

Given the importance of both the mean and asymptotic tail of the response time, we study the following question.

Question 3.1. *Does any scheduling policy* simultaneously *optimize the mean and asymptotic tail of response time?*

For a wide class of job size distributions, prior work answers Question 3.1 when job sizes are known to the scheduler. In this setting, the *Shortest Remaining Processing Time* (SRPT) policy, which preemptively serves the job of least remaining size, always minimizes mean response time [105]. However, SRPT's tail performance depends on the (known) job size distribution.

- If the job size distribution is *heavy-tailed* (Definition 2.2.1), then SRPT is *tail-optimal*, meaning it has the best possible asymptotic tail decay (Definition 2.2.2).
- If the job size distribution is *light-tailed* (Definition 3.3.2), then SRPT is *tail-pessimal*, meaning it has the worst possible asymptotic tail decay (Definition 3.3.3).

When the job size distribution is in one of these two classes, this answers Question 3.1 for known job sizes: "yes, namely SRPT" in the heavy-tailed case, "no" in the light-tailed case.

Unfortunately, in practice, the scheduler often does not know job sizes, and thus one cannot implement SRPT. Instead, the scheduler often only knows the job size *distribution*. We study Question 3.1 in this unknown-size setting.

The question of minimizing mean response time with unknown job sizes was settled by Gittins [52]. He introduced a policy, now known as the *Gittins* policy, which leverages the job size distribution to minimize mean response time. Roughly speaking, Gittins uses each job's *age*, namely the amount of time each job has been served so far, to figure out which job is most likely to complete after a small amount of service, then serves that job. For some job size distributions, Gittins reduces to a simpler policy, such as *First-Come, First-Served* (FCFS) or *Foreground-Background* (FB) [2, 3].

In the unknown-size setting, given that Gittins minimizes mean response time, Question 3.1 reduces to the following.

Question 3.2. *For which job size distributions is Gittins tail-optimal for response time?*

Unfortunately, the asymptotic tail behavior of Gittins is understood in only a few special cases.

- In the heavy-tailed case, Corollary 2.3.2 states that Gittins is tail-optimal, but only under an assumption on the job size distribution's hazard rate.
- In the light-tailed case, Gittins sometimes reduces to FCFS or FB [2, 3]. For light-tailed job sizes, FCFS is tail-optimal [29, 120], but FB is tail-pessimal [80].

This prior work leaves Question 3.2 largely open. We do not know whether Gittins is always tail-optimal in the heavy-tailed case, or whether it is sometimes suboptimal, or even tail-pessimal. Moreover, we do not understand Gittins's asymptotic tail at all in the light-tailed case, aside from when Gittins happens to reduce to a simpler policy.

The prior work above does tell us an important fact: Gittins *can* be tail-pessimal. This prompts another question.

Question 3.3. *For job size distributions for which Gittins is tail-pessimal, is there another policy that has* near-optimal *(arbitrarily close to optimal) mean response time while not being tail-pessimal?*

In this chapter, we answer Questions 3.1–3.3 for the $M/G/1$ queue with unknown job sizes, covering wide classes of heavy- and light-tailed job size distributions. The key tool we use to analyze Gittins's asymptotic response time tail is the *SOAP* framework (see the introduction of Part I). SOAP gives a universal $M/G/1$ response time analysis of all *SOAP policies*, which are scheduling policies where a job's priority level is a function of its age (Definition I.1). Underlying our Gittins results is a general tail analysis of SOAP policies.

3.1.1 Prior work

Before describing our results explicitly, we place our work in context with existing literature.

Asymptotic tail analysis of classic scheduling policies

Due to their frequent occurrence, light-tailed job size distributions have received a great amount of attention by queueing theorists. The performance of policies under light-tailed job sizes is generally measured in terms of the decay rate of the response time tail. In this sense FCFS has proven to be optimal among all service policies [120]. Conversely, *Foreground-Background* has the worst possible decay rate of the response time tail [80].

On the other hand, it is shown that heavy-tailed job sizes can have a large impact on the performance characteristics of the queue. For this reason also heavy-tailed job sizes have been thoroughly investigated in the literature. Remarkably, and contrary to the light-tailed case, FB is optimal in the heavy-tailed case whereas FCFS has the worst possible response time tail [20]. This dichotomy between light and heavy tails is not limited to FCFS and FB [29].

Other noteworthy literature highlighting both light and heavy tails includes delicate asymptotic results for a two-class priority policy [4] and robust optimization using a limited PS policy [87].

With one exception, discussed below, the literature on this subject has in common that only a few, relatively simple, policies are considered. This chapter considers policies in which the priority of a job can vary essentially arbitrarily with its age. This generality is needed to analyze the Gittins policy, which can be non-monotonic [3].

Tail optimality of certain SOAP policies in the heavy-tailed case

We mention particularly the relation between the current chapter and the previous chapter. Both chapters study the response time tail behavior of arbitrary SOAP policies, including the Gittins policy. There are two main factors that distinguish this chapter from Chapter 2.

- Chapter 2 only concerns heavy-tailed job size distributions. In contrast, this chapter studies both the heavy- and light-tailed cases.

- In Chapter 2 it is shown that Gittins is tail-optimal subject to a condition on the job size distribution's hazard rate (Corollary 2.3.2). However, the analysis is not sharp enough to completely characterize under which (heavy-tailed) job size distributions Gittins is tail-optimal. In contrast, the analysis in this chapter is sharper, allowing us to identify Gittins's tail performance under any job size distribution.

With this said, Chapter 2 lays an important technical foundation that we build upon to derive our heavy-tailed results. See Remark 3.2.1 for a more technical discussion of what aspects of Chapter 2 we use and what aspects are new in this Chapter.

Beyond asymptotic optimality

It is well known that FCFS has optimal tail decay rate under light-tailed job sizes. However, decay rate is a relatively crude tail performance measure, as it does not take into account the constant (or non-exponential term) in front of the exponent. Although this chapter focuses just on decay rates, we mention that very recently a policy was introduced that has a better

leading constant than FCFS [54]. An open question remains what is the best possible leading constant in the response time tail. A by-product of our results, namely that FCFS is the only SOAP policy with optimal decay rate, partially answers this question. Specifically, it follows that no SOAP policy is tail-optimal up to the leading constant.

Mean response time of modified Gittins policies

A recent study [109, Theorem 7.2] shows that if one slightly modifies the prioritization rules of SRPT, then the mean response time of the resulting policy is only slightly worse than that of unmodified SRPT (which is optimal in case job sizes are known). It turns out, as shown in this chapter, that a similar result holds for an approximate version of the Gittins policy, and that result can thus be seen as the unknown-job-sizes counterpart of [109, Theorem 7.2].

3.1.2 Contributions

Our main contributions, which we describe in more detail later (Section 3.2 for heavy tails and Section 3.3 for light tails), are as follows:

- *Heavy-tailed case:* We give a sufficient condition under which an arbitrary SOAP policy is tail-optimal (Section 3.4).
- *Heavy-tailed case:* We show that the above condition always applies to Gittins, implying it is always tail-optimal (Section 3.5).
- *Light-tailed case:* We characterize when an arbitrary SOAP policy is tail-optimal, tail-pessimal, or in between (Section 3.6).
- *Light-tailed case:* We spell out how the above characterization applies to Gittins and show how to modify Gittins to avoid tail pessimality (Section 3.7).
- *General case:* At the core of our modification of Gittins which avoids tail pessimality is a general result which states that slightly perturbing the Gittins rank function only slightly affects its mean response time (Theorem 3.3.3 and Appendix 3.A).

The rest of the chapter introduces definitions and notation (Section 3.1.3), and concludes with some remarks about our motivating questions (Section 3.8).

3.1.3 Model description and the Gittins policy

We consider an $M/G/1$ queue with arrival rate λ, job sizes distributed as X, and load $\rho = \lambda\mathbb{E}(X)$. For the tail of the job size distribution, we write $\overline{F}(t) = \mathbb{P}(X > t)$. We denote the maximum job size by $x_{\max} = \inf\{t \geqslant 0 \mid \overline{F}(t) = 0\}$, allowing $x_{\max} = \infty$. We assume that the scheduling policy is a SOAP policy (Definition I.1), and we write T_π for the response time under π. Additionally, we denote by $r_\pi : [0, x_{\max}) \to \mathbb{R}$ the rank function of the SOAP policy π. When the policy being discussed is clear from context, we often omit the subscript and

simply write $r(a)$. Recall that at every moment in time, a SOAP policy serves the job of minimum rank, breaking ties in FCFS order.

In this chapter, special attention is given to the Gittins policy. It assigns each job a rank based on the job's age, so it naturally falls within the SOAP framework.

Definition 3.1.1. *The* Gittins *policy, denoted "*Gtn*" in subscripts for brevity, is the SOAP policy with rank function*

$$r_{\mathrm{Gtn}}(a) = \inf_{b>a} \frac{\int_a^b \overline{F}(t)\,\mathrm{d}t}{\overline{F}(a) - \overline{F}(b)}.$$

Note that the Gittins rank function depends on the job size distribution by way of $\overline{F}$*.*

As Definition I.1 suggests, we consider rank functions that are piecewise-continuous and piecewise-monotonic. This holds for Gittins under very mild conditions on the job size distribution. For example, Aalto et al. [3] show that it holds when the hazard rate of X is continuous and piecewise-monotonic.

3.2 Heavy tails

The two main results in the heavy-tailed case are presented in Section 3.2.1:

- Theorem 3.2.1 gives a sufficient condition under which a SOAP policy is tail-optimal for heavy-tailed job sizes.
- Theorem 3.2.2 shows that for heavy-tailed job sizes, Gittins always satisfies this sufficient condition, and is thus always tail-optimal.

The class of heavy-tailed job sizes we consider coincides with that of Chapter 2. That is, we consider the class of HT distributions (Definition 2.2.1). Similarly, we work with the same tail-optimality criterion as in the previous chapter (Definition 2.2.2).

3.2.1 Results for the heavy-tailed case

Let us focus on a tagged job of size x. For determining whether or not it will be delayed by other jobs, it is important to know the worst (highest) rank that it will ever have. To that end, recall from Definition 2.A.1 that the *worst ever rank* of a job of size x is defined by $w_x = \sup_{0 \leqslant a < x} r(a)$. A second object of interest regarding the delay of the tagged job is the set of ages at which other jobs will have rank lower than w_x.

Definition 3.2.1. *A* w-interval *is an interval* (b, c) *with* $0 \leqslant b < c \leqslant x_{\max}$ *such that* $r(a) \leqslant w$ *for all* $a \in (b, c)$*.*

Note that the tagged job of size x always has priority over jobs having rank higher than w_x. Therefore, it can only be delayed by another job if that other job's age is in a w_x-interval. See Figure 3.1 for an illustration. To ensure that the tagged job does not wait too long behind other

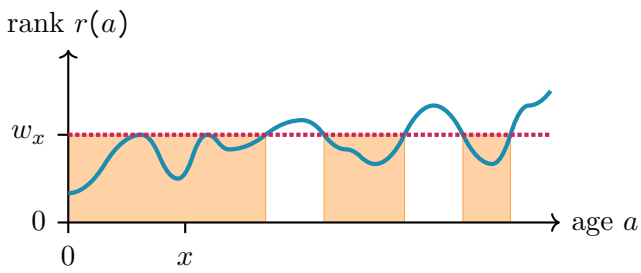

Figure 3.1: Illustration of worst ever rank w_x (purple dotted line) and w_x-intervals (orange regions) for a SOAP policy given by rank function r (cyan curve). Additionally, any sub-interval of a w_x-interval is also a w_x-interval.

jobs, the w_x-intervals must be relatively short. We use the following condition to characterize the length of w_x-intervals.

Condition 3.2.1. *There exist $\zeta, \theta \in [0, \infty)$ and $\eta \in [\max\{1, \zeta+\theta\}, \infty]$ such that the following hold for any w_x-interval (b, c):*

(i) If $b \geqslant x$, then $c - b \leqslant O(b^{\zeta} x^{\theta})$.

(ii) $c \leqslant O(x^{\eta})$.

Here, in (i), the multivariable $O(\cdot)$ notation is defined as follows. Suppose $x_1, \ldots, x_n$ are non-negative variables. The notation $O(f(x_1, \ldots, x_n))$ stands for an unspecified expression $g(x_1, \ldots, x_n) \geqslant 0$ for which there exist constants $C, y_0, \ldots, y_n \geqslant 0$ such that for all $x_1 \geqslant y_1, \ldots, x_n \geqslant y_n$, we have $g(x_1, \ldots, x_n) \leqslant C f(x_1, \ldots, x_n)$. The multivariable $\Omega(\cdot)$ notation is defined analogously, with the inequality reversed.

At an intuitive level, we can think of Condition 3.2.1 as saying the following. Condition (i) refers to the maximum amount of time that jobs older than x can delay the tagged job of size x:

- If θ is small, then whenever the rank function dips below w_x, it does so for only a short time.
- If $\zeta + \theta$ is small, then adjacent "peaks" of the rank function are not too far apart. Here, a "peak" is an age a that is both a local maximum and a "running maximum", meaning $r(a) > r(b)$ for all $b \in [0, a)$.

Condition (ii) refers to the threshold age after which a job can no longer delay the tagged job:

- If η is small, then not too long after age x, the rank function never dips below w_x again.

Thus, the smaller the parameters θ, ζ and η, the shorter the tagged job of size x has to wait for other jobs. Our main theorem in the heavy-tailed setting, stated below, characterizes how small these parameters should be in order to achieve tail optimality.

Theorem 3.2.1. *Consider an $M/G/1$ queue with any HT job size distribution under a SOAP policy. Condition 3.2.1 implies the policy is tail-optimal if*

$$\zeta + (\theta - 1)^+ - \frac{(1-\theta)^+}{\eta} < \frac{\alpha - 1}{\beta}. \tag{3.1}$$

We can apply Theorem 3.2.1 to show that Gittins is tail-optimal. Specifically, we will show that Gittins satisfies Condition 3.2.1 with $\zeta = 0$, $\theta = 1$, and $\eta = \infty$. As such, it achieves a value of 0 on the left-hand side of (3.1), so Gittins is tail-optimal regardless of the values of α and β.

Theorem 3.2.2. *The Gittins policy is tail-optimal for any HT job size distribution.*

Remark 3.2.1 (Comparison to Chapter 2). Having formally stated our sufficient condition for tail optimality in the heavy-tailed case, we may now compare it in more detail to that of Chapter 2. Assumption 2.3.1 and the corresponding result Theorem 2.3.2 are the same as this Chapter's Condition 3.2.1 and Theorem 3.2.1, respectively, but restricted to the $\theta = 0$ case. This means, roughly speaking, that in Chapter 2 only the lengths between "peaks" of the rank function are considered, rather than the lengths of w_x-intervals.

Unfortunately, looking only at distances between peaks of the rank function is not enough to prove Gittins is always tail-optimal in the heavy-tailed case, because Gittins's peaks can be too far apart. Specifically, to make Gittins satisfy Condition 3.2.1 with $\theta = 0$, one must in general set $\zeta = 1$, which turns out to be too large. However, it turns out that even though Gittins's peaks can be far apart, they have "gentle slopes", so w_x-intervals starting at very large ages are not asymptotically larger than w_x-intervals starting at smaller ages. Condition 3.2.1 is sharp enough to capture this by setting $\zeta = 0$ and $\theta = 1$.

Underlying the tail-optimality results of Chapter 2 is a busy period analysis combined with asymptotic response time bounds. We make use of this busy period analysis, which we distill into a simple statement (Theorem 3.4.1), but we replace the asymptotic response time bounds of Chapter 2 with a sharper analysis that accounts for the $\theta > 0$ possibility (Sections 3.4.1 and 3.4.2). ◇

3.3 Light tails

Similarly to the previous section, we first define the class of light-tailed distributions and state the corresponding tail-optimality criterion in Section 3.3.1. The main results in the light-tailed case, presented in Section 3.3.2, are summarized as follows:

- Theorem 3.3.1 classifies SOAP policies into tail-optimal, tail-intermediate, and tail-pessimal for light-tailed job sizes.
- Theorem 3.3.2 shows that for light-tailed job sizes, Gittins can be any of tail-optimal, tail-intermediate, or tail-pessimal.

- Theorem 3.3.3 shows that making a small change to the Gittins rank function results in only a small change to mean response time.
- Theorem 3.3.4 shows that for a wide class of light-tailed job size distributions for which Gittins is tail-pessimal, making a small change to Gittins's rank function results in a tail-optimal or -intermediate policy with mean response time arbitrarily close to Gittins's.

3.3.1 Background on light-tailed job sizes

Definition 3.3.1. *The* decay rate *of random variable V, denoted $d(V)$, is*

$$d(V) := \lim_{t\to\infty} \frac{-\log \mathbb{P}(V > t)}{t}.$$

That is, if the decay rate $d(V)$ is finite, then $\mathbb{P}(V > t) = \exp(-d(V)t \pm o(t))$. Higher decay rates thus correspond to asymptotically lighter tails.

Roughly speaking, the light-tailed job size distributions we study are those with positive decay rate. Our main tool for investigating the decay rate of a random variable V is via its Laplace-Stieltjes transform (LST),

$$\tilde{V}(s) := \mathbb{E}(\exp(-sV)).$$

Under mild conditions on V [82, 89, 90], we can determine its decay rate in terms of the convergence of its LST:

$$d(V) = -\sup\{s \leqslant 0 \mid \tilde{V}(s) < \infty\}. \tag{3.2}$$

The specific class of light-tailed job size distributions we consider, described below, are those which allow us to use (3.2) throughout this work (Appendix 3.B). The class includes essentially all light-tailed distributions of practical interest such as finite-support, phase-type, and Gaussian-tailed distributions. In the terminology of Abate and Whitt [4], we consider all "Class I" distributions.

Definition 3.3.2 (Light-Tailed Job Size Distribution)**.** *Given a job size X, let*

$$s^* = \inf\{s \leqslant 0 \mid \tilde{X}(s) < \infty\}.$$

We say that the distribution of X is LT *if $s^* = -\infty$ or $s^* \in (-\infty, 0)$ and $\tilde{X}(s^*) = \infty$.*

Remark 3.3.1. Our results can be generalized to some "Class II" distributions [4], which are also light-tailed. We comment on this in Appendix 3.B.3. However, working with Class II distributions generally requires additional regularity or smoothness assumptions [4, Section 5], so for simplicity of presentation, we focus on Class I distributions. ◇

Definition 3.3.3 (Tail Optimality in Light-Tailed Case)**.** *Consider an $M/G/1$ queue with an LT job size distribution. We say a scheduling policy π is*

- log-tail-optimal *if π maximizes $d(T_\pi)$,*

- log-tail-pessimal *if* π *minimizes* $d(T_\pi)$*, and*
- log-tail-intermediate *otherwise.*

In each case, we mean minimizing or maximizing over preemptive work-conserving policies. In informal discussion, we omit "log-".

3.3.2 Results for the light-tailed case

We have seen that a job's worst ever rank plays an important role in the heavy-tailed setting. When the job size distribution is LT, we are interested in the age at which the rank function's *global* maximum occurs.

Definition 3.3.4. *The* worst age*, denoted* a^**, is the earliest age at which a job has the global maximum rank:*

$$a^* = \inf\{a \in [0, x_{\max}) \mid \forall b \in [0, x_{\max}), r(a) \geqslant r(b)\}.$$

If the rank function has no maximum, we define $a^* = x_{\max}$*.*

As an example, FCFS has $a^* = 0$, because a job's priority is the lowest before it starts service. In contrast, FB has $a^* = x_{\max}$, because a job's priority gets strictly lower with age.

We already know that FCFS and FB are tail-optimal and tail-pessimal, respectively. The theorem below fills in the gaps for all other SOAP policies, showing that the performance of the response time tail is completely determined by the worst age a^*.

Theorem 3.3.1. *Consider an* $M/G/1$ *queue with any LT job size distribution under a SOAP policy. Let* $x_{\max} = \inf\{x \geqslant 0 \mid \mathbb{P}(X > x) = 0\}$*. The policy is*

- *log-tail-optimal if* $a^* = 0$*,*
- *log-tail-intermediate if* $0 < a^* < x_{\max}$*, and*
- *log-tail-pessimal if* $a^* = x_{\max}$*.*

To apply Theorem 3.3.1 to the Gittins policy, we need to characterize how the job size distribution affects Gittins's worst age a^*.

Definition 3.3.5. *We define two classes of distributions:* NBUE *and* ENBUE*.*

- *We say* X *is in the* New Better than Used in Expectation *(see for instance [121]), writing* $X \in \mathsf{NBUE}$*, if for all ages* $a \in [0, x_{\max})$*,*

$$\mathbb{E}(X) \geqslant \mathbb{E}(X - a \mid X > a).$$

- *We say* X *is* Eventually New Better than Used in Expectation*, writing* $X \in \mathsf{ENBUE}$*, if there exists* $a_0 \geqslant 0$ *such* $(X - a_0 \mid X > a_0) \in \mathsf{NBUE}$*. That is,* $X \in \mathsf{ENBUE}$ *if there exists* $a_0 \geqslant 0$ *such that for all* $a \geqslant a_0$*,*

$$\mathbb{E}(X - a_0 \mid X > a_0) \geqslant \mathbb{E}(X - a \mid X > a).$$

Results of Aalto et al. [2, 3] connect the classes NBUE and ENBUE to Gittins's worst age a^*, implying the following characterization.

Theorem 3.3.2. *Consider an $M/G/1$ queue with any LT job size distribution. Gittins is*

- *log-tail-optimal if $X \in \mathsf{NBUE}$,*
- *log-tail-intermediate if $X \in \mathsf{ENBUE} \setminus \mathsf{NBUE}$, and*
- *log-tail-pessimal if $X \notin \mathsf{ENBUE}$.*

The fact that Gittins can be log-tail-pessimal is intriguing, considering that it is optimal for mean response time, and tail-optimal under HT job sizes. Fortunately, in most cases where Gittins is log-tail-pessimal, slightly tweaking Gittins yields a log-tail-intermediate policy without sacrificing much mean response time performance.

Definition 3.3.6. *A SOAP policy π is a q-*approximate Gittins policy *if there exists a constant $m > 0$ such that for all ages $a \in [0, x_{\max})$,*

$$\frac{r_\pi(a)}{r_{\mathrm{Gtn}}(a)} \in [m, mq].$$

We may assume without loss of generality that $m = 1$, because the policy π' with rank function $r_{\pi'}(a) = r_\pi(a)/m$ has identical behavior to policy π.

Theorem 3.3.3. *Consider an $M/G/1$ queue with any job size distribution. For any $q \geqslant 1$ and any q-approximate Gittins policy π,*

$$\mathbb{E}(T_\pi) \leqslant q\mathbb{E}(T_{\mathrm{Gtn}}).$$

An important observation is that a q-approximate Gittins policy has near-optimal mean response time for q approaching one. At the same time, changing the Gittins rank function even within a small factor $q \geqslant 1$ can decrease the worst age, and therefore improve the tail performance.

Theorem 3.3.4. *Consider an $M/G/1$ queue with an LT job size distribution $X \notin \mathsf{ENBUE}$. Suppose that the expected remaining size of a job at all ages is uniformly bounded, meaning*

$$\sup_{a \in [0, x_{\max})} \mathbb{E}(X - a \mid X > a) < \infty.$$

Then for all $\varepsilon > 0$, there exists a $(1+\varepsilon)$-approximate Gittins policy that is log-tail-optimal or log-tail-intermediate.

The remainder of this chapter is organized as follows. We prove our results for heavy tails, Theorems 3.2.1 and 3.2.2, respectively in Sections 3.4 and 3.5. Similarly, proofs of our results for the light-tailed case are given in Sections 3.6 (Theorem 3.3.1) and 3.7 (Theorems 3.3.2 and 3.3.4). The remaining main result, Theorem 3.3.3, requires substantially more technical machinery for its proof, which is why we defer it to Appendix 3.A. Finally, Section 3.8 describes how our results answer the questions posed in Section 3.1.

Throughout the proofs, we regularly use the definitions and lemmas of Appendix 2.A related to SOAP.

3.4 Heavy-tailed job sizes: tail asymptotics of SOAP policies

In this section we prove Theorem 3.2.1. Throughout this section, we use a "polynomially strict" version of little-o notation, which we write as $\check{o}(\cdot)$. Given $p > 0$, the notation $\check{o}(x^p)$ stands for $O(x^{p-\varepsilon})$ for some unspecified $\varepsilon > 0$.

Our proof works by building upon intermediate results of Chapter 2. Specifically, since it was shown that Lemmas 2.6.3 and 2.6.5 together imply tail optimality, we can formulate these statements as conditions here.

Condition 3.4.1. *There exists $q > \beta$ such that for all $p \in (0, q]$,*

$$\sum_{k=0}^{K[w_x]} \mathbb{E}(X_k[w_x]^{p+1}) \leqslant \check{o}(x^p).$$

Condition 3.4.2.

$$\int_0^x \frac{1}{1 - \lambda \mathbb{E}(X_0[w_x(a)-])} \, \mathrm{d}a \geqslant \frac{x}{1-\rho} - \check{o}(x).$$

Theorem 3.4.1 (Chapter 2)**.** *Consider an $M/G/1$ queue with an HT job size distribution under a SOAP policy. Conditions 3.4.1 and 3.4.2 together imply tail optimality.*

The proof of Theorem 3.2.1 thus amounts to showing Conditions 3.4.1 and 3.4.2 hold under (3.1). We do so by way of the following two lemmas, both of which are novel.

Lemma 3.4.1. *Condition 3.2.1 implies Condition 3.4.1 if* (3.1) *holds.*

Lemma 3.4.2. *Condition 3.2.1 implies Condition 3.4.2 if $\zeta < 1$ or $\eta < \infty$.*

Proof of Theorem 3.2.1. It suffices to show that (3.1) implies $\zeta < 1$ or $\eta < \infty$. This is true because if $\eta = \infty$, then (3.1) implies $\zeta < (\alpha - 1)/\beta < 1$. □

The remainder of this section is devoted to proving these two lemmas. Section 3.4.1 proves Lemma 3.4.1, and Section 3.4.2 proves Lemma 3.4.2.

3.4.1 Upper bound

Definition 3.4.1. *Suppose Condition 3.2.1 holds. We say $p \geqslant 0$ is* good *if*

$$\zeta + (\theta - 1)^+ - \frac{(1-\theta)^+}{\eta} < \frac{\alpha - 1}{p}.$$

Proving Lemma 3.4.1 amounts to proving the following two lemmas.

Lemma 3.4.3. *Suppose Condition 3.2.1 holds. For all $p \geqslant 0$, if p is good, then*

$$\mathbb{E}(X_0[w_x]^{p+1}) \leqslant \check{o}(x^p).$$

Lemma 3.4.4. *Suppose Condition 3.2.1 holds. For all $p \geqslant 0$, if p is good, then*

$$\sum_{k=1}^{K[w_x]} \mathbb{E}(X_k[w_x]^{p+1}) \leqslant \check{o}(x^p).$$

Proof of Lemma 3.4.1. Our goal is to show that if β is good, then Condition 3.4.1 holds. Because Definition 3.4.1 can be written as a strict upper bound on p, if β is good, then there exists $q > \beta$ such that all $p \in (0, q]$ are good. The result therefore follows from Lemmas 3.4.3 and 3.4.4. □

We devote the rest of this section to proving Lemmas 3.4.3 and 3.4.4. The first step of each proof is computing $(p+1)$th moments of w_x-relevant job segments, which we do in Lemmas 3.4.5 and 3.4.6, respectively.

Lemma 3.4.5. *Suppose Condition 3.2.1 holds. For all $p \geqslant 0$,*

$$\mathbb{E}(X_0[w_x]^{p+1}) \leqslant \begin{cases} O(1) & \text{if } p < \alpha - 1 \\ O(\log x) & \text{if } p = \alpha - 1 \\ O(x^{\max\{1,\zeta+\theta\}(p-\alpha+1)}) & \text{if } p > \alpha - 1. \end{cases}$$

Proof. Recall from Definition 2.A.9 that $z_x = c_0[w_x]$. Because (x, z_x) is a w_x-interval, Condition 3.2.1 implies

$$z_x - x = O(x^{\zeta+\theta}). \tag{3.3}$$

We compute

$$\begin{aligned} \mathbb{E}(X_0[w_x]^{p+1}) &= \int_0^{z_x} (p+1)t^p \overline{F}(t)\,\mathrm{d}t && \text{[by Lemma 2.A.4 and Definition 2.A.9]} \\ &\leqslant \int_0^{O(x^{\max\{1,\zeta+\theta\}})} O(t^{p-\alpha})\,\mathrm{d}t && \text{[by Lemma 2.2.1 and (3.3)]} \\ &= \begin{cases} O(1) & \text{if } p < \alpha - 1 \\ O(\log x) & \text{if } p = \alpha - 1 \\ O(x^{\max\{1,\zeta+\theta\}(p-\alpha+1)}) & \text{if } p > \alpha - 1. \end{cases} \end{aligned}$$

□

Proof of Lemma 3.4.3. We apply Lemma 3.4.5, splitting into cases depending on how p compares to $\alpha - 1$.

Case 1 ($p \leqslant \alpha - 1$)**.** *By Lemma 3.4.5, we have* $\mathbb{E}(X_0[w_x]^{p+1}) \leqslant O(\log x) \leqslant \check{o}(x^p)$.

Case 2 ($p > \alpha - 1$)**.** *By Lemma 3.4.5, it suffices to show*

$$\max\{1, \zeta + \theta\}(p - \alpha + 1) < p.$$

Dividing by $p \max\{1, \zeta + \theta\}$ *gives*

$$1 - \frac{1}{\max\{1, \zeta + \theta\}} < \frac{\alpha - 1}{p}.$$

If $\zeta + \theta \leqslant 1$, then this holds because $\alpha > 1$. If instead $\zeta + \theta > 1$, then by the standard equality $\zeta + \theta - 1 = \zeta + (\theta - 1)^+ - (1 - \theta)^+$, it suffices to show

$$\frac{\zeta + (\theta - 1)^+}{\zeta + \theta} - \frac{(1 - \theta)^+}{\zeta + \theta} < \frac{\alpha - 1}{p}.$$

Because $1 \leqslant \zeta + \theta \leqslant \eta$, this holds by Definition 3.4.1. □

Lemma 3.4.6. *Suppose Condition 3.2.1 holds. For all $p \geqslant 0$,*

$$\sum_{k=1}^{K[w_x]} \mathbb{E}(X_k[w_x]^{p+1}) \leqslant \begin{cases} O(x^{\theta p + \zeta p - \alpha + 1}) & \text{if } \zeta p < \alpha - 1 \\ O(x^{\theta p} \log x^{\eta}) & \text{if } \zeta p = \alpha - 1 \\ O(x^{\theta p + \eta(\zeta p - \alpha + 1)}) & \text{if } \zeta p > \alpha - 1. \end{cases}$$

Proof. Note that Definitions 2.A.1 and 2.A.3 together imply

$$b_k[w_x] \geqslant x \quad \text{for all } k \geqslant 1. \tag{3.4}$$

We compute

$$\begin{aligned}
&\sum_{k=1}^{K[w_x]} \mathbb{E}(X_k[w_x]^{p+1}) \\
&\quad = \sum_{k=1}^{K[w_x]} \int_{b_k[w_x]}^{c_k[w_x]} (p+1)(t - b_k[w_x])^p \overline{F}(t)\,\mathrm{d}t && \text{[by Lemma 2.A.4]} \\
&\quad \leqslant \sum_{k=1}^{K[w_x]} \int_{b_k[w_x]}^{c_k[w_x]} (p+1)(c_k[w_x] - b_k[w_x])^p \overline{F}(t)\,\mathrm{d}t \\
&\quad \leqslant \sum_{k=1}^{K[w_x]} \int_{b_k[w_x]}^{c_k[w_x]} (p+1)(c_k[w_x] - b_k[w_x])^p \cdot O(t^{-\alpha})\,\mathrm{d}t && \text{[by Lemma 2.2.1]} \\
&\quad \leqslant \sum_{k=1}^{K[w_x]} \int_{b_k[w_x]}^{c_k[w_x]} O(x^{\theta p} t^{-\alpha} b_k[w_x]^{\zeta p})\,\mathrm{d}t && \text{[by Condition 3.2.1 and (3.4)]} \\
&\quad \leqslant \sum_{k=1}^{K[w_x]} \int_{b_k[w_x]}^{c_k[w_x]} O(x^{\theta p} t^{\zeta p - \alpha})\,\mathrm{d}t \\
&\quad \leqslant O(x^{\theta p}) \int_{x}^{c_{K[w_x]}[w_x]} O(t^{\zeta p - \alpha})\,\mathrm{d}t && \text{[by (3.4)]} \\
&\quad \leqslant O(x^{\theta p}) \int_{x}^{O(x^{\eta})} O(t^{\zeta p - \alpha})\,\mathrm{d}t && \text{[by Condition 3.2.1]} \\
&\quad = \begin{cases} O(x^{\theta p + \zeta p - \alpha + 1}) & \text{if } \zeta p < \alpha - 1 \\ O(x^{\theta p} \log x^{\eta}) & \text{if } \zeta p = \alpha - 1 \\ O(x^{\theta p + \eta(\zeta p - \alpha + 1)}) & \text{if } \zeta p > \alpha - 1. \end{cases}
\end{aligned}$$

□

Proof of Lemma 3.4.4. We apply Lemma 3.4.6, splitting into cases depending on how ζp compares to $\alpha - 1$.

Case 1 ($\zeta p < \alpha - 1$)**.** *By Lemma 3.4.6, it suffices to show*

$$\theta p + \zeta p - \alpha + 1 < p.$$

If $\theta \leqslant 1$*, then this holds because* $\zeta p < \alpha - 1$*. Otherwise, if* $\theta > 1$*, we can rearrange it to*

$$\zeta + \theta - 1 < \frac{\alpha - 1}{p},$$

in which case it also holds by Definition 3.4.1.

Case 2 ($\zeta p = \alpha - 1$)**.** *By Lemma 3.4.6, it suffices to show* $\eta < \infty$ *and* $\theta < 1$*. Because* $\zeta p = \alpha - 1$*, we can simplify Definition 3.4.1 to*

$$(\theta - 1)^+ - \frac{(1-\theta)^+}{\eta} < 0,$$

which implies $\eta < \infty$ *and* $\theta < 1$*.*

Case 3 ($\zeta p > \alpha - 1$)**.** *By Lemma 3.4.6, it suffices to show*

$$\theta p + \eta(\zeta p - \alpha + 1) < p.$$

Since $\theta - 1 = (\theta - 1)^+ - (1 - \theta)^+$*, dividing by* ζp *gives*

$$\zeta + \frac{(\theta - 1)^+}{\eta} - \frac{(1-\theta)^+}{\eta} < \frac{\alpha - 1}{p}.$$

Because $\eta \geqslant 1$*, this holds by Definition 3.4.1.* □

3.4.2 Lower bound

The proof of Lemma 3.4.2 requires a lower bound on $c_0[w_x(a)-]$.

Lemma 3.4.7. *Suppose Condition 3.2.1 holds, and let* $\kappa = 2\max\{\alpha - 1, \theta\}$*. For all* $x \geqslant 0$ *and* $a \in (y_x, x)$*,*

$$c_0[w_x(a)-] \geqslant \Omega\left(\left(\frac{x-a}{x^\zeta}\right)^{1/\kappa}\right).$$

Proof. Because $\kappa > \theta \geqslant 0$, by Condition 3.2.1 and (3.4), for all $x' \geqslant 0$ and $k \geqslant 1$,

$$x' \geqslant \Omega\left(\left(\frac{c_k[w_{x'}] - b_k[w_{x'}]}{b_k[w_{x'}]^\zeta}\right)^{1/\kappa}\right). \tag{3.5}$$

We now plug in $x' = c_0[w_x(a)-]$ and make the following observations.

- By Definition 2.A.3, we know $x' = c_0[w_x(a)-]$ is the earliest age at which a job has rank at least $w_x(a)$, so $w_{x'} = w_x(a)$.
- By Definition 2.A.1, a job's rank is at most $w_x(a)$ between ages a and x, so there exists $k \geqslant 1$ such that
$$b_k[w_x(a)] \leqslant a < x \leqslant c_k[w_x(a)].$$
In particular, $x > b_k[w_x(a)]$ and $x - a \leqslant c_k[w_x(a)] - b_k[w_x(a)]$.

Applying these observations to (3.5) with $x' = c_0[w_x(a)-]$ yields the desired bound. □

Proof of Lemma 3.4.2. The $\eta < \infty$ case follows from Lemma 2.6.3, so we address only the $\zeta < 1$ case.

We first observe that for all $\rho' \in [0, \rho]$,

$$\frac{1}{1-\rho'} \geqslant \frac{1}{1-\rho} - \frac{\rho - \rho'}{(1-\rho)^2}.$$

This means Condition 3.4.2 holds if

$$\int_0^x (\mathbb{E}(X) - \mathbb{E}(X_0[w_x(a)-]))\,\mathrm{d}a \leqslant \check{o}(x). \tag{3.6}$$

We can rewrite the integrand as

$$\begin{aligned}
\mathbb{E}(X) - \mathbb{E}(X_0[w_x(a)-]) &= \int_0^\infty \overline{F}(t)\,\mathrm{d}t - \int_0^{c_0[w_x(a)-]} \overline{F}(t)\,\mathrm{d}t && \text{[by Lemma 2.A.4]}\\
&\leqslant \int_{c_0[w_x(a)-]}^\infty O(t^{-\alpha})\,\mathrm{d}t && \text{[by Lemma 2.2.1]}\\
&\leqslant O(c_0[w_x(a)-]^{-(\alpha-1)}).
\end{aligned}$$

Of course, the integrand is also bounded above by $\mathbb{E}(X)$, so

$$\int_0^x (\mathbb{E}(X) - \mathbb{E}(X_0[w_x(a)-]))\,\mathrm{d}a \leqslant \int_0^x O(\min\{1, c_0[w_x(a)-]^{-(\alpha-1)}\})\,\mathrm{d}a. \tag{3.7}$$

A job of size x attains its worst ever rank w_x at age y_x. This means that for all $a < y_x$, we have $w_x(a) = w_x$, which by Definition 2.A.9 implies $c_0[w_x(a)-] = y_x$. Splitting the integral in (3.7) at $a = y_x$ yields

$$\begin{aligned}
\int_0^x (\mathbb{E}(X) - \mathbb{E}(X_0[w_x(a)-]))\,\mathrm{d}a &\leqslant \int_0^{y_x} O(y_x^{-(\alpha-1)})\,\mathrm{d}a + \int_{y_x}^x O(\min\{1, c_0[w_x(a)-]^{-(\alpha-1)}\})\,\mathrm{d}a\\
&\leqslant O(y_x^{(2-\alpha)^+}) + \int_{y_x}^x O(\min\{1, c_0[w_x(a)-]^{-(\alpha-1)}\})\,\mathrm{d}a.
\end{aligned} \tag{3.8}$$

Because $y_x \leqslant x$ and $\alpha > 1$, it suffices to show the integral in (3.8) is $\check{o}(x)$.

The remaining obstacle is bounding $c_0[w_x(a)-]$. Plugging the expression of Lemma 3.4.7 into (3.8) and substituting $u = x^{-\zeta}(x-a)$ gives

$$\begin{aligned}
\int_{y_x}^x O(\min\{1, c_0[w_x(a)-]^{-(\alpha-1)}\})\,\mathrm{d}a &\leqslant \int_0^x O\left(\min\left\{1, \left(\frac{x-a}{x^\zeta}\right)^{-(\alpha-1)/\kappa}\right\}\right)\mathrm{d}a\\
&\leqslant \int_{x-x^\zeta}^x O(1)\,\mathrm{d}a + x^\zeta \int_1^{x^{1-\zeta}} O(u^{-(\alpha-1)/\kappa})\,\mathrm{d}u\\
&= O(x^\zeta) + O(x^{\zeta+(1-\zeta)(1-(\alpha-1)/\kappa)}).
\end{aligned}$$

Because $\zeta < 1$ and $(\alpha-1)/\kappa \in (0, 1/2]$, this is $\check{o}(x)$, as desired. □

3.5 Heavy-tailed job sizes: Gittins is tail-optimal

In this section we prove Theorem 3.2.2. The proof boils down to a series of lemmas providing various bounds on the Gittins rank function and a related function, defined below.

Definition 3.5.1. *The* time-per-completion *function is*

$$\varphi(b,c) = \frac{\int_b^c \overline{F}(t)\,\mathrm{d}t}{\overline{F}(b) - \overline{F}(c)}.$$

Note that one can write the Gittins rank function as $r_{\mathrm{Gtn}}(a) = \inf_{c>a} \varphi(a,c)$.

We begin by stating (and restating) several bounds on r_{Gtn} and φ from prior work (and Chapter 2).

Lemma 3.5.1 (Section 2.3.2)**.** *Under Gittins with any HT job size distribution,* $r_{\mathrm{Gtn}}(a) = O(a)$*, and therefore* $w_x = O(x)$.

Lemma 3.5.2 (Scully et al. [108, Theorem 6.4])**.** *Under Gittins with any HT job size distribution,* $z_x - y_x = O(x)$.

Lemma 3.5.3 (Lemma 2.A.6)**.** *Under any SOAP policy, for any* w*-interval* (b,c)*, if* $x \in (b,c)$*, then*

$$y_x \leqslant b < c \leqslant z_x.$$

Lemma 3.5.4 (Scully et al. [108, Lemma 6.8])**.** *For any HT job size distribution, the time-per-completion function is bounded by*

$$\varphi(b,c) \geqslant \Omega\left(\frac{b}{c}(c-b)\right).$$

Combining these with one new property of the time-per-completion function, stated below and proven in Appendix 3.A, suffices to prove Theorem 3.2.2. While many similar results have been stated in prior work [2, 3], to the best of our knowledge, the following property is new.

Lemma 3.5.5. *Under Gittins, for any* w*-interval* (b,c)*, if* $r(c) \geqslant w$*, then* $\varphi(b,c) \leqslant w$.

We prove Lemma 3.5.5 in Appendix 3.A.2. With it in hand, we are ready to prove the main result of this section.

Theorem 3.2.2. *The Gittins policy is tail-optimal for any HT job size distribution.*

Proof. We show that Gittins satisfies each item of Condition 3.2.1, with $\zeta = 0$, $\theta = 1$, and $\eta = \infty$. We have $\eta = \infty$, so (ii) holds trivially. It thus remains only to show (i). Consider a w_x-interval (b,c) with $b \geqslant x$. Because $\zeta = 0$ and $\theta = 1$, our goal is to show $c - b = O(x)$.

Suppose first that $r(c) \geqslant w_x$, which allows us to apply Lemma 3.5.5. We compute

$$\begin{aligned} c - b &\leqslant \frac{z_{(b+c)/2}}{y_{(b+c)/2}} \cdot \frac{b}{c}(c-b) && \text{[by Lemma 3.5.3]} \\ &\leqslant O(1) \cdot \frac{b}{c}(c-b) && \text{[by Lemma 3.5.2]} \\ &\leqslant O(\varphi(b,c)) && \text{[by Lemma 3.5.4]} \\ &\leqslant O(w_x) && \text{[by Lemma 3.5.5]} \\ &= O(x). && \text{[by Lemma 3.5.1]} \end{aligned} \tag{3.9}$$

We conclude by reducing the $r(c) < w_x$ case to an $r(c') \geqslant w_x$ case. Suppose $r(c) < w_x$. Let c' be the minimum age after c with rank at least w_x (and $c' = \infty$ if such an age does not exist). By Definition 3.2.1, (b, c') is a w-interval, so Lemmas 3.5.2 and 3.5.3 together imply c' is finite. This means $r(c') \geqslant w_x$, so (3.9) implies $c - b \leqslant c' - b \leqslant O(x)$. □

3.6 Light-tailed job sizes: tail performance of SOAP policies

In this section prove our main theorem for LT job size distributions.

Theorem 3.3.1. *Consider an $M/G/1$ queue with any LT job size distribution under a SOAP policy. Let $x_{\max} = \inf\{x \geqslant 0 \mid \mathbb{P}(X > x) = 0\}$. The policy is*

- *log-tail-optimal if $a^* = 0$,*
- *log-tail-intermediate if $0 < a^* < x_{\max}$, and*
- *log-tail-pessimal if $a^* = x_{\max}$.*

Proof. The result follows from Lemmas 3.6.1, 3.6.2, and 3.6.7, which we prove in the rest of this section. □

3.6.1 Tail-optimal and tail-pessimal cases

Lemma 3.6.1. *Consider an $M/G/1$ queue with any LT job size distribution under a SOAP policy. The policy is log-tail-optimal if $a^* = 0$.*

Proof. Suppose that $a^* = 0$. In this case, due to the FCFS tiebreaking (Definition I.1), the oldest job in the system always has priority over all other jobs. Therefore the SOAP policy is exactly FCFS, which is known to be log-tail-optimal [29, 120]. □

Lemma 3.6.2. *Consider an $M/G/1$ queue with any LT job sizedistribution under a SOAP policy. The policy is log-tail-pessimal if $a^* = x_{\max}$.*

Proof. Assume that $a^* = x_{\max}$, i.e. the rank function has no global maximum. An important example of such a rank function is $r(a) = a$, which is the FB policy. Mandjes and Nuyens [80]

prove that FB is log-tail-pessimal, and our aim is to generalize this result to an arbitrary rank function without a global maximum.

Our proof uses the same strategy as Mandjes and Nuyens [80]. Recall that y_x denotes the (first) age of the maximum rank in the interval $[0, x]$. Since $T(x)$ is stochastically increasing in x, it holds that $\mathbb{P}(T(x) > t) \geqslant \mathbb{P}(T(y_x) > t)$ for all $t \geqslant 0$. Additionally we have that $\mathbb{P}(T(y_x) > t) \geqslant \mathbb{P}(T_{\mathrm{FB}}(y_x) > t)$ for all $t, x \geqslant 0$, where $T_{\mathrm{FB}}(x)$ is the response time for a job of size x under FB. The reason for this last inequality is that a job of size y_x must wait for all other jobs to receive up to y_x units of service before completing. As a result, and recalling Definition 3.3.1 for the decay rate, [80, Proposition 8] implies that

$$d(T(x)) \leqslant d(T(y_x)) \leqslant d(T_{\mathrm{FB}}(y_x)) = d(B_{y_x}), \tag{3.10}$$

with B_x denoting the duration of a busy period with job sizes $\min\{X, x\}$.

Additionally, up to its last line the proof of [80, Lemma 9] is valid for arbitrary service policies. If $x_0 > 0$ is such that $\mathbb{P}(X \geqslant x_0) > 0$, we thus find

$$d(T) \leqslant \mathbb{P}(X \geqslant x_0)^{-1} \int_{x_0}^{x_{\max}} d(T(x))\, dF(x). \tag{3.11}$$

Combining (3.10) and (3.11) yields

$$d(T) \leqslant \mathbb{P}(X \geqslant x_0)^{-1} \int_{x_0}^{x_{\max}} d(B_{y_x})\, dF(x).$$

Now let $c(x) = d(B_x)$ and $\widehat{c}(x) = c(y_x) = d(B_{y_x})$. Furthermore, we define the function $h_x(\theta) = \theta - \lambda\left(\mathbb{E}(e^{\theta \min\{X,x\}}) - 1\right)$. Due to [80, Equation (3)] we have the identity $c(x) = \sup_\theta h_x(\theta)$. Our next goal is to show that $\lim_{x \to x_{\max}} \widehat{c}(x) = \widehat{c}(x_{\max})$. Let $\varepsilon > 0$. By [80, Lemma 10], there exists x_0 such that $|c(x) - c(x_{\max})| < \varepsilon$ for all $x > x_0$. Because $a^* = x_{\max}$ there exists $x_1 > x_0$ such that $\widehat{c}(x_1) = c(x_1)$, and therefore $|\widehat{c}(x_1) - c(x_{\max})| < \varepsilon$. We now mention two properties of the function $\widehat{c}(x)$. On one hand $\widehat{c}(x)$ is decreasing because y_x is increasing in x and B_x is stochastically increasing in x. On the other hand, since $y_x \leqslant x$, we have that $\widehat{c}(x) \geqslant c(x)$. Combining these two properties we conclude that $|\widehat{c}(x) - c(x_{\max})| < \varepsilon$ for all $x > x_1$, and hence that $\lim_{x \to x_{\max}} \widehat{c}(x) = \widehat{c}(x_{\max})$.

The final step is to use the arguments of [80, Proposition 11] to show that $d(T) \leqslant d(B)$. Since no work-conserving policy has response time decay rate lower than the busy period decay rate [80, Corollary 6], we conclude that the policy is log-tail-pessimal. □

3.6.2 Tail-intermediate case

The final part of this section is devoted to the case that $0 < a^* < x_{\max}$, where we will show that the corresponding service policy is log-tail-intermediate. To this end, we define two policies, referred to as the *step* and *spike* policies, and relate the tail performance of any SOAP policy with $0 < a^* < x_{\max}$ to the performance of step and spike.

Definition 3.6.1. *The* step *and* spike *policies are the SOAP policies given by the following rank functions:*

$$r_{\text{step}}(a) = \min\{a, a^*\},$$
$$r_{\text{spike}}(a) = \mathbb{1}\{a = a^*\}.$$

We divide all jobs in two classes: those with size at most a^* and those with size larger than a^*. These classes will be called class 1 and class 2 respectively. For each class $i \in \{1, 2\}$, let

- $\lambda^{(i)}$ be the arrival rate of class i jobs,
- $T^{(i)}$ the response time of class i jobs.

Also, let W be the stationary amount of work in the queue and let $X_{a^*} := \min\{X, a^*\}$.

It turns out that only class 2 jobs affect the asymptotic decay rate of response time. Furthermore, we will show in the proof of Lemma 3.6.7 that for any SOAP policy π with a given worst age a^*,

$$T^{(2)}_{\text{spike}} \leqslant_{\text{st}} T^{(2)}_{\pi} \leqslant_{\text{st}} T^{(2)}_{\text{step}}.$$

It therefore suffices to show that both step and spike are tail-intermediate, and in particular to analyze $d(T^{(2)}_{\text{step}})$ and $d(T^{(2)}_{\text{spike}})$.

Our approach makes heavy use of LSTs (Section 3.3.1). Recall from (3.2) that one can determine a random variable's decay rate by determining when its LST converges. For functions $f : \mathbb{R} \to \mathbb{R} \cup \{-\infty, \infty\}$ which diverge below a certain value and converge above it, let

$$\gamma(f) := \sup\{s \in \mathbb{R} : |f(s)| = \infty\} = \inf\{s \in \mathbb{R} : |f(s)| < \infty\}$$

be the value at which f switches from diverging to converging. Recalling that $\tilde{V}(\cdot)$ denotes the LST of V, we can thus rewrite (3.2) as

$$d(V) = -\gamma(\tilde{V}). \tag{3.12}$$

With $B := B_\infty$ denoting a busy period, let $\sigma(s)$ be defined as $\sigma(s) := s + \lambda(1 - \tilde{B}(s))$, and similarly $\sigma_{a^*}(s) := s + \lambda(1 - \tilde{B}_{a^*}(s))$. Since $\tilde{B}(s) = \tilde{X}(\sigma(s))$ (and $\tilde{B}_{a^*}(s) = \tilde{X}_{a^*}(\sigma(s))$), we have that $\sigma^{-1}(s) = s - \lambda(1 - \tilde{X}(s))$ and $\sigma^{-1}_{a^*}(s) = s - \lambda(1 - \tilde{X}_{a^*}(s))$.

Lemma 3.6.3. *The function γ satisfies*

$$\gamma(\tilde{X}) < \gamma(\tilde{W}) < \sigma(\gamma(\sigma)) < \gamma(\sigma).$$

Proof. We rewrite each of the four quantities in terms of the function σ^{-1}. Note that $\gamma(\tilde{X}) = \gamma(\sigma^{-1})$. We also have that $\tilde{W}(s) = (1 - \rho)s/\sigma^{-1}(s)$ for $s \neq 0$, so $\gamma(\tilde{W})$ is the rightmost non-zero root of σ^{-1}. It can be seen that σ^{-1} is continuous and convex, with one negative root and one root at zero. Particularly, the negative root of σ^{-1} is larger than its singularity (if there is one), and it follows that $\gamma(\tilde{X}) < \gamma(\tilde{W})$.

For the third inequality we observe that $(\sigma^{-1})'(0) \in (0, 1)$. If $\widehat{s}$ minimizes σ^{-1}, then it follows by convexity that $\widehat{s} < \sigma^{-1}(\widehat{s})$. In terms of its inverse σ, this is equivalent to $\gamma(\sigma) > \sigma(\gamma(\sigma))$. Finally, $\widehat{s}$ is clearly larger than the negative root of σ^{-1} and therefore $\sigma(\gamma(\sigma)) > \gamma(\tilde{W})$. □

Let $f \circ g$ denote the function $s \mapsto f(g(s))$. The following lemma characterizes the decay rates of FCFS, FB, and the step and spike policies.

Lemma 3.6.4. *It holds that $d(T_{\text{FCFS}}) = -\gamma(\tilde{W})$ and $d(T_{\text{FB}}) = -\gamma(\tilde{W} \circ \sigma)$. Furthermore, if π is either the step or spike policy, then $d(T_\pi) = -\gamma(\tilde{W} \circ \sigma_{a^*})$.*

Proof. First, Equation (3.12) and Lemma 3.6.3 imply that

$$d(T_{\text{FCFS}}) = d(W + X) = \min\{d(W), d(X)\} = \min\{-\gamma(\tilde{W}), -\gamma(\tilde{X})\} = -\gamma(\tilde{W}).$$

The decay rate of T_{FB} equals the decay rate of the excess of a busy period [80], i.e. a busy period with initial work W. Therefore $d(T_{\text{FB}}) = -\gamma(T_{\text{FB}}) = -\gamma(\tilde{W} \circ \sigma)$ by (3.12).

In the remainder of the proof, let π be the step or spike policy. We write $\tilde{T}_\pi(s) = \frac{\lambda^{(1)}}{\lambda}\tilde{T}_\pi^{(1)}(s) + \frac{\lambda^{(2)}}{\lambda}\tilde{T}_\pi^{(2)}(s)$, where we note that $T_\pi^{(2)} \geqslant_{\text{st}} T_\pi^{(1)}$. It follows that $d(T_\pi) = d(T_\pi^{(2)})$, so we restrict ourselves to jobs with sizes larger than a^*. We tag such a job, and we split its response time in two parts $T_\pi^{(2)} = U_\pi^{(2)} + V_\pi^{(2)}$. Here, $U_\pi^{(2)}$ is the time it takes for all jobs currently in the system to finish and for all jobs arriving afterwards to reach age a^*, and $V_\pi^{(2)}$ is the remaining time it takes for the tagged job to finish. Observe that $U_\pi^{(2)}$ is a busy period with initial work W and job sizes X_{a^*}, so that

$$\tilde{U}_\pi^{(2)}(s) = \tilde{W}(s + \lambda(1 - \tilde{B}_{a^*}(s))) = \tilde{W}(\sigma_{a^*}(s)).$$

As a consequence,

$$d(T_\pi^{(2)}) \leqslant d(U_\pi^{(2)}) \tag{3.13}$$

for both policies.

First consider the case that π is the step policy. Note that $V_{\text{step}}^{(2)}$ is a busy period with initial work X and job sizes X_{a^*}. Therefore we have that

$$\tilde{V}_{\text{step}}^{(2)}(s) = \tilde{X}(s + \lambda(1 - \tilde{B}_{a^*}(s))) = \tilde{X}(\sigma_{a^*}(s)).$$

Because $\gamma(\tilde{X}) < \gamma(\tilde{W})$ by Lemma 3.6.3, it follows that $\gamma(\tilde{X} \circ \sigma_{a^*}) \leqslant \gamma(\tilde{W} \circ \sigma_{a^*})$ and thus that $d(U_{\text{step}}^{(2)}) \leqslant d(V_{\text{step}}^{(2)})$. The decay rate of the sum of two independent variables equals the minimum of the individual decay rates, leading to $d(T_{\text{step}}^{(2)}) = d(U_{\text{step}}^{(2)})$.

Now consider the spike policy. Due to its lower rank at ages after a^*, we have $V_{\text{spike}}^{(2)} \leqslant_{\text{st}} V_{\text{step}}^{(2)}$, so also $d(T_{\text{spike}}^{(2)}) \geqslant d(T_{\text{step}}^{(2)}) = d(U_{\text{step}}^{(2)})$. In combination with $d(T_{\text{spike}}^{(2)}) \leqslant d(U_{\text{spike}}^{(2)}) = d(U_{\text{step}}^{(2)})$, we therefore conclude with (3.12) that $d(T_\pi) = d(T_\pi^{(2)}) = d(U_\pi^{(2)}) = -\gamma(\tilde{U}_\pi^{(2)}) = -\gamma(\tilde{W} \circ \sigma_{a^*})$ for both the step and spike policies. □

The next lemma states a few properties of σ and σ_{a^*} that we will use later in this section.

Lemma 3.6.5. *For all $s < 0$, $\sigma(s) < \sigma_{a^*}(s) < s$ and $\sigma^{-1}(s) > \sigma_{a^*}^{-1}(s) > s$.*

Proof. We only prove the second inequality, because the first will follow directly. Note that $\sigma_{a^*}^{-1}(s) - s = \lambda(\tilde{X}_{a^*}(s) - 1)$, which is strictly positive since X_{a^*} is non-negative and $s < 0$. Note also that $\sigma^{-1}(s) - \sigma_{a^*}^{-1}(s) = \tilde{X}(s) - \tilde{X}_{a^*}(s)$, which is strictly positive since $X \geqslant_{\text{st}} X_{a^*}$ and $s < 0$. □

Lemma 3.6.6. *Let T be the response time under the step or spike policies. Then*

$$d(T_{\mathrm{FB}}) < d(T) < d(T_{\mathrm{FCFS}}).$$

Proof. In light of Lemma 3.6.4 we have to show that $\gamma(\tilde{W}) < \gamma(\tilde{W} \circ \sigma_{a^*}) < \gamma(\tilde{W} \circ \sigma)$. We focus first on the first inequality, and define $\gamma^* = \gamma(\tilde{W} \circ \sigma_{a^*})$ as a shorthand notation. There can be two reasons for $\tilde{W}(\sigma_{a^*}(\gamma))$ to diverge. The first possibility is that $\tilde{W}$ diverges at the point $\sigma_{a^*}(\gamma^*)$, that is, $\sigma_{a^*}(\gamma^*) = \gamma(\tilde{W})$. Then by Lemma 3.6.5 it holds that

$$\gamma^* = \sigma_{a^*}^{-1}(\gamma(\tilde{W})) > \gamma(\tilde{W}).$$

However, it could also be the case that σ_{a^*} diverges at γ^*. If so, then it follows from Lemmas 3.6.3 and 3.6.5 that

$$\gamma^* = \gamma(\sigma_{a^*}) > \sigma_{a^*}(\gamma(\sigma_{a^*})) > \gamma(\tilde{W}),$$

where we remark that Lemma 3.6.3 also holds if σ is replaced by σ_{a^*} (with analogous proof).

It remains to show that $\gamma(\tilde{W} \circ \sigma_{a^*}) < \gamma(\tilde{W} \circ \sigma)$. To do this we use the exact same argument, where $\tilde{W} \circ \sigma_{a^*}$ plays the role of $\tilde{W}$, and $\sigma_{a^*}^{-1} \circ \sigma$ plays the role of σ_{a^*}. $d(T_{\mathrm{FB}}) < d(U_\pi^{(2)})$. □

Lemma 3.6.7. *Consider an $M/G/1$ queue with any LT job size distribution under a SOAP policy. The policy is log-tail-intermediate if $0 < a^* < x_{\max}$.*

Proof. Let π be a policy with $0 < a^* < x_{\max}$. Once again we write $\tilde{T}_\pi(s) = \frac{\lambda^{(1)}}{\lambda}\tilde{T}_\pi^{(1)}(s) + \frac{\lambda^{(2)}}{\lambda}\tilde{T}_\pi^{(2)}(s)$, noting that $T_\pi^{(2)} \geqslant_{\mathrm{st}} T_\pi^{(1)}$. Therefore $d(T_\pi) = d(T_\pi^{(2)})$, so we restrict ourselves to jobs with size larger than a^*. We tag such a job, and we split its response time in two parts $T_\pi^{(2)} = U_\pi^{(2)} + V_\pi^{(2)}$. Here we recall that $U_\pi^{(2)}$ is the time it takes for all jobs currently in the system to finish and for all jobs arriving afterwards to reach age a^*, and $V_\pi^{(2)}$ is the remaining time it takes for the tagged job to finish. Observe that during the age interval $[0, a^*]$, the tagged job is delayed by all work in the system upon arrival, and all jobs arriving afterwards up to age a^*. This holds regardless of the behavior of the rank function in $[0, a^*]$, so $U_\pi^{(2)} = U_{\mathrm{step}}^{(2)} = U_{\mathrm{spike}}^{(2)}$. Furthermore, note that on one hand $V_\pi^{(2)}$ is minimized if $r_\pi(a)$ is minimal for all $a > a^*$ (which holds for r_{spike}). On the other hand, $V_\pi^{(2)}$ is maximized if $r_\pi(a) = r(a^*)$ for all $a \geqslant a^*$ (which holds for r_{step}). These observations lead to $V_{\mathrm{spike}}^{(2)} \leqslant_{\mathrm{st}} V_\pi^{(2)} \leqslant_{\mathrm{st}} V_{\mathrm{step}}^{(2)}$, and hence also to $T_{\mathrm{spike}}^{(2)} \leqslant_{\mathrm{st}} T_\pi^{(2)} \leqslant_{\mathrm{st}} T_{\mathrm{step}}^{(2)}$. The result now follows from Lemma 3.6.6. □

3.7 Light-tailed job sizes: Gittins can be tail-optimal, tail-pessimal, or in between

Theorem 3.3.2. *Consider an $M/G/1$ queue with any LT job size distribution. Gittins is*

- *log-tail-optimal if $X \in \mathsf{NBUE}$,*
- *log-tail-intermediate if $X \in \mathsf{ENBUE} \setminus \mathsf{NBUE}$, and*
- *log-tail-pessimal if $X \notin \mathsf{ENBUE}$.*

Proof. By Theorem 3.3.1, it suffices to determine the worst age a^*. Combining the following two prior results characterizes a^* in terms of whether X is in each of NBUE and ENBUE.

- A result of Aalto et al. [2, Proposition 7] implies $a^* = 0$ if and only if $X \in \mathsf{NBUE}$.
- A result of Aalto et al. [3, Proposition 9] implies $a^* < x_{\max}$ if and only if $X \in \mathsf{ENBUE}$. □

Theorem 3.3.4. *Consider an $M/G/1$ queue with an LT job size distribution $X \notin \mathsf{ENBUE}$. Suppose that the expected remaining size of a job at all ages is uniformly bounded, meaning*

$$\sup_{a \in [0, x_{\max})} \mathbb{E}(X - a \mid X > a) < \infty.$$

Then for all $\varepsilon > 0$, there exists a $(1+\varepsilon)$-approximate Gittins policy that is log-tail-optimal or log-tail-intermediate.

Proof. Suppose $\mathbb{E}(X - a \mid X > a)$ is uniformly bounded. Definition 3.1.1 implies

$$r_{\mathrm{Gtn}}(a) \leqslant \frac{\int_a^\infty \overline{F}(t)\,\mathrm{d}t}{\overline{F}(a)} = \mathbb{E}(X - a \mid X > a),$$

so $r_{\mathrm{Gtn}}(a)$ is also uniformly bounded. This means that for any $\varepsilon > 0$, there exists some sufficiently large age $a(\varepsilon)$ such that increasing the rank at age $a(\varepsilon) < x_{\max}$ from $r_{\mathrm{Gtn}}(a(\varepsilon))$ to $(1+\varepsilon) r_{\mathrm{Gtn}}(a(\varepsilon))$ and leaving all other ranks unchanged yields a new SOAP policy with worst age $a^* = a(\varepsilon)$. By construction, the new policy is a $(1+\varepsilon)$-approximate Gittins policy, and because its worst age is $a^* < x_{\max}$, Theorem 3.3.1 implies it is log-tail-optimal or log-tail-intermediate. □

Recall from Theorem 3.3.3 that a $(1+\varepsilon)$-approximate Gittins policy achieves mean response time within a factor of $1+\varepsilon$ of optimal. We defer its proof to Appendix 3.A. This means that Theorem 3.3.4, whose precondition applies to non-pathological LT job size distributions, gives a non-tail-pessimal policy with near-optimal mean response time.

Remark 3.7.1. The *Shortest Expected Remaining Processing Time* (SERPT) policy, which has rank function $r_{\mathrm{SERPT}}(a) = \mathbb{E}(X - a \mid X > a)$, is sometimes considered as a simpler alternative to Gittins [113, 114]. Our results imply that SERPT has the same tail-optimality properties as Gittins.

- In Chapter 2 it is shown that SERPT is always tail-optimal in the heavy-tailed case, which matches our result for Gittins.
- Theorem 3.3.1 and Definition 3.3.5 imply that in the light-tailed case, SERPT is tail-optimal, tail-intermediate, and tail-pessimal under the same conditions as we show for Gittins in Theorem 3.3.2. In fact, one can show a stronger property: SERPT's and Gittins's response time distributions have the same decay rate. This follows from the fact that SERPT and Gittins have the same worst age a^* [2, 3]. ◇

3.8 Conclusion

In this chapter we have characterized the asymptotic tail performance of the response time in an $M/G/1$ queue under very broad conditions, namely for every SOAP policy, and for both HT and LT job size distributions.

In the heavy-tailed case, we characterize tail-optimal policies by a sufficient condition on the rank function (Theorem 3.2.1). This condition holds for a wide range of SOAP policies, and specifically for the Gittins policy (Theorem 3.2.2), providing a proof of its tail optimality under general conditions.

In the light-tailed case, we classify policies' performance as tail-optimal, tail-pessimal, or tail-intermediate. We show that the performance of a SOAP policy depends on the age at which the maximal rank is attained (Theorem 3.3.1). It turns out that the Gittins policy may belong to any of the three categories, depending on the job size distribution (Theorem 3.3.2). Finally, when Gittins has pessimal tail performance, boundedness of the expected remaining job size implies that there exists a slight modification of Gittins that has optimal or intermediate tail while maintaining near-optimal mean response time (Theorem 3.3.4).

Returning to the motivating questions

We conclude by returning to Questions 3.1–3.3, restated below for convenience.

Question 3.1. *Does any scheduling policy* simultaneously *optimize the mean and asymptotic tail of response time?*

Question 3.2. *For which job size distributions is Gittins tail-optimal for response time?*

Question 3.3. *For job size distributions for which Gittins is tail-pessimal, is there another policy that has* near-optimal *(arbitrarily close to optimal) mean response time while not being tail-pessimal?*

Our characterization of Gittins's tail asymptotics (Theorems 3.2.2 and 3.3.2) answers Question 3.2, and our modification in the case where Gittins is tail-pessimal (Theorem 3.3.4) answers Question 3.3 affirmatively. This leaves only Question 3.1. In cases where we have shown that Gittins is tail-optimal, the answer is clearly affirmative. We might hope to conclude that the answer is negative in cases where we have shown that Gittins is tail-pessimal or tail-intermediate, but the situation is still slightly unclear. As we explain in more detail below, the remaining ambiguity is due to the fact that we have restricted our attention to FCFS tiebreaking when two jobs have the same rank (Definition I.1).

It has been shown that the Gittins policy minimizes mean response time with *arbitrary tiebreaking* between jobs of the same rank [52, 53, 111]. Moreover, at least one of these proofs [111] may be easily extended to show that any "non-Gittins" policy is *strictly suboptimal* for mean response time, where a "non-Gittins" policy is one that for a non-vanishing fraciton of time serves a job other than one of minimal Gittins rank. Therefore, to fully answer Question 3.1,

one would have to consider Gittins under arbitrary tiebreaking rules. We conjecture that using a different tiebreaking rule cannot improve the asymptotic decay rate of Gittins's response time in the light-tailed case.

Finally, we note that we have of course only answered Questions 3.1–3.3 for the classes of heavy-tailed and light-tailed job size distributions we consider in this work (Definitions 2.2.1 and 3.3.2). Practically speaking, we believe the classes of distributions we consider are likely broad enough to draw useful conclusions. But it remains an open question whether we may extend our theory to broader classes of distributions. In particular, it seems likely that our proofs may hold mostly unchanged for additional light-tailed job size distributions, as discussed in Appendix 3.B.3.

Appendix 3.A Properties of the Gittins policy via the "Gittins game"

The goal of this section is to prove two key remaining properties of the Gittins policy, Theorem 3.3.3 and Lemma 3.5.5. To prove both of these properties, we will use a different perspective on the Gittins policy called the "Gittins game" [107]. The Gittins game gives an alternative way to define the Gittins rank function. While it is less direct than the definitions we have used so far (Definitions 3.1.1 and 3.5.1), the intermediate steps it introduces turn out to be crucial for proving Theorem 3.3.3 and Lemma 3.5.5.

Aside from Theorem 3.3.3 and Lemma 3.5.5, most of the definitions and results in this section are due to Scully et al. [107], who actually study a much more general job model than ours. For simplicity, we restate the key definitions and results in our setting. However, the statements and proofs of Theorem 3.3.3 and Lemma 3.5.5 are straightforward to translate to the more general job model of Scully et al. [107].

3.A.1 The Gittins game

The Gittins game is an optimization problem. Its inputs are a job at some age b and a *penalty* w. During the game, we serve the job for as long as we like. If the job completes, the game ends. At any moment before the job completes, we may choose to give up, in which case we pay the penalty w and the game immediately ends. The goal of the game is to minimize the expected sum of the time spent serving the job plus the penalty paid.

We can think of the Gittins game with penalty w as an optimal stopping problem whose state is the age b of the job. Standard optimal stopping theory [99, 117] implies that the optimal strategy thus has the following form: serve the job until it reaches some age $c \geqslant b$, then give up. A possible policy here is never giving up, which is represented by $c = \infty$.

Suppose we start serving a job at age b and stop if it reaches age c. The expected amount of

time we spend serving the job is

$$\mathsf{service}(b,c) := \mathbb{E}(\min\{X,c\} \mid X > b) = \int_b^c \frac{\overline{F}(t)}{\overline{F}(b)}\,\mathrm{d}t,$$

and the probability the job finishes before reaching age c is

$$\mathsf{done}(b,c) := \mathbb{P}(S \leqslant c \mid S > b) = 1 - \frac{\overline{F}(c)}{\overline{F}(b)}.$$

We can write the time-per-completion function (Definition 3.5.1) as the ratio of these two quantities: $\varphi(b,c) = \mathsf{service}(b,c)/\,\mathsf{done}(b,c)$.

Suppose we employ the stop-at-age-c policy in the Gittins game starting from age b with penalty w. The expected cost of the Gittins game with this policy is

$$\mathsf{game}(w;b,c) := \mathsf{service}(b,c) + w(1 - \mathsf{done}(b,c)).$$

The *optimal* cost of the Gittins game is therefore

$$\mathsf{game}^*(w;b) := \inf_{c \geqslant b} \mathsf{game}(w;b,c).$$

The lemma below follows immediately from the definition of $\mathsf{game}^*(w;b)$ as an infimum of $\mathsf{game}(w;b,c)$, each of which is a linear function of w [107, Lemmas 5.2 and 5.3].

Lemma 3.A.1. *For all ages b, the optimal cost $\mathsf{game}^*(w;b)$ is increasing and concave as a function of w. Because giving up immediately is always a possible policy, it is also bounded above by $\mathsf{game}^*(w;b) \leqslant w$.*

3.A.2 Relating the Gittins game to the Gittins rank function

The Gittins game is intimately connected to the Gittins rank function, and it is this connection that is important for proving Lemma 3.5.5. The following lemmas state two such connections. They are the same or very similar to many previous results in the literature on Gittins in the $M/G/1$ queue [2, 52, 53, 107, 112], but we sketch their proofs for completeness.

Lemma 3.A.2. *The Gittins rank function can be expressed in terms of the Gittins game as*

$$r(a) = \inf(w \geqslant 0 \mid \mathsf{game}^*(w;a) < w) = \max(w \geqslant 0 \mid \mathsf{game}^*(w;a) = w).$$

Proof. The infimum and maximum are equivalent by Lemma 3.A.1. The infimum is equal to the rank $r(a) = \inf_{c>a} \varphi(a,c)$ because, by the fact that we can write $\mathsf{game}(w;b,c)$ as

$$\mathsf{game}(w;b,c) = w - (w - \varphi(b,c))\,\mathsf{done}(b,c), \tag{3.14}$$

we have $\mathsf{game}(w;b,c) < w$ if and only if $\varphi(b,c) < w$. □

Lemma 3.A.3. *In the Gittins game with penalty w with the job currently at age a, it is optimal to continue serving the job if and only if $r(a) \leqslant w$, and it is optimal to give up if and only if $r(a) \geqslant w$.*

Proof. Giving up incurs cost w, so by the maximum in Lemma 3.A.2, it is optimal to give up if and only if $r(a) \geqslant w$. This means it is optimal to continue serving the job if $r(a) < w$. The fact that serving is optimal in the $r(a) = w$ edge case follows from the fact that if $\varphi(a, c) = w$ for some $c > a$, then by (3.14), we have $\mathsf{game}(w; a, c) = w$. □

We are now ready to prove Lemma 3.5.5, which we restate below. Recall that a w-interval is one in which the Gittins rank is bounded above by w. The key to the proof is that Lemma 3.A.3 relates w-intervals to optimally playing the Gittins game.

Lemma 3.5.5. *Under Gittins, for any w-interval (b, c), if $r(c) \geqslant w$, then $\varphi(b, c) \leqslant w$.*

Proof. Consider playing the Gittins game starting from age b. By Lemma 3.A.3, giving up if the job reaches age c is an optimal policy. Specifically, because (b, c) is a w-interval, it is optimal to continue serving the job until at least age c; and because $r(c) \geqslant w$, it is optimal to give up if the job reaches age c. This means $\mathsf{game}^*(w; b) = \mathsf{game}(w; b, c)$. Combining Lemma 3.A.1 and (3.14) implies $\varphi(b, c) \leqslant w$. □

We note that Lemma 3.5.5 is similar, but not identical, to properties of Gittins in the $M/G/1$ queue studied by Aalto et al. [2, 3]. Related properties have also been shown for versions of Gittins in discrete-time settings [42, 52, 53].

3.A.3 Relating the Gittins game to mean response time

It remains only to prove Theorem 3.3.3, which bounds the mean response time of q-approximate Gittins policies. To do so, we use a result of Scully et al. [107] that relates the Gittins game to a system's mean response time.

Definition 3.A.1. *Let $r : [0, x_{\max}) \to \mathbb{R}$ be the rank function of some SOAP policy, and let $w \in \mathbb{R}$.*

(i) The (r, w)-relevant work of a job *is the amount of service the job requires to either complete or reach rank at least w according to r, meaning reaching an age a satisfying $r(a) \geqslant w$.*

(ii) The (r, w)-relevant work of the system *is the total (r, w)-relevant work of all jobs present. We denote the steady-state distribution of (r, w)-relevant work under policy π by $W_\pi(r, w)$. Note that r need not be the rank function of policy π.*

The (r_{Gtn}, w)-relevant work of a job is related to the Gittins game via Lemma 3.A.3: it is the amount of time we would serve the job when optimally playing the Gittins game with penalty w. It turns out that mean (r_{Gtn}, w)-relevant work directly translates into mean response time.

Lemma 3.A.4 (Scully et al. [107, Theorem 6.3])**.** *Under any scheduling policy π that does not use information on exact job sizes, the mean response time can be written in terms of (r_{Gtn}, w)-relevant work as*

$$\mathbb{E}(T_\pi) = \frac{1}{\lambda} \int_0^\infty \frac{\mathbb{E}(W_\pi(r_{\text{Gtn}}, w))}{w^2} \, \mathrm{d}w.$$

The proof of Lemma 3.A.4 relies on Little's law, a relation involving $\mathsf{service}(a, w)$, and the observation that $W_\pi(r_{\text{Gtn}}, w) = \sum_{i=1}^{n} \mathsf{service}(a_i, w)$, where n is the number of jobs in the system and a_i is the age of job i.

With Lemma 3.A.4 in hand, the proof of Theorem 3.3.3, restated below, reduces to bounding the mean amount of (r_{Gtn}, w)-relevant work under q-approximate Gittins policies.

Theorem 3.3.3. *Consider an $M/G/1$ queue with any job size distribution. For any $q \geqslant 1$ and any q-approximate Gittins policy π,*

$$\mathbb{E}(T_\pi) \leqslant q\mathbb{E}(T_{\text{Gtn}}).$$

Proof. Recall from Definition 3.3.6 that we may assume $r_\pi(a)/r_{\text{Gtn}}(a) \in [1, q]$ for all ages a without loss of generality. We will prove

$$\mathbb{E}(W_\pi(r_{\text{Gtn}}, w)) \leqslant \mathbb{E}(W_\pi(r_\pi, qw)) \leqslant \mathbb{E}(W_{\text{Gtn}}(r_{\text{Gtn}}, qw)), \tag{3.15}$$

from which the theorem follows by the computation below:

$$\begin{aligned}
\mathbb{E}(T_\pi) &= \frac{1}{\lambda}\int_0^\infty \frac{\mathbb{E}(W_\pi(r_{\text{Gtn}}, w))}{w^2}\,\mathrm{d}w && \text{[by Lemma 3.A.4]}\\
&\leqslant \frac{1}{\lambda}\int_0^\infty \frac{\mathbb{E}(W_{\text{Gtn}}(r_{\text{Gtn}}, qw))}{w^2}\,\mathrm{d}w && \text{[by (3.15)]}\\
&= \frac{1}{\lambda}\int_0^\infty \frac{\mathbb{E}(W_{\text{Gtn}}(r_{\text{Gtn}}, w'))}{(w'/q)^2}\,\mathrm{d}(w'/q) && \text{[by substituting } w' = qw\text{]}\\
&= q\mathbb{E}(T_{\text{Gtn}}). && \text{[by Lemma 3.A.4]}
\end{aligned}$$

To show the left-hand inequality of (3.15), it suffices to show that an arbitrary job's (r_{Gtn}, w)-relevant work is upper bounded by its (r_π, qw)-relevant work (Definition 3.A.1). This is indeed the case: $r_{\text{Gtn}}(a) \leqslant w$ implies $r_\pi(a) \leqslant qr_{\text{Gtn}}(a) \leqslant qw$, so the job will reach rank w under Gittins after at most as much service as it needs to reach rank qw under π.

To show the right-hand inequality of (3.15) we use a property of SOAP policies due to Scully et al. [113, proof of Lemma 5.2]. The property implies that for any rank w and SOAP policy π, we can express $\mathbb{E}(W_\pi(r_\pi, w))$ in terms of just the job size X, arrival rate λ, and the set of ages $A_\pi[w] = \{a \in [0, x_{\max}) \mid r_\pi(a) < w\}$. In particular, for any fixed job size distribution, arrival rate, and rank w, $\mathbb{E}(W_\pi(r_\pi, w))$ is a nondecreasing function of $A_\pi[w]$, where we order sets by the usual subset partial ordering. We have $r_\pi(a) \geqslant r_{\text{Gtn}}(a)$, which means $A_\pi[w] \subseteq A_{\text{Gtn}}[w]$, which implies the right-hand inequality of (3.15), as desired. □

We note that one can use the techniques of Scully and Harchol-Balter [110] to generalize the statement and proof of Theorem 3.3.3 beyond SOAP policies. It turns out that Theorem 3.3.3 still holds even if we allow q-approximate Gittins policies to *adversarially* assign ranks to jobs, provided that the assigned ranks are still within a factor-q window around the rank Gittins would assign.

Appendix 3.B Relationship between decay rate and LST

The goal of this section is to justify our computation of decay rates (Definition 3.3.1) by means of LST convergence (Section 3.6.2). Our specific goal is to justify our use of (3.12), which states $d(V) = -\gamma(\tilde{V})$. As a reminder,

$$d(V) = \lim_{t\to\infty} \frac{-\log \mathbb{P}(V > t)}{t},$$
$$\gamma(f) = \inf\{s \in \mathbb{R} \mid |f(s)| < \infty\}.$$

3.B.1 Sufficient condition for computing decay rates

Our main tool for translating between $d(V)$ and $\gamma(\tilde{V})$ is a result of Mimica [82], restated as Lemma 3.B.1 below, which gives a sufficient condition for $d(V) = -\gamma(\tilde{V})$. The result rests on the following definition.

Definition 3.B.1. *We say a function $f : \mathbb{R} \to \mathbb{R} \cup \{-\infty, \infty\}$ is* regularly varying from the right at s^* with negative index*, or simply "regularly varying at s^*", if there exists $\alpha > 0$ such that for all $c > 0$,*

$$\lim_{s\downarrow 0} \frac{f(s^* + cs)}{f(s^* + s)} = c^{-\alpha}.$$

In particular, f having a pole of finite order at s^* suffices.

It turns out being regularly varying at the singularity is the condition we need to express decay rate in terms of LST convergence.

Lemma 3.B.1 (special case of Mimica [82, Corollary 1.3])**.** *Let V be a non-negative random variable with $\gamma(\tilde{V}) > -\infty$. If either $\tilde{V}(s)$ or its derivative $\tilde{V}'(s)$ is regularly varying at $\gamma(\tilde{V})$, then*

$$d(V) = -\gamma(\tilde{V}).$$

3.B.2 Showing that the sufficient condition for computing decay rates holds

It remains to show that the precondition of Lemma 3.B.1 holds whenever we apply (3.12) in Section 3.6.2. It turns out that all of the LSTs to which we apply (3.12) have a common form, so we will show that Lemma 3.B.1 applies to all functions of that form. To describe the form, we need the following definition.

Definition 3.B.2. *Consider an $M/G/1$ queue with arrival rate λ, job sizes distributed as X, and load $\rho = \lambda\mathbb{E}(X)$.*

(i) We define the function

$$\sigma_X^{-1}(s) = s - \lambda(1 - \tilde{X}(s)).$$

Note that $\sigma_X^{-1}(s) = \infty$ if and only if $\tilde{X}(s) = \infty$.

(ii) *We define σ_X to be the inverse of σ_X^{-1}, choosing the branch that passes through the origin. That is, for $s \geqslant \inf_r \sigma_X^{-1}(r)$, we define $\sigma_X(s)$ to be the greatest real solution to*

$$\sigma_X(s) = s + \lambda(1 - \tilde{X}(\sigma_X(s))).$$

If $s < \inf_r \sigma_X^{-1}(r)$, then no such solution exists, so we define $\sigma_X(s) = -\infty$.

(iii) *We define the* work-in-system transform

$$\tilde{W}_X(s) = \frac{s(1-\rho)}{\sigma_X^{-1}(s)}.$$

Note that all of the above definitions depend on both λ and X, However, because the following discussion considers a fixed arrival rate λ and varies only the job size distribution, we keep λ implicit to reduce clutter. Additionally, we assume in all uses of the above definitions that $\rho < 1$.

One may recognize the functions defined in Definition 3.B.2 as core to the theory of the $M/G/1$ queue with job sizes distributed as X [57].

- The work-in-system transform is, as suggested by its name, the LST of the equilibrium distribution W_X of the total workload in the $M/G/1$.
- The function σ_X is related to busy periods in the $M/G/1$. Specifically, the length of a busy period started by initial workload V has LST $\tilde{V}(\sigma_X(s))$.

It turns out that throughout Section 3.6.2, all of the LSTs to which we apply (3.12) are of the form $\tilde{W}_X$ or $\tilde{W}_X \circ \sigma_Y$ (the latter meaning $s \mapsto \tilde{W}_X(\sigma_Y(s))$), for LT job sizes X and Y (Definition 3.3.2). Specifically, X is the system's job size, and Y is either X or a truncation $\min\{X, a^*\}$. Therefore, to justify the uses of (3.12) using Lemma 3.B.1, it suffices to prove Theorems 3.B.1 and 3.B.2 below.

Theorem 3.B.1. *For any LT job size X,*

(i) $\gamma(\tilde{W}_X) \in (-\infty, 0)$*; and*

(ii) *$\tilde{W}_X$ has a first-order pole at $\gamma(\tilde{W}_X)$, so it is regularly varying at $\gamma(\tilde{W}_X)$.*

Theorem 3.B.2. *For any LT job sizes X and Y,*

(i) $\gamma(\tilde{W}_X \circ \sigma_Y) \in (-\infty, 0)$*, and*

(ii) *either $\tilde{W}_X \circ \sigma_Y$ or $(\tilde{W}_X \circ \sigma_Y)'$ is regularly varying at $\gamma(\tilde{W}_X \circ \sigma_Y)$.*

Our approach is as follows. We first prove Theorem 3.B.1. We then prove a lemma characterizing σ_X, which we use in conjunction with Theorem 3.B.1 to prove Theorem 3.B.2.

Proof of Theorem 3.B.1. Recall from Definition 3.B.2 that $\tilde{W}_X(s) = s(1-\rho)/\sigma_X^{-1}(s)$, so we focus on σ_X^{-1}. Because $\tilde{X}$ is a mixture of exponentials, σ_X^{-1} is convex, so it has at most two real roots. It is well-known that under the assumption on X made in Definition 3.3.2, σ_X^{-1} has a first-order root at 0 and a negative first-order root [4, 80], the latter of which is $\gamma(\tilde{W}_X)$, but

we give a brief proof for completeness. One can compute $\sigma_X^{-1}(0) = 0$ and $(\sigma_X^{-1})'(0) = 1 - \rho$, so σ_X^{-1} has a first-order root at 0. Definition 3.3.2 implies $\tilde{X}(\gamma(\tilde{X})) = \infty$, so $\sigma_X^{-1}(\gamma(\tilde{X})) = \infty$. This means σ_X^{-1} has another first-order root in $(\gamma(\tilde{X}), 0)$. □

Lemma 3.B.2. *For any job size X with an LT distribution,*

(i) $\gamma(\sigma_X) \in (-\infty, 0)$*;*

(ii) $\sigma_X(\gamma(\sigma_X)) \in (-\infty, 0)$*; and*

(iii) there exist $C_0, C_1 > 0$ such that in the $s \downarrow 0$ limit,

$$\sigma_X(\gamma(\sigma_X) + s) = \sigma_X(\gamma(\sigma_X)) + C_0\sqrt{s} \pm \Theta(s),$$
$$\sigma_X'(\gamma(\sigma_X) + s) = \frac{C_1}{\sqrt{s}} \pm \Theta(1),$$

so σ_X' is regularly varying at $\gamma(\sigma_X)$.

Proof. As in the proof of Theorem 3.B.1, we again use the fact that σ_X^{-1} is convex, has roots at a negative number and at zero, and is negative between its roots. Specifically, this fact implies that σ_X^{-1} has a finite negative global minimum. By Definition 3.B.2, this minimum is $\gamma(\sigma_X)$, and the value at which the minimum is attained is $\sigma_X(\gamma(\sigma_X))$ proving (i) and (ii).

It remains only to prove (iii). The fact that LSTs are analytic in the interior of their domains of convergence implies that σ_X^{-1} can be written as a Taylor series about $\gamma(\sigma_X)$ whose first nonzero coefficient is quadratic, i.e. for some constant $K > 0$,

$$\sigma_X^{-1}(s) = Ks^2 \pm \Theta(s^3).$$

An extension of the Lagrange inversion theorem [93, §1.10(vii)] implies that the inverse of σ_X^{-1}, namely σ_X, may thus be written in the desired form. The desired form for σ_X', which completes (iii), then follows from

$$(\sigma_X^{-1})'(s) = 2Ks \pm \Theta(s^2),$$
$$\sigma_X'(s) = \frac{1}{(\sigma^{-1})'(\sigma_X(s))}.$$

□

Proof of Theorem 3.B.2. There are three cases to consider:

- $\gamma(\tilde{W}_X) > \sigma_Y(\gamma(\sigma_Y))$,
- $\gamma(\tilde{W}_X) < \sigma_Y(\gamma(\sigma_Y))$, and
- $\gamma(\tilde{W}_X) = \sigma_Y(\gamma(\sigma_Y))$.

For an intuitive grasp of these cases, it is helpful to imagine decreasing s starting at $s = 0$, tracking the behavior of $\tilde{W}_X(\sigma_Y(s))$ as s decreases.

If $\gamma(\tilde{W}_X) > \sigma_Y(\gamma(\sigma_Y))$, then at some point before $s = s^*$ reaches $\gamma(\sigma_Y)$, meaning for some $s^* \in (-\gamma(\sigma_Y), 0)$, we have $\gamma(\tilde{W}_X) = \sigma_Y(s^*)$. This means $\gamma(\tilde{W}_X \circ \sigma_Y) = s^*$. The Lagrange

inversion theorem [93, §1.10(vii)] and the fact that $s > \gamma(\sigma_Y)$ imply that σ_Y can be linearly approximated near s^*, so the result follows from Theorem 3.B.1.

If $\gamma(\tilde{W}_X) < \sigma_Y(\gamma(\sigma_Y))$, then in contrast to the previous case, s reaches $\gamma(\sigma_Y)$, the last point at which $\sigma_Y(s)$ is finite, before $\sigma(s)$ reaches the pole of $\tilde{W}_x$. This means $\gamma(\tilde{W}_X \circ \sigma_Y) = \gamma(\sigma_Y)$. Similarly to the previous case, we can linearly approximate $\tilde{W}_x$ near $\gamma(\sigma_Y)$, so the result follows from Lemma 3.B.2.

If $\gamma(\tilde{W}_X) = \sigma_Y(\gamma(\sigma_Y))$, then roughly speaking, both of the previous cases' events happen simultaneously: just as s reaches $\gamma(\sigma_Y)$, the last point at which $\sigma_Y(s)$ is finite, $\sigma_Y(s)$ reaches the pole of $\tilde{W}_X$. Combining Theorem 3.B.1 and Lemma 3.B.2 implies that in the $s \downarrow \gamma(\sigma_Y)$ limit, we can approximate $\tilde{W}_X(\sigma_Y(s))$ as

$$\tilde{W}_X(\sigma_Y(s)) = \frac{K_0}{\sigma_Y(\gamma(s))} \pm \Theta(1) = \frac{K_1}{\sqrt{s - \gamma(\sigma_Y)}} \pm \Theta(1),$$

for some constants $K_0, K_1 > 0$, from which the result follows. □

3.B.3 Expanding our definition of light-tailed job size distributions

The class of light-tailed distributions we consider in Definition 3.3.2, namely what Abate and Whitt [4] call "Class I" distributions, is well behaved enough for Theorems 3.B.1 and 3.B.2 to hold. More generally, our results apply to any job size distribution with positive decay rate for which one can show Theorems 3.B.1 and 3.B.2. In particular, this includes many distributions that Abate and Whitt [4] call "Class II". These are job sizes X such that $\tilde{X}(\gamma(\tilde{X})) < \infty$.

In order to prove Theorems 3.B.1 and 3.B.2 for Class II job size distributions, one would need to assume a regularity condition. We believe it would suffice to assume that $\tilde{X}'$ is regularly varying at $\gamma(\tilde{X})$. The main change to the proofs would be additional casework. For example, it may be that $\tilde{W}_X$ still has a first-order pole, or it may be that it diverges without a pole because $\tilde{X}$ does. See Abate and Whitt [4] and references therein for additional discussion.

More generally, it likely suffices to assume that some higher-order derivative $\tilde{X}^{(n)}$ is regularly varying at $\gamma(\tilde{X})$, as the result of Mimica [82, Corollary 1.3] we use applies to higher derivatives as well. Other results of Mimica [82] may allow one to relax the assumption even further.

Part II

A scaling approach for queueing networks

Introduction

Where in Part I the focus was on a queue in isolation, this part of the thesis considers *queueing networks*. That is, models in which the output of some queues contributes to the input of other queues. As one might expect, such links between queues result in the corresponding queues to exhibit dependent behavior. For instances in which the number of queues is large, this dependence can cause issues regarding numerical tractability. Chapter 4 focuses on an alternative, asymptotic, approach: scaling.

We introduce the work in Chapter 4 by means of two applications of stochastic models. In the first application, analyzed in [122], we consider a manufacturing model that combines the Erlang loss model and the machine-repair model. That is, the machines, which are the servers in the Erlang loss model, may become customers of a repair queue (see Figure II.1). Note that the Erlang loss feature entails that arrivaling products are lost when no machine is idle, and that the machine-repair feature entails that the service provided by the machines is interrupted when it breaks down, only to continue once repaired. In addition, it is assumed that only busy servers may break down.

While both the Erlang loss model and the machine-repair model have incited much research effort, the described combination of the two is significantly harder to analyze. This is due to the dual role of the machines: the serving of products and the breakdown of machines interfere, creating dependence of the two parts of the model. It turns out, however, that when representing the state of the model as the number of idle, busy, and broken machines, it can be viewed as a closed network consisting of three stations (queues) [122, Section 2]. Queue-length analysis of (a generalization of) such a network is the main subject of Chapter 4.

The second application we consider is a vehicle sharing system with multiple pick-up/drop-off locations (in the sequel we omit the term drop-off), as considered in [50]. The vehicles in the system are modeled as customers of a closed queueing network, visualized in Figure II.2. Each pick-up location is represented by a single-server station, and we have an infinite-server station for each ordered pair of pick-up locations (representing vehicles traveling between two pick-up locations).

An unfortunate aspect of the vehicle sharing system is that the number of infinite-server stations is the square of the number of pick-up locations. This leads to a large state space (where each state represents one way in which customers are distributed over the stations), in turn rendering exact computations impossible. For example, the model corresponding to a vehicle sharing system with 50 vehicles and 3 pick-up locations has over $2 \cdot 10^{12}$ states.

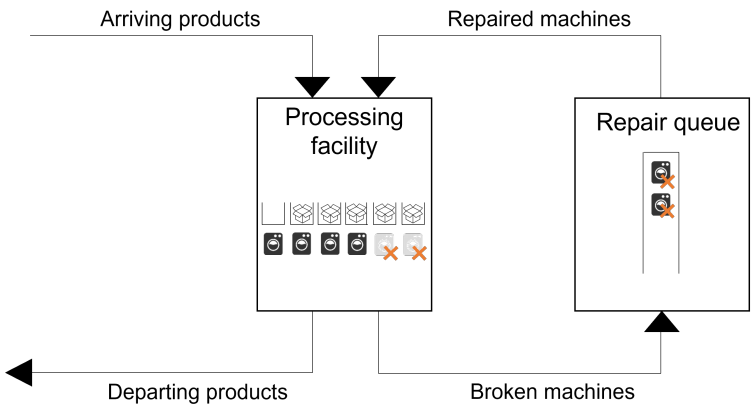

Figure II.1: Manufacturing model where the machines in the processing facility are simultaneously servers (left part of the network) and customers (right part of the network).

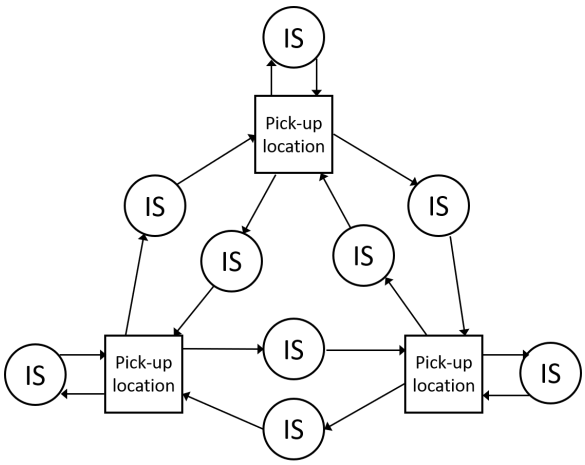

Figure II.2: Vehicle sharing system illustrated as a closed queueing network with single-server stations (pick-up locations) and infinite-server stations (IS).

The problems mentioned in the two examples (dependencies between queues and large state spaces) are exemplary for closed queueing networks in general. To overcome these problems, we consider in Chapter 4 a scaling approach to analyze the queue lengths in these types of networks. The limit result is easy to interpret and can be used as an approximation for the queue-length distributions in the unscaled system. We also show that such approximations are typically quite accurate. Finally, we mention in Section 4.5 the specific implication of our results for the manufacturing model and vehicle sharing system.

4 Scaling limits for closed product-form queueing networks

4.1 Introduction

Queues are often part of larger systems. Aiming to evaluate their performance, a substantial research effort has focused on the analysis of queueing networks, a prominent complication being that the individual queue lengths in the network are often dependent. A central role is played by the class of networks obeying a product-form stationary distribution, where its components correspond to the numbers of customers in the individual queues (also referred to as stations). These product-form networks were first studied in the seminal papers by R. Jackson [63] and J. Jackson [62] in the 1950s, triggering much research in this area. Most notably, a large class of product-form networks, so-called BCMP networks [14], was identified in the 1970s, covering queueing networks consisting of single-server, multi-server and infinite-server stations. Since the discovery of the BCMP class, many further results have been obtained. On one hand, it has been shown that introducing features such as batch routing [24], loss dynamics [12], discrete-time dynamics [25, Chapter 6] and negative customers [49] does not necessarily break the product-form nature of the stationary distribution. On the other hand, general properties of product-form networks have been revealed, such as the arrival theorem [76] and aggregation theorems [23]. For an overview of the queueing-network literature we refer to [25].

Within the study of queueing networks a distinction has been made between open and closed networks. In open networks, i.e., networks with external arrivals and departures, if the stationary distribution factorizes into components corresponding to individual queues, then the queue lengths are mutually independent. Closed networks, however, have the additional constraint that the sum of the queue lengths must equal the population size at any point in time, rendering the individual queue lengths dependent. As a consequence, analytical and numerical difficulties arise when one aims at evaluating performance measures. Closed-form expressions are often beyond reach because the population size constraint complicates the evaluation of terms summing over all possible queue-length vectors, which appear in for instance the normalization constant. Additionally, for large networks numerical approaches face computational challenges, such as the need to evaluate summations over a large set of states. In addition, there is the risk of running into computer-precision related problems, a challenge that has been addressed by Lam [75], but only in relation to the evaluation of the normalization constant.

To remedy the above-mentioned issues arising when analyzing closed product-form queueing networks, in various papers the use of scaling limits has been advocated. Here the objective is to

obtain closed-form distributional results in specific asymptotic regimes. A prominent approach relies on integral representations for the generating function of the stationary distribution, which have been used to asymptotically evaluate the normalization constant [19, 83]. In [79] strong approximation theory is applied to produce limit theorems for a large class of queueing networks. We also refer to the exact-order asymptotic analysis developed in [51].

The current chapter can be seen as part of the research area discussed in the previous paragraph, in that it addresses the analytical and numerical difficulties of closed queueing networks by proposing a scaling method. The general idea of such a method is to make some of the model parameters depend on a number n in a certain way, and let n tend to infinity. When done in a suitable way, one can obtain insightful asymptotic results, revealing a tractable approximation for the behavior of the more complex unscaled system. A prime example of the effectiveness of scaling is the widely-recognized Halfin-Whitt regime [56]. For an Erlang loss queue, this regime scales the workload and the number of servers in a quality-and-efficiency driven way: the utilization of the servers approaches 100%, while the blocking probability remains close to zero. Halfin and Whitt prove that the number of customers in the scaled system tends to a truncated normal random variable in the limit.

For our model, i.e., a general closed product-form network consisting of both single-server stations and infinite-server stations, we define a new scaling regime inspired by the Halfin-Whitt scaling. Indeed, we extend the Halfin-Whitt scaling in such a way that it is applicable to queueing networks rather than individual stations in isolation. This we do by letting the traffic load at all stations become large and choosing the population size in such a way that the joint queue-length distribution has a non-degenerate limit. We remark that for finite-capacity open networks, our regime has the same quality-and-efficiency property as the Halfin-Whitt regime. This study can be considered as an extension of [122], where a similar approach has been followed for a specific three-station closed network representing an extended machine-repair model. We substantially generalize the results from [122], in that we establish similar asymptotic results for a more general class of closed product-form networks.

The contributions of this chapter are the following. Under the Halfin-Whitt inspired scaling, we obtain the asymptotic stationary joint distribution of all queue lengths in the closed product-form network. This specifically entails that, appropriately normalized, the queue lengths of the single-server stations behave as (possibly truncated) exponential random variables, whereas the queue lengths of the infinite-server stations behave as (possibly truncated) normal random variables. Whether the truncation needs to be imposed, depends on whether the queue under consideration is a dominant queue, i.e. the station with largest queue-length variance. In the typical case that there is a single dominant station, the queue lengths are asymptotically independent. Importantly, although the pre-limit stationary distribution is relatively involved, it considerably simplifies in the scaling limit. Furthermore, we observe by means of numerical experiments that for a reasonably sized system, the queue-length distributions are well approximated by their limit distributions.

The chapter is organized as follows. In Section 4.2 we describe our model in detail, analyze the normalization constant and introduce our scaling regime. The main results are then stated and discussed in Section 4.3. The proof of our main theorem is given in Section 4.4, where we

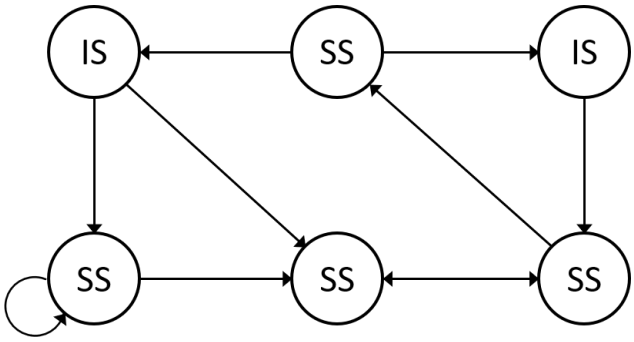

Figure 4.1: Closed queueing network with infinite-server (IS) stations and single-server (SS) stations.

leave some technical details for Appendix 4.A. Subsequently, the practical relevance of our model is discussed in greater detail in Section 4.5, while numerical results in Section 4.6 show that the limiting queue-length distributions are able to yield accurate approximations. We conclude and provide pointers for further research in Section 4.7.

4.2 Model and preliminaries

This section presents the model description and a number of key concepts that play an important role in this chapter. First, Section 4.2.1 introduces the product-form stationary distribution and describes which networks satisfy it. We discuss the normalization constant of this stationary distribution in Section 4.2.2. Subsequently, Section 4.2.3 gives the precise definition of the scaling regime that we study in this chapter.

4.2.1 Model description

We consider a closed queueing network with C customers. Each station can be of two types: R stations are infinite-server queues, while the remaining $K+1$ are single-server queues. At a later stage we omit one single-server station, as its queue length equals C minus the sum of the other queue lengths. It is worth noting that the total service rate provided to all customers in any of the R infinite-server stations is linear in the number of customers present, since all customers can be served simultaneously in an infinite-server queue. In any of the single-server stations, however, the service rate provided is constant whenever the number of customers present is positive, and zero otherwise. See Figure 4.1 for an example of such a network. Let $B_1, \ldots, B_R$ be the stationary numbers of customers at the infinite-server stations, and $D_1, \ldots, D_{K+1}$ their counterparts at the single-server stations.

The only further assumption we impose on our model is that it has the following product-form stationary distribution: for $b_1, \ldots, b_R, d_1, \ldots, d_{K+1}$ such that $b_1 + ... + b_R + d_1 + ... + d_{K+1} = C$,

$$\mathbb{P}(B_1 = b_1, ..., B_R = b_R, D_1 = d_1, ..., D_{K+1} = d_{K+1}) = \tilde{p}_0 \prod_{r=1}^{R} \frac{\eta_r^{b_r}}{b_r!} \prod_{k=1}^{K+1} \theta_k^{d_k}. \tag{4.1}$$

Here $\eta_r, \theta_k \geqslant 0$ are parameters representing traffic loads of individual stations, and $\tilde{p}_0$ is the normalization constant, which ensures that all probabilities sum to 1.

The stationary distribution (4.1) applies under fairly broad conditions. A precise specification providing all instances that yield this particular product form is rather challenging (see [25, Section 5.7] for a detailed discussion). However, an important sufficient network property for this is *quasi-reversibility*. This property, concerning individual stations, states that if a station has a Poisson arrival process, its departure process is also Poisson and the queue length is independent of past departures (see e.g. [70, Section 6] for more background). If all stations of the network are quasi-reversible when considered in isolation, then the stationary distribution is guaranteed to obey the product form (4.1). To give an example, quasi-reversibility holds for infinite-server stations under arbitrary service time distributions, and for single-server stations when service times are exponential under FCFS, or when the server applies processor sharing or the LCFS pre–emptive resume discipline. Quasi-reversibility is a sufficient condition for a product-form stationary distribution, but it is not necessary (cf. [32]).

Although the network described above is assumed to be closed, there is also a class of finite-capacity open networks with Poisson arrivals satisfying (4.1). Such networks are technically open, but behave exactly like a closed network. This concept is explained in great detail in [102], in the context of computer systems with window flow control. A similar reasoning applies to an open queueing network with $R + K$ stations, where external arrivals are blocked when there are already C customers present in the network. To see why this system can be interpreted as a closed network, suppose that all departures out of the system join an artificial single-server station (which we identify with station $K + 1$), and that all external admitted arrivals form the departure process of this station. The system is now closed, and with the total number of customers being equal to C, the behavior at the original stations is the same. Indeed, the original situation where a customer is blocked is equivalent to the situation where no customers are present in station $K + 1$.

We proceed with a few notational issues. Throughout this chapter we will denote vectors by bold symbols. This for instance means that we denote by $\boldsymbol{b}$ and $\boldsymbol{d}$ the vectors $(b_1, ..., b_R)$ and $(d_1, ..., d_K)$, respectively. We also introduce specific notation related to the truncation of such vectors to their first entries: for instance for $r \leqslant R$ we mean by $\boldsymbol{b}_r$ the vector $(b_1, ..., b_r)$, so that $\boldsymbol{b}_R = \boldsymbol{b}$. For the sum of the entries of vectors we use the well-known norm notation, e.g. $||\boldsymbol{b}|| := \sum_{r=1}^{R} b_r$. Furthermore, we use the convention that a geometric random variable has support $\{0, 1, 2, ...\}$; if the success probability is p, we write $\mathscr{G}(p)$. With $\mathscr{P}(\mu)$ we denote a Poisson random variable with mean μ. For sequences f_n and g_n, we write $f_n \sim g_n$ if $f_n/g_n \to 1$ as $n \to \infty$. Additionally, '$\leqslant_{\mathrm{st}}$', '$=_{\mathrm{d}}$' and '$\to_{\mathrm{d}}$', respectively, denote stochastically bounded, equality in distribution and convergence in distribution.

In this chapter, we will work with a normalized version of (4.1). We assume without loss of generality that $\theta_{K+1} = \max_{k=1,...,K+1}\{\theta_k\}$. Due to the closed nature of the network, rescaling all station parameters through division by θ_{K+1} leads to a different normalization constant, but otherwise this has no effect on the stationary joint distribution. Therefore, (4.1) can be rewritten as, with $||\boldsymbol{b}|| + ||\boldsymbol{d}|| \leqslant C$,

$$p_{\boldsymbol{b},\boldsymbol{d}} := \mathbb{P}(\boldsymbol{B} = \boldsymbol{b}, \boldsymbol{D} = \boldsymbol{d}) = p_0 \prod_{r=1}^{R} \frac{\rho_r^{b_r}}{b_r!} \prod_{k=1}^{K} \sigma_k^{d_k}, \tag{4.2}$$

where $p_0 = \theta_{K+1}^C \tilde{p}_0$, $\rho_r = \eta_r/\theta_{K+1}$ for all r and $\sigma_k = \theta_k/\theta_{K+1} \leqslant 1$ for all k. The parameters $\rho_1, ..., \rho_R$ and $\sigma_1, ..., \sigma_K$ can be interpreted as the traffic loads of the corresponding stations. The joint stationary distribution (4.2) is the starting point of the scaling analysis presented in this chapter, and we view $\rho_1, ..., \rho_R$ and $\sigma_1, ..., \sigma_K$ as system parameters. The results in the rest of the chapter are valid for every queueing network that satisfies (4.2).

Remark 4.2.1. In this chapter we consider the system's behavior in a specific regime in which the total number of customers C grows, according to a scaling that we will specify later. By (4.2), there is dependence between the stations: the individual stationary queue lengths are correlated due to the constraint $\|\boldsymbol{B}\| + \|\boldsymbol{D}\| \leqslant C$. However, it also implies that when C grows large, this dependence becomes weaker, and in the limit as $C \to \infty$, vanishes.

For the infinite-server station indexed by $r \leqslant R$, the probability of b_r customers at the station is proportional to $\rho_r^{b_r}/b_r!$. For this reason, the queue length at infinite-server station r is approximately distributed as $\mathscr{P}(\rho_r)$ as C grows large.

Likewise, for a single-server station index by $k \leqslant K$, the probability of d_k customers at the station is proportional to $\sigma_k^{d_k}$. If $\sigma_k < 1$ this implies that the queue-length distribution approximately behaves as $\mathscr{G}(\sigma_k)$ as C grows large.

The remaining station, single-server station $K + 1$, has the highest traffic load among the single-server stations and thus acts as a bottleneck of the network. This means that its queue length becomes arbitrarily large with C. ◇

4.2.2 Normalization constant

We now turn our attention to the calculation of the normalization constant p_0 in (4.2). Note that since $\|\boldsymbol{B}\| + \|\boldsymbol{D}\| \leqslant C$, we have

$$p_0^{-1} = \sum_{\boldsymbol{b},\boldsymbol{d}:\, \|\boldsymbol{b}\|+\|\boldsymbol{d}\| \leqslant C} \prod_{r=1}^{R} \frac{\rho_r^{b_r}}{b_r!} \prod_{k=1}^{K} \sigma_k^{d_k}. \tag{4.3}$$

Observe that the normalization constant involves summation over terms that are products of R Poisson-type factors, and K geometric-type factors. Because of the $R + K$ indices, the number of terms in the summation in (4.3) is $\binom{C+R+K}{R+K}$, making direct calculation of the normalization constant beyond reach for large networks. Alternative approaches have been proposed, such as the method of generating functions [103], for devising efficient algorithms to compute the normalization constant and marginal queue-length distributions. More recently, a generalized method of moments has been studied for closed product-form networks [31].

To provide a different way of decreasing its numerical complexity, we now focus on a simplified representation of the normalization constant, given in Lemma 4.2.2 below. The proofs of the two lemmas in this subsection are not essential for understanding the main results of this chapter, and could be skipped at first reading. We still include them however, since we use the same approach in the proof of our main theorem.

To obtain the different representation for p_0, we evaluate the summation over the K geometric-type indices, and subsequently use a probabilistic argument for the summation over the R

Poisson-type factors. For this purpose, it is useful to define

$$S_j(x) := \sum_{\boldsymbol{b}:\,\|\boldsymbol{b}\|\leqslant C} \prod_{r=1}^{R} \frac{(\rho_r/x)^{b_r}}{b_r!} \sum_{\boldsymbol{d}:\,\|\boldsymbol{d}\|\leqslant C-\|\boldsymbol{b}\|} \prod_{k=1}^{j} \left(\frac{\sigma_k}{x}\right)^{d_k}, \tag{4.4}$$

so that $p_0^{-1} = S_K(1)$. To evaluate the inner geometric sum, we wish to express $p_0^{-1} = S_K(1)$ in terms of $S_0(x)$ for certain x. The following recursion is a key element in this derivation.

Lemma 4.2.1. *For $x \neq \sigma_j$, $S_j(x)$ satisfies the recursion*

$$S_j(x) = \frac{1}{1-\sigma_j/x} S_{j-1}(x) - \frac{(\sigma_j/x)^{C+1}}{1-\sigma_j/x} S_{j-1}(\sigma_j), \qquad j = 1, ..., K.$$

Proof. The recursion follows from an evaluation of the geometric series. Taking the sum over d_j, we see that for $x \neq \sigma_j$,

$$\begin{aligned} S_j(x) &= \sum_{\boldsymbol{b}:\,\|\boldsymbol{b}\|\leqslant C} \prod_{r=1}^{R} \frac{(\rho_r/x)^{b_r}}{b_r!} \sum_{\boldsymbol{d}_{j-1}:\,\|\boldsymbol{d}_{j-1}\|\leqslant C-\|\boldsymbol{b}\|} \prod_{k=1}^{j-1} \left(\frac{\sigma_k}{x}\right)^{d_k} \cdot \frac{1-(\sigma_j/x)^{C-\|\boldsymbol{b}\|-\|\boldsymbol{d}_{j-1}\|+1}}{1-\sigma_j/x} \\ &= \frac{1}{1-\sigma_j/x}\Bigg(\sum_{\boldsymbol{b}:\,\|\boldsymbol{b}\|\leqslant C} \prod_{r=1}^{R} \frac{(\rho_r/x)^{b_r}}{b_r!} \sum_{\boldsymbol{d}_{j-1}:\,\|\boldsymbol{d}_{j-1}\|\leqslant C-\|\boldsymbol{b}\|} \prod_{k=1}^{j-1} \left(\frac{\sigma_k}{x}\right)^{d_k} \\ &\quad - (\sigma_j/x)^{C+1} \sum_{\boldsymbol{b}:\,\|\boldsymbol{b}\|\leqslant C} \prod_{r=1}^{R} \frac{(\rho_r/\sigma_j)^{b_r}}{b_r!} \sum_{\boldsymbol{d}_{j-1}:\,\|\boldsymbol{d}_{j-1}\|\leqslant C-\|\boldsymbol{b}\|} \prod_{k=1}^{j-1} \left(\frac{\sigma_k}{\sigma_j}\right)^{d_k} \Bigg) \\ &= \frac{1}{1-\sigma_j/x} S_{j-1}(x) - \frac{(\sigma_j/x)^{C+1}}{1-\sigma_j/x} S_{j-1}(\sigma_j), \end{aligned} \tag{4.5}$$

thus proving the claim. □

The lemma shows that $S_j(x)$ can be split into two terms, each involving $S_{j-1}(\cdot)$. We exploit this recursion to derive an alternative expression for p_0, which is presented in the following lemma. As an aside, we remark that here and in the sequel, the cases $\sigma_k = 1$ for some k and $\sigma_j = \sigma_l$ for some j, l can be resolved using L'Hôpital's rule.

Lemma 4.2.2. *The normalization constant equals*

$$p_0 = \left(\prod_{k=1}^{K} \frac{1}{1-\sigma_k} \left(S_0(1) - \sum_{l=1}^{K} \sigma_l^{C+1} \prod_{j=1,\, j\neq l}^{K} \frac{1-\sigma_j}{1-\sigma_j/\sigma_l} S_0(\sigma_l) \right) \right)^{-1}, \tag{4.6}$$

where

$$S_0(x) = \sum_{i=0}^{C} \frac{(\|\boldsymbol{\rho}\|/x)^i}{i!}.$$

Proof. Note that by (4.3) and Lemma 4.2.1,

$$p_0^{-1} = S_K(1) = \frac{1}{1-\sigma_K} S_{K-1}(1) - \frac{\sigma_K^{C+1}}{1-\sigma_K} S_{K-1}(\sigma_K).$$

Applying Lemma 4.2.1 another $K-1$ times leads to an expression of the form

$$p_0^{-1} = a\, S_0(1) + \sum_{l=1}^{K} u_l\, S_0\left(\sigma_l\right), \tag{4.7}$$

where a and $u_1, ..., u_K$ are coefficients depending on $\sigma_1, ..., \sigma_K$. To find a, observe that the only term with $S_0(1)$ results from the first term of Lemma 4.2.1 of all K iterations. Therefore, $a = \prod_{k=1}^{K}(1-\sigma_k)^{-1}$. Similarly, observe that the only term with $S_0\left(\sigma_K\right)$ follows from the second term in the first iteration and then the first term in all remaining iterations. Therefore, $u_K = -\sigma_K^{C+1}(1-\sigma_K)^{-1}\prod_{j=1}^{K-1}(1-\sigma_j/\sigma_K)^{-1}$. Note that the single-server stations $1, ..., K$ are identical in (4.3) up to their parameters $\sigma_1, ..., \sigma_K$. By symmetry, we conclude that, for any $l = 1, \ldots, K$,

$$u_l = -\frac{\sigma_l^{C+1}}{1-\sigma_l}\prod_{j=1,\, j\neq l}^{K}\frac{1}{1-\sigma_j/\sigma_l}.$$

Thus, it holds that

$$\begin{aligned} p_0^{-1} &= S_0(1)\prod_{k=1}^{K}\frac{1}{1-\sigma_k} - \sum_{l=1}^{K} S_0\left(\sigma_l\right)\frac{\sigma_l^{C+1}}{1-\sigma_l}\prod_{j=1,\, j\neq l}^{K}\frac{1}{1-\sigma_j/\sigma_l} \\ &= \left(\prod_{k=1}^{K}\frac{1}{1-\sigma_k}\right)\cdot\left(S_0(1) - \sum_{l=1}^{K} S_0\left(\sigma_l\right)\sigma_l^{C+1}\prod_{j=1,\, j\neq l}^{K}\frac{1-\sigma_j}{1-\sigma_j/\sigma_l}\right). \end{aligned} \tag{4.8}$$

To prove Lemma 4.2.2, it remains to show that $S_0(x) = \sum_{i=0}^{C}\left(\|\boldsymbol{\rho}\|/x\right)^i/i!$. This is done by expressing $S_0(x)$ in terms of cumulative Poisson probabilities. That is, using (4.4),

$$\begin{aligned} S_0(x) &= \sum_{\boldsymbol{b}:\|\boldsymbol{b}\|\leqslant C}\prod_{r=1}^{R}\frac{(\rho_r/x)^{b_r}}{b_r!} \\ &= e^{\|\boldsymbol{\rho}\|/x}\sum_{\boldsymbol{b}:\|\boldsymbol{b}\|\leqslant C}\mathbb{P}\left(\mathscr{P}\left(\frac{\rho_1}{x}\right) = b_1, ..., \mathscr{P}\left(\frac{\rho_R}{x}\right) = b_R\right) \\ &= e^{\|\boldsymbol{\rho}\|/x}\mathbb{P}\left(\sum_{r=1}^{R}\mathscr{P}\left(\frac{\rho_r}{x}\right) \leqslant C\right) = e^{\|\boldsymbol{\rho}\|/x}\mathbb{P}\left(\mathscr{P}\left(\frac{\|\boldsymbol{\rho}\|}{x}\right) \leqslant C\right) = \sum_{i=0}^{C}\frac{\left(\|\boldsymbol{\rho}\|/x\right)^i}{i!}, \end{aligned}$$

which concludes the proof. □

The number of numerical operations needed to evaluate the normalization constant using (4.6) is $O(K^2 + KC)$. The algorithm of [103] has a complexity of $O(RC^2 + KC)$, and has the benefit that marginal queue lengths follow without much additional computational effort. However, when knowledge of the joint distribution is required, one cannot avoid computing the individual probabilities of all $\binom{C+R+K}{R+K}$ states, which is tractable only for very small networks. We resolve this issue by working in a scaling regime, that will be introduced in the next subsection, and in which the stationary distribution exhibits easy-to-interpret behavior. The asymptotic findings can be used to devise approximations for the unscaled system, as will be pointed out in Section 4.6.

4.2.3 The scaling regime

When distributions do not allow a closed-form analysis, a commonly used approach in applied probability is to resort to scaling limits. The main idea is to parametrize (a subset of) the system parameters by n, with the objective to arrive at an explicit limiting distribution as $n \to \infty$. It is often not *a priori* clear how this parametrization should be done; finding a scaling that leads to useful and meaningful results in the limit is an art on its own. The resulting limiting distribution can be used to produce approximations for the pre-limit system.

In their celebrated 1981 paper, Halfin and Whitt [56] introduced an important new scaling for many-server queues. The asymptotic regime considered corresponds to letting the workload ρ and the number of servers C grow to infinity in such a way that $(C-\rho)/\sqrt{\rho}$ converges to a constant $\bar{\beta} > 0$. In the specific context of the Erlang loss model, an appropriately normalized version of the queue length then asymptotically behaves as a normal random variable truncated at $\bar{\beta}$ [56, Theorem 3]. The scaling we impose in our network setting is inspired by the Halfin-Whitt regime, in that we also scale the parameter ρ and the total number of customers C.

We now give a precise definition of the scaling we impose. Let $\nu_1, ..., \nu_R, \alpha_1, ..., \alpha_K \in \mathbb{R}$ and $w_1, ..., w_R, c_1, ..., c_K > 0$ be scaling parameters for individual stations. It proves useful to assume without loss of generality that $\nu_1 \geqslant ... \geqslant \nu_R$ and that $\alpha_1 \leqslant ... \leqslant \alpha_K$. We scale the system parameters as follows:

- we replace ρ_r by $\rho_r^{(n)}$ for each $r \in \{1, ..., R\}$,
- we replace σ_k by $\sigma_k^{(n)}$ for each $k \in \{1, ..., K\}$,
- we replace C by C_n,

where

$$\rho_r^{(n)} = w_r\, n^{\nu_r}, \quad \sigma_k^{(n)} := \frac{n}{n + c_k n^{\alpha_k}},$$

which can be interpreted as scaled traffic loads, and where for $\beta > 0$ and $\gamma := \max\{1-\alpha_1, \frac{1}{2}\nu_1\}$, we let the total number of customers be defined as

$$C_n := \left\lfloor \left\| \boldsymbol{\rho}^{(n)} \right\| + \beta n^{\gamma} \right\rfloor. \tag{4.9}$$

The definitions for $\rho_r^{(n)}$, $\sigma_k^{(n)}$ and C_n may seem restrictive, but any network satisfying (4.2) can be constructed with the right choices of the scaling parameters. For a detailed discussion on fitting these parameters to a system in practice, we refer to Section 4.6.2.

Observe that in this scaling regime, the traffic loads of the infinite-server stations become arbitrarily large as $n \to \infty$ (provided $\nu_1, ..., \nu_R > 0$), and the traffic loads of the single-server stations tend to 1 as $n \to \infty$ (provided $\alpha_1, ..., \alpha_K < 1$). To account for the large queue lengths that are inherent for these traffic loads, also the population size C_n grows (at a suitable pace) as $n \to \infty$.

Our precise choice (4.9) for C_n can be motivated as follows. It turns out that, to get non-degenerate limits, the total number of customers should be picked such that it equals the

mean of $\|\boldsymbol{B}\|$ increased by a constant β times the largest of the standard deviations of all queue lengths. As argued in Remark 4.2.1, when the total number of customers is large, B_r behaves as $\mathscr{P}(\rho_r) = \mathscr{P}(\rho_r^{(n)})$, which has standard deviation $(\rho_r^{(n)})^{\frac{1}{2}} = \sqrt{w_r}\, n^{\frac{1}{2}\nu_r}$. In addition, when $\sigma_k < 1$, D_k behaves as $\mathscr{G}(1-\sigma_k) = \mathscr{G}(1-\sigma_k^{(n)})$, which has standard deviation

$$\frac{\sqrt{\sigma_k^{(n)}}}{1-\sigma_k^{(n)}} \sim \frac{1}{c_k} n^{1-\alpha_k}.$$

Since $\nu_1 \geqslant \ldots \geqslant \nu_R$ and $\alpha_1 \leqslant \ldots \leqslant \alpha_K$, the largest asymptotic standard deviation is attained by either $B_1^{(n)}$ or $D_1^{(n)}$. Note that the first case applies if $1-\alpha_1 < \frac{1}{2}\nu_1$, and the second if $1-\alpha_1 > \frac{1}{2}\nu_1$. These observations, and the fact that the population size must be an integer, intuitively explain our choice (4.9) for C_n.

In line with the observations above, we say that a station is *dominant* if its asymptotic queue-length variance has the largest power of n out of all the stations (excluding single-server station $K+1$). Thus, infinite-server station $r \in \{1, \ldots, R\}$ is dominant if $\frac{1}{2}\nu_r = \gamma$, and single-server station $k \in \{1, \ldots, K\}$ is dominant if $1-\alpha_k = \gamma$. Later it will turn out that in the limit as $n \to \infty$, it is precisely the dominant stations that are affected by the population size constraint.

Under our scaling the queue lengths $\boldsymbol{B}$ at the infinite-server stations and the queue lengths $\boldsymbol{D}$ at the single-server stations depend on n. In the sequel we let $\boldsymbol{B}^{(n)}$ and $\boldsymbol{D}^{(n)}$, respectively, denote the corresponding random vectors. Our objective is to analyze their behavior as $n \to \infty$. Since their means may become arbitrarily large with n, we consider normalized versions: for $r \in \{1, \ldots, R\}$ and $k \in \{1, \ldots, K\}$,

$$\bar{B}_r^{(n)} := \frac{B_r^{(n)} - \rho_r^{(n)}}{\sqrt{\rho_r^{(n)}}}, \qquad \text{and} \qquad \bar{D}_k^{(n)} := (1-\sigma_k^{(n)}) D_k^{(n)}. \tag{4.10}$$

Note that in both cases we scale by a factor of the order of the standard deviation. For $\bar{B}_r^{(n)}$, a term representing the mean is subtracted, similarly to the central limit theorem.

To help the reader understand the key concepts in this chapter, a table of the most important scaling parameters is included in Appendix 4.D.

4.3 Results

In this section, we derive the asymptotic joint distribution of $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$. The remaining queue length, $D_{K+1}^{(n)}$, can then be obtained from the identity

$$\left\|\boldsymbol{B}^{(n)}\right\| + \left\|\boldsymbol{D}^{(n)}\right\| + D_{K+1}^{(n)} = C_n. \tag{4.11}$$

Our main result, which holds for $\nu_1, \ldots, \nu_R > 0$ and $\alpha_1, \ldots, \alpha_K < 1$, is presented in Section 4.3.1. Section 4.3.2 presents an adaptation of our main result for networks with only single-server stations. The remaining cases, where $\nu_R \leqslant 0$ or $\alpha_K \geqslant 1$, are investigated in Section 4.3.3.

4.3.1 Main result

We first consider the case where the traffic loads are large at all stations. That is, we assume that $\nu_1, ..., \nu_R > 0$ (the traffic loads at infinite-server stations tend to infinity) and $\alpha_1, ..., \alpha_K < 1$ (the traffic loads at single-server stations tend to 1). We study the normalized queue lengths $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$ by means of their joint Laplace-Stieltjes transform (LST). We define this LST using (4.10):

$$\begin{aligned} P_n(\boldsymbol{s},\boldsymbol{t}) &:= \mathbb{E}\left(\prod_{r=1}^{R} e^{-s_r \bar{B}_r^{(n)}} \prod_{k=1}^{K} e^{-t_k \bar{D}_k^{(n)}}\right) \\ &= \sum_{\boldsymbol{b},\boldsymbol{d}:\, \|\boldsymbol{b}\|+\|\boldsymbol{d}\| \leqslant C_n} \left(\prod_{r=1}^{R} e^{-s_r \frac{b_r-\rho_r^{(n)}}{\sqrt{\rho_r^{(n)}}}}\right) \cdot \left(\prod_{k=1}^{K} e^{-t_k(1-\sigma_k^{(n)})d_k}\right) \mathbb{P}(\boldsymbol{B}^{(n)} = \boldsymbol{b}, \boldsymbol{D}^{(n)} = \boldsymbol{d}). \end{aligned} \tag{4.12}$$

It is noted that strictly speaking $P_n(\boldsymbol{s},\boldsymbol{t})$ is not an LST, as the random variables $\bar{B}_r^{(n)}$ may attain negative values. This feature does not affect the upcoming analysis, including the application of Lévy's convergence theorem, and hence we will stick to the term LST.

Our main theorem, to be proved in Section 4.4, gives an explicit expression for the limit of $P_n(\boldsymbol{s},\boldsymbol{t})$ as $n \to \infty$, from which we can directly derive the asymptotic distribution of $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$.

A bit of notation is necessary for the statement of the main theorem. First, for a standard-normally distributed random variable, we write

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

for its density function and $\Psi(x) := \Phi(x)/\varphi(x)$ for its *Mills ratio*, i.e. its distribution function divided by its density function. Secondly, concerning the highest-loaded stations, let $R^- \leqslant R$ be the largest integer such that $\nu_1 = \nu_2 = ... = \nu_{R^-}$ and let $K^- \leqslant K$ be the largest integer such that $\alpha_1 = \alpha_2 = ... = \alpha_{K^-}$. Observe that the number of dominant stations can now be expressed as

$$R^- \mathbb{1}_{\{1-\alpha_1 \leqslant \frac{1}{2}\nu_1\}} + K^- \mathbb{1}_{\{1-\alpha_1 \geqslant \frac{1}{2}\nu_1\}}.$$

We then denote

$$W := \sum_{r=1}^{R^-} w_r \quad \text{and} \quad \lambda(\boldsymbol{s}) := \frac{\beta + \sum_{r=1}^{R^-} s_r \sqrt{w_r}}{\sqrt{W}}.$$

In addition, define

$$\begin{aligned} \kappa_{jl}(\boldsymbol{t}) &:= \frac{c_j(1+t_j)}{c_j(1+t_j) - c_l(1+t_l)}, \quad \zeta(\boldsymbol{t}) := 1 - \sum_{l=1}^{K^-} \left(\prod_{j=1,\, j\neq l}^{K^-} \kappa_{jl}(\boldsymbol{t})\right) e^{-\beta c_l(1+t_l)}, \\ \eta(\boldsymbol{s},\boldsymbol{t}) &:= \Psi\left(\lambda(\boldsymbol{s})\right) - \sum_{l=1}^{K^-} \left(\prod_{j=1,\, j\neq l}^{K^-} \kappa_{jl}(\boldsymbol{t})\right) \Psi\left(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\right), \\ \xi(\boldsymbol{s}) &:= \Phi\left(\lambda(\boldsymbol{s})\right) \quad \text{and} \quad \chi(\boldsymbol{s}) := \varphi(\lambda(\boldsymbol{s})). \end{aligned}$$

With this notation, we can state the main theorem as follows.

Theorem 4.3.1. *Consider a queueing network with stationary distribution* (4.2)*. Assume that* $\nu_R > 0$ *and that* $\alpha_K < 1$*. Then the joint LST* $P_n(\boldsymbol{s},\boldsymbol{t})$ *of* $(\bar{\boldsymbol{B}}^{(n)},\bar{\boldsymbol{D}}^{(n)})$ *satisfies*

$$\lim_{n\to\infty} P_n(\boldsymbol{s},\boldsymbol{t}) = \left(\prod_{r=1}^{R} e^{\frac{1}{2}s_r^2}\right)\left(\prod_{k=1}^{K}\frac{1}{1+t_k}\right)\cdot U(\boldsymbol{s},\boldsymbol{t}), \tag{4.13}$$

where

$$U(\boldsymbol{s},\boldsymbol{t}) := \begin{cases} \dfrac{\zeta(\boldsymbol{t})}{\zeta(\boldsymbol{0})} & \text{if } 1-\alpha_1 > \frac{1}{2}\nu_1, \\[2ex] \dfrac{\chi(\boldsymbol{s})}{\chi(\boldsymbol{0})}\cdot\dfrac{\eta(\boldsymbol{s},\boldsymbol{t})}{\eta(\boldsymbol{0},\boldsymbol{0})} & \text{if } 1-\alpha_1 = \frac{1}{2}\nu_1, \\[2ex] \dfrac{\xi(\boldsymbol{s})}{\xi(\boldsymbol{0})} & \text{if } 1-\alpha_1 < \frac{1}{2}\nu_1. \end{cases}$$

In all three cases we recognize known distributions from the joint LST. Let $\mathscr{N}_1,\ldots,\mathscr{N}_R,\mathscr{E}_1,\ldots,\mathscr{E}_K$ be independent random variables, the first R having standard-normal distributions and the last K having unit-rate exponential distributions. In the corollary below we claim that the right-hand side of (4.13) is the LST of the $(R+K)$-tuple $(\boldsymbol{\mathscr{N}},\boldsymbol{\mathscr{E}})$ conditioned on $Z(\boldsymbol{\mathscr{N}}_{R^-},\boldsymbol{\mathscr{E}}_{K^-}) \leqslant \beta$, where

$$Z(\boldsymbol{\mathscr{N}}_{R^-},\boldsymbol{\mathscr{E}}_{K^-}) := \mathbb{1}_{\{1-\alpha_1\leqslant\frac{1}{2}\nu_1\}}\sum_{r=1}^{R^-}\sqrt{w_r}\mathscr{N}_r + \mathbb{1}_{\{1-\alpha_1\geqslant\frac{1}{2}\nu_1\}}\sum_{k=1}^{K^-}\frac{1}{c_k}\mathscr{E}_k. \tag{4.14}$$

To provide some intuition why this condition makes sense, consider the population size constraint $\left\|\boldsymbol{B}^{(n)}\right\| + \left\|\boldsymbol{D}^{(n)}\right\| \leqslant C_n$. Subtracting $\left\|\boldsymbol{\rho}^{(n)}\right\|$ and dividing by n^γ on both sides, we have

$$\frac{\left\|\boldsymbol{B}^{(n)}\right\| - \left\|\boldsymbol{\rho}^{(n)}\right\|}{n^\gamma} + \frac{\left\|\boldsymbol{D}^{(n)}\right\|}{n^\gamma} \leqslant \frac{C_n - \left\|\boldsymbol{\rho}^{(n)}\right\|}{n^\gamma}. \tag{4.15}$$

Recalling the definition $\gamma := \max\{1-\alpha_1, \frac{1}{2}\nu_1\}$ and the scaled queue lengths (4.10), if $\bar{B}_r^{(n)} \to_{\mathrm{d}} \mathscr{N}_r$ for each r and $\bar{D}_k^{(n)} \to_{\mathrm{d}} \mathscr{E}_k$ for each k, the inequality (4.15) would converge, as $n\to\infty$, to $Z(\boldsymbol{\mathscr{N}}_{R^-},\boldsymbol{\mathscr{E}}_{K^-}) \leqslant \beta$.

To summarize, Theorem 4.3.1 leads to the following corollary.

Corollary 4.3.1. *As* $n\to\infty$*,*

$$(\bar{\boldsymbol{B}}^{(n)},\bar{\boldsymbol{D}}^{(n)}) \to_{\mathrm{d}} \Big(\boldsymbol{\mathscr{N}},\boldsymbol{\mathscr{E}} \mid Z(\boldsymbol{\mathscr{N}}_{R^-},\boldsymbol{\mathscr{E}}_{K^-}) \leqslant \beta\Big). \tag{4.16}$$

Consequently, the random variables $\bar{B}_1^{(n)},\ldots,\bar{B}_R^{(n)},\bar{D}_1^{(n)},\ldots,\bar{D}_K^{(n)}$ *are asymptotically independent if the number of dominant stations is one.*

Proof. With standard integration techniques one can check that the joint Laplace-Stieltjes transform of the tuple $(\boldsymbol{\mathscr{N}},\boldsymbol{\mathscr{E}} \mid Z(\boldsymbol{\mathscr{N}}_{R^-},\boldsymbol{\mathscr{E}}_{K^-}) \leqslant \beta)$ is precisely as given in Theorem 4.3.1 (see Appendix 4.C), so that the stated follows by Lévy's convergence theorem. The independence statement follows from the fact that if there is only one dominant station, then Z depends on just one random variable. □

It can be seen from (4.14) that the condition $Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \leqslant \beta$, relating to the condition that there are at most C_n customers at the $R+K$ stations, only involves indices corresponding to the dominant stations. Hence in the limit, the population size constraint only affects the dominant stations. In addition, suppose that the number of dominant stations is one, which happens precisely if $\frac{1}{2}\nu_1 > \max\{\nu_2, 1-\alpha_1\}$ or $1-\alpha_1 > \max\{\nu_1, 1-\alpha_2\}$. In that case, the condition applies to only one random variable, which amounts to a truncation of that variable.

Remark 4.3.1. We mention the consequences of our main result for single-server station $K+1$. As described earlier, this is the single-server station with the highest traffic load, the queue length of which follows from the remaining queue lengths by the population size constraint. Corollary 4.3.1 thus implicitly provides the asymptotic distribution of $D_{K+1}^{(n)}$. Dividing the population size constraint (4.11) by n^γ, we have that

$$\frac{D_{K+1}^{(n)}}{n^\gamma} = \frac{C_n - \left\|\boldsymbol{\rho}^{(n)}\right\|}{n^\gamma} - \frac{\left\|\boldsymbol{B}^{(n)}\right\| - \left\|\boldsymbol{\rho}^{(n)}\right\|}{n^\gamma} - \frac{\left\|\boldsymbol{D}^{(n)}\right\|}{n^\gamma}.$$

By (4.9), (4.10) and Corollary 4.3.1, it thus holds that

$$\frac{D_{K+1}^{(n)}}{n^\gamma} \to_{\mathrm{d}} \beta - \left(Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \mid Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \leqslant \beta\right)$$

as $n \to \infty$. ◇

4.3.2 Scaling result for $R = 0$

So far we have omitted networks consisting of single-server stations only, because Theorem 4.3.1 relies on the value of ν_1. In the case that $R = 0$, however, this parameter does not exist. With a slight modification, we can establish the counterpart of Corollary 4.3.1 for single-server networks.

Corollary 4.3.2. *Suppose $R = 0$ and $\alpha_K < 1$. As $n \to \infty$,*

$$\bar{\boldsymbol{D}}^{(n)} \to_{\mathrm{d}} \left(\mathscr{E} \mid Z(\mathbf{0}, \mathscr{E}_{K^-}) \leqslant \beta\right). \tag{4.17}$$

The variables $\bar{D}_1^{(n)}, \ldots, \bar{D}_K^{(n)}$ are thus asymptotically independent if $K^- = 1$.

Proof. The result follows from Corollary 4.3.1 by setting $\nu_1 = -\infty$. □

4.3.3 Scaling results for $\nu_R \leqslant 0$ and/or $\alpha_K \geqslant 1$

Since Theorem 4.3.1 only covers the case where $\nu_R > 0$ and $\alpha_K < 1$, it remains to analyze its complement in which $\nu_R \leqslant 0$ and/or $\alpha_K \geqslant 1$. Recall that in Section 4.2.3, we introduced normalized versions of $\boldsymbol{B}^{(n)}$ and $\boldsymbol{D}^{(n)}$ in order to preserve finite mean. Note however that for infinite-server stations r with $\nu_r \leqslant 0$ and for single-server stations k with $\alpha_k \geqslant 1$, the *unnormalized* queue length converges to a finite-mean random variable. Because of this, it is no longer necessary to normalize in these cases. For all $r \in \{1, \ldots, R\}$ for which $\nu_r \leqslant 0$, we will

therefore consider the distribution of the random variable $B_r^{(n)}$ instead of $\bar{B}_r^{(n)}$. Likewise, for all $k \in \{1, ..., K\}$ for which $\alpha_k \geqslant 1$, we will consider the distribution of the random variable $D_k^{(n)}$ instead of $\bar{D}_k^{(n)}$.

In this regime, a statement similar to Corollary 4.3.1 holds, which is given in the following corollary.

Corollary 4.3.3. *Assume that $\nu_1 > 0$ or $\alpha_1 < 1$. Let I be the smallest integer such that $\nu_I \leqslant 0$, and let J be the smallest integer such that $\alpha_J \geqslant 1$. As $n \to \infty$,*

$$\left(\bar{B}_1^{(n)}, ..., \bar{B}_{I-1}^{(n)}, \bar{D}_1^{(n)}, ..., \bar{D}_{J-1}^{(n)}\right) \to_{\mathrm{d}} \left(\mathscr{N}_1, ..., \mathscr{N}_{I-1}, \mathscr{E}_1, ..., \mathscr{E}_{J-1} \mid Z\left(\boldsymbol{\mathscr{N}}_{R^-}, \boldsymbol{\mathscr{E}}_{K^-}\right) \leqslant \beta\right).$$

Proof. See Appendix 4.B. □

Remark 4.3.2. The remaining random variables, i.e. $B_I^{(n)}, ..., B_R^{(n)}$ and $D_J^{(n)}, ..., D_K^{(n)}$, all have finite mean because $\nu_I, ..., \nu_R \leqslant 0$ and $\alpha_J, ..., \alpha_K \geqslant 1$. In the proof of Corollary 4.3.3 we will see that they behave as Poisson random variables with means $\rho_I^{(n)}, ..., \rho_R^{(n)}$ and geometric random variables with parameters $1 - \sigma_J^{(n)}, ..., 1 - \sigma_K^{(n)}$, respectively. This implies in particular that, for each r such that $\nu_r < 0$ and for each k such that $\alpha_k > 1$, the random variables $B_r^{(n)}$ and $D_k^{(n)}$ become degenerate with value 0 as $n \to \infty$. ◇

Remark 4.3.3. Corollary 4.3.3 assumes that $\nu_1 > 0$ or $\alpha_1 < 1$ because our scaling would not make sense otherwise. If $\nu_1 \leqslant 0$ and $\alpha_1 \geqslant 1$, the stations' traffic loads no longer increase with n. ◇

4.4 Proof of Theorem 4.3.1

In this section we present a proof of our main theorem, Theorem 4.3.1, which consists of two parts. First, in Section 4.4.1, we derive a structured expression for $P_n(\boldsymbol{s}, \boldsymbol{t})$ (Lemma 4.4.1) relying on techniques similar to those used in the proof of Lemma 4.2.2 (the derivation of the normalization constant). Then, we discuss this expression piece by piece, already recognizing some known LSTs and providing intuition.

In the second part of the proof (Section 4.4.2), we asymptotically analyze in Lemmas 4.4.3-4.4.6 all parts of the expression obtained from Lemma 4.4.1. In most cases, we can build on a version of the central limit theorem (Lemma 4.4.2) to find the asymptotics. One particular case, however, requires more subtle asymptotic bounds, and this case is treated in Lemma 4.4.6. We finish the proof by substituting the asymptotically analyzed parts back into the expression for $P_n(\boldsymbol{s}, \boldsymbol{t})$.

In the proofs, some mathematical expressions will repeatedly appear in our calculations. To keep these calculations readable, we use the following notation.

- The adapted traffic load for infinite-server station r:

$$\zeta_r^{(n)}(s_r) := \rho_r^{(n)} \exp\left(-\frac{s_r}{\sqrt{\rho_r^{(n)}}}\right), \quad \zeta^{(n)}(\boldsymbol{s}) := \sum_{r=1}^{R} \zeta_r^{(n)}(s_r).$$

- The adapted traffic load for single-server station l:

$$\delta_l^{(n)}(t_l) := \sigma_l^{(n)} e^{-t_l(1-\sigma_l^{(n)})}.$$

- A frequently occurring quantity related to the single-server stations j and l:

$$y_{jl}^{(n)}(t_j, t_l) := \frac{1-\delta_j^{(n)}(t_j)}{1-\delta_j^{(n)}(t_j)/\delta_l^{(n)}(t_l)}.$$

- A Poisson probability related to $S_0(1)$:

$$f^{(n)}(\boldsymbol{s}) := \mathbb{P}\left(\mathscr{P}\left(\zeta^{(n)}(\boldsymbol{s})\right) \leqslant C_n\right).$$

- A Poisson probability related to $S_0(\delta_l^{(n)}(t_l))$:

$$g_l^{(n)}(\boldsymbol{s}, t_l) := \mathbb{P}\left(\mathscr{P}\left(\zeta^{(n)}(\boldsymbol{s})/\delta_l^{(n)}(t_l)\right) \leqslant C_n\right).$$

- A quantity appearing in $P_n(\boldsymbol{s}, \boldsymbol{t})$:

$$h_l^{(n)}(\boldsymbol{s}, t_l) := \exp\left(\zeta^{(n)}(\boldsymbol{s})\left(\frac{1}{\delta_l^{(n)}(t_l)} - 1\right)\right) \cdot \left(\delta_l^{(n)}(t_l)\right)^{C_n+1}.$$

4.4.1 Structured form of LST

The following lemma gives an exact expression for the LST $P_n(\boldsymbol{s}, \boldsymbol{t})$ in terms of the new notation that was introduced above, and forms the backbone of the proof of Theorem 4.3.1.

Lemma 4.4.1. *The LST of* $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$ *satisfies*

$$\begin{aligned} P_n(\boldsymbol{s}, \boldsymbol{t}) = &\left(\prod_{r=1}^{R} e^{-\rho_r^{(n)} + s_r\sqrt{\rho_r^{(n)}} + \zeta_r^{(n)}(s_r)}\right) \times \left(\prod_{k=1}^{K} \frac{1-\sigma_k^{(n)}}{1-\delta_k^{(n)}(t_k)}\right) \\ &\times \frac{f^{(n)}(\boldsymbol{s}) - \sum_{l=1}^{K}\left(\prod_{j=1, j\neq l}^{K} y_{jl}^{(n)}(t_j, t_l)\right) g_l^{(n)}(\boldsymbol{s}, t_l)\, h_l^{(n)}(\boldsymbol{s}, t_l)}{f^{(n)}(\boldsymbol{0}) - \sum_{l=1}^{K}\left(\prod_{j=1, j\neq l}^{K} y_{jl}^{(n)}(0,0)\right) g_l^{(n)}(\boldsymbol{0}, 0)\, h_l^{(n)}(\boldsymbol{0}, 0)}. \end{aligned} \tag{4.18}$$

Proof. Denote by $p_{\boldsymbol{b},\boldsymbol{d}}^{(n)}$ and $p_0^{(n)}$, respectively, the stationary distribution and normalization constant of the scaled system. Then, we can rewrite the joint LST of $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$ in (4.12) as

$$\begin{aligned} P_n(\boldsymbol{s}, \boldsymbol{t}) &= \sum_{\boldsymbol{b},\boldsymbol{d}:\, \|\boldsymbol{b}\|+\|\boldsymbol{d}\| \leqslant C_n} \left(\prod_{r=1}^{R} e^{-s_r \frac{b_r - \rho_r^{(n)}}{\sqrt{\rho_r^{(n)}}}}\right) \cdot \left(\prod_{k=1}^{K} e^{-t_k(1-\sigma_k^{(n)})d_k}\right) p_{\boldsymbol{b},\boldsymbol{d}}^{(n)} \\ &= p_0^{(n)} \left(\prod_{r=1}^{R} e^{s_r\sqrt{\rho_r^{(n)}}}\right) \sum_{\boldsymbol{b}:\, \|\boldsymbol{b}\| \leqslant C_n} \left(\prod_{r=1}^{R} \frac{\left(\zeta_r^{(n)}(s_r)\right)^{b_r}}{b_r!}\right) \sum_{\boldsymbol{d}:\, \|\boldsymbol{d}\| \leqslant C_n - \|\boldsymbol{b}\|} \left(\prod_{k=1}^{K} \left(\delta_k^{(n)}(t_k)\right)^{d_k}\right) \\ &= p_0^{(n)} \left(\prod_{r=1}^{R} e^{s_r\sqrt{\rho_r^{(n)}}}\right) S_K^{(n)}(1), \end{aligned}$$

where $S_j^{(n)}(x)$ is obtained from $S_j(x)$ when (ρ_r, σ_k, C) is replaced by $(\zeta_r^{(n)}(s_r), \delta_k^{(n)}(t_k), C_n)$. Therefore we have by (4.8) that

$$\begin{aligned}
P_n(\boldsymbol{s},\boldsymbol{t}) &= p_0^{(n)}\left(\prod_{r=1}^{R} e^{s_r\sqrt{\rho_r^{(n)}}}\right)\cdot\left(\prod_{k=1}^{K}\frac{1}{1-\delta_k^{(n)}(t_k)}\right)\\
&\quad\times\left(S_0^{(n)}(1)-\sum_{l=1}^{K}\left(\delta_l^{(n)}(t_l)\right)^{C_n+1}\left(\prod_{j=1,\,j\neq l}^{K} y_{jl}^{(n)}(t_j,t_l)\right)S_0^{(n)}\left(\delta_l^{(n)}(t_l)\right)\right)\\
&= p_0^{(n)}\left(\prod_{r=1}^{R} e^{s_r\sqrt{\rho_r^{(n)}}}\right)\cdot\left(\prod_{k=1}^{K}\frac{1}{1-\delta_k^{(n)}(t_k)}\right)\cdot\Bigg(e^{\zeta^{(n)}(\boldsymbol{s})}\mathbb{P}\left(\mathscr{P}\left(\zeta^{(n)}(\boldsymbol{s})\right)\leqslant C_n\right)\\
&\quad-\sum_{l=1}^{K}\left(\delta_l^{(n)}(t_l)\right)^{C_n+1}\left(\prod_{j=1,\,j\neq l}^{K} y_{jl}^{(n)}(t_j,t_l)\right)e^{\frac{1}{\delta_l^{(n)}(t_l)}\zeta^{(n)}(\boldsymbol{s})}\mathbb{P}\left(\mathscr{P}\left(\frac{1}{\delta_l^{(n)}(t_l)}\zeta^{(n)}(\boldsymbol{s})\right)\leqslant C_n\right)\Bigg)\\
&= p_0^{(n)}\left(\prod_{r=1}^{R} e^{s_r\sqrt{\rho_r^{(n)}}+\zeta_r^{(n)}(s_r)}\right)\cdot\left(\prod_{k=1}^{K}\frac{1}{1-\delta_k^{(n)}(t_k)}\right)\\
&\quad\times\left(f^{(n)}(\boldsymbol{s})-\sum_{l=1}^{K}\left(\prod_{j=1,\,j\neq l}^{K} y_{jl}^{(n)}(t_j,t_l)\right)g_l^{(n)}(\boldsymbol{s},t_l)\,h_l^{(n)}(\boldsymbol{s},t_l)\right).
\end{aligned}\tag{4.19}$$

Using $P_n(\mathbf{0},\mathbf{0}) = 1$ we find that $p_0^{(n)}$ equals

$$\left(\prod_{r=1}^{R} e^{-\rho_r^{(n)}}\right)\cdot\left(\prod_{k=1}^{K}\left(1-\sigma_k^{(n)}\right)\right)\cdot\left(f^{(n)}(\mathbf{0})-\sum_{l=1}^{K}\left(\prod_{j=1,\,j\neq l}^{K} y_{jl}^{(n)}(0,0)\right)g_l^{(n)}(\mathbf{0},0)\,h_l^{(n)}(\mathbf{0},0)\right)^{-1},$$

and after substituting this back in (4.19), the proof is completed. □

The expression for $P_n(\boldsymbol{s},\boldsymbol{t})$ in Lemma 4.4.1 is a product of three factors (separated by the $\times$-symbols). These factors, say $u_1^{(n)}, u_2^{(n)}$, and $u_3^{(n)}$, each play an intuitively appealing role in relation to Corollary 4.3.1. More specifically, our analysis below reveals that as $n \to \infty$ the first two factors $u_1^{(n)}$ and $u_2^{(n)}$ correspond to the transforms of the normal and exponential distribution, respectively. In addition, we show that as $n \to \infty$ the factor $u_3^{(n)}$ (which is the second line of (4.18)) immediately relates to the condition $Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \leqslant \beta$. As will become clear in the proofs, the factor $u_3^{(n)}$ is significantly more subtle to analyze than the factors $u_1^{(n)}$ and $u_2^{(n)}$.

Let us start with $u_1^{(n)}$. By applying a standard Taylor expansion to $\exp(-s_r/\sqrt{\rho_r}^{(n)})$ around zero, we obtain

$$\begin{aligned}
&\exp\left(-\rho_r^{(n)}+s_r\sqrt{\rho_r^{(n)}}+\zeta_r^{(n)}(s_r)\right)\\
&\quad=\exp\left(-\rho_r^{(n)}+s_r\sqrt{\rho_r^{(n)}}+\rho_r^{(n)}\left(1-\frac{s_r}{\sqrt{\rho_r^{(n)}}}+\frac{s_r^2}{2\rho_r^{(n)}}+o\left(\frac{1}{\rho_r^{(n)}}\right)\right)\right).
\end{aligned}\tag{4.20}$$

As $n \to \infty$, Expression (4.20) converges to $\exp(\frac{1}{2}s_r^2)$, which can be recognized as the transform $\mathbb{E}(\exp(-s_r\mathscr{N}))$ of a standard-normal random variable $\mathscr{N}$. From this we can conclude that, as $n \to \infty$, $u_1^{(n)}$ converges to a product of R standard-normal LSTs.

For $u_2^{(n)}$, we can apply the same strategy: for each k and $\alpha_k < 1$, as $n \to \infty$, we have $\sigma_k^{(n)} \to 1$ so that we can apply a Taylor expansion to $\exp(-t_k(1-\sigma_k^{(n)}))$ around zero. Therefore,

$$\begin{aligned}\frac{1-\sigma_k^{(n)}}{1-\delta_k^{(n)}(t_k)} &= \frac{1-\sigma_k^{(n)}}{1-\sigma_k^{(n)}(1-t_k(1-\sigma_k^{(n)})+o(1-\sigma_k^{(n)}))} \\ &= \frac{1-\sigma_k^{(n)}}{(1-\sigma_k^{(n)})(1+\sigma_k^{(n)}t_k)+o(1-\sigma_k^{(n)})}. \end{aligned} \tag{4.21}$$

As $n \to \infty$, Expression (4.21) converges to $(1+t_k)^{-1}$, which is the LST of an exponentially distributed random variable with rate 1. This implies that, as $n \to \infty$, $u_2^{(n)}$ converges to a product of K unit-rate exponential LSTs.

With the asymptotic behavior of $u_1^{(n)}$ and $u_2^{(n)}$ at hand, to prove Theorem 4.3.1 it remains to analyze $u_3^{(n)}$. To this end, we inspect the behavior of the functions $f^{(n)}(\boldsymbol{s})$, $g_l^{(n)}(\boldsymbol{s},t_l)$ and $h_l^{(n)}(\boldsymbol{s},t_l)$ in the limiting regime for $n \to \infty$, distinguishing different values of α_1 and α_l. This analysis is covered by the next subsection.

In addition to $f^{(n)}(\boldsymbol{s})$, $g_l^{(n)}(\boldsymbol{s},t_l)$ and $h_l^{(n)}(\boldsymbol{s},t_l)$, the sequence $u_3^{(n)}$ also contains the coefficients $\prod_{j=1,j\neq l}^K y_{jl}^{(n)}(t_j,t_l)$ for each single-server station $l = 1,...,K$. Multiplying the numerator and denominator of

$$y_{jl}^{(n)}(t_j,t_l) = \frac{1-\delta_j^{(n)}(t_j)}{1-\delta_j^{(n)}(t_j)/\delta_l^{(n)}(t_l)}$$

by $n/\delta_j^{(n)}(t_j)$, it follows that

$$y_{jl}^{(n)}(t_j,t_l) = \frac{n\big(e^{t_j(1-\sigma_j^{(n)})}-1\big)+c_j n^{\alpha_j}e^{t_j(1-\sigma_j^{(n)})}}{n\big(e^{t_j(1-\sigma_j^{(n)})}-e^{t_l(1-\sigma_l^{(n)})}\big)+c_j n^{\alpha_j}e^{t_j(1-\sigma_j^{(n)})}-c_l n^{\alpha_l}e^{t_l(1-\sigma_l^{(n)})}}. \tag{4.22}$$

As $\alpha_j, \alpha_l < 1$, we have

$$y_{jl}^{(n)}(t_j,t_l) \sim \frac{c_j(1+t_j)}{c_j(1+t_j)-c_l n^{\alpha_l-\alpha_j}(1+t_l)} \to \begin{cases} 0 & \text{if} \quad \alpha_j < \alpha_l, \\ \dfrac{c_j(1+t_j)}{c_j(1+t_j)-c_l(1+t_l)} & \text{if} \quad \alpha_j = \alpha_l, \\ 1 & \text{if} \quad \alpha_l < \alpha_j, \end{cases} \tag{4.23}$$

as $n \to \infty$. In particular, it holds that

$$\lim_{n\to\infty} \prod_{j=1,j\neq l}^K y_{jl}^{(n)}(t_j,t_l) = 0 \iff \alpha_l > \alpha_1.$$

This fact has an intuitive backing: we expect the condition in Corollary 4.3.1 to apply only to the stations with largest variability in queue lengths. For single-server stations, these are the stations l for which α_l is minimal.

4.4.2 Asymptotic analysis of $f^{(n)}(\cdot)$, $g_l^{(n)}(\cdot,\cdot)$ and $h_l^{(n)}(\cdot,\cdot)$

In Lemmas 4.4.3-4.4.6 we explicitly analyze the asymptotic behavior of the functions $f^{(n)}(\boldsymbol{s})$, $g_l^{(n)}(\boldsymbol{s},t_l)$ and $h_l^{(n)}(\boldsymbol{s},t_l)$ for all values of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\nu}$. To evaluate the cumulative

Poisson probabilities appearing in $f^{(n)}(\boldsymbol{s})$ and $g_l^{(n)}(\boldsymbol{s}, t_l)$, we require an additional central-limit type result given in Lemma 4.4.2. All lemmas in this subsection are proved in Appendix 4.A.

Lemma 4.4.2. *Suppose $x_n \to \infty$ as $n \to \infty$. If $(C_n - x_n)/\sqrt{x_n} \to Q$ with $Q \in [-\infty, \infty]$, then $\mathbb{P}(\mathscr{P}(x_n) \leqslant C_n) \to \Phi(Q)$ as $n \to \infty$.*

We proceed by determining the asymptotic behavior of the functions $f^{(n)}(\boldsymbol{s})$, $g_l^{(n)}(\boldsymbol{s}, t_l)$ and $h_l^{(n)}(\boldsymbol{s}, t_l)$ as $n \to \infty$. This behavior is highly dependent on the values of ν_1, α_1 and α_l, so that it is necessary to distinguish various cases. In most cases standard asymptotic methods suffice (Lemmas 4.4.3–4.4.5), but one particular case requires a more refined approach (Lemma 4.4.6).

Lemma 4.4.3. *As $n \to \infty$,*

$$f^{(n)}(\boldsymbol{s}) \to \begin{cases} 1 & \text{if} \quad 1 - \alpha_1 > \frac{1}{2}\nu_1, \\ \Phi(\lambda(\boldsymbol{s})) & \text{if} \quad 1 - \alpha_1 \leqslant \frac{1}{2}\nu_1. \end{cases}$$

Lemma 4.4.4. *As $n \to \infty$,*

$$g_l^{(n)}(\boldsymbol{s}, t_l) \to \begin{cases} 1 & \text{if} \quad 1 - \alpha_1 = 1 - \alpha_l > \frac{1}{2}\nu_1, \\ \Phi\big(\lambda(\boldsymbol{s}) - c_l(1 + t_l)\sqrt{W}\big) & \text{if} \quad 1 - \alpha_1 = 1 - \alpha_l = \frac{1}{2}\nu_1. \end{cases}$$

Lemma 4.4.5. *As $n \to \infty$,*

$$h_l^{(n)}(\boldsymbol{s}, t_l) \to \begin{cases} 0 & \text{if} \quad 1 - \alpha_1 > 1 - \alpha_l \text{ and } 1 - \alpha_1 \geqslant \nu_1, \\ 0 & \text{if} \quad \nu_1 > 1 - \alpha_1 > 1 - \alpha_l \geqslant \frac{1}{2}\nu_1, \\ \exp(-\beta c_l(1 + t_l)) & \text{if} \quad 1 - \alpha_1 = 1 - \alpha_l > \frac{1}{2}\nu_1, \\ \dfrac{\varphi(\lambda(\boldsymbol{s}))}{\varphi\big(\lambda(\boldsymbol{s}) - c_l(1 + t_l)\sqrt{W}\big)} & \text{if} \quad 1 - \alpha_1 = 1 - \alpha_l = \frac{1}{2}\nu_1. \end{cases}$$

The above lemmas treat all cases where either $1 - \alpha_1 \geqslant \nu_1$ or $1 - \alpha_l \geqslant \frac{1}{2}\nu_1$. Although $g_l^{(n)}(\boldsymbol{s}, t_l)$ is not evaluated in all these cases, observe that in (4.18) this function only occurs as the product $g_l^{(n)}(\boldsymbol{s}, t_l)\, h_l^{(n)}(\boldsymbol{s}, t_l)$. Since $g_l^{(n)}(\boldsymbol{s}, t_l) \in [0, 1]$, it follows that its specific value is irrelevant as long as $h_l^{(n)}(\boldsymbol{s}, t_l) \to 0$. We conclude that only the case where both $1 - \alpha_1 < \nu_1$ and $1 - \alpha_l < \frac{1}{2}\nu_1$ remains.

This last case requires a more subtle reasoning. Since $g_l^{(n)}(\boldsymbol{s}, t_l) \to 0$ and $h_l^{(n)}(\boldsymbol{s}, t_l) \to \infty$ as $n \to \infty$, we must analyze the product of the two functions before taking the limit. The proof of Lemma 4.4.6 relies on a change-of-measure argument.

Lemma 4.4.6. *If $1 - \alpha_1 < \nu_1$ and $1 - \alpha_l < \frac{1}{2}\nu_1$, then $g_l^{(n)}(\boldsymbol{s}, t_l)\, h_l^{(n)}(\boldsymbol{s}, t_l) \to 0$ as $n \to \infty$.*

We have now collected all the ingredients to establish the asymptotic expression for $P_n(\boldsymbol{s}, \boldsymbol{t})$ as presented in Theorem 4.3.1.

Proof of Theorem 4.3.1. The result is a consequence of Lemma 4.4.1 when substituting Equations (4.20), (4.21), and (4.23), in combination with the functions that we asymptotically evaluated in Lemmas 4.4.3–4.4.6 (both for general $\boldsymbol{s}, \boldsymbol{t}$ and for $\boldsymbol{s} = \boldsymbol{t} = \boldsymbol{0}$, that is). □

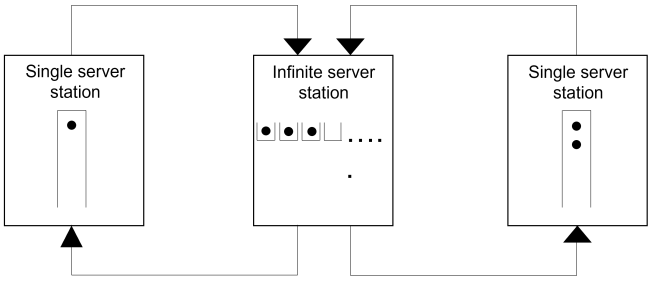

Figure 4.2: Closed network equivalent to the manufacturing model of Figure II.1.

4.5 Applications

Corollary 4.3.1 describes the asymptotic joint queue-length distribution under our scaling. This result may serve as the basis for approximations of the pre-limit distribution, which can be used e.g. when designing the network. Closed queueing networks can be broadly applied, as they can be used to represent for instance hospital units, computer systems, communication networks and manufacturing systems. In this section we discuss the implication of our results for the two examples mentioned in the introduction of Part II.

Example 1 (Extended machine-repair model)**.** In the extended machine-repair model, products that require processing arrive at a facility with C machines. If all machines are occupied, products are blocked and immediately leave the system upon arrival. An occupied machine may break down, and resumes processing only after it has been repaired by a single repairer. It is hereby assumed that a product remains assigned to the same machine for the duration of its service, even if the machine breaks down intermediately. An in-depth analysis of this system can be found in [122].

The queueing dynamics of this facility are visualized in Figure II.1. The network is open, but by the discussion in Section 4.2.1 it is equivalent to a closed network with two single-server stations (the external station and the repair station) and an infinite-server station (processing station). This closed network is depicted in Figure 4.2, and under the conditions described in Section 4.2 it obeys a product-form distribution (4.2).

Corollary 4.3.1 then states that the normalized numbers of occupied and broken machines tend to a normal and exponential distribution respectively, with a condition depending on the values of the chosen scaling parameters. If $1 - \alpha_1 > \frac{1}{2}\nu_1$, the limiting distribution of the number of broken machines is truncated at $c_1\beta$. If $1 - \alpha_1 < \frac{1}{2}\nu_1$, the limiting distribution of the number of occupied machines is truncated at $\beta/\sqrt{w_1}$. Finally, if $1 - \alpha_1 = \frac{1}{2}\nu_1$, the condition amounts to $\sqrt{w_1}\mathscr{N}_1 + \frac{1}{c_1}\mathscr{E}_1 \leqslant \beta$, which in particular implies dependence between the queue lengths. ◇

Example 2 (Vehicle sharing system)**.** In modern society the demand for flexible transportation has led to the development of vehicle sharing systems. In such systems, a number of vehicles is scattered among a fixed number of locations. Users may pick up a vehicle at any location (if available) and drop it off at any, possibly different, location. To accurately describe the behavior of such a system, a well-fitting model is important.

Closed queueing networks are often used in modeling vehicle sharing systems, see e.g. George and Xia [50]. In this model, the population size C is the total number of vehicles across the network. The pick-up (and drop-off) locations are modeled by single-server queues, and between each ordered pair of pick-up locations, an infinite-server queue is used to describe the time spent by a user between these locations. See Figure II.2 for an example with three pick-up locations.

Notice that the number of stations used to model the network grows quadratically in the number of pick-up locations. For this reason, vehicle-sharing systems quickly become analytically and numerically intractable when the number of pick-up locations increases. The result of Corollary 4.3.1, however, does not become more complex as the number of stations grows. Under typical circumstances R^- and K^- are low and the asymptotic queue-length distributions are therefore (asymptotically) tractable. ◇

4.6 Numerical illustrations

In Section 4.4 we have established a convergence-in-distribution result for the random vector $(\bar{\boldsymbol{B}}^{(n)}, \bar{\boldsymbol{D}}^{(n)})$. In this section we will discuss the performance of approximations based on this scaling limit. In Section 4.6.1, we assess the pre-limit distributions by means of numerical experiments, and compare them to the limiting distributions of Corollary 4.3.1. Importantly, the number of scaling parameters (relating to the vectors $\boldsymbol{w}$, $\boldsymbol{\nu}$, $\boldsymbol{c}$ and $\boldsymbol{\alpha}$ and the scalar β) exceeds the number of parameters of our pre-limit model (i.e., the vectors $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ and the scalar C). This leaves us with some freedom to choose the scaling parameters; using an example network, we show in Section 4.6.2 how this can be done.

Computation of individual queue-length distributions directly from the stationary distribution (4.2) is hard for networks with many stations, as the state space grows quickly with the size of the network. We therefore choose to use an acceptance-rejection simulation [58] for all numerical experiments in this section. This entails sampling from the stationary distribution without the population size constraint (which comes down to separately sampling from Poisson and geometric distributions), and rejecting the samples that fail to satisfy the population size constraint. The set of accepted samples is then stochastically equivalent to a set of equally many independent samples from (4.2). An estimate for the marginal queue-length distributions can thus be found by counting the number of accepted samples with each possible queue-length value. Throughout this section we constantly use 10^7 samples for all simulation results (of which at least half gets accepted).

4.6.1 Accuracy of approximation

We start by considering networks consisting of a large number of stations — for instance, one can think of a vehicle-sharing system from Section 4.5 with ten pick-up locations, which has more than a hundred stations. We consider a setting with $R^- = 1$ and $K^- \leqslant 2$ (such that there are at most three dominant stations). This entails by Corollary 4.3.1 that the

	Values of $\nu_1, \alpha_1, \alpha_2$	Condition on the joint distribution of $(\bar{B}_1^{(n)}, \bar{D}_1^{(n)}, \bar{D}_2^{(n)})$
Case 1	$1-\alpha_1 > 1-\alpha_2$, $1-\alpha_1 > \frac{1}{2}\nu_1$	$\frac{1}{c_1}\bar{D}_1^{(n)} \leqslant \beta$
Case 2	$1-\alpha_1 = 1-\alpha_2 > \frac{1}{2}\nu_1$	$\frac{1}{c_1}\bar{D}_1^{(n)} + \frac{1}{c_2}\bar{D}_2^{(n)} \leqslant \beta$
Case 3	$1-\alpha_1 = 1-\alpha_2 = \frac{1}{2}\nu_1$	$\sqrt{w_1}\bar{B}_1^{(n)} + \frac{1}{c_1}\bar{D}_1^{(n)} + \frac{1}{c_2}\bar{D}_2^{(n)} \leqslant \beta$
Case 4	$1-\alpha_1 = \frac{1}{2}\nu_1 > 1-\alpha_2$	$\sqrt{w_1}\bar{B}_1^{(n)} + \frac{1}{c_1}\bar{D}_1^{(n)} \leqslant \beta$
Case 5	$1-\alpha_1 < \frac{1}{2}\nu_1$	$\sqrt{w_1}\bar{B}_1^{(n)} \leqslant \beta$

Table 4.1: Five cases for the values of $\nu_1, \alpha_1, \alpha_2$ and the corresponding condition on the joint distribution of $(\bar{B}_1^{(n)}, \bar{D}_1^{(n)}, \bar{D}_2^{(n)})$.

	ν_1	$\nu_2, ..., \nu_6$	α_1	α_2	$\alpha_3, ..., \alpha_7$	$w_1, ..., w_6$	c_1	c_2	$c_3, ..., c_7$	β	n
Case 1	1	0.5	-1	0	0.5	1	1	1	1	1	25
Case 2	1	0.5	0	0	0.5	1	1	2	1	1	100
Case 3	1	0.5	0.5	0.5	0.9	1	1	2	1	1	100
Case 4	1	0.5	0.5	0.8	0.9	1	1	1	1	1	100
Case 5	2	0.5	0.9	0.9	0.9	1	1	1	1	1	100

Table 4.2: Scaling parameter values for the plots in Figure 4.3.

variables $\bar{B}_2^{(n)}, ..., \bar{B}_R^{(n)}$ converge to independent standard-normal random variables, and that $\bar{D}_3^{(n)}, ..., \bar{D}_K^{(n)}$ converge to independent unit-rate exponential random variables. The asymptotic distributions of $\bar{B}_1^{(n)}$, $\bar{D}_1^{(n)}$ and $\bar{D}_2^{(n)}$ are less trivial because they may be affected by the condition $Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \leqslant \beta$. Figure 4.3 therefore focuses on these random variables: it shows their density functions, estimated by simulation. Strictly speaking, the variables $\bar{B}_1^{(n)}$, $\bar{D}_1^{(n)}$ and $\bar{D}_2^{(n)}$ have probability masses instead of densities for finite n since they are discrete. However, we consider the scaled mass functions of these variables and refer to them as densities in the sequel, so as to facilitate comparison with their limits as $n \to \infty$.

How the condition $Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \leqslant \beta$ impacts the distributions of $\bar{B}_1^{(n)}$, $\bar{D}_1^{(n)}$ and $\bar{D}_2^{(n)}$ depends mainly on the values of ν_1, α_1 and α_2, see Table 4.1. The five different cases are visible in the rows of Figure 4.3, in which simulation results are shown for a network with $R = 6$ and $K = 7$. In cases 1 and 5, the condition applies to only one random variable, which causes the associated density function to be truncated at β.

In all of the cases the density of $\bar{B}_1^{(n)}$ resembles the normal density, whereas the densities of $\bar{D}_1^{(n)}$ and $\bar{D}_2^{(n)}$ resemble the exponential density. At a more detailed level, Figure 4.3 also shows the impact of the different conditions (i.e. the five cases that were displayed in Table 4.2). In cases where $\bar{B}_1^{(n)}$ (or $\bar{D}_1^{(n)}, \bar{D}_2^{(n)}$) is not part of the condition, its density function is simply a slightly perturbed version of that of a standard normal (or unit-rate exponential). On the other hand, in cases where $\bar{B}_1^{(n)}$ (or $\bar{D}_1^{(n)}, \bar{D}_2^{(n)}$) *is* part of the condition, we see that

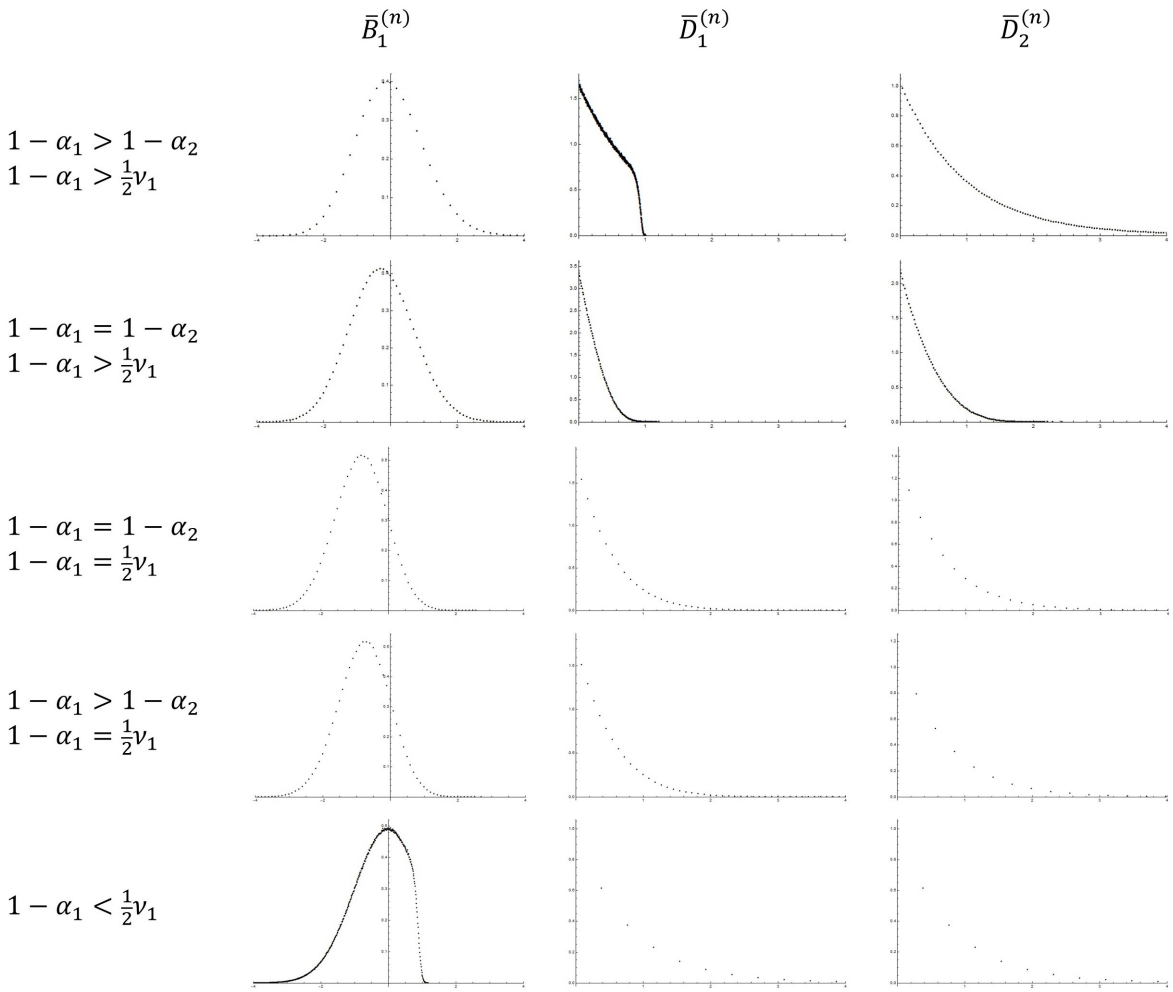

Figure 4.3: Density functions of $\bar{B}_1^{(n)}$, $\bar{D}_1^{(n)}$ and $\bar{D}_2^{(n)}$, estimated by simulation, depending on the values of the scaling parameters $\nu_1, \alpha_1, \alpha_2$. The exact parameter values for the five cases (top to bottom) can be found in Table 4.2.

the corresponding random variable is less likely to assume larger values. We conclude that the structure of the limit distributions, as identified in Corollary 4.3.1, carries over to the pre-limit setting.

4.6.2 Fitting scaling parameters

In the remainder of this section we show how to use our scaling regime in a concrete queueing network model. Particularly, we show how the scaling parameters can be chosen to appropriately reflect the model at hand. The extended machine-repair model described in Section 4.5 will serve as an example. For this system we compare the actual queue-length distributions (obtained by acceptance-rejection simulation) to the limiting distributions in the scaling regime (as stated in Corollary 4.3.1). Although the acceptance-rejection simulation gives sufficiently accurate results and is consistent with the previous subsection, we remark that for the small machine-repair network, direct calculation of (4.2) is also numerically tractable.

To compare the behavior of the queue lengths under a given set of model parameters with our

limit results, we have to choose appropriate scaling parameters. Since we have $2R + 2K + 1$ scaling parameters (the entries of the vectors $\boldsymbol{w}$, $\boldsymbol{\nu}$, $\boldsymbol{c}$, and $\boldsymbol{\alpha}$ and the scalar β) compared to only $R + K + 1$ model parameters (the entries of the vectors $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ and the scalar C), this can be done in many different ways. Choosing appropriate scaling parameters is important, because not all choices lead to accurate approximations. The following intuitive procedure may serve as a guideline.

1. First select which stations are dominant, i.e. the stations whose queue lengths will be incorporated in the condition $Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \leqslant \beta$. These should correspond to the queue lengths with largest variance, as these are affected most by the population size constraint. These variances are, respectively, $\rho_1, ..., \rho_R, \sigma_1/(1-\sigma_1)^2, \ldots, \sigma_K/(1-\sigma_K)^2$ (if we drop the population size constraint). There is some freedom in choosing the number of dominant stations. Working with more dominant stations yields a limit result that is more accurate but whose evaluation is more complex.

2. Choose values for $\boldsymbol{\nu}$ and $\boldsymbol{\alpha}$ such that the dominant stations are indeed affected by the condition, and the remaining stations are not.

3. Fix n, and choose values for $\boldsymbol{w}, \boldsymbol{c}, \beta$ such that the scaled network coincides with the network under consideration.

With these underlying ideas, one may choose scaling parameters as follows.

1. Let A denote the set of dominant stations, which we construct as follows. Order the variances $\rho_1, ..., \rho_R, \sigma_1/(1-\sigma_1)^2, \ldots, \sigma_K/(1-\sigma_K)^2$ from high to low. Include in A the smallest number of stations with highest variance, such that

$$\sum_{r \in A} \rho_r + \sum_{k \in A} \frac{\sigma_k}{(1-\sigma_k)^2} \geqslant T \cdot \left(\sum_{r=1}^{R} \rho_r + \sum_{k=1}^{K} \frac{\sigma_k}{(1-\sigma_k)^2} \right)$$

   for some threshold fraction $T \leqslant 1$. That is, choose the largest variances such that the sum of these variances makes up for at least a fraction T of the total variance. The best value of T may depend on the model at hand and the trade-off between accuracy and complexity discussed above. We empirically observed that picking $T = 0.8$ provides rather accurate approximations in most cases.

2. For $r, k \in A$ and some $\gamma > 0$, choose values for ν_r, α_k such that $\frac{1}{2}\nu_r = 1 - \alpha_k = \gamma := \max\{\frac{1}{2}\nu_1, 1 - \alpha_1\}$. It turns out that the value γ may be chosen arbitrarily, since each value leads to the same limit result (see the discussion below). For $r, k \notin A$, choose values for ν_r, α_k such that $\frac{1}{2}\nu_r < \gamma$ and $1 - \alpha_k < \gamma$ (the exact values are again irrelevant).

3. Pick any value for n. Then choose values for $w_1, ..., w_R$ such that $\boldsymbol{\rho}^{(n)} = \boldsymbol{\rho}$, values for $c_1, ..., c_K$ such that $\boldsymbol{\sigma}^{(n)} = \boldsymbol{\sigma}$ and β such that $C_n = C$.

Despite the freedom in choosing scaling parameters, we underline that the decision for the set of dominant stations A completely determines the limit result. This can be verified with the following argument. For non-dominant stations, observe that the queue lengths converge to

	C	ρ_1	σ_1
Parameter set 1	100	40	0.99
Parameter set 2	100	90	0.90
Parameter set 3	100	90	0.66

Table 4.3: Three parameter sets of the extended machine-repair model.

	ν_1	α_1	w_1	c_1	β	n
Scaling parameter set 1	1	-0.24567	1	1	0.61	40
Scaling parameter set 2	1	0.5	1	1.054	1.1	90
Scaling parameter set 3	1	0.8526	1	1	1.1	90

Table 4.4: Scaling parameter sets generating the same system as the parameter sets in Table 4.3.

standard-normal and unit-rate exponential variables regardless of the scaling parameters. The queue-length distributions of the dominant stations on the other hand, depend on the condition $\sum_{r\in A}\sqrt{w_r}\mathscr{N}_r + \sum_{k\in A}\frac{1}{c_k}\mathscr{E}_k \leqslant \beta$. It therefore seems like the distributions of the dominant queue lengths depend on the values of $\boldsymbol{w}$, $\boldsymbol{c}$ and β. However, the identities $\boldsymbol{\rho}^{(n)} = \boldsymbol{\rho}$, $\boldsymbol{\sigma}^{(n)} = \boldsymbol{\sigma}$ and $C_n = C$ imply that $w_r = \rho_r n^{-2\gamma}$ for $r \in A$, $c_k = (\sigma_k^{-1} - 1)n^{\gamma}$ for $k \in A$ and $\beta = (C - \|\boldsymbol{\rho}\|)n^{-\gamma}$. With these scaling parameter values, the factor $n^{-\gamma}$ cancels out of the condition, which makes the values of n and γ irrelevant for the limit result.

We now move to the extended machine-repair model (described in Section 4.5) for a concrete numerical example of the steps above. Recall that in this model, the queue lengths of interest are the number of occupied machines and the number of broken machines. We denote these random variables by B_1 and D_1 respectively. The model has three parameters: the total number of machines C, the traffic load of the processing station ρ_1, and the traffic load of the repair station σ_1. Consider the three sets of parameter values shown in Table 4.3.

Following the steps above, we find the scaling parameters given in Table 4.4. Note that by definition of our scaling regime, taking scaling parameter set 1 induces precisely the machine-repair model with parameter set 1. Hence, for this parameter set, comparing the actual queue-length distributions and our limit results amounts to comparing $n = 40$ and $n \to \infty$. The same holds for $n = 90$ and parameter sets 2 and 3.

In Figures 4.4–4.6 we show plots of the density functions of $\bar{B}_1^{(n)}$ and $\bar{D}_1^{(n)}$ for each parameter set, obtained by simulation. For comparison, the densities are plotted against the limit results of Corollary 4.3.1.

Parameter set 1. Observe that $1 - \alpha_1 > \frac{1}{2}\nu_1$. Corollary 4.3.1 states in this case that

$$\bar{B}_1^{(n)} \to_{\mathrm{d}} \mathscr{N}_1, \qquad \bar{D}_1^{(n)} \to_{\mathrm{d}} (\mathscr{E}_1 \mid \mathscr{E}_1 \leqslant \beta)$$

as $n \to \infty$. Therefore, we have plotted in Figure 4.4 the densities of $\bar{B}_1^{(40)}$ and $\bar{D}_1^{(40)}$ (obtained through simulation) against respectively a standard-normal density and a unit-rate exponential

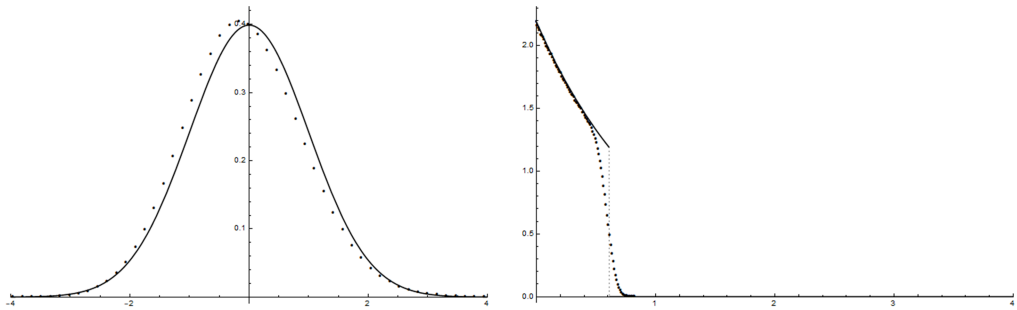

Figure 4.4: Density functions of $\bar{B}_1^{(n)}$ (left) and $\bar{D}_1^{(n)}$ (right) for parameter set 1, both for $n = 40$ (dots) and for $n \to \infty$ (line).

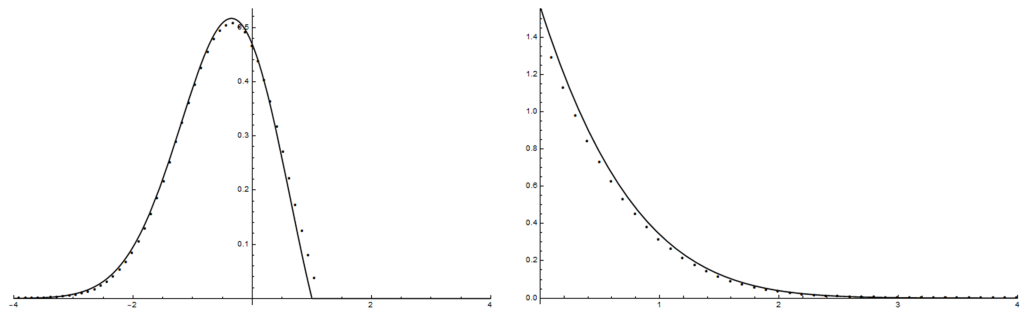

Figure 4.5: Density functions of $\bar{B}_1^{(n)}$ (left) and $\bar{D}_1^{(n)}$ (right) for parameter set 2, both for $n = 90$ (dots) and for $n \to \infty$ (line).

density truncated at β.

Parameter set 2. Observe that $1 - \alpha_1 = \frac{1}{2}\nu_1$. Corollary 4.3.1 states for $1 - \alpha_1 = \frac{1}{2}\nu_1$ that

$$(\bar{B}_1^{(n)}, \bar{D}_1^{(n)}) \to_{\mathrm{d}} (\mathscr{N}_1, \mathscr{E}_1 | \mathscr{N}_1 + \mathscr{E}_1 \leqslant \beta)$$

as $n \to \infty$. Therefore, Figure 4.5 plots the densities of $\bar{B}_1^{(90)}$ and $\bar{D}_1^{(90)}$ against respectively the densities of $(\mathscr{N}_1 \mid \mathscr{N}_1 + \mathscr{E}_1 \leqslant \beta)$ and $(\mathscr{E}_1 \mid \mathscr{N}_1 + \mathscr{E}_1 \leqslant \beta)$.

Parameter set 3. Observe that $1 - \alpha_1 < \frac{1}{2}\nu_1$. Corollary 4.3.1 states in this case that

$$\bar{B}_1^{(n)} \to_{\mathrm{d}} (\mathscr{N}_1 \mid \mathscr{N}_1 \leqslant \beta), \qquad \bar{D}_1^{(n)} \to_{\mathrm{d}} \mathscr{E}_1$$

as $n \to \infty$. Therefore, Figure 4.6 plots the densities of $\bar{B}_1^{(90)}$ and $\bar{D}_1^{(90)}$ against respectively a standard-normal density truncated at β and a unit-rate exponential density.

Figures 4.4–4.6 show that for a network with $C = 100$, Corollary 4.3.1 provides rather accurate approximations of the queue-length densities. In relatively small networks there is the obvious alternative of direct evaluation of the product-form density. For larger networks this will lead to computational issues, whereas the complexity of our asymptotic results is just mildly affected by the network size.

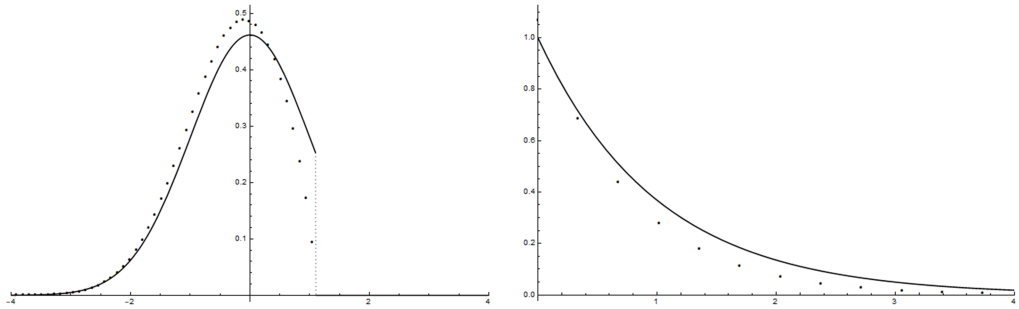

Figure 4.6: Density functions of $\bar{B}_1^{(n)}$ (left) and $\bar{D}_1^{(n)}$ (right) for parameter set 3, both for $n = 90$ (dots) and for $n \to \infty$ (line).

4.7 Discussion and further research

For a broad class of queueing networks, such as those of BCMP type, the joint queue-length distribution has a product-form structure. It may seem to lend itself well to numerical evaluation, but in case of closed networks the population size constraint makes this a non-trivial task. To overcome such computational issues, we have proposed a scaling regime, inspired by the Halfin-Whitt scaling. The corresponding limiting joint stationary queue-length distribution is transparent, numerically tractable and provides insight into the dependencies between the individual queue lengths. We have pointed out how to map our scaling parameters on those of the queueing network under consideration. A series of numerical experiments shows that the resulting approximations are close to the true (pre-limit, that is) values.

Scaling methods in queueing networks form a rich research area in which there is still ample room to extend our current results. One option is to include multi-server stations in the network. As the queue lengths become very large in our scaling regime, we expect that such a station would effectively behave as a single-server station, with the service rate multiplied by the number of servers.

Another model extension preserving product form relates to multiclass networks. In these models customers may be of different classes, where each class may have its specific routing and service requirements. The product form of the stationary distribution is preserved under class-dependent routing probabilities and, for certain station types, under class-dependent service requirements. Many queueing network results apply to multiclass networks, but scaling analysis becomes more involved, primarily because each customer class now has its own population size.

Further research efforts could focus on exploiting our scaling results for design and optimization purposes. In addition, as we have indicated, our scaling method provides freedom in relation to the choice of the scaling parameters, which raises the question how to choose the entries of $\boldsymbol{w}, \boldsymbol{\nu}, \boldsymbol{c}, \boldsymbol{\alpha}$ and β so as to maximally accurately represent the underlying queueuing network.

As a final remark, we discuss the recent related work by Jelenković et al. [66, 67]. Similar to this chapter, their work concerns scaling analysis of a closed product-form queueing network with infinite-server and single-server stations. A notable feature in the scaling proposed by

Jelenković et al. is that the number of single-server stations is also scaled (together with the workloads and the population size). Limiting queue length distributions are derived in a number of scaling regimes where the relative value of the workloads, population size and number of heaviest-loaded stations is varied.

Interestingly, as opposed to our Laplace-Stieltjes transform approach, Jelenković et al. work with a probabilistic analysis method. Their idea is based on a particular probabilistic representation of the stationary queue-length distribution. For example, when substituting $R = K = 1$, one can write the right-hand side of (4.2) as

$$\frac{\mathbb{P}(\mathscr{P}(\rho_1) = b_1) \cdot \mathbb{P}(\mathscr{G}(\sigma_1) = d_1)}{\mathbb{P}(\mathscr{P}(\rho_1) + \mathscr{G}(\sigma_1) \leqslant C)}. \tag{4.24}$$

Using the convergence results of Poisson to normal distributions and Geometric to exponential distributions, our main results can alternatively be proven with this probabilistic representation: for the $R = K = 1$ example, Corollary 4.3.1 could be proven as follows. Let s, t be such that $Z(s,t) \leqslant \beta$. By (4.10) and (4.24),

$$\begin{aligned}\mathbb{P}\left(\bar{B}_1^{(n)} \leqslant s, \bar{D}_1^{(n)} \leqslant t\right) &= \mathbb{P}\left(\frac{B_1^{(n)} - \rho_1^{(n)}}{\sqrt{\rho_1^{(n)}}} \leqslant s, (1-\sigma_1^{(n)})D_1^{(n)} \leqslant t\right) \\ &= \frac{\mathbb{P}\left(\frac{\mathscr{P}(\rho_1^{(n)})-\rho_1^{(n)}}{\sqrt{\rho_1^{(n)}}} \leqslant s\right) \cdot \mathbb{P}\left((1-\sigma_1^{(n)})\mathscr{G}(\sigma_1^{(n)}) \leqslant t\right)}{\mathbb{P}\left(\mathscr{P}(\rho_1^{(n)}) + \mathscr{G}(\sigma_1^{(n)}) \leqslant C_n\right)}. \end{aligned} \tag{4.25}$$

Applying standard convergence results of the Poisson and geometric distributions, it can be seen that the numerator of (4.25) converges to $\mathbb{P}(\mathscr{N} \leqslant s) \cdot \mathbb{P}(\mathscr{E} \leqslant t)$ as $n \to \infty$ (assuming that $\nu_1 > 0$ and $\alpha_1 < 1$). For the inequality inside the probability in the denominator, we subtract $\rho_1^{(n)}$ and divide by n^γ on both sides. The definition of C_n then implies that the denominator of (4.25) is asymptotically equal to

$$\mathbb{P}\left(\frac{\mathscr{P}(\rho_1^{(n)}) - \rho_1^{(n)}}{n^\gamma} + \frac{\mathscr{G}(\sigma_1^{(n)})}{n^\gamma} \leqslant \beta\right).$$

As $n \to \infty$, this probability converges to $\mathbb{P}(Z(\mathscr{N}, \mathscr{E}) \leqslant \beta)$.

When we combine the expressions of the numerator and denominator of (4.25), we have

$$\mathbb{P}\left(\bar{B}_1^{(n)} \leqslant s, \bar{D}_1^{(n)} \leqslant t\right) \to \frac{\mathbb{P}(\mathscr{N} \leqslant s) \cdot \mathbb{P}(\mathscr{E} \leqslant t)}{\mathbb{P}(Z(\mathscr{N}, \mathscr{E}) \leqslant \beta)}$$

as $n \to \infty$, from which Corollary 4.3.1 follows. In light of this alternative proof, it can be worthwhile exploring which other scaling approaches to product-form networks could benefit from a probabilistic representation similar to the one in (4.24).

Appendix 4.A Proofs of Section 4.4

Lemma 4.4.2. *Suppose $x_n \to \infty$ as $n \to \infty$. If $(C_n - x_n)/\sqrt{x_n} \to Q$ with $Q \in [-\infty, \infty]$, then $\mathbb{P}(\mathscr{P}(x_n) \leqslant C_n) \to \Phi(Q)$ as $n \to \infty$.*

Proof. Observe that a Poisson random variable with mean $m \in \mathbb{N}$ can be written as a sum of m Poisson random variables with mean 1. Therefore, with $(X_i)_{i\in\mathbb{N}}, (Y_i)_{i\in\mathbb{N}}$ i.i.d. copies of $\mathscr{P}(1)$,

$$\sum_{i=1}^{\lfloor x_n \rfloor} X_i =_{\mathrm{d}} \mathscr{P}(\lfloor x_n \rfloor) \leqslant_{\mathrm{st}} \mathscr{P}(x_n) \leqslant_{\mathrm{st}} \mathscr{P}(\lceil x_n \rceil) =_{\mathrm{d}} \sum_{i=1}^{\lceil x_n \rceil} Y_i.$$

Substracting x_n and dividing by $\sqrt{x_n}$ yields

$$\frac{1}{\sqrt{x_n}} \sum_{i=1}^{\lfloor x_n \rfloor} (X_i - 1) - \frac{x_n - \lfloor x_n \rfloor}{\sqrt{x_n}} \leqslant_{\mathrm{st}} \frac{\mathscr{P}(x_n) - x_n}{\sqrt{x_n}} \leqslant_{\mathrm{st}} \frac{1}{\sqrt{x_n}} \sum_{i=1}^{\lceil x_n \rceil} (Y_i - 1) + \frac{\lceil x_n \rceil - x_n}{\sqrt{x_n}}.$$

Appropriately rewritten as

$$\frac{1}{\sqrt{\lfloor x_n \rfloor} + O(1)} \sum_{i=1}^{\lfloor x_n \rfloor} (X_i - 1) - \frac{O(1)}{\sqrt{x_n}} \leqslant_{\mathrm{st}} \frac{\mathscr{P}(x_n) - x_n}{\sqrt{x_n}} \leqslant_{\mathrm{st}} \frac{1}{\sqrt{\lceil x_n \rceil} - O(1)} \sum_{i=1}^{\lceil x_n \rceil} (Y_i - 1) + \frac{O(1)}{\sqrt{x_n}},$$

we may apply the central limit theorem to conclude that $(\mathscr{P}(x_n) - x_n)/\sqrt{x_n}$ converges to a standard-normal random variable as $x_n \to \infty$. Using this observation the result immediately follows from the fact that

$$\mathbb{P}\left(\mathscr{P}(x_n) \leqslant C_n\right) = \mathbb{P}\left(\frac{\mathscr{P}(x_n) - x_n}{\sqrt{x_n}} \leqslant \frac{C_n - x_n}{\sqrt{x_n}}\right) \to \Phi(Q)$$

as $x_n \to \infty$. □

In the following proofs we write ρ_r, ζ_r, σ_l and δ_l for $\rho_r^{(n)}$, $\zeta_r^{(n)}(s_r)$, $\sigma_l^{(n)}$ and $\delta_l^{(n)}(t_l)$ to simplify the notation.

Lemma 4.4.3. *As $n \to \infty$,*

$$f^{(n)}(\boldsymbol{s}) \to \begin{cases} 1 & \text{if} \quad 1 - \alpha_1 > \frac{1}{2}\nu_1, \\ \Phi\left(\lambda(\boldsymbol{s})\right) & \text{if} \quad 1 - \alpha_1 \leqslant \frac{1}{2}\nu_1. \end{cases}$$

Proof. We start with the case $1 - \alpha_1 > \frac{1}{2}\nu_1$. Note that in this case

$$C_n = \left\lfloor \sum_{r=1}^{R} \rho_r + \beta n^{1-\alpha_1} \right\rfloor.$$

To prove $f^{(n)}(\boldsymbol{s}) \to 1$, we apply Lemma 4.4.2 with

$$x_n = \sum_{r=1}^{R} \zeta_r = \sum_{r=1}^{R} \left(\rho_r - s_r\sqrt{\rho_r} + o\left(\sqrt{\rho_r}\right)\right)$$

(so that $Q = \infty$).

For the limit of $f^{(n)}(\boldsymbol{s})$ in case $1 - \alpha_1 \leqslant \frac{1}{2}\nu_1$, an application of Lemma 4.4.2 with

$$x_n = \sum_{r=1}^{R} \zeta_r = \sum_{r=1}^{R} \left(\rho_r - s_r\sqrt{\rho_r} + o\left(\sqrt{\rho_r}\right)\right)$$

and

$$C_n = \left\lfloor \sum_{r=1}^{R} \rho_r + \beta n^{\frac{1}{2}\nu_1} \right\rfloor$$

leads to

$$Q = \lim_{n\to\infty} \frac{C_n - x_n}{\sqrt{x_n}} = \lim_{n\to\infty} \frac{\beta n^{\frac{1}{2}\nu_1} + \sum_{r=1}^{R} \left(s_r\sqrt{\rho_r} + o\left(\sqrt{\rho_r}\right)\right)}{\sqrt{\sum_{r=1}^{R} \left(\rho_r + o\left(\rho_r\right)\right)}} = \frac{\beta + \sum_{r=1}^{R^-} s_r\sqrt{w_r}}{\sqrt{\sum_{r=1}^{R^-} w_r}} = \lambda(\boldsymbol{s}).$$

Recall that R^- is defined as the largest integer such that $\nu_1 = \ldots = \nu_{R^-}$. Hence, if $1-\alpha_1 \leqslant \frac{1}{2}\nu_1$, then $f^{(n)}(\boldsymbol{s}) \to \Phi\left(\lambda(\boldsymbol{s})\right)$. □

Lemma 4.4.4. *As $n \to \infty$,*

$$g_l^{(n)}(\boldsymbol{s}, t_l) \to \begin{cases} 1 & \text{if} \quad 1-\alpha_1 = 1-\alpha_l > \frac{1}{2}\nu_1, \\ \Phi\big(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\big) & \text{if} \quad 1-\alpha_1 = 1-\alpha_l = \frac{1}{2}\nu_1. \end{cases}$$

Proof. The proof for $g_l^{(n)}(\boldsymbol{s}, t_l)$ is similar to the proof for $f^{(n)}(\boldsymbol{s})$. Suppose first that $1-\alpha_1 = 1-\alpha_l > \frac{1}{2}\nu_1$. Note that with

$$x_n = \sum_{r=1}^{R} \zeta_r/\delta_l = \sum_{r=1}^{R} \rho_r \frac{n + c_l n^{\alpha_l}}{n} \exp(-s_r/\sqrt{\rho_r})\, e^{t_l(1-\sigma_l)},$$

we have

$$\frac{C_n - x_n}{\sqrt{x_n}} = \frac{\sum_{r=1}^{R} \rho_r + \frac{\beta}{c_1} n^{1-\alpha_1} - \sum_{r=1}^{R} \left(\rho_r + O(n^{\nu_r+\alpha_l-1}) - O(n^{\frac{1}{2}\nu_r})\right)}{\sqrt{\sum_{r=1}^{R} \left(\rho_r + o(\rho_r)\right)}}.$$

When $1-\alpha_1 = 1-\alpha_l > \frac{1}{2}\nu_1$ we take $Q = \infty$ in Lemma 4.4.2, concluding that $g_l^{(n)}(\boldsymbol{s}, t_l) \to 1$. Next, for $g_l^{(n)}(\boldsymbol{s}, t_l)$ as $1-\alpha_1 = 1-\alpha_l = \frac{1}{2}\nu_1$, an application of Lemma 4.4.2 with

$$x_n = \sum_{r=1}^{R} \zeta_r/\delta_l = \sum_{r=1}^{R} \rho_r \frac{n + c_l n^{\alpha_l}}{n} \exp(-s_r/\sqrt{\rho_r})\, e^{t_l(1-\sigma_l)}$$

and

$$C_n = \left\lfloor \sum_{r=1}^{R} \rho_r + \beta n^{\frac{1}{2}\nu_1} \right\rfloor$$

leads to

$$Q = \lim_{n\to\infty} \frac{C_n - x_n}{\sqrt{x_n}} = \lim_{n\to\infty} \frac{\sum_{r=1}^{R} \rho_r + \beta n^{\frac{1}{2}\nu_1} - \sum_{r=1}^{R} \left(\rho_r + c_l(1+t_l)n^{\alpha_l - 1}\rho_r - s_r\sqrt{\rho_r}\right) + o(\sqrt{\rho_r})}{\sqrt{\sum_{r=1}^{R} (\rho_r + o(\rho_r))}}$$

$$= \frac{\beta + \sum_{r=1}^{R^-} \left(s_r\sqrt{w_r} - c_l(1+t_l)w_r\right)}{\sqrt{\sum_{j=1}^{R^-} w_r}} = \lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}.$$

We conclude that if $1-\alpha_1 = 1-\alpha_l = \frac{1}{2}\nu_1$, then $g_l^{(n)}(\boldsymbol{s}, t_l) \to \Phi(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W})$. □

Lemma 4.4.5. *As $n \to \infty$,*

$$h_l^{(n)}(\boldsymbol{s}, t_l) \to \begin{cases} 0 & \text{if} \quad 1-\alpha_1 > 1-\alpha_l \text{ and } 1-\alpha_1 \geqslant \nu_1, \\ 0 & \text{if} \quad \nu_1 > 1-\alpha_1 > 1-\alpha_l \geqslant \frac{1}{2}\nu_1, \\ \exp\left(-\beta c_l(1+t_l)\right) & \text{if} \quad 1-\alpha_1 = 1-\alpha_l > \frac{1}{2}\nu_1, \\ \dfrac{\varphi(\lambda(\boldsymbol{s}))}{\varphi\left(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\right)} & \text{if} \quad 1-\alpha_1 = 1-\alpha_l = \frac{1}{2}\nu_1. \end{cases}$$

Proof. Let $H_l^{(n)}(\boldsymbol{s}, t_l) := \log(h_l^{(n)}(\boldsymbol{s}, t_l))$ and observe that $H_l^{(n)}(\boldsymbol{s}, t_l)$ is a sum of the two components $\zeta^{(n)}(\boldsymbol{s})(1/\delta_l(t_l) - 1)$ and $(C_n + 1)\log(\delta_l(t_l))$. We explicitly consider these two components separately. In the following calculations, terms irrelevant as $n \to \infty$ will be dealt with using the '$\sim$' symbol. From the first component of $H_l^{(n)}(\boldsymbol{s}, t_l)$ we extract the leading terms by applying Taylor expansions. As $n \to \infty$, this component can be rewritten as

$$\begin{aligned} &\sum_{r=1}^{R} \rho_r e^{-s_r/\sqrt{\rho_r}} \left(\frac{e^{t_l(1-\sigma_l)}}{\sigma_l} - 1\right) \\ &\quad \sim \sum_{r=1}^{R} \rho_r (1 - s_r/\sqrt{\rho_r}) \left(\frac{1}{\sigma_l} - 1 + \frac{t_l(1-\sigma_l)}{\sigma_l} + \frac{\frac{1}{2}t_l^2(1-\sigma_l)^2}{\sigma_l}\right) \\ &\quad \sim \sum_{r=1}^{R} w_r n^{\nu_r} (1 - s_r/\sqrt{w_r n^{\nu_r}}) \left(c_l n^{\alpha_l - 1} + t_l c_l n^{\alpha_l - 1} + \frac{1}{2} c_l^2 t_l^2 n^{2\alpha_l - 2}\right) \\ &\quad \sim \sum_{r=1}^{R} \left(w_r c_l (1+t_l) n^{\alpha_l - 1 + \nu_r} - \sqrt{w_r} c_l (1+t_l) s_r n^{\alpha_l - 1 + \frac{1}{2}\nu_r} + \frac{1}{2} w_r c_l^2 t_l^2 n^{2\alpha_l - 2 + \nu_r}\right). \end{aligned} \tag{4.26}$$

We continue by considering the second component of $H_l^{(n)}(\boldsymbol{s}, t_l)$. Defining $\tau_l(t_l) := 1 - \delta_l(t_l) = 1 - \sigma_l e^{-t_l(1-\sigma_l)} = (1 - e^{-t_l(1-\sigma_l)}) + (1-\sigma_l)e^{-t_l(1-\sigma_l)}$ and using that $1 - \sigma_l \sim c_l n^{\alpha_l - 1} - c_l^2 n^{2\alpha_l - 2}$, we have for this component that

$$\begin{aligned} (C_n + 1)\log(1 - \tau_l(t_l)) &= -(C_n + 1)\left(\tau_l(t_l) + \frac{1}{2}\tau_l(t_l)^2 + o(\tau_l(t_l)^2)\right) \\ &= -C_n\left(\tau_l(t_l) + \frac{1}{2}\tau_l(t_l)^2 + o(\tau_l(t_l)^2)\right) + o(1). \end{aligned}$$

Observing that $\tau_l(t_l) \sim (1+t_l)(1-\sigma_l) - (\frac{1}{2}t_l^2 + t_l)(1-\sigma_l)^2$, the second component is thus asymptotically equivalent to

$$\begin{aligned}
&-C_n\left((1+t_l)(1-\sigma_l) + \frac{1}{2}(1-\sigma_l)^2 + o(1-\sigma_l)^2\right) \\
&\quad \sim -C_n\left((1+t_l)(c_l n^{\alpha_l-1} - c_l^2 n^{2\alpha_l-2}) + \frac{1}{2}c_l^2 n^{2\alpha_l-2}\right) \\
&\quad = -C_n\left((1+t_l)c_l n^{\alpha_l-1} - (\frac{1}{2}+t_l)c_l^2 n^{2\alpha_l-2}\right). \qquad (4.27)
\end{aligned}$$

When adding the two components displayed in (4.26) and (4.27), we conclude that $h_l^{(n)}(\boldsymbol{s}, t_l) = \exp(H_l^{(n)}(\boldsymbol{s}, t_l))$ as $n \to \infty$, with

$$\begin{aligned}
H_l^{(n)}(\boldsymbol{s}, t_l) &\sim (1+t_l)c_l n^{\alpha_l-1}\left(\sum_{r=1}^{R}\left(w_r n^{\nu_r} - \sqrt{w_r}s_r n^{\frac{1}{2}\nu_r}\right) - C_n\right) \\
&\quad + c_l^2 n^{2\alpha_l-2}\left(\sum_{r=1}^{R}\frac{1}{2}w_r t_l^2 n^{\nu_r} + C_n(t_l + \frac{1}{2})\right) \\
&= -(1+t_l)c_l n^{\alpha_l-1}\left(\sum_{r=1}^{R}\sqrt{w_r}s_r n^{\frac{1}{2}\nu_r} + \left(C_n - \sum_{r=1}^{R}w_r n^{\nu_r}\right)\right) \\
&\quad + c_l^2 n^{2\alpha_l-2}\left(\frac{1}{2}(1+t_l)^2\sum_{r=1}^{R}w_r n^{\nu_r} + (\frac{1}{2}+t_l)\left(C_n - \sum_{r=1}^{R}w_r n^{\nu_r}\right)\right) \\
&= -(1+t_l)c_l n^{\alpha_l-1}\left(\sum_{r=1}^{R}\sqrt{w_r}s_r n^{\frac{1}{2}\nu_r} + \beta n^{\gamma}\right) \\
&\quad + c_l^2 n^{2\alpha_l-2}\left(\frac{1}{2}(1+t_l)^2\sum_{r=1}^{R}w_r n^{\nu_r} + (\frac{1}{2}+t_l)\beta n^{\gamma}\right). \qquad (4.28)
\end{aligned}$$

We consider (4.28) as a reference point from now on, and distinguish four cases:

1. Suppose that $1-\alpha_1 > 1-\alpha_l$ and $1-\alpha_1 \geqslant \nu_1$. Then $\gamma = 1-\alpha_1 \geqslant \nu_1$, so (4.28) has leading term $-c_l(1+t_l)\beta n^{\alpha_l-1+\gamma}$, which tends to $-\infty$ as $n \to \infty$. Hence, in this case $h_l^{(n)}(\boldsymbol{s}, t_l) \to 0$ as $n \to \infty$.

2. If $\nu_1 > 1-\alpha_1 > 1-\alpha_l \geqslant \frac{1}{2}\nu_1$, then $n^{\alpha_l-1+\frac{1}{2}\nu_r} = O(1)$, hence

$$\begin{aligned}
h_l^{(n)}(\boldsymbol{s}, t_l) &= \exp\left(-(1+t_l)\left(O(1) + \beta c_l n^{1-\alpha_1}n^{\alpha_l-1}\right)\right) \\
&\quad \times \exp\left(O(1) + \beta n^{1-\alpha_1}(\frac{1}{2}+t_l)c_l^2 n^{2\alpha_l-2}\right) \\
&= \exp\left(-(1+t_l)\beta c_l n^{\alpha_l-\alpha_1} + O(1)\right) \to 0.
\end{aligned}$$

3. If $1-\alpha_1 = 1-\alpha_l > \frac{1}{2}\nu_1$, then $n^{\alpha_l-1+\frac{1}{2}\nu_1} = o(1)$, so with (4.28) we have

$$\begin{aligned}
h_l^{(n)}(\boldsymbol{s}, t_l) &= \exp\left(-(1+t_l)\left(O(n^{\alpha_l-1+\frac{1}{2}\nu_1}) + \beta n^{1-\alpha_1}c_l n^{\alpha_l-1}\right)\right) \\
&\quad \times \exp\left(\left(O(n^{2\alpha_l-2+\nu_1}) + O(n^{2\alpha_l-2-(\alpha_1-1)})\right)\right) \\
&= \exp\left(-(1+t_l)\beta c_l + o(1)\right) \to \exp\left(-\beta c_l(1+t_l)\right).
\end{aligned}$$

4. Finally, if $1-\alpha_1 = 1-\alpha_l = \frac{1}{2}\nu_1$, then with (4.28),

$$\begin{aligned}
h_l^{(n)}(\boldsymbol{s},t_l) &= \exp\Bigg(-c_l(1+t_l)\Bigg(\sum_{r=1}^{R}\sqrt{w_r}s_r n^{\alpha_l-1+\frac{1}{2}\nu_r} + \beta\Bigg) \\
&\qquad + \frac{1}{2}c_l^2(1+t_l)^2\sum_{r=1}^{R} w_r n^{2\alpha_l-2+\nu_r} + O(n^{2\alpha_l-2+\frac{1}{2}\nu_1})\Bigg) \\
&\to \exp\Bigg(-c_l(1+t_l)\Bigg(\beta + \sum_{r=1}^{R^-} s_r\sqrt{w_r}\Bigg) + \frac{1}{2}c_l^2\,(1+t_l)^2\sum_{r=1}^{R^-} w_r\Bigg) \\
&= \exp\Bigg(-c_l(1+t_l)\sqrt{W}\lambda(\boldsymbol{s}) + \frac{1}{2}c_l^2\,(1+t_l)^2\,W\Bigg) \\
&= \exp\Bigg(\frac{1}{2}\Big(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\Big)^2 - \frac{1}{2}\lambda(\boldsymbol{s})^2\Bigg) \\
&= \frac{\varphi\left(\lambda(\boldsymbol{s})\right)}{\varphi\big(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\big)}.
\end{aligned}$$

This completes the proof of Lemma 4.4.5. □

Lemma 4.4.6. *If* $1-\alpha_1 < \nu_1$ *and* $1-\alpha_l < \frac{1}{2}\nu_1$, *then* $g_l^{(n)}(\boldsymbol{s},t_l)\,h_l^{(n)}(\boldsymbol{s},t_l) \to 0$ *as* $n\to\infty$.

Proof. Let

$$x_l^{(n)} = \zeta^{(n)}(\boldsymbol{s})/\delta_l^{(n)}(t_l) = \sum_{r=1}^{R}\frac{\rho_r^{(n)}}{\sigma_l^{(n)}}\exp(-s_r/\sqrt{\rho_r^{(n)}})\,e^{t_l(1-\sigma_l^{(n)})};$$

in the sequel we write just x_n for brevity. In this proof, our first objective is to identify the asymptotics of $g_l^{(n)}(\boldsymbol{s},t_l) = \mathbb{P}(\mathscr{P}(x_n)\leqslant C_n)$. To this end, let $P_n =_{\mathrm{d}} \mathscr{P}(x_n)$, and let $\mathbb{Q}$ be an alternative measure under which this Poisson random variable has mean C_n, such that

$$g_l^{(n)}(\boldsymbol{s},t_l) = \mathbb{E}_{\mathbb{P}}\left(\mathbb{1}\{P_n\leqslant C_n\}\right) = \mathbb{E}_{\mathbb{Q}}\left(L\,\mathbb{1}\{P_n\leqslant C_n\}\right),$$

with L denoting the likelihood ratio or Radon-Nikodym derivative

$$L = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = \left(\frac{e^{-x_n}(x_n)^{P_n}}{P_n!}\right)\Bigg/\left(\frac{e^{-C_n}(C_n)^{P_n}}{P_n!}\right) = e^{C_n-x_n}\left(\frac{x_n}{C_n}\right)^{P_n}.$$

We thus arrive at

$$g_l^{(n)}(\boldsymbol{s},t_l) = e^{C_n-x_n}\mathbb{E}_{\mathbb{Q}}\Big((x_n/C_n)^{P_n}\;\mathbb{1}\{P_n\leqslant C_n\}\Big).$$

Define $\bar{P}_n := (P_n - C_n)/\sqrt{C_n}$ and recall that, by the central limit theorem, the distribution of $\bar{P}_n$ converges to a standard-normal distribution. In terms of this new random variable, we have

$$g_l^{(n)}(\boldsymbol{s},t_l) = e^{C_n-x_n}\left(\frac{x_n}{C_n}\right)^{C_n} q_n, \tag{4.29}$$

where

$$q_n := \mathbb{E}_{\mathbb{Q}}\left(\left((x_n/C_n)^{\sqrt{C_n}}\right)^{\bar{P}_n}\;\mathbb{1}\{\bar{P}_n\leqslant 0\}\right) = \int_{-\infty}^{0}\left(\left(\frac{x_n}{C_n}\right)^{\sqrt{C_n}}\right)^{y}\mathrm{d}F_{\bar{P}_n}(y), \tag{4.30}$$

with $F_{\bar{P}_n}(y)$ being the distribution function of $\bar{P}_n$. The idea is to show that $F_{\bar{P}_n}(y)$ behaves as a standard-normal distribution for sufficiently large C_n, and hence that

$$q_n \sim \int_{-\infty}^{0} \left(\left(\frac{x_n}{C_n} \right)^{\sqrt{C_n}} \right)^y \varphi(y)\,\mathrm{d}y,$$

where $\varphi(y)$ is the standard-normal density function in y. To formally achieve this, we bound $F_{\bar{P}_n}(y)$ using the Berry-Esseen theorem. This states that for all C_n large enough,

$$\sup_y \left| F_{\bar{P}_n}(y) - \Phi(y) - \frac{m_3}{6\sqrt{C_n}}(1-y^2)\varphi(y) - \varphi(y)l(y) \right| = O\left(\frac{1}{\sqrt{C_n}} \right),$$

where m_3 is the third moment of a Poisson(1) random variable and $l(\cdot)$ is a function that is bounded by a constant times $1/\sqrt{C_n}$.

We proceed by analyzing q_n using the Berry-Esseen theorem. Observe that (4.30) contains the density of $\bar{P}_n$, whereas 'Berry-Esseen' concerns a bound in terms of the corresponding distribution function. Therefore, we apply integration by parts, yielding

$$\begin{aligned} q_n &= \int_{-\infty}^{0} \left(\left(\frac{x_n}{C_n} \right)^{\sqrt{C_n}} \right)^y \mathrm{d}F_{\bar{P}_n}(y) = \int_{-\infty}^{0} e^{a_n y} \frac{\mathrm{d}}{\mathrm{d}y} \left(F_{\bar{P}_n}(y) - F_{\bar{P}_n}(0) \right) \mathrm{d}y \\ &= -\int_{-\infty}^{0} a_n e^{a_n y} \left(F_{\bar{P}_n}(y) - F_{\bar{P}_n}(0) \right) \mathrm{d}y, \end{aligned} \tag{4.31}$$

where $a_n := \sqrt{C_n}\log(x_n/C_n)$. Now applying the Berry-Esseen bound in (4.31),

$$\begin{aligned} q_n = &-\int_{-\infty}^{0} a_n e^{a_n y} \left(\Phi(y) - \Phi(0) \right) \mathrm{d}y + \int_{-\infty}^{0} a_n e^{a_n y} \frac{m_3}{6\sqrt{C_n}} \left((1-y^2)\varphi(y) - \varphi(0) \right) \mathrm{d}y \\ &+ \int_{-\infty}^{0} a_n e^{a_n y} \left(\varphi(y)l(y) - \varphi(0)l(0) \right) \mathrm{d}y + \int_{-\infty}^{0} a_n e^{a_n y} \cdot O\left(\frac{1}{\sqrt{C_n}} \right) \mathrm{d}y. \end{aligned}$$

Recall that $l(y) = O\left(1/\sqrt{C_n}\right)$, so the last three integrals contain a term of that order. For the first integral, we integrate by parts once more to obtain

$$q_n = \int_{-\infty}^{0} e^{a_n y} \varphi(y)\,dy + \frac{1}{\sqrt{C_n}} \int_{-\infty}^{0} a_n e^{a_n y} r(y)\,dy,$$

where $r(y)$ is bounded by a quadratic function. Observe that the second integral converges, so the second term is $O(1/\sqrt{C_n})$. Therefore, completing the square in the exponent,

$$\begin{aligned} q_n &= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left(a_n y - \frac{y^2}{2} \right) \mathrm{d}y + O\left(\frac{1}{\sqrt{C_n}} \right) \\ &= \exp\left(\frac{a_n^2}{2} \right) \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-a_n)^2 \right) \mathrm{d}y + O\left(\frac{1}{\sqrt{C_n}} \right) \\ &= \exp\left(\frac{a_n^2}{2} \right) \Phi(-a_n) + O\left(\frac{1}{\sqrt{C_n}} \right) \\ &= \exp\left(\frac{a_n^2}{2} \right) (1 - \Phi(a_n)) + O\left(\frac{1}{\sqrt{C_n}} \right), \end{aligned} \tag{4.32}$$

using the symmetry of the normal distribution. A known property of the tail of the normal distribution is

$$e^{x^2/2}\,(1-\Phi(x)) \sim \frac{1}{x\sqrt{2\pi}} \tag{4.33}$$

as $x \to \infty$ (cf. [45, p. 175]). To apply this property to the first term on the right-hand side of (4.32), it is necessary to verify that a_n goes to ∞ as $n \to \infty$. This can be seen by relying on a Taylor expansion, and recalling that $1-\alpha_1 < \nu_1$ and $1-\alpha_l < \frac{1}{2}\nu_1$:

$$\begin{aligned} a_n = \sqrt{C_n}\log\frac{x_n}{C_n} &= \sqrt{\sum_{r=1}^{R} w_r n^{\nu_r} + o(n^{\nu_1})}\cdot\log\left(\frac{\sum_{r=1}^{R} w_r n^{\nu_r}(1+c_l n^{\alpha_l-1})e^{-\frac{s_r}{\sqrt{w_r n^{\nu_r}}}}e^{t_l(1-\sigma_l^{(n)})}}{\sum_{r=1}^{R} w_r n^{\nu_r} + o(n^{\nu_1})}\right) \\ &= \Omega(n^{\frac{1}{2}\nu_1})\cdot\log\left(1+\Omega\left(n^{\alpha_l-1}\right)\right) = \Omega(n^{\frac{1}{2}\nu_1-(1-\alpha_l)}) \to \infty \end{aligned}$$

as $n \to \infty$, where we use the $\Omega(\cdot)$-notation of Section 1.4. Using property (4.33) in (4.32), and substituting the result in (4.29), we thus obtain that, as $n \to \infty$,

$$g_l^{(n)}(\boldsymbol{s}, t_l) = e^{C_n - x_n}\left(\frac{x_n}{C_n}\right)^{C_n}\left(\frac{1}{a_n\sqrt{2\pi}} + O\left(\frac{1}{\sqrt{C_n}}\right)\right) = e^{C_n - x_n}\left(\frac{x_n}{C_n}\right)^{C_n}\cdot o(1).$$

Multiplying with $h_l^{(n)}(\boldsymbol{s}, t_l)$, and using $\delta_l^{(n)}(t_l)x_n = \zeta^{(n)}(\boldsymbol{s})$, it holds that

$$\begin{aligned} g_l^{(n)}(\boldsymbol{s}, t_l)\, h_l^{(n)}(\boldsymbol{s}, t_l) &= e^{C_n - x_n}\left(\frac{x_n}{C_n}\right)^{C_n}\cdot o(1)\cdot e^{x_n - \zeta^{(n)}(\boldsymbol{s})}\left(\delta_l^{(n)}(t_l)\right)^{C_n+1} \\ &= o(1)\cdot e^{C_n - \zeta^{(n)}(\boldsymbol{s})}\left(\frac{\zeta^{(n)}(\boldsymbol{s})}{C_n}\right)^{C_n}. \end{aligned}$$

The stated result now follows from writing all terms as exponentials and applying the Taylor expansion to the logarithm:

$$\begin{aligned} g_l^{(n)}(\boldsymbol{s}, t_l)\, h_l^{(n)}(\boldsymbol{s}, t_l) &= o(1)\exp\left(C_n - \zeta^{(n)}(\boldsymbol{s}) + C_n\log\left(1+\left(\frac{\zeta^{(n)}(\boldsymbol{s}) - C_n}{C_n}\right)\right)\right) \\ &= o(1)\exp\left(C_n - \zeta^{(n)}(\boldsymbol{s}) + C_n\frac{\zeta^{(n)}(\boldsymbol{s}) - C_n}{C_n} + O(1)\right) \to 0, \end{aligned}$$

as $n \to \infty$. □

Appendix 4.B Proof of Corollary 4.3.3

Proof of Corollary 4.3.3. This proof mimics the proof of Theorem 4.3.1. For convenience we write

$$T_n(\boldsymbol{s}, \boldsymbol{t}) := \mathbb{E}\left(\prod_{r=1}^{I-1} e^{-s_r \bar{B}_r^{(n)}}\prod_{m=I}^{R} e^{-s_m B_m^{(n)}}\prod_{k=1}^{J-1} e^{-t_k \bar{D}_k^{(n)}}\prod_{l=J}^{K} e^{-t_l D_l^{(n)}}\right),$$

which differs from $P_n(\boldsymbol{s}, \boldsymbol{t})$ in the fact that $B_I^{(n)}, \ldots, B_R^{(n)}$ and $D_J^{(n)}, \ldots, D_K^{(n)}$ are unscaled. We now follow the line of the proof of Lemma 4.4.1, with a few adjustments for the unscaled random variables:

- the factors $e^{s_m\sqrt{\rho_m^{(n)}}}$ are removed, for $m = I, \ldots, R$,
- the variables $\zeta_m^{(n)}(s_m)$ are defined as $\rho_m^{(n)} e^{-s_m}$ rather than $\rho_m^{(n)} e^{-s_m/\sqrt{\rho_m^{(n)}}}$, for $m = I, \ldots, R$,
- $t_l(1-\sigma_l)$ is replaced by t_l, for $l = J, \ldots, K$.

Hence, $T_n(\boldsymbol{s}, \boldsymbol{t})$ equals the right-hand side of (4.18) subject to the adjustments above. The first term of $T_n(\boldsymbol{s}, \boldsymbol{t})$ then equals

$$\prod_{r=1}^{I-1} e^{-\rho_r^{(n)} + s_r\sqrt{\rho_r^{(n)}} + \zeta_r^{(n)}(s_r)} \prod_{m=I}^{R} e^{\rho_m^{(n)}(e^{-s_m}-1)}.$$

In this expression, the first $I-1$ factors are as in (4.20) and converge to standard-normal LSTs as $n \to \infty$. We recognize the latter $R - I + 1$ factors as LSTs of Poisson random variables with mean $\rho_m^{(n)}$.

The second term of $T_n(\boldsymbol{s}, \boldsymbol{t})$ equals

$$\prod_{k=1}^{J-1} \frac{1-\sigma_k^{(n)}}{1-\delta_k^{(n)}(t_k)} \prod_{l=J}^{K} \frac{1-\sigma_l^{(n)}}{1-\sigma_l^{(n)} e^{-t_l}}.$$

With (4.21), the first $J-1$ factors of this expression converge to LSTs of unit-rate exponential random variables as $n \to \infty$, and the second $K - J + 1$ are easily identified as geometric LSTs with parameter $1 - \sigma_l^{(n)}$.

Following the proofs of Lemmas 4.4.3–4.4.6, it can be seen that the adjustments do not change the asymptotic behavior of the last term of $T_n(\boldsymbol{s}, \boldsymbol{t})$. Therefore, the result follows from Theorem 4.3.1 and recognizing known LSTs in the first two terms of $T_n(\boldsymbol{s}, \boldsymbol{t})$ as described above. □

Appendix 4.C Proof of Corollary 4.3.1

Proof of Corollary 4.3.1. This proof amounts to verifying that the LST corresponding to $(\mathscr{N}, \mathscr{E} \mid Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \leqslant \beta)$ equals the right-hand side of (4.13). This can be done with standard integration techniques. In this section we illustrate the proof for the case that $1 - \alpha_1 = \frac{1}{2}\nu_1$. We leave out the other two cases, as these can be verified using the precise same steps.

Rather than the LST of $(\mathscr{N}, \mathscr{E} \mid Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \leqslant \beta)$, we consider in this section the LST of $(\mathscr{N}_{R^-}, \mathscr{E}_{K^-} \mid Z(\mathscr{N}_{R^-}, \mathscr{E}_{K^-}) \leqslant \beta)$, which we call $Q(\boldsymbol{s}, \boldsymbol{t})$. The former LST can simply be obtained through multiplying $Q(\boldsymbol{s}, \boldsymbol{t})$ by $R - R^-$ standard-normal LSTs and $K - K^-$ unit-rate exponential LSTs. For $1 - \alpha_1 = \frac{1}{2}\nu_1$, observe that the right-hand side of (4.13) equals

$$\left(\prod_{r=1}^{R^-} e^{\frac{1}{2}s_r^2}\right)\left(\prod_{k=1}^{K^-} \frac{1}{1+t_k}\right) \cdot \frac{\varphi(\lambda(\boldsymbol{s}))}{\varphi(\lambda(\boldsymbol{0}))} \cdot \frac{\Psi(\lambda(\boldsymbol{s})) - \sum_{l=1}^{K^-}\left(\prod_{j=1,\, j\neq l}^{K^-} \kappa_{jl}(\boldsymbol{t})\right)\Psi\left(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\right)}{\Psi(\lambda(\boldsymbol{0})) - \sum_{l=1}^{K^-}\left(\prod_{j=1,\, j\neq l}^{K^-} \kappa_{jl}(\boldsymbol{0})\right)\Psi\left(\lambda(\boldsymbol{0}) - c_l\sqrt{W}\right)}. \tag{4.34}$$

Our objective is to show that $Q(\boldsymbol{s},\boldsymbol{t})$ equals (4.34).

To simplify the notation, we write

$$p := \mathbb{P}\left(\sum_{r=1}^{R^-}\sqrt{w_r}\mathscr{N}_r + \sum_{k=1}^{K^-}\frac{1}{c_k}\mathscr{E}_k \leqslant \beta\right),$$
$$\widehat{b} := \sum_{r=1}^{R^-} b_r\sqrt{w_r}, \quad \text{and}$$
$$\widehat{d}_j := \sum_{k=1}^{j} d_k/c_k.$$

By definition of the LST, and with $\mathrm{d}(\boldsymbol{b}_{R^-},\boldsymbol{d}_k)$ denoting an abbreviation for $\mathrm{d}b_1...\mathrm{d}b_{R^-}\,\mathrm{d}d_1...\mathrm{d}d_k$, it follows that

$$\begin{aligned}
Q(\boldsymbol{s},\boldsymbol{t}) &= \mathbb{E}\left(\left(\prod_{r=1}^{R^-} e^{-s_r\mathscr{N}_r}\right)\left(\prod_{k=1}^{K^-} e^{-t_k\mathscr{E}_k}\right) \,\Big|\, \sum_{r=1}^{R^-}\sqrt{w_r}\mathscr{N}_r + \sum_{k=1}^{K^-}\frac{1}{c_k}\mathscr{E}_k \leqslant \beta\right)\\
&= \int\limits_{\boldsymbol{b}_{R^-},\boldsymbol{d}_{K^-}:\,\widehat{b}+\widehat{d}_{K^-}\leqslant\beta} \left(\prod_{r=1}^{R^-} e^{-s_r b_r}\varphi(b_r)\right)\cdot\left(\prod_{k=1}^{K^-} e^{-t_k d_k}e^{-d_k}\right)\cdot\frac{1}{p}\,\mathrm{d}(\boldsymbol{b}_{R^-},\boldsymbol{d}_{K^-})\\
&= \frac{1}{p}\left(\prod_{r=1}^{R^-} e^{\frac{1}{2}s_r^2}\right)\int\limits_{\boldsymbol{b}_{R^-},\boldsymbol{d}_{K^-}:\,\widehat{b}+\widehat{d}_{K^-}\leqslant\beta}\left(\prod_{r=1}^{R^-}\varphi\left(b_r+s_r\right)\right)\cdot\left(\prod_{k=1}^{K^-} e^{-(t_k+1)d_k}\right)\,\mathrm{d}(\boldsymbol{b}_{R^-},\boldsymbol{d}_{K^-}).
\end{aligned}$$

This expression may be compared to Expression (4.3), where we encountered a large summation containing products of Poisson-type and geometric-type factors. Here, we have its continuous version: an integral containing products of normal and exponential densities. This effectively means that the proof steps are similar to those of Lemmas 4.2.1 and 4.2.2: we give a recursive argument to evaluate the integrals over exponential densities, and a probabilistic approach is used for the integrals over normal densities.

For intermediate steps where $j \leqslant K^-$ integrals over exponential densities are left, define

$$\begin{aligned}
V_j(x) := e^{-\lambda(\boldsymbol{s})\sqrt{W}x+\frac{1}{2}Wx^2} &\int\limits_{\boldsymbol{b}_{R^-}:\,\widehat{b}\leqslant\beta}\left(\prod_{r=1}^{R^-}\varphi\left(b_r+s_r-\sqrt{w_r}x\right)\right)\\
\times \int\limits_{d_1=0}^{c_1\left(\beta-\widehat{b}\right)} e^{\left(\frac{x}{c_1}-(t_1+1)\right)d_1}\cdots &\int\limits_{d_j=0}^{c_j\left(\beta-\widehat{b}-\widehat{d}_{j-1}\right)} e^{\left(\frac{x}{c_j}-(t_j+1)\right)d_j}\,\mathrm{d}(\boldsymbol{b}_{R^-},\boldsymbol{d}_j),
\end{aligned} \tag{4.35}$$

and notice that $Q(\boldsymbol{s},\boldsymbol{t}) = p^{-1}\left(\prod_{r=1}^{R^-} e^{\frac{1}{2}s_r^2}\right)V_{K^-}(0)$.

Lemma 4.C.1. *$V_j(x)$ satisfies the recursion*

$$V_j(x) = \frac{c_j}{c_j(1+t_j)-x}\left(V_{j-1}(x) - V_{j-1}(c_j(1+t_j))\right).$$

Proof. Integrating (4.35) over d_j yields

$$V_j(x) = e^{-\lambda(\boldsymbol{s})\sqrt{W}x+\frac{1}{2}Wx^2} \int\limits_{\boldsymbol{b}_{R^-}:\hat{b}\leqslant\beta} \left(\prod_{r=1}^{R^-} \varphi\left(b_r + s_r - \sqrt{w_r}x\right)\right)$$
$$\times \int\limits_{d_1=0}^{c_1\left(\beta-\hat{b}\right)} e^{\left(\frac{x}{c_1}-(1+t_1)\right)d_1} \cdots \int\limits_{d_{j-1}=0}^{c_{j-1}\left(\beta-\hat{b}-\hat{d}_{j-2}\right)} e^{\left(\frac{x}{c_{j-1}}-(1+t_{j-1})\right)d_{j-1}} \,\mathrm{d}(\boldsymbol{b}_{R^-}, \boldsymbol{d}_{j-1})$$
$$\times \frac{c_j}{c_j(1+t_j)-x}\left(1 - e^{(x-c_j(1+t_j))\left(\beta-\hat{b}-\hat{d}_{j-1}\right)}\right).$$

Observe that the last exponential contains the indices $b_1, ..., b_{R^-}$ and $d_1, ..., d_{j-1}$. Carefully distributing these indices over the corresponding integrals gives

$$V_j(x) = \frac{c_j}{c_j(1+t_j)-x}\Bigg(V_{j-1}(x)$$
$$- e^{-\lambda(\boldsymbol{s})\sqrt{W}c_j(1+t_j)+\frac{1}{2}\left(c_j(1+t_j)\sqrt{W}\right)^2} \int\limits_{\boldsymbol{b}_{R^-}:\hat{b}\leqslant\beta} \left(\prod_{r=1}^{R^-} \varphi\left(b_r + s_r - \sqrt{w_r}c_j(1+t_j)\right)\right)$$
$$\int\limits_{d_1=0}^{c_1\left(\beta-\hat{b}\right)} e^{\left(\frac{c_j(1+t_j)}{c_1}-(1+t_1)\right)d_1} \cdots \int\limits_{d_{j-1}=0}^{c_{j-1}\left(\beta-\hat{b}-\hat{d}_{j-2}\right)} e^{\left(\frac{c_j(1+t_j)}{c_{j-1}}-(1+t_{j-1})\right)d_{j-1}} \,\mathrm{d}(\boldsymbol{b}_{R^-}, \boldsymbol{d}_{j-1})\Bigg)$$
$$= \frac{c_j}{c_j(1+t_j)-x}\left(V_{j-1}(x) - V_{j-1}(c_j(1+t_j))\right),$$

yielding the stated. □

We are finally ready to show that $Q(\boldsymbol{s},\boldsymbol{t})$ is given by (4.34), which proves Corollary 4.3.1. We proceed as follows: first, we use Lemma 4.C.1 to write $Q(\boldsymbol{s},\boldsymbol{t})$ in terms of $V_0(x)$, for certain x. A probabilistic argument subsequently gives an expression for the integrals in $V_0(x)$. Some rearrangements then lead to the equality of $Q(\boldsymbol{s},\boldsymbol{t})$ and (4.34).

Since $Q(\boldsymbol{s},\boldsymbol{t}) = p^{-1}\prod_{r=1}^{R^-} e^{\frac{1}{2}s_r^2} V_{K^-}(0)$, we are interested in the value of $V_{K^-}(0)$. For this variable, Lemma 4.C.1 implies

$$V_{K^-}(0) = \frac{1}{1+t_{K^-}}\left(V_{K^--1}(0) - V_{K^--1}(c_{K^-}(1+t_{K^-}))\right).$$

Iterating K^- times gives an expression of the form

$$V_{K^-}(0) = a\,V_0(0) + \sum_{l=1}^{K^-} u_l\,V_0\left(c_l(1+t_l)\right),$$

where a and $u_1, ..., u_{K^-}$ are coefficients depending on $c_1, ..., c_{K^-}$ and $t_1, ..., t_{K^-}$. To find a, observe that the only term with $V_0(0)$ results from repeatedly taking the left term of all K^- iterations. Therefore, $a = \prod_{k=1}^{K^-}(1+t_k)^{-1}$. Similarly, observe that the only term with $V_0\left(c_{K^-}(1+t_{K^-})\right)$

results from taking the right term in the first iteration and then repeatedly taking the left term of the remaining iterations. Therefore,

$$u_{K^-} = -\frac{1}{1+t_{K^-}} \prod_{j=1}^{K^--1} \frac{c_j}{c_j(1+t_j) - c_{K^-}(1+t_{K^-})}.$$

By symmetry, we conclude that, for any $l = 1, ..., K^-$,

$$u_l = -\frac{1}{1+t_l} \prod_{j=1,\, j\neq l}^{K^-} \frac{c_j}{c_j(1+t_j) - c_l(1+t_l)}.$$

Thus, it holds that

$$\begin{aligned} V_{K^-}(0) &= \left(\prod_{k=1}^{K^-} \frac{1}{1+t_k}\right) V_0(0) - \sum_{l=1}^{K^-} \frac{1}{1+t_l} \left(\prod_{j=1,\, j\neq l}^{K^-} \frac{c_j}{c_j(1+t_j) - c_l(1+t_l)}\right) V_0\left(c_l(1+t_l)\right) \\ &= \left(\prod_{k=1}^{K^-} \frac{1}{1+t_k}\right) \cdot \left(V_0(0) - \sum_{l=1}^{K^-} \left(\prod_{j=1,\, j\neq l}^{K^-} \kappa_{jl}(\boldsymbol{t})\right) V_0\left(c_l(1+t_l)\right)\right). \end{aligned} \tag{4.36}$$

With expression (4.36) at hand, $V_0(x)$ still needs to be analyzed. Using (4.35) and observing that its integral can be written as a probability involving R^- normal random variables, we have

$$\begin{aligned} V_0(x) &= e^{-\lambda(\boldsymbol{s})\sqrt{W}x + \frac{1}{2}Wx^2} \int\limits_{\boldsymbol{b}_{R^-}:\, \hat{b} \leqslant \beta} \left(\prod_{r=1}^{R^-} \varphi\left(b_r + s_r - \sqrt{w_r}x\right)\right) \mathrm{d}\boldsymbol{b}_{R^-} \\ &= e^{\frac{1}{2}\left(\lambda(\boldsymbol{s}) - \sqrt{W}x\right)^2} e^{-\frac{1}{2}\lambda(\boldsymbol{s})^2} \mathbb{P}\left(\sum_{r=1}^{R^-} \sqrt{w_r}\mathcal{N}\left(\sqrt{w_r}x - s_r, 1\right) \leqslant \beta\right) \\ &= \frac{\varphi(\lambda(\boldsymbol{s}))}{\varphi\left(\lambda(\boldsymbol{s}) - \sqrt{W}x\right)} \mathbb{P}\left(\mathcal{N}\left(w_r x - \sqrt{w_r}s_r, W\right) \leqslant \beta\right) \\ &= \frac{\varphi(\lambda(\boldsymbol{s}))}{\varphi\left(\lambda(\boldsymbol{s}) - \sqrt{W}x\right)} \Phi\left(\lambda(\boldsymbol{s}) - \sqrt{W}x\right) = \varphi(\lambda(\boldsymbol{s}))\Psi\left(\lambda(\boldsymbol{s}) - \sqrt{W}x\right). \end{aligned} \tag{4.37}$$

Substituting (4.37) into (4.36), we conclude that

$$\begin{aligned} Q(\boldsymbol{s}, \boldsymbol{t}) = \frac{1}{p}\left(\prod_{r=1}^{R^-} e^{\frac{1}{2}s_r^2}\right) V_{K^-}(0) &= \frac{1}{p}\left(\prod_{r=1}^{R^-} e^{\frac{1}{2}s_r^2}\right)\left(\prod_{k=1}^{K^-} \frac{1}{1+t_k}\right) \varphi(\lambda(\boldsymbol{s})) \\ &\times \left(\Psi\left(\lambda(\boldsymbol{s})\right) - \sum_{l=1}^{K^-} \left(\prod_{j=1,\, j\neq l}^{K^-} \kappa_{jl}(\boldsymbol{t})\right) \Psi\left(\lambda(\boldsymbol{s}) - c_l(1+t_l)\sqrt{W}\right)\right). \end{aligned}$$

We now find the value of p by using that $Q(\mathbf{0}, \mathbf{0}) = 1$. We then indeed have that $Q(\boldsymbol{s}, \boldsymbol{t})$ equals (4.34), which completes the proof. □

Appendix 4.D Table of definitions

Object	Definition
w_r $(r = 1, ..., R)$	positive parameter affecting the traffic load at station r
c_k $(k = 1, ..., K)$	positive parameter affecting the traffic load at station k
ν_r $(r = 1, ..., R)$	real parameter affecting the traffic load at station r, where $\nu_1 \geqslant ... \geqslant \nu_R$
α_k $(k = 1, ..., K)$	real parameter affecting the traffic load at station k, where $\alpha_1 \leqslant ... \leqslant \alpha_K$
β	positive parameter affecting the population size
γ	parameter affecting the population size, defined as $\gamma = \max\{1 - \alpha_1, \frac{1}{2}\nu_1\}$
$\rho_r^{(n)}$ $(r = 1, ..., R)$	traffic load at station r, defined as $\rho_r^{(n)} = w_r n^{\nu_r}$
$\sigma_k^{(n)}$ $(k = 1, ..., K)$	traffic load at station k, defined as $\sigma_k^{(n)} = n/(n + c_k n^{\alpha_k})$
C_n	population size of the network, defined as $C_n = \lfloor \lVert \boldsymbol{\rho}^{(n)} \rVert + \beta n^{\gamma} \rfloor$
$\overline{B}_r^{(n)}$ $(r = 1, ..., R)$	scaled queue length at station r, defined as $\overline{B}_r^{(n)} = (B_r^{(n)} - \rho_r^{(n)})/\sqrt{\rho_r^{(n)}}$
$\overline{D}_k^{(n)}$ $(k = 1, ..., K)$	scaled queue length at station k, defined as $\overline{D}_k^{(n)} = (1 - \sigma_k^{(n)})D_k^{(n)}$
R^-	number of heaviest loaded infinite-server stations (i.e. largest integer such that $\nu_1 = \nu_2 = ... = \nu_{R^-}$)
K^-	number of heaviest loaded single-server stations (i.e. largest integer such that $\alpha_1 = \alpha_2 = ... = \alpha_{K^-}$)

Table 4.5: Definitions of the main scaling parameters and random variables.

Part III

Extremes of Markov additive processes

Introduction

The *Markov additive process* (throughout the manuscript abbreviated to MAP) can be seen as a generalized Markov-modulated version of the Lévy process. Indeed, when an independently evolving continuous-time Markov chain on a finite state space, usually referred to as the *background process*, is in state i, the MAP behaves as a Lévy process $X_i(\cdot)$. Additionally, a MAP allows for jumps at transition epochs of the background process. As such, MAPs offer a natural modeling framework to study stochastic processes of which the behavior changes over time, with broad applications in e.g. credit and risk theory, inventory management, finance and queueing. This last application may seem unrelated, but MAPs come into play when studying the workload process of Markov-modulated queues. Early references on MAPs include [33, 92].

A key object of study concerns the *extreme values* attained by the MAP over a finite or infinite horizon. With $Y(\cdot)$ denoting the MAP under consideration, the focus of this study is on the analysis of the distribution of its *running maximum* process $\overline{Y}(t) := \sup_{s\in[0,t]} Y(s)$ (as well as the corresponding running minimum process). Besides being interesting in its own right, the running maximum process can be directly translated in terms of the first-passage process $\tau(y) := \inf\{s \geqslant 0 : Y(s) > y\}$, due to the known duality between the events $\{\overline{Y}(t) > y\}$ and $\{\tau(y) < t\}$. Building upon related results for Lévy processes, a wide range of characterizations has been derived, typically in terms of Laplace transforms or so-called scale functions. We refer to Ivanovs [59, Chapter 2] for an extensive account of the main results on extremes of MAPs as well as the corresponding first-passage process. Particularly noteworthy are the results obtained by Asmussen and Kella [11], who used martingale methods to effectively extend the Pollaczek-Khinchine formula for a general class of Lévy processes to the MAP setting. We in addition mention the work by Dieker and Mandjes [40] as well as D'Auria et al. [43] on the joint distribution of the maximum and the epoch at which it occurs (the latter work being predominantly in terms of the first-passage process).

This part of the thesis addresses two gaps in the existing literature on extremes of MAPs. In the first place, the analysis of most papers requires the background process to have a unique stationary distribution, and thus the assumption is made that it is irreducible. Sometimes an absorbing state is added through the concept of "killing" (see e.g. [59, Section 2.5]), but this allows for only one transient class and a single recurrent state. To the best of our knowledge, the only work considering multiple transient classes is [38], in which the background chain is essentially a counting process. In contrast, we make no assumption on the structure of the background chain in Chapter 5, hence allowing for any number of transient and recurrent classes. This generalization is particularly useful when modeling processes of which the behavior

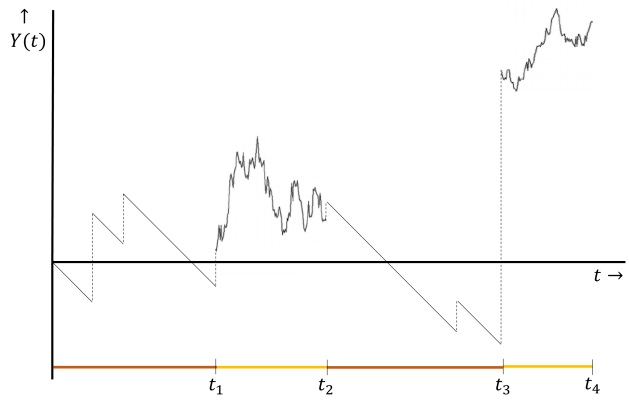

Figure III.1: Example of a MAP with two background states. Here, during $[0, t_1) \cup [t_2, t_3)$ the MAP behaves as the Lévy process $X_1(\cdot)$ and during $[t_1, t_2) \cup [t_3, t_4)$ it behaves as the Lévy process $X_2(\cdot)$.

may change irreversibly, such as models with catastrophes.

In the second place, general results on the distribution of a spectrally-positive MAP's maximum are usually formulated in terms of Laplace transforms [11, 30, 40, 59, 74]. Although there exist various numerical techniques to retrieve the cumulative distribution function from such a transform, none of these remain accurate further along the tail. Employing a change of measure, we derive transform-free results in Chapter 6. We specifically find an expression for the asymptotic tail probability of the maximum, which is suitable for insurance risk applications.

We proceed by discussing a number of commonalities between Chapters 5 and 6, including a description of the model, two preliminary results, and a sketch of the analysis approach.

Model

We give the formal definition of a MAP. Let the background process $(J(t))_{t\geq 0}$ be a continuous-time Markov chain with $d \in \mathbb{N}$ states, generator matrix $Q := (q_{ij})_{i,j=1}^{d}$, and $q_i := -q_{ii} > 0$. Associated with every state i let $(X_i(t))_{t\geq 0}$ be a Lévy process, and assume that these Lévy processes are mutually independent. Let t_n denote the time of the n-th transition of the background chain. If $J(0) = j_0$, we have that $Y(t) = X_{j_0}(t)$ for $t \in [0, t_1)$, and, if the transition at t_n is from (say) i to j (where $i \neq j$), then

$$Y(t) := Y(t_n-) + L_{ij}^n + \left(X_j(t) - X_j(t_n)\right),$$

where $t \in [t_n, t_{n+1})$, $n \in \mathbb{N}$, and $\left(L_{ij}^n\right)_{n\in\mathbb{N}}$ is a sequence of independent copies of the random variable L_{ij}, representing the size of the jump at the time of a transition from background state i to background state j (where $i \neq j$). We assume that these jumps are independent of the underlying Lévy processes and of the background process. Since jumps at self-transitions, say from background state i to itself, can be incorporated in the Lévy process $X_i(\cdot)$, we assume without loss of generality that there are no such self-transitions. An example of a MAP is shown in Figure III.1.

Preliminaries and method

We now mention two important results that are essential in Chapters 5 and 6. The first of these results, the Wiener-Hopf decomposition, shows that the value of the Lévy process at an exponentially distributed epoch can be written as the difference between two independent non-negative quantities. In case the Lévy process is spectrally one-sided, these distributions can be characterized explicitly in terms of the model primitives; notably, one of the two quantities is exponentially distributed.

The second result, referred to as the Number of Zeroes proposition in the sequel, concerns a property of the *matrix exponent* of a spectrally one-sided MAP. The matrix exponent should be interpreted as a matrix function that has a one-on-one correspondence with a MAP, similarly to how the Laplace-Stieltjes transform has a one-on-one correspondence with a random variable. The Number of Zeroes proposition quantifies the number of zeroes of the matrix exponent's determinant.

These two results are formally stated in Chapters 5 and 6 (in the specific form suitable for that chapter), and constitute the main ingredients in the common recipe followed in both chapters. First, the Wiener-Hopf decomposition is used to relate the values of the MAP at two successive transition epochs. The resulting expression is then turned into a system of linear equations involving the matrix exponent, which still contains a number of unknown constants. These unknowns are then identified by combining Cramer's rule and the Number of Zeroes proposition.

5 Extreme values of Markov additive processes with a non-irreducible background process

5.1 Introduction

In the literature on extremes for MAPs one typically assumes that the background process is *irreducible*. This assumption is convenient, as it guarantees existence and uniqueness of its invariant distribution, and allows the use of time-reversal arguments. On the other hand, the assumed irreducibility obviously restricts the general applicability of the model. There may be events, such as a crash of a stock market, that cause permanent changes in the fluctuations of the price levels. The resulting process can naturally be modeled using a non-irreducible background process. Furthermore, in the context of credit risk, one could think of companies paying interest to an obligor until they go into default, thus causing a loss to the obligor, after which they effectively leave the system — another setting that can be modelled using a non-irreducible background process. This credit-related example motivated Delsing and Mandjes [38] to consider the extreme values attained by a MAP of a specific Cramér-Lundberg type, endowed with a non-irreducible background process of a specific structure. Importantly, however, the topic of extremes of more general MAPs under a non-irreducible background process was so far largely unexplored.

The main objective of this chapter is to extend existing results for extremes of MAPs to the case of an arbitrary Markovian background process, thus covering the situation that the background process is non-irreducible. Concretely, we describe the distribution of the maximum attained by the MAP until the background process reaches an absorbing state (which covers the running maximum over an exponentially distributed horizon, and in particular also the all-time maximum). With i representing the initial state of the background chain, we let this maximum be denoted by Z_i. The main results in this chapter provide the distribution of Z_i for all $i = 1, ..., d$ (where d is the number of background states), for a spectrally one-sided MAP (i.e., the direction of all jumps is either positive or negative). As a by-product of this result, we also succeed in deriving the distribution of the maximum of a spectrally one-sided Lévy process over a phase-type distributed time interval.

The way in which our results are obtained is transparent and remarkably straightforward. Our approach is based on the two key results described in the introduction of Part III. Firstly, the Wiener-Hopf decomposition for spectrally one-sided Lévy processes [37, 73] allows us to describe the dynamics of the MAP between two successive transition epochs of the background process. Standard analytic techniques are then used to transform the resulting expressions

into a linear system of equations from which the distribution of the MAP's maximum follows. The second key result, as was established by Ivanovs et al. [61], characterizes the number of singularities with positive real part of the matrix exponent corresponding to a spectrally one-sided MAP. Using this result we find a procedure for obtaining the solution to the system of equations, consequently determining the distribution of Z_i for all i. Notably, the approach does not rely on the use of martingale methods.

This chapter is structured as follows. In Section 5.2 we describe our model, introduce the necessary notation, and state the two important preliminary results that were mentioned above. For the spectrally-positive case (no downward jumps, that is), as treated in Section 5.3, we derive a system of linear equations for the transforms of Z_i, $i = 1, ..., d$. To solve these equations we introduce an ordering on the communicating classes of the background chain, which allows us to recursively determine the distribution of Z_i for all initial states $i = 1, ..., d$. In Section 5.4, covering the spectrally-negative case (no upward jumps, that is), we derive a similar system of equations for the complementary distribution function of Z_i, $i = 1, ..., d$. By using the fact that the solution is of a specific form, we show how to solve this system. Where in most of the literature on extreme values of Lévy processes one considers extreme values over an exponentially distributed interval, in Section 5.5 it is pointed out how we can use our MAP-based framework to extend such results to extreme values of Lévy processes over a phase-type time interval. Section 5.6 describes a series of numerical experiments illustrating the use of our results, and presents some implementation guidelines. Concluding remarks are made in Section 5.7.

5.2 Model and preliminaries

In this section we first describe the model that we consider in this chapter. This is followed by the introduction of some notation and a description of our key objective. We then discuss the Wiener-Hopf decomposition for spectrally one-sided Lévy processes and the Number of Zeroes proposition, i.e., the results mentioned in the introduction of Part III, which play a crucial role in our reasoning. We conclude this section by explicitly outlining the approach that will be followed.

Model

In this chapter, we consider a spectrally one-sided MAP $Y(\cdot)$. The MAP is characterized by its underlying Lévy processes $X_1(\cdot), ..., X_d(\cdot)$, its background process $J(\cdot)$ with generator matrix $Q := (q_{ij})_{i,j=1}^d$ (with $q_i := -q_{ii}$), and its jumps distributed as L_{ij} whenever a transition from state i to state j occurs. We assume by convention that $Y(0) = 0$.

Additionally, we incorporate (state-dependent) *killing*, which happens with rate $\vartheta_i \geqslant 0$ when the background process is in state i. At the moment the MAP is killed, it remains constant indefinitely, such that the running maximum becomes the all-time maximum of the process. Alternatively, killing can be thought of as reaching an absorbing background state that

corresponds to a Lévy process that is identical to 0. Various specific choices of the killing rates ϑ_i are of interest. When choosing $\vartheta_i = \vartheta > 0$ for all i, for instance, we consider the running maximum over an exponentially distributed horizon with mean $1/\vartheta$. In addition, the choice $\vartheta_i = 0$ for all i corresponds to the all-time maximum. We also note that with a specific choice of the rates ϑ_i we can analyze the maximum of a Lévy process over a phase-type interval, as argued in Section 5.5.

Denoting by Δ the killing time of the MAP, its distribution is characterized by the following system of equations:

$$\mathbb{E}\big(e^{-\alpha\Delta} \mid J(0) = i\big) = \frac{q_i + \vartheta_i}{q_i + \vartheta_i + \alpha}\left(\sum_{j\neq i} \frac{q_{ij}}{q_i + \vartheta_i}\mathbb{E}\big(e^{-\alpha\Delta} \mid J(0) = j\big) + \frac{\vartheta_i}{q_i + \vartheta_i}\right),$$

for $\alpha \geqslant 0$. This system of equations follows by observing that the time till the first event (being either a transition of the background process $J(\cdot)$ or killing) is exponentially distributed with rate $q_i + \vartheta_i$; then one needs to distinguish between the background state becoming j (for $j \neq i$) and killing.

Notation

Each spectrally-positive Lévy process $X_i(\cdot)$ is characterized by its *Laplace exponent* $\varphi_i(\cdot)$, as given by

$$\varphi_i(\alpha) := \log \mathbb{E}\big(e^{-\alpha X_i(1)}\big)$$

for $\alpha \geqslant 0$, where its right inverse is denoted by $\psi_i(\cdot)$. Similarly, for a spectrally-negative Lévy process $X_i(\cdot)$, we consider the *cumulant generating function*

$$\Phi_i(\alpha) := \log \mathbb{E}\big(e^{\alpha X_i(1)}\big)$$

for $\alpha \geqslant 0$, with $\Psi_i(\cdot)$ denoting its right inverse. To be interpreted as the MAP-counterpart of the Laplace exponent defined above, let $M(\alpha) = (m_{ij}(\alpha))_{i,j=1,\ldots,d}$ (referred to as the matrix exponent) be the matrix with coefficients

$$m_{ij}(\alpha) := q_{ij}\,\mathbb{E}\big(e^{-\alpha L_{ij}}\big) + \varphi_i(\alpha)\,\mathbb{1}_{\{i=j\}} - \vartheta_i\,\mathbb{1}_{\{i=j\}};$$

here all $X_i(\cdot)$ are assumed spectrally positive, and the jumps L_{ij} are assumed non-negative almost surely. Later, we will also work with a similar object for the spectrally-negative case (in which the jumps L_{ij} are assumed non-positive almost surely); this MAP-counterpart of the cumulant generating function will be introduced at the beginning of Section 5.4.

Objective

Our aim is to analyze the distribution of Z_i, i.e., the maximum of the MAP under state-dependent killing, conditional on the initial background state being i:

$$Z_i := \sup\{Y(s) : s \in [0, \Delta] \mid J(0) = i\}.$$

As mentioned before, we wish to do this without a priori assuming that $J(\cdot)$ corresponds to an *irreducible* continuous-time Markov chain. A central role is played by the probabilities, for $u \geqslant 0$,

$$p_i(u) := \mathbb{P}(Z_i \geqslant u).$$

Preliminaries

To formally state the Wiener-Hopf decomposition, denote for a given Lévy process $X(\cdot)$ its running maximum process by $(\overline{X}(t))_{t\geqslant 0}$, and its running minimum process by $(\underline{X}(t))_{t\geqslant 0}$. Let T_ν be an exponentially distributed random variable with mean ν^{-1}, sampled independently of anything else. In this chapter we use the Wiener-Hopf decomposition for both spectrally-positive and spectrally-negative Lévy processes, as stated below.

Proposition 5.2.1 (Wiener-Hopf decomposition)**.** *Let $(X(t))_{t\geqslant 0}$ be a Lévy process. Then $X(T_\nu)$ can be decomposed as the sum of the two independent quantities $\overline{X}(T_\nu)$ and $X(T_\nu) - \overline{X}(T_\nu)$. Moreover, the second component $X(T_\nu) - \overline{X}(T_\nu)$ is distributed as $\underline{X}(T_\nu)$.*

If $X(\cdot)$ is spectrally positive with Laplace exponent $\varphi(\cdot)$ and corresponding right inverse $\psi(\cdot)$, then $\underline{X}(T_\nu)$ is distributed as $-T_{\psi(\nu)}$ and

$$\mathbb{E}\left(e^{-\gamma\overline{X}(T_\nu)}\right) = \frac{\nu}{\nu - \varphi(\gamma)}\left(1 - \frac{\gamma}{\psi(\nu)}\right).$$

If $X(\cdot)$ is spectrally negative with cumulant generating function $\Phi(\cdot)$ and corresponding right inverse $\Psi(\cdot)$, then $\overline{X}(T_\nu)$ is distributed as $T_{\Psi(\nu)}$ and

$$\mathbb{E}\left(e^{\gamma\underline{X}(T_\nu)}\right) = \frac{\nu}{\nu - \Phi(\gamma)}\left(1 - \frac{\gamma}{\Psi(\nu)}\right).$$

This decomposition shows that, when $X(\cdot)$ is spectrally one-sided, the (transforms of the) two components can be expressed explicitly in terms of the underlying Laplace exponent (in the spectrally-positive case) or cumulant generating function (in the spectrally-negative case), and their right inverses. For the proof behind this decomposition and more background we refer to e.g. Kyprianou [73, Chapter 6].

As mentioned in the introduction of Part III, the Number of Zeroes proposition concerns a characterization of the zeroes of the determinant of the matrix exponent $M(\alpha)$ of a spectrally-positive MAP. A special role is played by Lévy processes $X_i(\cdot)$ that are monotone a.s. (also referred to as *subordinators*). Let $S^\uparrow$ ($S^\downarrow$) represent the set of background states corresponding to increasing (decreasing, respectively) subordinators. The following result is a slight restatement of [61, Theorem 1 & Remark 2.1].

Proposition 5.2.2 (Number of Zeroes)**.** *Let $\gamma \geqslant 0$, and suppose the background chain $J(\cdot)$ consists of a single (hence recurrent) class.*

If $X_1(\cdot), \ldots, X_d(\cdot)$ are spectrally-positive Lévy processes and the jump sizes at transition epochs L_{ij} are non-negative a.s. for all i, j, then the equation $\det M(\gamma) = 0$ has $d - |S^\uparrow|$ solutions in $\mathbb{C}$ with positive real part.

If $X_1(\cdot), ..., X_d(\cdot)$ are spectrally-negative Lévy processes and the jump sizes at transition epochs L_{ij} are non-positive a.s. for all i, j, then the equation $\det M(-\gamma) = 0$ has $d - |S^{\downarrow}|$ solutions in $\mathbb{C}$ with positive real part.

Next to these two results, we often exploit the following standard relation between two transform types. For $\alpha \geqslant 0$ and a non-negative random variable Y, let $\beta_Y(\alpha) := \int_0^\infty e^{-\alpha x}\mathbb{P}(Y > x)\,\mathrm{d}x$ and $\eta_Y(\alpha) := \int_0^\infty e^{-\alpha x}\mathbb{P}(Y \in \mathrm{d}x)$, the latter representing the Laplace-Stieltjes transform of Y. The two transforms are related through the identity

$$\beta_Y(\alpha) = \frac{1 - \eta_Y(\alpha)}{\alpha}. \tag{5.1}$$

The relation trivially follows as an application of integration by parts.

Approach

Now that we have the essential notation and previous results at our disposal, we proceed by summarizing our approach. In both spectrally one-sided cases the starting point is to use Proposition 5.2.1 to find a relationship between characteristics of the MAP at two successive transition epochs of the background chain. This relationship can be transformed to a system of equations for (transforms related to) $p_1(u), \ldots, p_d(u)$, involving the matrix $M(\cdot)$; here we recall that we defined $p_i(u) := \mathbb{P}(Z_i > u)$. In the spectrally-positive case (Section 5.3), this system contains unknown constants which can be determined exploiting Proposition 5.2.2. In the spectrally-negative case (Section 5.4), the solution is directly expressed in terms of the zeroes of $\det M(\cdot)$, entailing that again Proposition 5.2.2 can be used.

5.3 Spectrally-positive case

Throughout this section we assume that the MAP $Y(\cdot)$ is spectrally positive. As pointed out above this entails that, for each $i, j \in \{1, ..., d\}$, $X_i(\cdot)$ has no downward jumps and the random variable L_{ij} is non-negative a.s.

We will point out how to identify

$$P_i(\gamma) := \int_0^\infty e^{-\gamma u} p_i(u)\,\mathrm{d}u,$$

which is the transform of the tail distribution of Z_i, and the corresponding Laplace-Stieltjes transform

$$\zeta_i(\gamma) := \mathbb{E}\big(e^{-\gamma Z_i}\big).$$

Note that by (5.1), $\zeta_i(\gamma) = 1 - \gamma P_i(\gamma)$, so that either of these transforms uniquely characterizes the distribution of Z_i, the random variable of our interest. Once the Laplace-Stieltjes transform $\zeta_i(\gamma)$ is evaluated, a numerical inversion algorithm, e.g. the one developed in [5], can be used to obtain the distribution of Z_i.

A few observations can be made.

- Recalling that T_γ denotes an exponentially distributed random variable with rate γ, it holds that $\gamma P_i(\gamma) = \mathbb{P}(Z_i > T_\gamma)$. In other words, $\gamma P_i(\gamma)$ can be interpreted as the probability of $Y(\cdot)$ reaching an exponentially distributed level (with mean γ^{-1}) before the process is killed.
- Furthermore, bearing in mind that killing occurs at rate ϑ_i when the background process $J(\cdot)$ is in state i, it is worth noting that when $\vartheta_i = \vartheta$ for all i, (numerical) inversion of

$$\begin{aligned}\frac{P_i(\gamma)}{\vartheta} &= \frac{1}{\vartheta}\int_0^\infty e^{-\gamma u}\mathbb{P}(\overline{Y}(T_\vartheta) > u \mid J(0) = i)\,\mathrm{d}u \\ &= \int_0^\infty\int_0^\infty e^{-\gamma u}e^{-\vartheta t}\mathbb{P}(\overline{Y}(t) > u \mid J(0) = i)\,\mathrm{d}t\,\mathrm{d}u\end{aligned}$$

  with respect to both γ and ϑ yields $\mathbb{P}(\overline{Y}(t) > u \mid J(0) = i)$, i.e., the tail probability of the running maximum of the unkilled MAP at time t.
- Finally, we note that $P_i(0) = \mathbb{E}(Z_i)$, the expected maximum that the MAP attains.

Throughout this section, we analyze the behavior of $P_i(\gamma)$ for a fixed initial state i. Since the running maximum of a non-decreasing process necessarily equals the current value of the process, the analysis turns out to be slightly different depending on whether or not the Lévy process $X_i(\cdot)$ is a non-decreasing subordinator. We consider in Section 5.3.1 the case in which the fixed state i does not correspond to a non-decreasing subordinator, and in Section 5.3.2 the case in which it does. In the analysis of these sections, unknown constants appear; Section 5.3.3 points out how to determine these constants.

5.3.1 Non-subordinator case

In this subsection we focus on the case that $J(0) = i$, where $i \notin S^\uparrow$, so that the spectrally-positive Lévy process $X_i(\cdot)$ may move downwards in any interval with positive probability. Recall that $\vartheta_i + q_i$ is the rate of the exponentially distributed time until the first event; this first event corresponds to killing with probability $\pi_i^\circ := \vartheta_i/(\vartheta_i + q_i)$, and to a transition of the background process to state j with probability $\pi_{ij} := q_{ij}/(\vartheta_i + q_i)$. We decompose $p_i(u)$ by distinguishing between the case that the value of the MAP's running maximum $\overline{X}_i(T_{\vartheta_i+q_i})$ at time $T_{\vartheta_i+q_i}$ is above or below u. In the former case we have that $Z_i > u$, so that we are done; in the latter case with probability π_i° we will not exceed u before killing, whereas with probability π_{ij} we are left with the probability of Z_j exceeding level $u - X_i(T_{\vartheta_i+q_i}) - L_{ij}$ before killing. Formalizing this reasoning, and applying Proposition 5.2.1 to decompose $X_i(T_{\vartheta_i+q_i})$ into the non-negative independent random variables $\overline{X}_i(T_{\vartheta_i+q_i})$ and $X_i(T_{\vartheta_i+q_i}) - \overline{X}_i(T_{\vartheta_i+q_i})$, we obtain

$$\begin{aligned}P_i(\gamma) = &\int_{u=0}^\infty e^{-\gamma u}\int_{w=u}^\infty \mathbb{P}\big(\overline{X}_i(T_{\vartheta_i+q_i}) \in \mathrm{d}w\big)\,\mathrm{d}u + \\ &\int_{u=0}^\infty e^{-\gamma u}\int_{w=0}^u \mathbb{P}\big(\overline{X}_i(T_{\vartheta_i+q_i}) \in \mathrm{d}w\big)\sum_{j\neq i}\frac{q_{ij}}{\vartheta_i + q_i}\mathbb{P}\big(\underline{X}_i(T_{\vartheta_i+q_i}) + L_{ij} + Z_j \geqslant u - w\big)\,\mathrm{d}u,\end{aligned}$$

where we have also used that $\underline{X}_i(T_{\vartheta_i+q_i})$ has the same distribution as $X_i(T_{\vartheta_i+q_i}) - \overline{X}_i(T_{\vartheta_i+q_i})$. We continue by evaluating these two terms, which in the sequel we refer to by $P_i^+(\gamma)$ and

$P_i^-(\gamma)$, separately. Evaluation of the first term is relatively straightforward; an interchange of the order of integration readily yields

$$\begin{aligned} P_i^+(\gamma) &= \int_{w=0}^{\infty} \int_{u=0}^{w} e^{-\gamma u} \mathrm{d}u \, \mathbb{P}\big(\overline{X}_i(T_{\vartheta_i+q_i}) \in \mathrm{d}w\big) \\ &= \int_{w=0}^{\infty} \frac{1-e^{-\gamma w}}{\gamma} \, \mathbb{P}\big(\overline{X}_i(T_{\vartheta_i+q_i}) \in \mathrm{d}w\big) = \frac{1-\kappa_i(\gamma)}{\gamma}, \end{aligned}$$

where

$$\kappa_i(\gamma) := \mathbb{E}\big(e^{-\gamma \overline{X}_i(T_{\vartheta_i+q_i})}\big) = \frac{\vartheta_i + q_i}{\vartheta_i + q_i - \varphi_i(\gamma)} \left(1 - \frac{\gamma}{\psi_i(\vartheta_i + q_i)}\right) \tag{5.2}$$

by virtue of Proposition 5.2.1. We now focus on the evaluation of the second term, which is considerably more involved. As a first step, we interchange the order of the sum and the integrals:

$$P_i^-(\gamma) = \sum_{j \neq i} \frac{q_{ij}}{\vartheta_i + q_i} P_{ij}^-(\gamma),$$

where

$$P_{ij}^-(\gamma) := \int_{u=0}^{\infty} e^{-\gamma u} \int_{w=0}^{u} \mathbb{P}\big(\overline{X}_i(T_{\vartheta_i+q_i}) \in \mathrm{d}w\big) \, \mathbb{P}\big(\underline{X}_i(T_{\vartheta_i+q_i}) + L_{ij} + Z_j \geqslant u - w\big) \, \mathrm{d}u.$$

The quantities $P_{ij}^-(\gamma)$ can be evaluated separately, as follows. Realize that by Proposition 5.2.1, $-\underline{X}_i(T_{\vartheta_i+q_i})$ is exponentially distributed with rate $\mu_i := \psi_i(\vartheta_i + q_i)$. We thus obtain the triple integral

$$P_{ij}^-(\gamma) = \int_{u=0}^{\infty} e^{-\gamma u} \int_{w=0}^{u} \int_{y=0}^{\infty} \mu_i e^{-\mu_i y} \, \mathbb{P}\big(\overline{X}_i(T_{\vartheta_i+q_i}) \in \mathrm{d}w\big) \, \mathbb{P}\big(L_{ij} + Z_j \geqslant u - w + y\big) \, \mathrm{d}y \, \mathrm{d}u,$$

which, after replacing y by $x - u + w$, can be rewritten as

$$\int_{u=0}^{\infty} e^{-\gamma u} \int_{w=0}^{u} \int_{x=u-w}^{\infty} \mu_i e^{-\mu_i (x-u+w)} \, \mathbb{P}\big(\overline{X}_i(T_{\vartheta_i+q_i}) \in \mathrm{d}w\big) \, \mathbb{P}\big(L_{ij} + Z_j \geqslant x\big) \, \mathrm{d}x \, \mathrm{d}u.$$

Our strategy is to interchange the order of the integrals, so as to be able to do the (easy) integration over u first. By first swapping the integrals over u and w, and then those over u and x, we find

$$\begin{aligned} P_{ij}^-(\gamma) &= \int_{w=0}^{\infty} \int_{x=0}^{\infty} \left(\int_{u=w}^{x+w} e^{-(\gamma-\mu_i)u} \mathrm{d}u \right) \mu_i e^{-\mu_i (x+w)} \, \mathbb{P}\big(\overline{X}_i(T_{\vartheta_i+q_i}) \in \mathrm{d}w\big) \, \mathbb{P}\big(L_{ij} + Z_j \geqslant x\big) \, \mathrm{d}x \\ &= \int_{w=0}^{\infty} \int_{x=0}^{\infty} \frac{\mu_i \, e^{-\gamma w}}{\gamma - \mu_i} \left(e^{-\mu_i x} - e^{-\gamma x}\right) \, \mathbb{P}\big(\overline{X}_i(T_{\vartheta_i+q_i}) \in \mathrm{d}w\big) \, \mathbb{P}\big(L_{ij} + Z_j \geqslant x\big) \, \mathrm{d}x, \end{aligned}$$

where the second equality follows by performing the integration over u and reorganizing the resulting expression. We can rewrite this expression as the difference of two terms, in each of which the double integral factorizes into the product of two single integrals. In particular, with

$$\eta_{ij}(\gamma) := \int_0^{\infty} e^{-\gamma x} \, \mathbb{P}\big(L_{ij} + Z_j \geqslant x\big) \, \mathrm{d}x,$$

some rearranging of terms leads to the expression

$$P_{ij}^-(\gamma) = \frac{\mu_i}{\gamma - \mu_i} \int_0^{\infty} e^{-\gamma w} \mathbb{P}\big(\overline{X}_i(T_{\vartheta_i+q_i}) \in \mathrm{d}w\big) \, (\eta_{ij}(\mu_i) - \eta_{ij}(\gamma)) = \frac{\mu_i \, \kappa_i(\gamma)}{\gamma - \mu_i} \, (\eta_{ij}(\mu_i) - \eta_{ij}(\gamma)).$$

To separate L_{ij} and Z_j in the expression for $\eta_{ij}(\gamma)$, we rely on a probabilistic argument for non-negative and independent random variables A and B. That is, using the memoryless property of the exponential distribution,

$$\begin{aligned}\int_0^\infty e^{-\gamma x}\mathbb{P}(A+B \geqslant x)\,\mathrm{d}x &= \frac{1}{\gamma}\mathbb{P}(A+B > T_\gamma)\\ &= \frac{1}{\gamma}\big(\mathbb{P}(A > T_\gamma) + \mathbb{P}(A < T_\gamma)\mathbb{P}(A+B > T_\gamma | A < T_\gamma)\big)\\ &= \frac{1}{\gamma}\big(\mathbb{P}(A > T_\gamma) + \mathbb{P}(A < T_\gamma)\mathbb{P}(B > T_\gamma)\big)\\ &= \frac{1-\mathbb{E}\left(e^{-\gamma A}\right)}{\gamma} + \mathbb{E}\big(e^{-\gamma A}\big)\int_0^\infty e^{-\gamma x}\mathbb{P}(B \geqslant x)\,\mathrm{d}x,\end{aligned} \tag{5.3}$$

where we use (5.1) in the last step. Furthermore, let $\lambda_{ij}(\cdot)$ be the Laplace-Stieltjes transform of L_{ij}, the size of the (non-negative) jump at a transition by the background chain from state i to state j (in other words, $\lambda_{ij}(\gamma) := \mathbb{E}(e^{-\gamma L_{ij}})$). Recalling the definition of $P_i(\gamma)$, we conclude from the identity (5.3) that, with $\Lambda_{ij}(\gamma) := (1-\lambda_{ij}(\gamma))/\gamma$,

$$\eta_{ij}(\gamma) = \Lambda_{ij}(\gamma) + \lambda_{ij}(\gamma)P_j(\gamma).$$

We now combine all the above findings, which enable us to express $P_i(\gamma)$ in terms of $P_j(\gamma)$, with $j \neq i$. Recalling that

$$P_i(\gamma) = P_i^+(\gamma) + \sum_{j\neq i}\frac{q_{ij}}{\vartheta_i + q_i}P_{ij}^-(\gamma)$$

and substituting the obtained expressions for $P_i^+(\gamma)$ and $P_{ij}^-(\gamma)$, we obtain, for any $i \notin S^\uparrow$,

$$P_i(\gamma) = \frac{1-\kappa_i(\gamma)}{\gamma} + \frac{\mu_i\,\kappa_i(\gamma)}{\gamma-\mu_i}\sum_{j\neq i}\frac{q_{ij}}{\vartheta_i+q_i}\Big(\Lambda_{ij}(\mu_i) + \lambda_{ij}(\mu_i)P_j(\mu_i) - \Lambda_{ij}(\gamma) - \lambda_{ij}(\gamma)P_j(\gamma)\Big).$$

By using (5.2), recalling that $\mu_i = \psi_i(\vartheta_i + q_i)$ and defining

$$\bar{\kappa}_i(\gamma) := \frac{1-\kappa_i(\gamma)}{\gamma} = \frac{1}{\vartheta_i + q_i - \varphi_i(\gamma)}\left(\frac{\vartheta_i+q_i}{\psi_i(\vartheta_i+q_i)} - \frac{\varphi_i(\gamma)}{\gamma}\right),$$

we can compactly summarize this result as follows.

Lemma 5.3.1. *For $i \notin S^\uparrow$ and any $\gamma \geqslant 0$, the transform of the tail probability $p_i(u)$ is given by*

$$P_i(\gamma) = \bar{\kappa}_i(\gamma) + \sum_{j\neq i}\frac{q_{ij}}{\vartheta_i+q_i-\varphi_i(\gamma)}\Big(\Lambda_{ij}(\gamma) + \lambda_{ij}(\gamma)P_j(\gamma) - \Lambda_{ij}(\mu_i) - \lambda_{ij}(\mu_i)P_j(\mu_i)\Big). \tag{5.4}$$

So far we have been working with the transform $P_i(\gamma)$ of the tail probability $p_i(u)$. In the remainder of this subsection, we rewrite the above lemma in terms of $\zeta_i(\gamma) := \mathbb{E}(e^{-\gamma Z_i})$, which will take a particularly nice form. To this end, first note that, as a consequence of (5.1),

$$P_i(\gamma) = \frac{1-\zeta_i(\gamma)}{\gamma}. \tag{5.5}$$

Substituting this in (5.4) and rewriting leads to, for $\gamma \geqslant 0$,

$$\frac{1-\zeta_i(\gamma)}{\gamma} = \frac{1-\kappa_i(\gamma)}{\gamma} + \sum_{j\neq i} \frac{q_{ij}}{\vartheta_i + q_i - \varphi_i(\gamma)} \Bigg(\frac{1-\lambda_{ij}(\gamma)}{\gamma} + \lambda_{ij}(\gamma)\frac{1-\zeta_j(\gamma)}{\gamma} - \frac{1-\lambda_{ij}(\mu_i)}{\mu_i} - \lambda_{ij}(\mu_i)\frac{1-\zeta_j(\mu_i)}{\mu_i} \Bigg),$$

or, equivalently,

$$\begin{aligned} \zeta_i(\gamma) &= \kappa_i(\gamma) + \sum_{j\neq i} \frac{q_{ij}}{\vartheta_i + q_i - \varphi_i(\gamma)} \Big(\lambda_{ij}(\gamma)\zeta_j(\gamma) - \lambda_{ij}(\mu_i)\zeta_j(\mu_i)\frac{\gamma}{\mu_i} \Big) \\ &\quad - \frac{q_i}{\vartheta_i + q_i - \varphi_i(\gamma)} \Big(1 - \frac{\gamma}{\mu_i}\Big) \\ &= \sum_{j\neq i} \frac{q_{ij}}{\vartheta_i + q_i - \varphi_i(\gamma)} \Big(\lambda_{ij}(\gamma)\zeta_j(\gamma) - \lambda_{ij}(\mu_i)\zeta_j(\mu_i)\frac{\gamma}{\mu_i} \Big) + \frac{\vartheta_i}{\vartheta_i + q_i - \varphi_i(\gamma)} \Big(1 - \frac{\gamma}{\mu_i}\Big). \end{aligned} \tag{5.6}$$

Multiplying (5.6) by $\vartheta_i + q_i + \varphi_i(\gamma)$ yields the identity

$$\big(\vartheta_i + q_i - \varphi_i(\gamma)\big)\,\zeta_i(\gamma) = \vartheta_i \Big(1 - \frac{\gamma}{\mu_i}\Big) + \sum_{j\neq i} q_{ij} \Big(\lambda_{ij}(\gamma)\zeta_j(\gamma) - \lambda_{ij}(\mu_i)\zeta_j(\mu_i)\frac{\gamma}{\mu_i} \Big). \tag{5.7}$$

We continue by considering the case that none of the states corresponds to a non-decreasing subordinator. Then the system of equations (5.7) that characterizes $\zeta_1(\gamma), \ldots, \zeta_d(\gamma)$ can be written in a considerably more compact form. To this end, recall that the matrix $M(\gamma) \equiv (m_{ij}(\gamma))_{i,j=1}^d$ is defined by

$$m_{ij}(\gamma) = q_{ij}\lambda_{ij}(\gamma) + \varphi_i(\gamma)\,\mathbb{1}_{\{i=j\}} - \vartheta_i\,\mathbb{1}_{\{i=j\}}. \tag{5.8}$$

Furthermore, using that $m_{ii}(\mu_i) = 0$, we define the quantity $b_i(\gamma)$ as follows:

$$\begin{aligned} b_i(\gamma) &:= \sum_{j\neq i} m_{ij}(\mu_i)\zeta_j(\mu_i)\frac{\gamma}{\mu_i} - \vartheta_i \Big(1 - \frac{\gamma}{\mu_i}\Big) \\ &= \Bigg(\sum_{j=1}^d m_{ij}(\mu_i)\zeta_j(\mu_i) + \vartheta_i \Bigg)\frac{\gamma}{\mu_i} - \vartheta_i = \gamma\,\frac{\omega_i + \vartheta_i}{\mu_i} - \vartheta_i, \end{aligned} \tag{5.9}$$

with the constants ω_i defined by

$$\omega_i := \sum_{j=1}^d m_{ij}(\mu_i)\,\zeta_j(\mu_i). \tag{5.10}$$

Upon combining the above, we obtain the equations $\sum_{j=1}^d m_{ij}(\gamma)\,\zeta_j(\gamma) = b_i(\gamma)$, for $i = 1, \ldots, d$. In evident vector/matrix notation, we have thus rewritten (5.7) as follows.

Theorem 5.3.1. *Suppose that no state corresponds to a non-decreasing subordinator (i.e., $i \notin S^{\uparrow}$ for all $i = 1, \ldots, d$). Then, for any $\gamma \geqslant 0$, the vectors $\boldsymbol{\zeta}(\gamma) := (\zeta_1(\gamma), \ldots, \zeta_d(\gamma))$ and $\boldsymbol{b}(\gamma) := (b_1(\gamma), \ldots, b_d(\gamma))$ satisfy*

$$M(\gamma)\,\boldsymbol{\zeta}(\gamma) = \boldsymbol{b}(\gamma). \tag{5.11}$$

It is important to note that, throughout the analysis, no assumptions on the chain structure of the background process $J(\cdot)$ have been imposed. Also observe that we still need to identify the constants ω_i that appear in (5.9), which we will do in Section 5.3.3.

As an aside we mention that the identity (5.6) can alternatively be derived using a probabilistic argumentation, which is for completeness provided in Appendix 5.A.1.

5.3.2 Subordinator case

The previous subsection dealt with the case that the initial state i was such that the spectrally-positive Lévy process $X_i(\cdot)$ does not correspond to a non-decreasing subordinator ($i \notin S^{\uparrow}$, that is). The analysis led to the matrix equation (5.11) for the case that no state corresponds to a non-decreasing subordinator. In the present subsection we address the case where $i \in S^{\uparrow}$, and point out how (5.11) should be adjusted if some of the background states correspond to non-decreasing subordinators.

To this end, let for a given $i = 1, \ldots, d$ the Lévy process $X_i(\cdot)$ be non-decreasing almost surely. It is important to note that in this case necessarily $\varphi_i(\gamma) \leqslant 0$ and $\psi_i(\gamma) = \infty$ for all $\gamma \geqslant 0$. Our method for analyzing $P_i(\gamma)$ is largely the same as in the previous subsection, but is somewhat simpler due to the evident fact that any non-decreasing process attains its maximum at the end of the interval under consideration. Concretely, we could mimic the approach of the previous subsection while replacing $\overline{X}(T_{\vartheta_i+q_i})$ by $X(T_{\vartheta_i+q_i})$, but it turns out to be convenient to condition on the value of $Y(\cdot)$ at the minimum of the killing time and the first transition of the background process. This yields

$$\begin{aligned} P_i(\gamma) = \int_0^\infty e^{-\gamma u} p_i(u)\,\mathrm{d}u \;=\; & \int_0^\infty e^{-\gamma u} \mathbb{P}\big(X_i(T_{\vartheta_i}) \geqslant u, T_{\vartheta_i} \leqslant T_{q_i}\big)\,\mathrm{d}u + \\ & \int_0^\infty e^{-\gamma u} \sum_{j\neq i} \frac{q_{ij}}{q_i} \mathbb{P}\big(X_i(T_{q_i}) + L_{ij} + Z_j \geqslant u, T_{\vartheta_i} > T_{q_i}\big)\,\mathrm{d}u. \end{aligned} \tag{5.12}$$

With $P_i^+(\gamma)$ and $P_i^-(\gamma)$, respectively, representing the two terms in the right-hand side of (5.12), we use Proposition 5.2.1 and (5.1) to obtain

$$\begin{aligned} P_i^+(\gamma) &= \int_0^\infty e^{-\gamma u} \frac{\vartheta_i}{\vartheta_i + q_i} \mathbb{P}\big(X_i(T_{\vartheta_i+q_i}) \geqslant u\big)\,\mathrm{d}u \\ &= \frac{\vartheta_i}{\vartheta_i + q_i} \frac{1}{\gamma}\Big(1 - \mathbb{E}\Big(e^{-\gamma X_i(T_{\vartheta_i+q_i})}\Big)\Big) = \frac{\vartheta_i}{\vartheta_i + q_i} \frac{1}{\gamma}\left(1 - \frac{\vartheta_i + q_i}{\vartheta_i + q_i - \varphi_i(\gamma)}\right) \end{aligned}$$

and

$$\begin{aligned} P_i^-(\gamma) &= \int_0^\infty e^{-\gamma u} \sum_{j\neq i} \frac{q_{ij}}{\vartheta_i + q_i} \mathbb{P}\big(X_i(T_{\vartheta_i+q_i}) + L_{ij} + Z_j \geqslant u\big)\,\mathrm{d}u \\ &= \sum_{j\neq i} \frac{q_{ij}}{\vartheta_i + q_i} \frac{1}{\gamma}\Big(1 - \mathbb{E}\Big(e^{-\gamma X_i(T_{\vartheta_i+q_i})}\Big)\lambda_{ij}(\gamma)\,\zeta_j(\gamma)\Big) \\ &= \sum_{j\neq i} \frac{q_{ij}}{\vartheta_i + q_i} \frac{1}{\gamma}\Big(1 - \frac{\vartheta_i + q_i}{\vartheta_i + q_i - \varphi_i(\gamma)} \lambda_{ij}(\gamma)\,\big(1 - \gamma P_j(\gamma)\big)\Big) \\ &= \frac{q_i}{\vartheta_i + q_i} \frac{1}{\gamma} - \frac{1}{\gamma} \sum_{j\neq i} \frac{q_{ij}}{\vartheta_i + q_i - \varphi_i(\gamma)} \big(1 - \gamma\Lambda_{ij}(\gamma) - \gamma\lambda_{ij}(\gamma) P_j(\gamma)\big), \end{aligned}$$

recalling the identities $\zeta_j(\gamma) = 1 - \gamma P_j(\gamma)$ and $\lambda_{ij}(\gamma) = 1 - \gamma\Lambda_{ij}(\gamma)$ in the last two steps.

After collecting the above intermediate results, we obtain the following characterization.

Lemma 5.3.2. *For $i \in S^{\uparrow}$ and any $\gamma \geqslant 0$,*

$$P_i(\gamma) = \frac{-\varphi_i(\gamma)}{\vartheta_i + q_i - \varphi_i(\gamma)}\frac{1}{\gamma} + \sum_{j \neq i} \frac{q_{ij}}{\vartheta_i + q_i - \varphi_i(\gamma)}\Big(\Lambda_{ij}(\gamma) + \lambda_{ij}(\gamma)P_j(\gamma)\Big). \tag{5.13}$$

Observe that (5.13) could also be obtained by taking the limit $\mu_i = \psi_i(\vartheta_i + q_i) \to \infty$ in Lemma 5.3.1, which is consistent with the fact that $\psi_i(\cdot) = \infty$ for subordinator processes $X_i(\cdot)$. Similar to the non-subordinator case, we can again present a vector/matrix version for the Laplace-Stieltjes transforms $\zeta_i(\gamma)$ of (the distributions of) the random variables Z_i. To this end, define $b_i^{\circ}(\gamma) := -\vartheta_i$ for $i \in S^{\uparrow}$ and $b_i^{\circ}(\gamma) := b_i(\gamma)$ for $i \notin S^{\uparrow}$, with $b_i(\gamma)$ as defined in (5.9). Then, using similar steps as before, we eventually find the following counterpart of Theorem 5.3.1.

Theorem 5.3.2. *The vectors $\boldsymbol{\zeta}(\gamma) = (\zeta_1(\gamma), \ldots, \zeta_d(\gamma))$ and $\boldsymbol{b}^{\circ}(\gamma) := (b_1^{\circ}(\gamma), \ldots, b_d^{\circ}(\gamma))$ satisfy*

$$M(\gamma)\,\boldsymbol{\zeta}(\gamma) = \boldsymbol{b}^{\circ}(\gamma) \tag{5.14}$$

for any $\gamma \geqslant 0$.

Note that also in this set of equations, the vector $\boldsymbol{b}^{\circ}(\gamma)$ still contains unknowns. These constants ω_i, one for each $i \notin S^{\uparrow}$, will be identified in the next subsection.

5.3.3 Evaluation of the unknowns

So far, we have established that in the spectrally-positive case, the Laplace-Stieltjes transforms of $Z_1, ..., Z_d$ are given by the solutions of (5.14) (which amounts to (5.11) in case none of the Lévy processes $X_i(\cdot)$ is a non-decreasing subordinator process). This subsection settles the complication that (5.11) contains unknown constants ω_i. As we have seen, the number of such constants equals the number of states that do not correspond to non-decreasing subordinators, which we denote by d° (i.e., $d^{\circ} := d - |S^{\uparrow}|$). To identify these d° unknowns, and ultimately the solution $\boldsymbol{\zeta}(\gamma)$ of (5.14), we subsequently analyze three cases:

- the background chain has no transient classes,
- the background chain has exactly one transient class, and
- the background chain has more than one transient class.

We proceed by studying each of these cases separately.

∘ *No transient classes.* In case the background chain has no transient classes, all classes of the chain are necessarily recurrent. To analyze Z_i it evidently suffices to restrict ourselves to the recurrent class that the background state i is in. As a consequence, without loss of generality we may assume that the background process $J(\cdot)$ is irreducible. In this case, which has been

studied extensively (see e.g. the results in [40, 43]), the following procedure can be used to identify the ω_i. Note that, using the linear equations given in (5.14), one may express the vector $\boldsymbol{\zeta}(\gamma)$ by relying on Cramer's rule. More concretely, with the matrix $M_{\boldsymbol{b},i}(\gamma)$ denoting the matrix $M(\gamma)$ in which the i-th column is replaced by the vector $\boldsymbol{b}^\circ(\gamma)$, we have that

$$\zeta_i(\gamma) = \frac{\det M_{\boldsymbol{b},i}(\gamma)}{\det M(\gamma)}.$$

Since $\zeta_i(\gamma)$ is finite, any zero of the denominator should be a zero of the numerator. According to Proposition 5.2.2, in case $J(\cdot)$ consists of a single class, $\det M(\gamma) = 0$ has d° zeroes in the right half of the complex plane. For ease of exposition, we make the assumption that these zeroes have multiplicity 1 (and we call them, say, $\gamma_1, \ldots, \gamma_{d^\circ}$). In the special case this assumption does not hold, a reasoning similar to the one below still applies, but one needs to resort to the concept of Jordan chains. We do not discuss this procedure in detail, but instead refer to the in-depth treatment in [43].

Having distinct zeroes guarantees that we have d° equations to identify the ω_i. That is, for $i = 1, \ldots, d$ and $j = 1, \ldots, d^\circ$,

$$\det M_{\boldsymbol{b},i}(\gamma_j) = 0; \tag{5.15}$$

in other words, the zeroes of $\det M$ (in the right half of the complex plane, that is) are also zeroes of $\det M_{\boldsymbol{b},i}$, for each $i = 1, \ldots, d$. For any given $j = 1, \ldots, d^\circ$, this seemingly yields d equations, but it can be seen easily that each of these d equations effectively provides the same information. Indeed, with $\boldsymbol{m}_k(\gamma)$ denoting the k-th column of $M(\gamma)$, suppose for any fixed i, that $\det M(\gamma) = 0$ and $\det M_{\boldsymbol{b},i}(\gamma) = 0$ for some $\gamma \geqslant 0$. This implies that both $M(\gamma)$ and $M_{\boldsymbol{b},i}(\gamma)$ are singular, and as a consequence there are non-trivial vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ such that

$$\sum_{j=1}^{d} \boldsymbol{m}_j(\gamma)\, v_j = \boldsymbol{0}, \quad \sum_{j \neq i} \boldsymbol{m}_j(\gamma)\, u_j + \boldsymbol{b}^\circ(\gamma)\, u_i = \boldsymbol{0}.$$

As a consequence, for any $i' \neq i$,

$$\begin{aligned} \boldsymbol{0} &= -u_{i'} \sum_{j=1}^{d} \boldsymbol{m}_j(\gamma)\, v_j + v_{i'} \sum_{j \neq i} \boldsymbol{m}_j(\gamma)\, u_j + v_{i'} \boldsymbol{b}^\circ(\gamma)\, u_i \\ &= -u_{i'} v_i \boldsymbol{m}_i(\gamma) + \sum_{j \neq i, i'} (v_{i'} u_j - u_{i'} v_j)\, \boldsymbol{m}_j(\gamma) + v_{i'} u_i \boldsymbol{b}^\circ(\gamma), \end{aligned}$$

entailing that there is a linear combination of the columns of $M_{\boldsymbol{b},i'}(\gamma)$ that equals $\boldsymbol{0}$. In other words, $M_{\boldsymbol{b},i'}(\gamma)$ is singular, and hence $\det M_{\boldsymbol{b},i'}(\gamma) = 0$ as well.

Now that we know that (5.15), for any given index $j = 1, \ldots, d^\circ$, provides us with just a single equation, we study this equation in more detail. Let us focus on $\det M_{\boldsymbol{b},1}(\gamma_j) = 0$, for $j = 1, \ldots, d^\circ$ (we take $i = 1$, that is). With $\bar{M}_{ij}(\gamma)$ representing the $(d-1) \times (d-1)$ matrix which results after deleting the i-th column and the j-th row from $M(\gamma)$, and recalling that $b_i^\circ(\gamma) = \gamma(\omega_i + \vartheta_i)/\mu_i - \vartheta_i$ for $i \notin S^\uparrow$ and $b_i^\circ(\gamma) = -\vartheta_i$ for $i \in S^\uparrow$, this equation can be rewritten as

$$\sum_{i \notin S^\uparrow} \left(\gamma_j \frac{\omega_i + \vartheta_i}{\mu_i} - \vartheta_i \right) (-1)^{1+i} \det \bar{M}_{i1}(\gamma_j) + \sum_{i \in S^\uparrow} \vartheta_i\, (-1)^i \det \bar{M}_{i1}(\gamma_j) = 0.$$

We thus obtain d° equations (one for each γ_j) that are linear in the unknowns $\omega_1, \ldots, \omega_{d^\circ}$, which can be dealt with in the standard manner, thus yielding a solution for the ω_i.

∘ *A single transient class.* We now consider the case in which the background chain has a single transient class, say $T \subset \{1, \ldots, d\}$, next to one or more recurrent classes. In this case, note that the $\zeta_i(\gamma)$ for all recurrent states i, i.e., $i \notin T$, can be computed by the procedure pointed out above. Subsequently, for $i \in T$ we rewrite the i-th equation of (5.14) as

$$\sum_{j \in T} m_{ij}(\gamma)\zeta_j(\gamma) = b_i^\circ(\gamma) - \sum_{j \notin T} m_{ij}(\gamma)\zeta_j(\gamma). \tag{5.16}$$

Observe that the right-hand side is known; we will denote it by $\bar{b}_i^\circ(\gamma)$. Define $\bar{d} := |T|$ as the number of transient states and $\bar{d}^\circ := |T \setminus S^\uparrow|$ as the number of transient states that do not correspond to non-decreasing subordinators. In addition, we define the $\bar{d} \times \bar{d}$ matrix $\bar{M}(\gamma) := (m_{ij}(\gamma))_{i,j \in T}$, and we let the $\bar{d}$-dimensional vector $\bar{\boldsymbol{\zeta}}(\gamma) := (\zeta_i(\gamma))_{i \in T}$ represent the entries of $\boldsymbol{\zeta}(\gamma)$ that correspond to the states in T. Likewise, $\bar{\boldsymbol{b}}^\circ(\gamma) = (\bar{b}_i^\circ(\gamma))_{i \in T}$ represents the vector of right-hand sides of (5.16). Using these definitions, (5.16) can be written as

$$\bar{M}(\gamma)\,\bar{\boldsymbol{\zeta}}(\gamma) = \bar{\boldsymbol{b}}^\circ(\gamma).$$

Clearly, suppose that we could prove that $\det \bar{M}(\gamma) = 0$ has $\bar{d}^\circ$ zeroes in the right half of the complex plane, then we could identify the constants ω_i by following the same approach as the one we developed for the case of no transient classes. This is why we now verify that the entries of $\bar{M}(\gamma)$ can be written in the form (5.8), with transition rates that correspond to a single recurrent class, so that we can apply Proposition 5.2.2 to establish the desired property for the number of zeroes of $\det \bar{M}(\gamma) = 0$ in the right half of the complex plane. By rewriting the diagonal elements of $\bar{M}(\gamma)$ as

$$m_{ii}(\gamma) = q_{ii} + \varphi_i(\gamma) - \vartheta_i = \bar{q}_{ii} + \varphi_i(\gamma) - \bar{\vartheta}_i, \tag{5.17}$$

with

$$\bar{q}_{ii} := -\sum_{j \in T \setminus \{i\}} q_{ij} \quad \text{and} \quad \bar{\vartheta}_i := \left(\vartheta_i + \sum_{j \notin T} q_{ij}\right), \tag{5.18}$$

we conclude that the row sums of transition rates $\bar{q}_{ii} + \sum_{j \in T \setminus \{i\}} q_{ij}$ equal zero for all $i \in T$. This means that $\bar{M}(\gamma)$ indeed has the desired form: the entries are of the form (5.8), with transition rates that correspond to a single recurrent class. Applying Proposition 5.2.2 we have that $\det \bar{M}(\gamma) = 0$ has $\bar{d}^\circ$ zeroes in the right half of the complex plane, so that we can identify the ω_i for $i \in T \setminus S^\uparrow$ (repeating the remark on roots with the multiplicity larger than 1, as made above in relation to the case with recurrent states only).

∘ *Multiple transient classes.* We now consider the case where there are $K > 1$ transient classes (say $T_1, \ldots, T_K$). We let R be the union of all remaining recurrent classes. Furthermore, we write $T_k \rightsquigarrow T_{k'}$ if there is a direct transition from a state in T_k to a state in $T_{k'}$, i.e., there is a state $i \in T_k$ and a state $j \in T_{k'}$ such that $q_{ij} > 0$. To handle the case of multiple transient

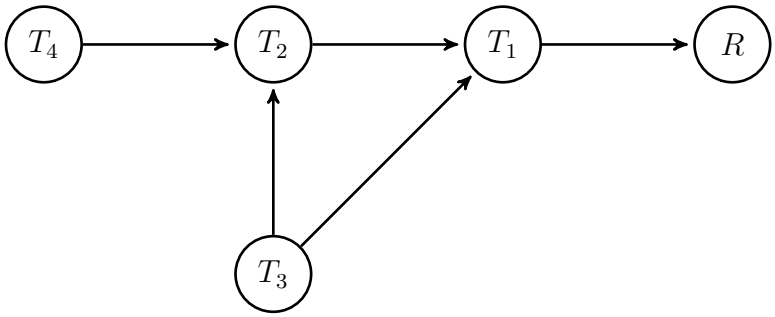

Figure 5.1: Example of layer sets, with $K = 4$ transient classes. In this case, $C_0 = R$, $C_1 = R \cup T_1$, $C_2 = R \cup T_1 \cup T_2$, and $C_3 = R \cup T_1 \cup T_2 \cup T_3 \cup T_4$. In this figure, an arrow between the classes U and V means that $U \rightsquigarrow V$.

classes, we order the transient classes in 'layers', as follows. Let $C_0 := R$, and for $n = 1, 2, \ldots$ let the n-th layer set be given by

$$C_n := C_{n-1} \cup \left\{ T_k : \text{ for all } k' \text{ such that } T_k \rightsquigarrow T_{k'} \text{ it holds that } k' \in C_{n-1} \right\}.$$

It is worth noting that if a background state i is element of the layer set C_j but not of C_{j-1}, then the background chain can reach a recurrent state in minimally j transitions. In addition, we can observe that the number of non-empty layer sets (including C_0) is at most $K + 1$. See Figure 5.1 for a pictorial illustration.

In the previous two cases, we already explained how to compute $\zeta_i(\gamma)$ for $i \in R$ and $i \in C_1$, respectively. We now point out how we can evaluate $\zeta_i(\gamma)$ for $i \in C_n$, having $\zeta_i(\gamma)$ for $i \in R, C_1, \ldots, C_{n-1}$ at our disposal, so that we can recursively determine all $\zeta_i(\gamma)$. Suppose that $T_k \subseteq C_n \setminus C_{n-1}$ (where it is noted that there are potentially multiple transient classes in $C_n \setminus C_{n-1}$). As states in T_k cannot have direct transitions to classes outside C_{n-1}, we have for $i \in T_k$ that

$$\sum_{j \in T_k} m_{ij}(\gamma)\zeta_j(\gamma) = b_i^\circ(\gamma) - \sum_{j \in C_{n-1}} m_{ij}(\gamma)\zeta_j(\gamma).$$

From this point, the analysis follows that of the case with a single transient class. More specifically, the number of zeroes of the determinant of the matrix $(m_{ij}(\gamma))_{i,j \in T_k}$ in the right half of the complex plane equals the number of states in T_k that do not correspond to non-decreasing subordinators, using the same argument as in the case of a single transient class. This allows us to identify the ω_i for $i \in T_k \setminus S^\uparrow$.

5.4 Spectrally-negative case

The model we analyze in this section can be seen as the spectrally-negative counterpart of the one considered in the previous section. This concretely means that now the Lévy processes $X_i(\cdot)$ are assumed to be spectrally negative, and that the jumps L_{ij} are non-positive. In addition, we replace our earlier definition of the entries of the matrix $M(\nu) = (m_{ij}(\nu))_{i,j=1}^d$ by

$$m_{ij}(\nu) := q_{ij}\lambda_{ij}(-\nu) + \Phi_i(\nu)\,\mathbb{1}_{\{i=j\}} - \vartheta_i\,\mathbb{1}_{\{i=j\}}, \tag{5.19}$$

for $\nu \geqslant 0$, to account for the non-positive jumps. As in the spectrally-positive case, the matrix $M(\nu)$ will be helpful in establishing the main result of this section.

Unlike in the previous section, throughout this section we focus directly on $p_i(u) = \mathbb{P}(Z_i \geqslant u)$ rather than on its Laplace transform $P_i(\gamma)$; as it turns out, Laplace transforms are not required in the analysis of the spectrally-negative case. A convenient feature, made more precise later, is that in the spectrally-negative setting the *form* of the distribution of the Z_i is known.

Somewhat comparably to the setup of Section 5.3, to make the presentation as transparent as possible we first treat the case where none of the Lévy processes $X_i(\cdot)$ is a non-increasing subordinator (Section 5.4.1), after which we point out how to adapt the analysis to the case where some of the $X_i(\cdot)$ are (Section 5.4.2).

The following claim plays a crucial role in this section.

Lemma 5.4.1. *The equation* $\det M(\nu) = 0$ *has* d° *zeroes with a positive real part, say* $\nu_1, \ldots, \nu_d$.

By Proposition 5.2.2 we already know that Lemma 5.4.1 holds if the background process is irreducible. In Section 5.4.3 we provide a proof for the case that the background process has a general chain structure.

5.4.1 Non-subordinator case

In this subsection we consider the situation that none of the states corresponds to a non-increasing subordinator. This means that, for all $i = 1, \ldots, d$, Proposition 5.2.1 implies that the running maximum $\overline{X}_i(T_{\vartheta_i+q_i})$ has an exponential distribution with rate $\mu_i := \Psi_i(\vartheta_i + q_i)$. Recalling that the time until either the process is killed or the background chain changes its state is distributed exponentially with rate $\vartheta_i + q_i$, we thus obtain the identity

$$p_i(u) = e^{-\mu_i u} + \sum_{j \neq i} \frac{q_{ij}}{\vartheta_i + q_i} p_{ij}^-(u), \tag{5.20}$$

with

$$p_{ij}^-(u) := \int_0^u \mu_i e^{-\mu_i w} \mathbb{P}\big(\underline{X}_i(T_{\vartheta_i+q_i}) + L_{ij} + Z_j \geqslant u - w\big)\, \mathrm{d}w. \tag{5.21}$$

To streamline our analysis, we impose (Property A) below. By Lemma 5.4.1 we know that in this setting without non-increasing subordinators the equation $\det M(\nu) = 0$ has d zeroes with a positive real part.

The d zeroes of $\det M(\nu)$ with a positive real part are distinct. (Property A)

Importantly, however, imposing (Property A) effectively does not impose any restriction: as discussed in Remark 5.4.2, the argumentation can be adapted to cover zeroes with higher multiplicities.

A crucial fact is that the first-passage process pertaining to $Y(\cdot)$ is a MAP itself, irrespective of whether or not the background process is irreducible; cf. the discussion in [59, Section 2.6].

This implies that the random variables Z_i have phase-type distributions; see e.g. [7, Section III.4] for background on this class of distributions. More specifically, one obtains their Laplace transforms by plugging in $\alpha = 0$ in the expression of the first statement of [59, Corollary 4.21]. The result concretely entails that, in our setting without non-increasing subordinators, for a $d \times d$ transition rate matrix Λ, a vector $\boldsymbol{\lambda} := -\Lambda \mathbf{1} \geqslant \mathbf{0}$ with at least one positive entry, and initial distributions $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d$,

$$\mathbb{E}\big(e^{-\gamma Z_i}\big) = \boldsymbol{a}_i(\gamma I - \Lambda)^{-1}\boldsymbol{\lambda}.$$

Recalling the definition of $M(\nu)$ in (5.19), the zeroes of $\det M(\nu)$ coincide with those of $\det(-\nu I - \Lambda)$; cf. [59, Theorem 4.7] and again the first statement of [59, Corollary 4.21]. Because of this, the matrix $(\gamma I - \Lambda)$ is singular in $\gamma = -\nu_1, ..., -\nu_d$, hence $\mathbb{E}\big(e^{-\gamma Z_i}\big)$ can be written as a linear combination of the terms $1/(\gamma + \nu_1), ..., 1/(\gamma + \nu_d)$. This means that under (Property A) we can write, for $k = 1, \ldots, d$ and $u \geqslant 0$,

$$p_i(u) = \sum_{k=1}^{d} c_{ik}\, e^{-\nu_k u}, \tag{5.22}$$

where $C = (c_{ik})_{i,k=1}^{d}$ is a matrix of unknown coefficients whose rows add up to 1.

Remark 5.4.1. As mentioned above, the first statement of [59, Corollary 4.21] already provides a characterization of the distribution of the random variables Z_i (for $i = 1, \ldots, d$) under a possibly non-irreducible background chain. It is noted, though, that in [59, Corollary 4.21], the distribution of the Z_i is given in terms of a Laplace-Stieltjes transform, which contains unknown matrices (viz. in the terminology of [59], the matrices $\Lambda(q)$ and $\Pi(q)$) which can be numerically computed, e.g. using [59, Theorem 4.14]. Our contribution is that we obtain a more explicit result in Theorem 5.4.1 below: our result concerns the probabilities $p_i(u)$, corresponding to the tail of Z_i, rather than their transforms. For each i, we succeed in expressing $p_i(u)$ in terms of the solutions of an eigensystem. ◇

We now exploit the structure as given in (5.22) to generate equations by which we can determine the coefficients c_{ik}. To this end, we define $a_{ik} := c_{ik}\nu_k$. By conditioning on the value of Z_j in (5.21) using (5.22), we thus obtain

$$p_{ij}^{-}(u) = \int_0^u \mu_i e^{-\mu_i w} \int_0^\infty \sum_{k=1}^{d} a_{jk}\, e^{-\nu_k v}\, \mathbb{P}\big(\underline{X}_i(T_{\vartheta_i+q_i}) + L_{ij} \geqslant u - w - v\big)\, \mathrm{d}v\, \mathrm{d}w. \tag{5.23}$$

We then substitute v by $u - w - x$ and recall that $\underline{X}_i(T_{\vartheta_i+q_i})$ and L_{ij} are non-positive random variables, leading to

$$\begin{aligned} p_{ij}^{-}(u) &= \int_0^u \mu_i e^{-\mu_i w} \int_{-\infty}^{u-w} \sum_{k=1}^{d} a_{jk} e^{-\nu_k(u-w-x)}\, \mathbb{P}\big(\underline{X}_i(T_{\vartheta_i+q_i}) + L_{ij} \geqslant x\big)\, \mathrm{d}x\, \mathrm{d}w \\ &= \int_0^u \mu_i e^{-\mu_i w} \int_{-\infty}^{0} \sum_{k=1}^{d} a_{jk} e^{-\nu_k(u-w-x)}\, \mathbb{P}\big(\underline{X}_i(T_{\vartheta_i+q_i}) + L_{ij} \geqslant x\big)\, \mathrm{d}x\, \mathrm{d}w. \end{aligned}$$

Pulling the sum in front of the integrals leads to a sum in which the two integrals factorize. That is, we obtain

$$p_{ij}^{-}(u) = \sum_{k=1}^{d} a_{jk} \int_{-\infty}^{0} e^{\nu_k x} \Big(1 - \mathbb{P}\big(\underline{X}_i(T_{\vartheta_i+q_i}) + L_{ij} < x\big)\Big)\, \mathrm{d}x \cdot \int_0^u \mu_i e^{-\mu_i w} e^{-\nu_k(u-w)}\, \mathrm{d}w. \tag{5.24}$$

Now, we can rewrite the first integral in this expression using (5.1) and Proposition 5.2.1, yielding

$$\begin{aligned}\int_{-\infty}^{0} e^{\nu_k x}\Big(1-\mathbb{P}\big(\underline{X}_i(T_{\vartheta_i+q_i})+L_{ij}<x\big)\Big)\,\mathrm{d}x &= \frac{1}{\nu_k}\,\mathbb{E}\Big(e^{\nu_k \underline{X}_i(T_{\vartheta_i+q_i})}\Big)\,\lambda_{ij}(-\nu_k)\\ &= \frac{1}{\nu_k}\left(\frac{\Psi_i(\vartheta_i+q_i)-\nu_k}{\vartheta_i+q_i-\Phi_i(\nu_k)}\frac{\vartheta_i+q_i}{\Psi_i(\vartheta_i+q_i)}\right)\lambda_{ij}(-\nu_k)\\ &= \frac{1}{\nu_k}\left(\frac{\vartheta_i+q_i}{\vartheta_i+q_i-\Phi_i(\nu_k)}\frac{\mu_i-\nu_k}{\mu_i}\right)\lambda_{ij}(-\nu_k).\end{aligned}$$

Furthermore, we note for the second integral of (5.24) that

$$\int_0^u \mu_i e^{-\mu_i w} e^{-\nu_k(u-w)}\,\mathrm{d}w = \frac{\mu_i}{\mu_i-\nu_k}\big(e^{-\nu_k u}-e^{-\mu_i u}\big).$$

Combining the above, we conclude that

$$p_{ij}^{-}(u) = \sum_{k=1}^{d} c_{jk}\frac{\vartheta_i+q_i}{\vartheta_i+q_i-\Phi_i(\nu_k)}\big(e^{-\nu_k u}-e^{-\mu_i u}\big)\lambda_{ij}(-\nu_k). \tag{5.25}$$

It can be seen that the μ_i differ from the ν_k, because if they would be equal for some pair (i,k), then $p_i(u)$ would have a term that is constant in u, thus violating its form given in (5.22). In Appendix 5.A.2 an alternative, probabilistic proof of (5.25) is given.

We now focus on finding the values of the coefficients c_{ik} for $i,k=1,\dots,d$. Observe that we have two alternative ways of writing $p_i(u)$: the representation (5.22), and a representation based on (5.20) and (5.25). Note that both are linear combinations of $e^{-\mu_i u}$ and $e^{-\nu_1 u},\dots,e^{-\nu_d u}$. The weights corresponding to each of these $d+1$ exponentials should match, thus providing equations that impose constraints on the c_{ik}.

- Focusing on the terms corresponding to $e^{-\nu_k u}$, for $k=1,\dots,d$, we thus obtain the equations
$$c_{ik} = \sum_{j\neq i}\left(q_{ij}\frac{1}{\vartheta_i+q_i-\Phi_i(\nu_k)}\lambda_{ij}(-\nu_k)\right)c_{jk}, \tag{5.26}$$
where, as observed earlier, $\sum_{k=1}^{d} c_{ik}=1$.
- Regarding the terms corresponding to $e^{-\mu_i u}$, recalling that μ_i differs from all the ν_k, we should have
$$1-\sum_{k=1}^{d}\sum_{j\neq i}\left(q_{ij}\frac{1}{\vartheta_i+q_i-\Phi_i(\nu_k)}\lambda_{ij}(-\nu_k)\right)c_{jk}=0.$$
This equation holds true if (5.26) applies, which can be seen by recognizing the left-hand side as $1-\sum_{k=1}^{d} c_{ik}$ (as the obvious consequence of $\sum_{k=1}^{d} c_{ik}=1$). In other words, this equation does not provide any additional information.

We now observe that (5.26) is equivalent to, for $i,k=1,\dots,d$,

$$\sum_{j=1}^{d} m_{ij}(\nu_k)\,c_{jk}=0. \tag{5.27}$$

We reassuringly notice from (5.27) that the matrix $M(\nu_k)$ is singular for all $k = 1, ..., d$, and hence that the ν_k are indeed the solutions to $\det M(\nu) = 0$.

As was done in the spectrally-positive case, our result can be rewritten in a more compact vector/matrix form. In particular, to find the c_{jk}, it is enough to solve, for $k = 1, \dots, d$, the matrix-vector equation

$$M(\nu_k)\, \boldsymbol{c}_k = \mathbf{0},$$

where $\boldsymbol{c}_k := (c_{1k}, \dots, c_{dk})^\top$, subject to $C\,\mathbf{1} = \mathbf{1}$. The following theorem summarizes the findings above.

Theorem 5.4.1. *Under* (Property A), $p_i(u)$ *satisfies*

$$p_i(u) = \sum_{k=1}^{d} c_{ik}\, e^{-\nu_k u},$$

for $i = 1, \dots, d$. *Here, for* $k = 1, \dots, d$, *the vectors* $\boldsymbol{c}_k$ *solve* $M(\nu_k)\, \boldsymbol{c}_k = \mathbf{0}$ *subject to* $C\,\mathbf{1} = \mathbf{1}$.

Remark 5.4.2. We briefly comment on the case where some solutions of $\det M(\nu) = 0$ have multiplicity larger than one. For instance in the case of a root with multiplicity 2, suppose that, for some $k_1 \neq k_2$, $\nu_{k_1} = \nu_{k_2} = \nu$, giving rise to terms in (5.22) proportional to $e^{-\nu u}$ and $u\, e^{-\nu u}$. Finding the associated weights works effectively as pointed out above: use the identity (5.20) to find two alternative expressions for $p_i(u)$, and then equate the terms proportional to $u\, e^{-\nu u}$, so as to obtain linear equations for these coefficients (in addition to equating all terms proportional to $e^{-\nu_k u}$). For an in-depth treatment of these multiplicity issues, we again refer to [43]. ◇

5.4.2 Subordinator case

We now consider the case where some of the states of the background process correspond to a non-increasing subordinator. Let i be in the set of states corresponding to non-increasing subordinators, denoted by $S^{\downarrow}$. For $i \in S^{\downarrow}$, $Z_i = 0$ with positive probability.

The structure of this section is similar to that of the non-subordinator case, the main difference being that now the MAP cannot cross positive levels (in the upward direction, that is) while the background process is in $i \in S^{\downarrow}$. Therefore, the following decomposition applies, for $u > 0$:

$$p_i(u) = \sum_{j \neq i} \frac{q_{ij}}{\vartheta_i + q_i} p_{ij}^{-}(u), \tag{5.28}$$

with

$$p_{ij}^{-}(u) := \mathbb{P}\big(X_i(T_{\vartheta_i + q_i}) + L_{ij} + Z_j \geqslant u\big).$$

Regarding the zeroes of $\det M(\nu)$, we make a similar claim and assumption as in Section 5.4.1. Let $d^\circ := |S \setminus S^{\downarrow}|$ be the number of states that do not correspond to a non-increasing subordinator. Then by Lemma 5.4.1 we know that $\det M(\nu)$ has d° zeroes with a positive real part, say $(\nu_k)_{k \notin S^{\downarrow}}$. In our analysis we impose (Property A') below.

The d° zeroes of $\det M(\nu)$ with a positive real part are distinct. (Property A')

The case of zeroes with higher multiplicities can be dealt with as discussed in Remark 5.4.2, and the case that $J(\cdot)$ is not irreducible is covered by Section 5.4.3.

Relying on the same reasoning as in Section 5.4.1, under (Property A'), we again have due to the first-passage process being a MAP and the first statement of [59, Corollary 4.21], that $p_i(u)$ is a linear combination of exponential terms, where the number of such terms now equals d°. Concretely, for $u > 0$ and $i = 1, \ldots, d$,

$$p_i(u) = \sum_{k \notin S^\downarrow} c_{ik} e^{-\nu_k u}. \tag{5.29}$$

To identify the coefficients c_{ik}, it proves worthwhile to further study $p_{ij}^-(u)$. In particular, for any $j = 1, \ldots, d$, conditioning on the value of Z_j yields

$$p_{ij}^-(u) = \sum_{k \notin S^\downarrow} \int_u^\infty c_{jk} \nu_k e^{-\nu_k v} \, \mathbb{P}\big(X_i(T_{\vartheta_i + q_i}) + L_{ij} \geqslant u - v\big) \, \mathrm{d}v.$$

Then we subsequently use the relation (5.1) and Proposition 5.2.1 to obtain

$$p_{ij}^-(u) = \sum_{k \notin S^\downarrow} c_{jk} e^{-\nu_k u} \, \mathbb{E}\left(e^{\nu_k X_i(T_{\vartheta_i + q_i})}\right) \lambda_{ij}(-\nu_k) = \sum_{k \notin S^\downarrow} c_{jk} \frac{\vartheta_i + q_i}{\vartheta_i + q_i - \Phi_i(\nu_k)} e^{-\nu_k u} \lambda_{ij}(-\nu_k),$$

such that in combination with (5.28), we have

$$p_i(u) = \sum_{j \neq i} \sum_{k \notin S^\downarrow} c_{jk} \frac{q_{ij}}{\vartheta_i + q_i - \Phi_i(\nu_k)} e^{-\nu_k u} \lambda_{ij}(-\nu_k). \tag{5.30}$$

By equating (5.29) and (5.30) we thus obtain equations that the coefficients should satisfy. As it turns out, doing this for any $i \in S^\downarrow$ and $k \notin S^\downarrow$ we again obtain (5.26). Following the same steps as the ones leading to Theorem 5.4.1, we obtain the following result. Notably, the matrix C now consists of entries c_{ik} with $k \notin S^\downarrow$, while $\boldsymbol{c}_k := (c_{1k}, \ldots, c_{dk})^\top$ as before.

Theorem 5.4.2. *Under* (Property A'), *the tail probability* $p_i(u)$ *satisfies*

$$p_i(u) = \sum_{k \notin S^\downarrow} c_{ik} e^{-\nu_k u},$$

for $i = 1, \ldots, d$. *Here, for* $k \notin S^\downarrow$, *the vectors* $\boldsymbol{c}_k$ *solve* $M(\nu_k)\, \boldsymbol{c}_k = \boldsymbol{0}$ *subject to* $\sum_{k \notin S^\downarrow} c_{ik} = 1$ *for all* $i \notin S^\downarrow$.

We note that, due to the fact that the rows i of C such that $i \in S^\downarrow$ do not add up to one, we have that $\mathbb{P}(Z_i = 0) = 1 - \sum_{k \notin S^\downarrow} c_{ik} > 0$.

5.4.3 Number of roots with a positive real part

In the previous subsections we provided a recipe to compute the tail probabilities $p_i(u)$ using Lemma 5.4.1. The objective of this subsection is to prove this lemma. To this end, we partition the state space of the background chain in K transient classes (say $T_1, \ldots, T_K$) and L recurrent

classes (say $R_1, \dots, R_L$). We label the classes such that for $\ell \in \{1, \dots, K\}$ class ℓ refers to T_ℓ, and for $\ell \in \{K+1, \dots, K+L\}$ class ℓ refers to $R_{\ell-K}$. We also order the transient classes as was done in Section 5.3.3: for any ℓ, T_ℓ has no transitions to other classes $T_{\ell'}$ such that $\ell' \leqslant \ell$. Furthermore, we let d_ℓ° be the number of states in class ℓ that do not correspond to non-increasing subordinators, $\ell = 1, \dots, K+L$.

With the introduced ordering of class, the transition rate matrix of the background chain can now be written in the following form:

$$Q = \begin{pmatrix} \bar{Q}_1 & S_{12} & \cdots & S_{1K} & S_{1,K+1} & \cdots & S_{1,K+L} \\ 0 & \bar{Q}_2 & \cdots & S_{2K} & S_{2,K+1} & \cdots & S_{2,K+L} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \bar{Q}_K & S_{K,K+1} & \cdots & S_{K,K+L} \\ 0 & 0 & \cdots & 0 & Q_{K+1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & Q_{K+L} \end{pmatrix}. \tag{5.31}$$

The block matrices $Q_{K+1}, \dots, Q_{K+L}$ correspond to the recurrent classes, and can be interpreted as 'true' transition rate matrices of Markov chains of lower dimension, in that they have non-negative entries except on their diagonals, and their row sums are all zero. This does not hold for the block matrices $\bar{Q}_1, \dots, \bar{Q}_K$: since they correspond to transient classes, their off-diagonal entries are still non-negative, but they have at least one strictly negative row sum. The matrices $S_{k,\ell}$, with $k = 1, \dots, K$ and $\ell = K+1, \dots, K+L$, contain non-negative entries and correspond to transitions from T_k into a different class.

The next step is to construct the matrix $M(\nu)$ that corresponds to the 'rearranged transition matrix' Q. This matrix is 'block-upper-triangular', which is inherited from the matrix Q. It concretely means that, for appropriately constructed matrices $\bar{M}_1(\nu), \dots, \bar{M}_K(\nu)$ and $M_{K+1}(\nu), \dots, M_{K+L}(\nu)$ (based on $\bar{Q}_1(\nu), \dots, \bar{Q}_K(\nu)$ and $Q_{K+1}(\nu), \dots, Q_{K+L}(\nu)$, respectively),

$$\det M(\nu) = \det \bar{M}_1(\nu) \cdots \det \bar{M}_K(\nu) \det M_{K+1}(\nu) \cdots \det M_{K+L}(\nu); \tag{5.32}$$

here the matrices $\bar{M}_\ell(\nu)$ (for $\ell = 1, \dots, K$) correspond to transient classes, whereas the matrices $M_\ell(\nu)$ (for $\ell = K+1, \dots, K+L$) correspond to recurrent classes. It is clear that $\det M_\ell(\nu)$, where $\ell = K+1, \dots, K+L$, has d_ℓ° roots with a positive real part, as an immediate consequence of Proposition 5.2.2. This also holds for $\det \bar{M}_\ell(\nu)$, where $\ell = 1, \dots, K$, which follows by rewriting the diagonal entries as we did in (5.17) and (5.18) in such a way that $\bar{M}_\ell(\nu)$ has the desired form to apply Proposition 5.2.2. Upon combining the above, we conclude that $\det M(\nu) = 0$ has $\sum_{\ell=1}^{K+L} d_\ell^\circ = d^\circ$ zeroes, as desired.

Remark 5.4.3. As indicated above, the computation of the coefficients c_{ik} amounts to solving an eigensystem; see Theorems 5.4.1 and 5.4.2. However, using the structure of the background chain, in specific cases this computation can be simplified considerably. Appealing to the factorization of $\det M(\nu)$ provided in Equation (5.32), it can be argued that some of the c_{ik} are necessarily equal to 0. In the first place, let ν_k be a root of $\det M(\nu)$ such that $\det M_\ell(\nu_k) = 0$ for some $\ell \in \{K+1, \dots, K+L\}$ (i.e., ℓ corresponds to a recurrent class). If the roots of $\det M(\nu)$ are simple, this means that ν_k does not solve $\det M_{\ell'}(\nu) = 0$ for

$\ell' \in \{K+1,\ldots,K+L\}$ with $\ell' \neq \ell$, nor $\det \bar{M}_{\ell'}(\nu) = 0$ for $\ell' \in \{1,\ldots,K\}$. By virtue of the structure of the matrix $M(\nu)$, which is inherited from the transition rate matrix Q (as given in (5.31)), we thus conclude that $c_{ik} = 0$ for all states i from which the recurrent class ℓ cannot be reached. In the second place, analogously, if ν_i is such that $\det \bar{M}_\ell(\nu_i) = 0$ for some $\ell \in \{1,\ldots,K\}$ (i.e., ℓ corresponds to a transient class), then $c_{ik} = 0$ for all states k that cannot be reached from this transient class. This reduction procedure makes intuitive sense: informally, the distribution of Z_i cannot be affected by properties of the MAP that correspond to states that cannot be reached from state i. ◇

5.5 Maximum of a spectrally one-sided Lévy process over a phase-type period

In Lévy fluctuation theory, the focus is predominantly on the evaluation of the distribution of extreme values over exponentially distributed intervals; see for instance Proposition 5.2.1 for a key result in this context. In the present section, we use our results on the maximum of a killed MAP to determine the distribution of a spectrally one-sided Lévy process over a phase-type distributed time interval.

The practical relevance of working with the class $\mathscr{P}$ of phase-type distributions lies in the fact that any distribution on the positive half-line can be approximated arbitrarily closely by a distribution in $\mathscr{P}$ [7, Theorem III.4.2]. The proof of this property reveals that actually any distribution on the positive half-line can be approximated arbitrarily closely by elements from a smaller class, namely the class of mixtures of Erlang distributions. In particular, a deterministic positive number can be approximated by an Erlang-distributed random variable with a large number of phases.

This section has two main goals. In Section 5.5.1, we show how our results on the maximum of a killed spectrally one-sided MAP can be applied to derive the distribution of the maximum of a spectrally one-sided Lévy process over a phase-type distributed time interval. Then, in Section 5.5.2, we obtain more specific results for the practically relevant class of mixtures of Erlang distributions.

5.5.1 Translation into the MAP framework

We start our exposition by interpreting a phase-type distributed random variable as an absorption time in a continuous-time Markov chain. Each element in the class $\mathscr{P}$ is characterized by (i) a finite state space $\{1,...,d\}$, (ii) an initial distribution $\boldsymbol{\alpha} \in \mathbb{R}^d$, (iii) a $d \times d$ matrix $T = (t_{ij})_{i,j=1}^d$ with non-positive diagonal entries, non-negative off-diagonal entries and non-positive row sums, and (iv) a non-negative *exit vector* $\boldsymbol{t} := -T\mathbf{1}$. Note that the $(d+1)\times(d+1)$ matrix

$$\bar{T} := \begin{pmatrix} T & \boldsymbol{t} \\ \mathbf{0}^\top & 0 \end{pmatrix},$$

is a genuine transition rate matrix of a $(d+1)$-state Markov chain, in that its diagonal entries are non-positive, its off-diagonal entries are non-negative, and its row sums are equal to zero. The $(d+1)$-st column and row in this matrix correspond to a newly added state $d+1$, which we refer to as the *absorbing state*. Observe that this chain can hit state $d+1$ from any other state according to the exit vector $\boldsymbol{t}$. Now, the phase-type random variable corresponding to the above instance is the time it takes the expanded Markov chain (with transition rate matrix $\bar{T}$) to reach the absorbing state, where the initial state has been sampled according to the distribution $\boldsymbol{\alpha}$.

We now consider the distribution of the maximum of the spectrally one-sided Lévy process $X(\cdot)$ over a phase-type distributed time interval (being characterized by the initial distribution $\boldsymbol{\alpha}$ and the transition rate matrix $\bar{T}$). To use the MAP framework that we have been working with in the previous sections, we let $X_1(\cdot), ..., X_d(\cdot)$ be independent copies of a common spectrally one-sided Lévy process $X(\cdot)$, such that the resulting MAP evolves as this Lévy process. We write $\varphi(\cdot)$ for the Laplace exponent of $X(\cdot)$ in case it is spectrally positive, and we write $\Phi(\cdot)$ for the cumulant generating function of $X(\cdot)$ in case it is spectrally negative. In addition, we let $X_{d+1}(t) \equiv 0$ for all $t \geqslant 0$. Furthermore, we choose $Q = T + \mathrm{diag}(\boldsymbol{t})$ and $\boldsymbol{\vartheta} = \boldsymbol{t}$ such that absorption in state $d+1$ corresponds to killing. In addition, let the jumps of the MAP at transition epochs of the background process, as represented by the random variables L_{ij}, be equal to zero. Observe that under this construction, with $\bar{Z}$ denoting the maximum of the Lévy process $X(\cdot)$ over the phase-type interval,

$$\mathbb{P}(\bar{Z} \geqslant u) = \sum_{i=1}^{d} \alpha_i \mathbb{P}(Z_i \geqslant u),$$

where $\mathbb{P}(Z_i \geqslant u)$ can be analyzed using the techniques for extremes of MAPs, as developed earlier in the chapter.

5.5.2 Mixtures of Erlang distributions

We are particularly interested in the case where the time interval is a mixture of Erlang distributions, because with this distribution class we can approximate any non-negative random variable arbitrarily closely. A mixture of Erlang distributions concretely means that, for some $k \in \mathbb{N}$ and $i = 1, \ldots k$, with probability $p_i \in [0,1]$ we sample from an Erlang distribution with shape parameter $d_i \in \mathbb{N}$ and scale parameter $\tau_i > 0$ (obviously requiring $\sum_{i=1}^{k} p_i = 1$). It takes little thought to conclude that, in order to evaluate the maximum of the Lévy process of such an interval, it suffices to be able to evaluate its maximum over an Erlang-distributed time interval (say with parameters $d \in \mathbb{N}$ and $\tau > 0$). This requires us to extend the result of an example from Asmussen and Ivanovs [10] which focuses on the maximum of Brownian motion (with a given drift and variance parameter) over an Erlang(d, τ) distributed time interval. Specifically, we generalize this result to any spectrally one-sided Lévy process (where, to avoid trivial cases, we assume that the underlying Lévy process is not a subordinator). Related results on maxima over an Erlang horizon include [27, Section 5], [37, Section IV.1], and [119]. In the remainder of this subsection we treat both spectrally one-sided cases separately.

Maximum of a spectrally-positive Lévy process over an Erlang-distributed time interval

Recall from Theorem 5.3.1 that the Laplace-Stieltjes transform of the maximum, $\boldsymbol{\zeta}(\gamma)$, satisfies the linear system $M(\gamma)\boldsymbol{\zeta}(\gamma) = \boldsymbol{b}(\gamma)$, where the $(d \times d)$-dimensional matrix $M(\cdot)$ is given by

$$M(\gamma) = \begin{pmatrix} -\tau + \varphi(\gamma) & \tau & 0 & \cdots & 0 & 0 \\ 0 & -\tau + \varphi(\gamma) & \tau & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\tau + \varphi(\gamma) & \tau \\ 0 & 0 & 0 & \cdots & 0 & -\tau + \varphi(\gamma) \end{pmatrix}.$$

The direct implication of the matrix $M(\cdot)$ being upper triangular is that $\boldsymbol{\zeta}(\gamma)$ as well as the unknown constants ω_i can be solved recursively. A concrete recipe for this could be the following. Defining $m(\gamma) := -\tau + \varphi(\gamma)$, we first find $\zeta_d(\gamma) = b_d(\gamma)/m(\gamma)$. Note that by definition of $b_d(\gamma)$ (see (5.9)), the numerator contains the constant ω_d, which can be identified using the observation that the (single) zero, say $\gamma^\star := \psi(\tau)$, from the denominator should be a zero of the numerator as well. As a next step, we identify $\zeta_{d-1}(\gamma)$ observing that

$$\zeta_{d-1}(\gamma) = \frac{b_{d-1}(\gamma) - \tau\,\zeta_d(\gamma)}{m(\gamma)} = \frac{m(\gamma)\,b_{d-1}(\gamma) - \tau\,b_d(\gamma)}{m^2(\gamma)}.$$

This expression contains the (by now known) constant ω_d as well as the (still unknown) constant ω_{d-1} through the function $b_{d-1}(\gamma)$. However, ω_{d-1} can again be found noting that the double zero from the denominator (which is again $\gamma^\star$) is also a double zero of the numerator. Thus noting that $m(\gamma^\star) = 0$, we obtain

$$m'(\gamma^\star)\,b_{d-1}(\gamma^\star) - \tau\,b_d'(\gamma^\star) = 0,$$

from which we find the unknown constant ω_{d-1}. We can continue along these lines until we have identified all Laplace-Stieltjes transforms $\zeta_i(\gamma)$ and corresponding constants ω_i, for $i = 1, \ldots, d$. After a number of computations, one then obtains

$$\zeta_i(\gamma) = \left(1 - \frac{\gamma}{\psi(\tau)}\right)\left(-\frac{\tau}{m(\gamma)}\right)^{d-i+1} - \frac{\gamma}{\psi(\tau)} \sum_{j=1}^{d-i} \left(-\frac{\tau}{m(\gamma)}\right)^{j} \lim_{\nu \to \psi(\tau)} \zeta_{i+j}(\nu).$$

Since $\zeta_j(\psi(\tau))$ is not well defined for any $j = 1, ..., d$, we note that $\lim_{\nu \to \psi(\tau)} \zeta_j(\nu)$ can be derived from $\zeta_j(\gamma)$ by L'Hôpital's rule.

Maximum of a spectrally-negative Lévy process over an Erlang-distributed time interval

For the spectrally-negative case, we follow the line of reasoning used in Section 5.4. The first observation is that in this case $M(\nu)$ has a single positive root with multiplicity d, namely $\nu := \Psi(\tau)$. Because the geometric multiplicity of $M(\nu)$ is one, its exponent contains multiples of $u^k e^{-\nu u}$ for $k = 0, ..., d-1$ (cf. Remark 5.4.2). Let Z_i be the maximum of the process $X(\cdot)$ when starting in phase i, i.e., with still $d - i + 1$ phases ahead (which is equivalent to setting $\alpha_i = 1$). We can represent the density of Z_i by a mixture of $d - i + 1$ Erlang densities, i.e.,

$$\mathbb{P}(Z_i \in \mathrm{d}u) = \sum_{k=1}^{d-i+1} a_{ik} \frac{\nu^k}{(k-1)!} u^{k-1} e^{-\nu u} \mathrm{d}u, \tag{5.33}$$

where the a_{ik} are coefficients that we will determine below. Evidently, the quantity $\nu^k/(k-1)!$ could have been incorporated in the coefficient a_{ik}, but, as it turns out, not doing so makes the formulas slightly cleaner. We already mention that $a_{d1} = 1$, as the maximum of a spectrally-negative Lévy process over an exponentially distributed interval is exponentially distributed with parameter $\Psi(\tau)$ (see Proposition 5.2.1). As a consequence of (5.33), we find that

$$p_i(u) = \int_u^\infty \mathbb{P}(Z_i \in \mathrm{d}x) = \sum_{k=1}^{d-i+1} a_{ik}\, e^{-\nu u} \sum_{m=0}^{k-1} \frac{\nu^m u^m}{m!}. \tag{5.34}$$

Similar to the strategy followed in Section 5.4, we now derive a second expression for $p_i(u)$, also in terms of the a_{ik}, which, in combination with (5.34), allows us to determine these unknown coefficients. To this end, note that in our setting (5.20) implies, for $i = 1, \dots, d-1$, that

$$p_i(u) = e^{-\nu u} + p_{i+1}^-(u), \tag{5.35}$$

where

$$p_{i+1}^-(u) = \int_0^u \nu e^{-\nu w} \mathbb{P}\big(\underline{X}(T_\tau) + Z_{i+1} \geqslant u - w\big)\,\mathrm{d}w. \tag{5.36}$$

Conditioning on the value v of Z_{i+1}, using (5.33), and substituting $x = u - w - v$, we thus obtain

$$\begin{aligned} p_{i+1}^-(u) &= \int_0^u \nu e^{-\nu w} \int_{u-w}^\infty \sum_{k=1}^{d-i} a_{i+1,k} \frac{\nu^k}{(k-1)!} v^{k-1} e^{-\nu v} \mathbb{P}\big(\underline{X}(T_\tau) \geqslant u - w - v\big)\,\mathrm{d}v\,\mathrm{d}w \\ &= \int_0^u e^{-\nu w} \int_{-\infty}^0 \sum_{k=1}^{d-i} a_{i+1,k} \frac{\nu^{k+1}}{(k-1)!} (u-w-x)^{k-1} e^{-\nu(u-w-x)} \mathbb{P}\big(\underline{X}(T_\tau) \geqslant x\big)\,\mathrm{d}x\,\mathrm{d}w. \end{aligned}$$

It then follows from an application of the binomial theorem that

$$\begin{aligned} p_{i+1}^-(u) &= \int_0^u \int_{-\infty}^0 \sum_{k=1}^{d-i} a_{i+1,k} \frac{\nu^{k+1}}{(k-1)!} \sum_{\ell=0}^{k-1} \binom{k-1}{\ell} (u-w)^\ell (-x)^{k-1-\ell} e^{\nu(x-u)} \mathbb{P}\big(\underline{X}(T_\tau) \geqslant x\big)\,\mathrm{d}x\,\mathrm{d}w \\ &= \sum_{k=1}^{d-i} a_{i+1,k} \sum_{\ell=0}^{k-1} e^{-\nu u} \int_0^u \frac{\nu^{\ell+1}(u-w)^\ell}{\ell!}\mathrm{d}w \cdot \int_{-\infty}^0 \frac{\nu^{k-\ell}(-x)^{k-1-\ell}}{(k-1-\ell)!} e^{\nu x} \mathbb{P}\big(\underline{X}(T_\tau) \geqslant x\big)\,\mathrm{d}x \\ &= \sum_{k=1}^{d-i} a_{i+1,k} \sum_{\ell=0}^{k-1} G_\ell(u)\; H_{k-1-\ell}, \end{aligned} \tag{5.37}$$

with

$$G_m(u) := e^{-\nu u} \int_0^u \frac{\nu^{m+1} w^m}{m!}\,\mathrm{d}w \quad \text{and} \quad H_m := \int_{-\infty}^0 \frac{\nu^{m+1}(-x)^m}{m!} e^{\nu x} \mathbb{P}\big(\underline{X}(T_\tau) \geqslant x\big)\,\mathrm{d}x.$$

For an alternative, probabilistic derivation of (5.37), we refer to Appendix 5.A.3. We proceed by evaluating the objects $G_m(u)$ and H_m. The former can be rewritten as

$$G_m(u) = e^{-\nu u} \frac{\nu^{m+1}}{m!} \int_0^u w^m\,\mathrm{d}w = e^{-\nu u} \frac{\nu^{m+1}}{(m+1)!} u^{m+1}.$$

Deriving a closed-form expression for H_m is significantly more involved. The following lemma proves useful in this regard.

Lemma 5.5.1. *For any non-negative random variable A and $m = 1, 2, \ldots,$*

$$\int_{-\infty}^{0} (-x)^m e^{\nu x} \mathbb{P}\big(-A \geqslant x\big)\,\mathrm{d}x = m!\,\nu^{-m-1} - h_m,$$

where $h_m := \int_0^\infty x^m e^{-\nu x}\,\mathbb{P}(A > x)\,\mathrm{d}x$. In addition, h_m satisfies the recursion

$$h_m = \frac{m}{\nu} h_{m-1} - \frac{1}{\nu} h_m^\circ,$$

with $h_m^\circ := \mathbb{E}\big(A^m e^{-\nu A}\big)$ and $h_0 = \frac{1}{\nu}\big(1 - \mathbb{E}\big(e^{-\nu A}\big)\big)$.

Proof: The first claim is verified by observing that $\mathbb{P}(-A \geqslant x) = 1 - \mathbb{P}(A > -x)$. The second claim, the recursive relation for h_m, follows by applying integration by parts. Namely, for $m = 1, 2, \ldots,$

$$\begin{aligned} h_m &:= \int_0^\infty x^m e^{-\nu x}\,\mathbb{P}(A > x)\,\mathrm{d}x = -\frac{1}{\nu}\int_0^\infty x^m\,\mathbb{P}(A > x)\,\mathrm{d}\big(e^{-\nu x}\big) \\ &= \int_0^\infty \frac{e^{-\nu x}}{\nu}\,\mathrm{d}\big(x^m\,\mathbb{P}(A > x)\big) = \frac{m}{\nu} h_{m-1} - \frac{1}{\nu} h_m^\circ. \end{aligned}$$

Finally, the expression for h_0 results from the definition of h_m in combination with (5.1). This completes the proof. □

In our analysis, we apply this lemma to the case where $A = -\underline{X}(T_\tau)$, so that

$$h_m^\circ = \mathbb{E}\left(A^m e^{-\nu A}\right) = (-1)^m \frac{\mathrm{d}^m}{\mathrm{d}\nu^m}\mathbb{E}\left(e^{-\nu A}\right) = (-1)^m \frac{\mathrm{d}^m}{\mathrm{d}\nu^m}\left(\frac{\Psi(\tau) - \nu}{\tau - \Phi(\nu)}\frac{\tau}{\Psi(\tau)}\right).$$

In conclusion,

$$H_m = 1 - \frac{\nu^{m+1}}{m!} h_m,$$

where for $n = 1, ..., m$,

$$h_n = \frac{n}{\nu} h_{n-1} - \frac{1}{\nu} h_n^\circ, \qquad \text{and} \qquad h_0 = \frac{1}{\nu}\left(1 - \frac{\Psi(\tau) - \nu}{\tau - \Phi(\nu)}\frac{\tau}{\Psi(\tau)}\right).$$

At this point, we have obtained two expressions for $p_i(u)$, both in terms of a sum whose summands are proportional to $e^{-\nu u}$, $u\,e^{-\nu u}, \ldots, u^{d-i}\,e^{-\nu u}$. Equating these gives a linear system from which the coefficients a_{ik} can be solved. In this context it is noted that, conveniently, solving this linear system allows a *recursive* solution procedure. To see this, first recall that $a_{d1} = \nu$. Then $a_{d-1,1}$ and $a_{d-1,2}$, appearing in (5.34) for $i = d - 1$, are by (5.37) expressed in terms of a_{d1}. Then, along similar lines, $a_{d-2,1}$, $a_{d-2,2}$ and $a_{d-2,3}$ are expressed in terms of $a_{d-1,1}$ and $a_{d-1,2}$, and so on.

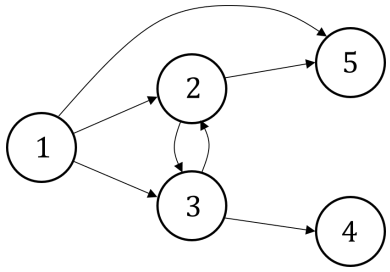

Figure 5.2: Example background chain.

5.6 Numerical experiments

In the previous sections, we developed theory on the distribution of the maximum of a spectrally one-sided MAP. We now discuss some practical issues concerning the implementation of our findings, the identification of the role played by each of the model parameters, and the application of our results in a practical context. So as to cover these three issues, we consider three experiments: the first highlights the impact of the structure of the background process, the second focuses on the maximum of a Lévy process in Erlang-distributed time intervals, and the third is motivated by a problem in risk theory.

5.6.1 Impact of the chain structure of the background process

In this first experiment we consider a spectrally-positive MAP in which the background chain $J(\cdot)$ has the structure shown in Figure 5.2 (where $q_{ij} > 0$ when there is an arrow from i to j and $q_{ij} = 0$ otherwise). With states 4 and 5 being absorbing, the background process is clearly not irreducible. As a consequence, we cannot use results from the existing literature, and have to rely on the results found in Section 5.3.

Our goal is to evaluate $\boldsymbol{\zeta}(\gamma)$, i.e., the vector of Laplace-Stieltjes transforms of the Z_i. This vector is the solution of the matrix equation (5.14), where we follow the procedure developed in Section 5.3.3 to determine the unknown constants ω_i. We first categorize the communicating classes of $J(\cdot)$ in layers: using the notation from Section 5.3.3, we have $C_0 = R = \{4, 5\}$, $C_1 = \{2, 3, 4, 5\}$ and $C_2 = \{1, 2, 3, 4, 5\}$. Note that even though the communicating class $\{1\}$ has a transition to the recurrent state 5, it belongs only to C_2 because it also has transitions into C_1. Following our procedure, we consecutively evaluate $\zeta_i(\gamma)$ for $i \in \{4, 5\}$, then for $i \in \{2, 3\}$, and finally for $i = 1$.

Using this approach, we now consider an example MAP with the background chain structure given in Figure 5.2. We let $X_1(\cdot), X_2(\cdot), X_3(\cdot), X_5(\cdot)$ correspond to standard Brownian motions, and $X_4(\cdot)$ to a gamma process (i.e. Lévy process with independent gamma distributed increments) with jump intensity 2 and jump size parameter 2, so that

$$\varphi_4(\alpha) = 2 \log\left(\frac{2}{\alpha + 2}\right).$$

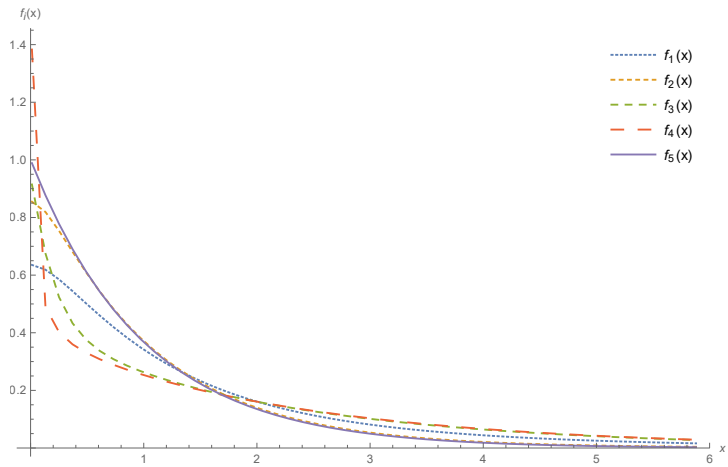

Figure 5.3: Probability density functions $f_1(\cdot), \ldots, f_5(\cdot)$, corresponding to $Z_1, ..., Z_5$, with the model parameters as specified in this section.

Additionally, we let the background be governed by the transition rate matrix

$$Q = \begin{pmatrix} -4 & 1 & 2 & 0 & 1 \\ 0 & -20 & 1 & 0 & 19 \\ 0 & 1 & -20 & 19 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and we set $L_{ij} = 0$ for all i, j, so there are no jumps at transition epochs of the background process. Finally, we consider the setting where $\boldsymbol{\vartheta} = (0, 0, 0, 0.5, 0.5)^\top$, meaning that killing only happens in states 4 and 5. For these model parameters, we plot in Figure 5.3 the density functions $f_i(\cdot)$ of the Z_i.

We comment on a few aspects pertaining to Figure 5.3 that illustrate the impact of the chain structure. First, from the given transition rate matrix it is clear that if $J(0) = 2$, then the process likely ends up in state 5. This explains why the densities of Z_2 and Z_5 behave similarly. A similar reasoning applies to Z_3 and Z_4. Also, Z_4 and Z_5 are 'closer to being killed' than Z_2 and Z_3, and therefore have more probability mass close to zero. Finally, notice that from initial state 1, absorption in state 4 or 5 is about equally likely, resulting in a density function that roughly behaves as the average of the two pairs mentioned above.

5.6.2 Maximum of a Lévy process in an Erlang-distributed time interval

This example focuses on the distribution of the maximum of a Lévy process over a time interval of Erlang-distributed length, applying the theory of Section 5.5. We choose the model parameters as pointed out in Section 5.5.1. That is, we let background state i represent the i-th phase of the interval. With n denoting the total number of phases, we let $q_{i,i+1} = \frac{1}{n}$ for all $i = 1, ..., n-1$ and $\boldsymbol{\vartheta} = (0, 0, ..., 0, \frac{1}{n})^\top$. This way, the mean interval length equals unity and killing occurs only in the last phase. Because phase transitions should not affect the Lévy process, we take $L_{i,i+1} = 0$ for all $i = 1, ..., n-1$. We particularly study the impact of the

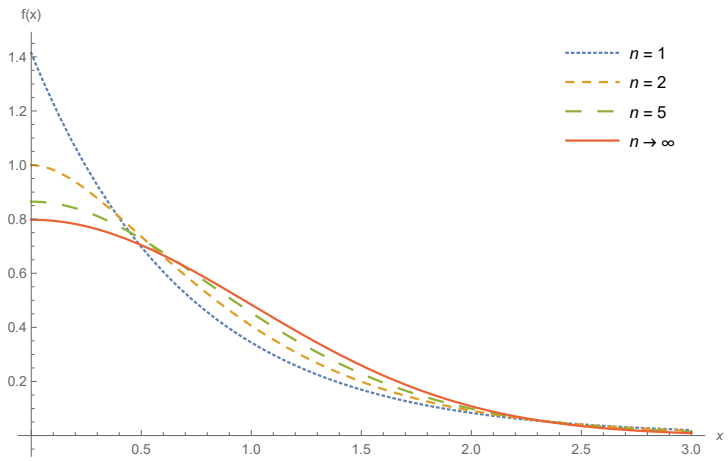

Figure 5.4: Density functions $f(\cdot)$ of the maximum of a standard Brownian motion over an n-phase Erlang-distributed interval with mean 1 (for $n = 1, 2, 5$), with the solid line representing their counterpart for a deterministic interval of length 1 (taking $n \to \infty$).

number of phases on the distribution of the maximum of the Lévy process, bearing in mind the Erlang distribution's capability of approximating a deterministic number. Indeed, our Erlang random variable converges to the deterministic value 1 as $n \to \infty$, and this experiment serves to get insight into the maximum of a Lévy process during a deterministic time interval.

We first consider the case of $X(\cdot)$ being a standard Brownian motion, noting that for this instance we know that its maximum in a deterministic interval has a half-normal distribution (i.e., the distribution of the absolute value of a normally distributed random variable). Figure 5.4 shows the corresponding density functions for $n = 1, 2, 5$ phases, as well as its limiting counterpart. The figure confirms that the densities converge to the limit, where the curve for $n = 5$ already produces a reasonable fit.

We proceed with an example in which the distribution of the maximum over a deterministic time horizon is not known. Let $X(\cdot)$ be the independent sum of (i) a standard Brownian motion that is increased by a positive drift 1, and (ii) a compound Poisson process with arrival rate 1 and Erlang$(2, 2)$ distributed jumps in the negative direction (thus rendering the process spectrally negative). Figure 5.5 illustrates the (fast) convergence of the density functions as n grows, thus providing us with a way to approximate the distribution of the maximum of $X(\cdot)$ evaluated over a deterministic interval.

5.6.3 Risk model

The last example is a special case of the model discussed in [38], and is motivated by applications in credit risk. The process of interest is the capital of an insurance company over time, with a finite number of obligors n. Each obligor independently goes into default after an exponentially distributed time with mean 1. When going into default, the obligor makes a claim of exponential size with mean 1, and immediately ends the contract with the insurance company (i.e., leaves the system). Each obligor not gone into default pays premiums at rate r per time unit. Figure

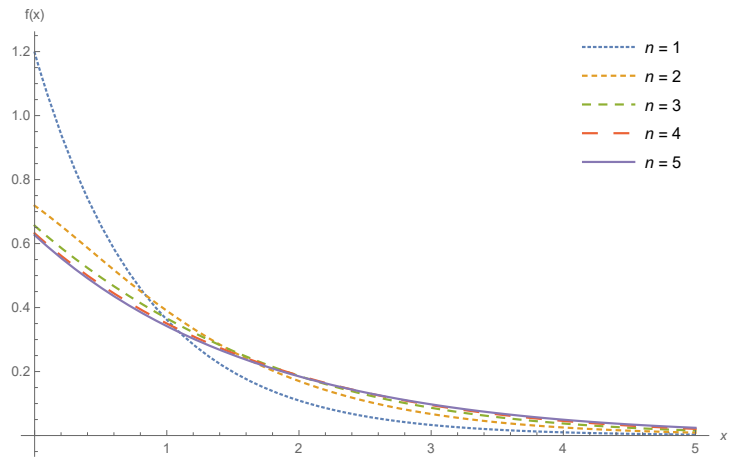

Figure 5.5: Density functions $f(\cdot)$ of the maximum of the Lévy process $X(\cdot)$ over an n-phase Erlang-distributed interval with mean 1 (for $n = 1, ..., 5$). Here, $X(\cdot)$ is the sum of a standard Brownian motion with positive drift, and a compound Poisson process with negative Erlang-distributed jump sizes.

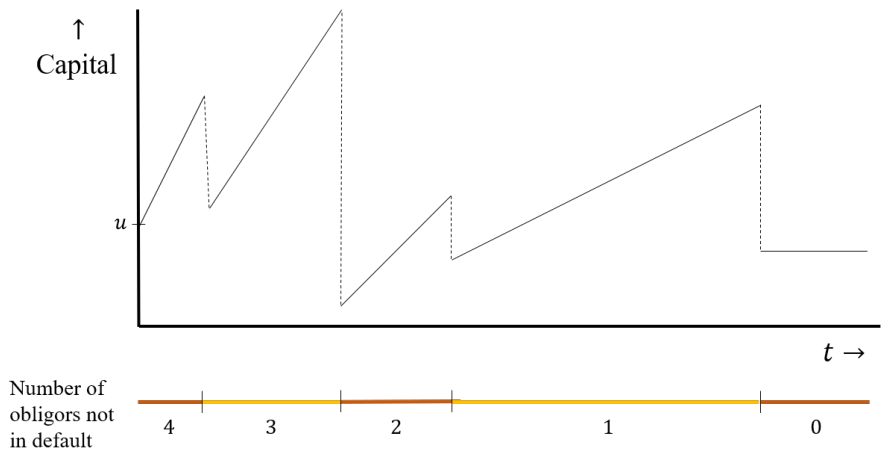

Figure 5.6: Sample path of the capital surplus of the insurance company, with four obligors.

5.6 shows a possible sample path of the process. We wish to quantify the ruin probability, i.e., the probability that the capital of the insurance company eventually hits zero, given some initial reserve $u \geqslant 0$.

This model can be cast in our framework as follows. Let background state i represent the number of obligors that have not yet gone into default. Then, the transition rates of the background chain are $q_{i,i-1} = i$ for $i = 1, \ldots, n$ (all other transition rates are 0). As we are interested in the all-time ruin probability, we let $\boldsymbol{\vartheta} = \mathbf{0}$. Observe that the ruin probability depends on the minimum of the process, where the results in this chapter are in terms of the maximum, but this is easily remedied by flipping the sign. Concretely, we choose $X_i(t) = -irt$ for $t \geqslant 0$, $i = 0, \ldots, n$, and we have positive jumps $L_{i,i-1}$ of exponentially distributed size with mean 1 for $i = 1, \ldots, n$. The ruin probability with initial capital u is now given by $\mathbb{P}(Z_n \geqslant u)$.

In an example with (initially) four obligors, Figure 5.7 shows how the ruin probability depends

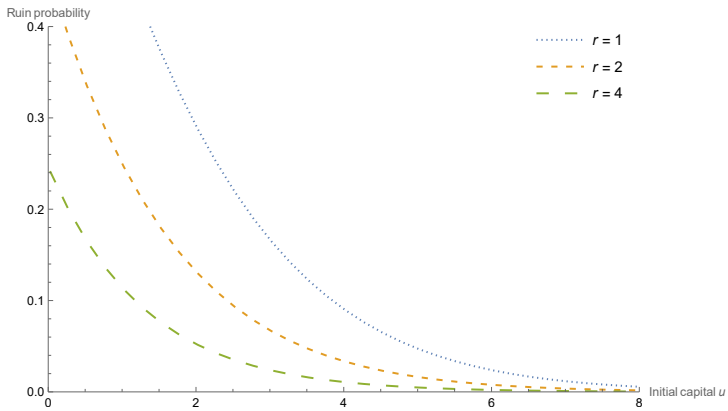

Figure 5.7: Ruin probability per initial capital, for a few different premium rates r.

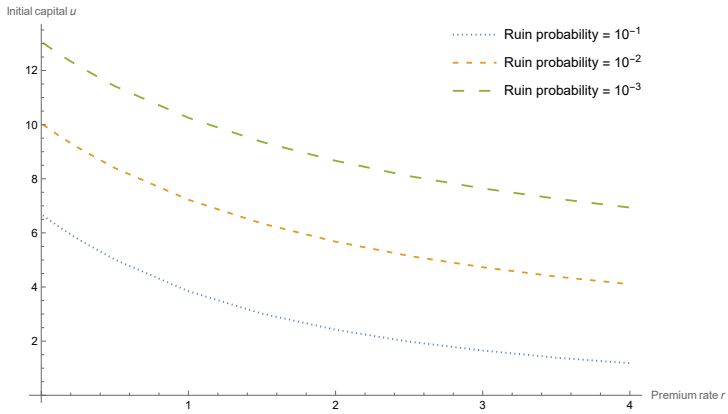

Figure 5.8: Relation between premium rate and initial capital, for a few fixed ruin probabilities.

on the initial capital u and premium rate r. The ruin probability is decreasing in both u and r, and our techniques can be used to assess the impact of these parameters. Clearly, one can trade off u and r: when reducing the premium rate r, a higher initial surplus u is needed to guarantee a given ruin probability. This trade-off is illustrated in Figure 5.8.

5.7 Directions for further research

In this final section, we discuss two directions for further research.

Adding jumps 'in the opposite direction'

A possible extension to our model could be the inclusion of phase-type distributed jumps in the 'opposite direction' to the model. More concretely, in our spectrally-positive setup we would allow phase-type negative jumps (both in the Lévy processes and in the jumps at transition epochs of the background process), and analogously in our spectrally-negative setup we would allow phase-type positive jumps. Alternatively, one could consider jumps 'in the opposite direction' whose distribution has a rational Laplace transform (rather than being a phase-type distribution). Extensions of this type are in line with earlier analyses, such as [78],

being conceptually relatively straightforward, but requiring a substantial amount of additional notation. Relative to the spectrally one-sided MAPs that we considered in the present chapter, models that include phase-type jumps 'in the opposite direction' are significantly more general; we recall that any positive random variable can be approximated arbitrarily closely by a phase-type distributed random variable [7, Theorem III.4.2].

Wiener-Hopf-type results under non-irreducibility

In the present chapter the focus has been on spectrally one-sided MAPs, the underlying aim being the derivation of computable quantities. As a result, our analysis provides expressions for the distribution of the maximum in terms of the model primitives. In [60], general MAPs (i.e., without any assumptions on the direction of the jumps, but still requiring that $J(\cdot)$ is irreducible) are considered, leading to a Wiener-Hopf-type decomposition. The price to be paid, however, is that the characterization of the distribution of the maximum is considerably less explicit than in the spectrally one-sided cases. An interesting question is to what extent the results in [60] for the extrema carry over to the case that $J(\cdot)$ is not irreducible.

Appendix 5.A Probabilistic arguments

In this section we present alternative, probabilistic derivations of some equations in this chapter.

5.A.1 Alternative derivation of Equation (5.6)

Starting from state i, we once again condition on whether the MAP is killed before leaving state i or the background process jumps to some state $j \neq i$. This leads to

$$Z_i = \begin{cases} \overline{X}_i(T_{\vartheta_i+q_i}) & \text{with probability } \dfrac{\vartheta_i}{\vartheta_i+q_i}, \\ \overline{X}_i(T_{\vartheta_i+q_i}) + [\underline{X}_i(T_{\vartheta_i+q_i}) + L_{ij} + Z_j]^+ & \text{with probability } \dfrac{q_{ij}}{\vartheta_i+q_i},\ j \neq i. \end{cases}$$

In terms of Laplace-Stieltjes transforms, this is equivalent to

$$\zeta_i(\gamma) = \frac{\vartheta_i}{\vartheta_i+q_i}\kappa_i(\gamma) + \sum_{j\neq i}\frac{q_{ij}}{\vartheta_i+q_i}\kappa_i(\gamma)\mathbb{E}\left(e^{-\gamma[\underline{X}_i(T_{\vartheta_i+q_i})+L_{ij}+Z_j]^+}\right), \tag{5.38}$$

where by Proposition 5.2.1, the random variable $-\underline{X}_i(T_{\vartheta_i+q_i})$ is exponentially distributed with rate μ_i. We now present a lemma enabling us to evaluate the rightmost Laplace-Stieltjes transform.

Lemma 5.A.1. *Let A be a non-negative random variable. Then for $\gamma, \mu \geqslant 0$, we have*

$$\mathbb{E}\left(e^{-\gamma[A-T_\mu]^+}\right) = \frac{\mu}{\mu-\gamma}\mathbb{E}\left(e^{-\gamma A}\right) - \frac{\gamma}{\mu-\gamma}\mathbb{E}\left(e^{-\mu A}\right).$$

Proof. Applying the standard identity $e^{-x^+} + e^{x^-} = e^{-x} + 1$ to $x = \gamma(A - T_\mu)$, it holds that, using the memoryless property of T_μ,

$$\begin{aligned}\mathbb{E}\left(e^{-\gamma[A-T_\mu]^+}\right) &= \mathbb{E}\left(e^{-\gamma(A-T_\mu)}\right) + 1 - \mathbb{E}\left(e^{\gamma[A-T_\mu]^-}\right)\\ &= \mathbb{E}\left(e^{-\gamma A}\right)\mathbb{E}\left(e^{\gamma T_\mu}\right) + 1 - \left(\mathbb{P}(A > T_\mu) + \mathbb{P}(A < T_\mu)\mathbb{E}\left(e^{\gamma T_\mu}\right)\right).\end{aligned}$$

The result follows from the fact that $\mathbb{P}(A < T_\mu) = \mathbb{E}\left(e^{-\mu A}\right)$. □

Equation (5.6) can now immediately be obtained from (5.38) by using the above lemma with $A = L_{ij} + Z_j$ and recalling (5.2).

5.A.2 Alternative derivation of Equation (5.25)

Observe that the definition (5.21) is equivalent to

$$p_{ij}^-(u) := \mathbb{P}\left(\overline{X}_i(T_{\vartheta_i+q_i}) < u\,,\, \overline{X}_i(T_{\vartheta_i+q_i}) + \underline{X}_i(T_{\vartheta_i+q_i}) + L_{ij} + Z_j \geqslant u\right).$$

For $k = 1, ..., d$, let Z_{jk} be independent exponentially distributed random variables with rate ν_k. An alternative formulation of (5.23) is that, recalling that $\overline{X}_i(T_{\vartheta_i+q_i})$ is exponentially distributed with parameter μ_i,

$$p_{ij}^-(u) = \sum_{k=1}^{d} c_{jk}\mathbb{P}\left(\overline{X}_i(T_{\vartheta_i+q_i}) < u\,,\, \overline{X}_i(T_{\vartheta_i+q_i}) + \underline{X}_i(T_{\vartheta_i+q_i}) + L_{ij} + Z_{jk} \geqslant u\right).$$

Because $\underline{X}_i(T_{\vartheta_i+q_i})$ and L_{ij} are non-positive, the memoryless property of Z_{jk} implies that

$$p_{ij}^-(u) = \sum_{k=1}^{d} c_{jk}\mathbb{P}\left(\overline{X}_i(T_{\vartheta_i+q_i}) < u\,,\, Z_{jk} + \overline{X}_i(T_{\vartheta_i+q_i}) \geqslant u\right)\cdot \mathbb{P}\left(Z_{jk} > -(\underline{X}_i(T_{\vartheta_i+q_i}) + L_{ij})\right).$$

Since Z_{jk} has an exponential distribution with rate ν_k, we have that $\mathbb{P}(Z_{jk} > A) = \mathbb{E}(e^{-\nu_k A})$ for any non-negative random variable A. Combining this observation with the fact that $\overline{X}_i(T_{\vartheta_i+q_i})$ is exponentially distributed with rate μ_i, we then arrive at

$$\begin{aligned}p_{ij}^-(u) &= \sum_{k=1}^{d} c_{jk}\frac{\mu_i}{\mu_i - \nu_k}(e^{-\nu_k u} - e^{-\mu_i u})\mathbb{E}\left(e^{\nu_k \underline{X}_i(T_{\vartheta_i+q_i})}\right)\lambda_{ij}(-\nu_k)\\ &= \sum_{k=1}^{d} c_{jk}\frac{\vartheta_i + q_i}{\vartheta_i + q_i - \Phi_i(\nu_k)}(e^{-\nu_k u} - e^{-\mu_i u})\lambda_{ij}(-\nu_k),\end{aligned}$$

where the transform of $\underline{X}_i(T_{\vartheta_i+q_i})$ is taken from Proposition 5.2.1. This completes the alternative derivation of Equation (5.25).

5.A.3 Alternative derivation of Equation (5.37)

Let U_k be an Erlang random variable with k phases of rate ν, and let $(V_n)_{n\leqslant k}$ denote the first n phases of U_k. Equation (5.36) can be written as

$$\begin{aligned} p_{i+1}^{-}(u) &= \mathbb{P}(\overline{X}_i(T_\tau) \leqslant u,\, \overline{X}_i(T_\tau) + \underline{X}_i(T_\tau) + Z_{i+1} > u) \\ &= \sum_{k=1}^{d-i} a_{i+1,k}\mathbb{P}(\overline{X}_i(T_\tau) \leqslant u,\, \overline{X}_i(T_\tau) + \underline{X}_i(T_\tau) + U_k > u) \end{aligned}$$

since Z_{i+1} is a mixture of $d-i$ Erlang random variables (see (5.33)). Suppose we need to add to $\overline{X}_i(T_\tau)$ exactly $\ell \in \{1, ..., k\}$ exponential phases of U_k in order to exceed u, that is, $u - \overline{X}_i(T_\tau) \in [V_{\ell-1}, V_\ell)$. The remainder of the ℓ-th phase is again exponential, so there are $k-\ell+1$ phases left to negate the non-positive variable $\underline{X}_i(T_\tau)$. Therefore we write

$$p_{i+1}^{-}(u) = \sum_{k=1}^{d-i} a_{i+1,k} \sum_{\ell=1}^{k} \mathbb{P}(\overline{X}_i(T_\tau)+V_{\ell-1} \leqslant u < \overline{X}_i(T_\tau)+V_\ell\,,\, \underline{X}_i(T_\tau)+(V_k-V_\ell+T_\nu) > 0), \quad (5.39)$$

where by Proposition 5.2.1, $\overline{X}_i(T_\tau)$ is exponentially distributed with rate ν. Now note that the two events in the above probability are independent, and their respective probabilities are

$$\begin{aligned} \mathbb{P}(\overline{X}_i(T_\tau) + V_{\ell-1} \leqslant u < \overline{X}_i(T_\tau) + V_\ell) &= \int_0^u \mathbb{P}(V_\ell - V_{\ell-1} > u - x)\mathbb{P}(\overline{X}_i(T_\tau) + V_{\ell-1} \in \mathrm{d}x) \\ &= \int_0^u e^{-\nu(u-x)} \frac{\nu^\ell x^{\ell-1}}{(\ell-1)!} e^{-\nu x}\,\mathrm{d}x = G_{\ell-1}(u) \end{aligned} \quad (5.40)$$

and

$$\begin{aligned} \mathbb{P}(\underline{X}_i(T_\tau) + (V_k - V_\ell + T_\nu) > 0) &= \int_0^\infty \mathbb{P}(\underline{X}_i(T_\tau) > -x)\mathbb{P}(V_k - V_\ell + T_\nu \in \mathrm{d}x) \\ &= \int_0^\infty \mathbb{P}(\underline{X}_i(T_\tau) > -x) \frac{\nu^{k-\ell+1} x^{k-\ell}}{(k-\ell)!} e^{-\nu x}\mathrm{d}x = H_{k-\ell}. \end{aligned} \quad (5.41)$$

Combining (5.39), (5.40) and (5.41) finally leads to Equation (5.37).

6 Ruin probabilities of Markov additive processes

6.1 Introduction

Primarily motivated by applications in insurance risk, much attention has been directed to evaluating ruin probabilities [8]. In this context, one typically assumes that the driving net cumulative claim process $X(\cdot)$ is of Lévy type (or a subclass thereof, such as the compound Poisson process). With the insurance firm's initial surplus being u, one is mainly interested in computing the probability $p(u)$ of the surplus level process $u - X(\cdot)$ dropping below 0. If this is beyond reach, which is often the case, one may settle for the (exact) *tail asymptotics*, the goal being to identify an explicit function $\bar{p}(u)$ such that $p(u)/\bar{p}(u) \to 1$ as $u \to \infty$.

Research into such exact asymptotics goes back about a century. In the regime that the driving Lévy process is of compound Poisson type, and in case the claim size distribution is *light-tailed*, the classical Cramér-Lundberg asymptotics entail that $p(u)$ decays, for u large, like $\bar{p}(u) = \alpha\, e^{-\theta^\star u}$ for positive constants α and $\theta^\star$ (where the decay rate $\theta^\star$ solves the so-called Lundberg equation). In the analysis one often relies on a duality with the stationary workload in the M/G/1 queue; see e.g. [8, Chapter IV.2] and [37]. Where in the literature the focus lies on finding the exact tail asymptotics for spectrally-positive (i.e., having only positive jumps) light-tailed Lévy processes, Bertoin and Doney [17] derive such asymptotics for their spectrally two-sided counterpart (i.e., having jumps in both directions).

A natural extension of the above body of work concerns exact tail asymptotics of ruin probabilities in the context of MAPs. Studying such an extension is one of the main objectives of this chapter, specifically for a spectrally-positive light-tailed MAP. We find that the exact asymptotics of the ruin probability $p(u)$ are of Cramér-Lundberg type, i.e., they are of the form $\bar{p}(u) = \alpha\, e^{-\theta^\star u}$ for positive α and $\theta^\star$. The present chapter complements Chapter 5 and its prior works, such as [40, 43, 59], which succeeded in identifying the Laplace transform of $p(u)$ for spectrally one-sided MAPs. With the expressions found there, one could in principle try to perform Laplace inversion in order to obtain numerical values for $p(u)$, but a complication is that these become inaccurate in the regime that ruin is rare. The findings of the present chapter remedy this issue: we provide an asymptotic expansion that becomes increasingly accurate as u grows.

Cramér-Lundberg asymptotics have been studied for various kinds of specific MAPs. Without providing an exhaustive overview, we discuss a few relevant contributions. Siegl and Tichy [118] consider a compound Poisson process with dividend barrier and two possible claim frequencies governed by a two-state Markov chain. In Jasiulewicz [65] a model is analyzed with the premium rate depending continuously on the reserve level and the Markov-modulated claim

frequency. More general MAPs of compound Poisson type are studied with constant premium rate [84] or constant claim frequency [128]. The case where all components of the compound Poisson process are allowed to change with the background chain is considered by Zhu and Yang [129]. Typically, for the case that the driving process is of MAP type, the prefactor α is characterized relatively implicitly; see e.g. [8, Theorem VII.3.7], where α is expressed in terms of random quantities at a stopping time under an alternative probability measure. An exception to this is the work of Miyazawa [85] on the Cramér-Lundberg asymptotics of a Markov-modulated compound Poisson process including jumps at transition epochs. By considering a ladder-height approach, that work determines an explicit expression for the prefactor α for this model.

In detail, the contributions of this chapter are the following:

- In the first place, we succeed in establishing exact asymptotics $\bar{p}(u) = \alpha\, e^{-\theta^\star u}$ for the ruin probability $p(u)$ in the model where the net cumulative claim process is represented by a spectrally-positive MAP $Y(\cdot)$ with light-tailed claim sizes and negative drift (such that $p(u) \to 0$ as $u \to \infty$). Through a delicate analysis of the overshoot of $Y(\cdot)$ over level u (under a specific alternative probability measure), we manage to express the constants α and $\theta^\star$ in terms of the model primitives. By lower bounding the overshoot by 0 we obtain, as a by-product, a version of the Lundberg inequality for light-tailed spectrally-positive MAPs.

- Secondly, we show that the ruin probability $p(u)$ can be efficiently computed by simulation; this is particularly useful in the regime that ruin is rare, in which direct simulation would be extremely time-consuming. To this end, we construct a generalized version of Siegmund's algorithm [9, Chapter VI.2], which amounts to performing importance sampling using the alternative measure that we identified when establishing the exact asymptotics. The resulting estimator can be considered optimal, in that we succeed in proving that it has bounded relative error. Numerical experiments reveal that the number of runs required to obtain an estimate with a given precision (i.e., the ratio of the width of the confidence interval and the estimate) is essentially constant in u; this is a huge improvement over direct simulation, under which this number of runs roughly grows as $e^{\theta^\star u}$. The experiments also show that $\bar{p}(u) = \alpha\, e^{-\theta^\star u}$ is an accurate approximation of the ruin probability, even for relatively small initial surplus levels u.

A crucial idea underlying our analysis is the observation that in order to identify whether the spectrally-positive MAP $Y(\cdot)$ has crossed level u, it is sufficient to monitor only the maxima of $Y(\cdot)$ between every two successive transition epochs of the background process; in other words, it is not needed to monitor $Y(\cdot)$ continuously in time. The resulting embedded process turns out to be conceptually substantially simpler than the original one. When deriving the exact asymptotics via the embedded process, two steps play a key role at the technical level. (i) The first concerns the introduction of a specific new probability measure $\mathbb{Q}$ under which the net cumulative claim process $Y(\cdot)$ has a positive drift, so that ruin occurs almost surely; see for instance [8, Chapter VII.3]. The main idea is then to write $p(u)$ in terms of the likelihood ratio of a path to ruin under the original measure $\mathbb{P}$, relative to the new measure $\mathbb{Q}$. This

reasoning leads to an expression for $p(u)$ in terms of the overshoot of $Y(\cdot)$ over u under $\mathbb{Q}$. (ii) In the second step, which combines application of the Wiener-Hopf decomposition with the Number of Zeroes proposition, this overshoot is analyzed in further detail (in the regime that $u \to \infty$).

The outline of this chapter is as follows. Section 6.2 defines our MAP $Y(\cdot)$, introduces the ruin probability $p(u)$, and highlights relevant preliminary results. The change-of-measure argument is carried out in Section 6.3, which leads, as mentioned above, to an expression for $p(u)$ in terms of the overshoot of $Y(\cdot)$ over u under $\mathbb{Q}$. As an intermediate step towards the identification of the exact asymptotics of $p(u)$, the purpose of Section 6.4 is to find the transform of the overshoot distribution under $\mathbb{Q}$. As it turns out, this overshoot transform can be expressed in terms of the solution to a set of linear equations. Combining the above findings, in Section 6.5 the exact asymptotics of the ruin probability are derived (i.e., the constants α and $\theta^\star$ in $\bar{p}(u) = \alpha\, e^{-\theta^\star u}$ are identified). A simulation algorithm of generalized Siegmund type, enabling fast estimation of $p(u)$, is given in Section 6.6. In addition, we give a proof for its bounded relative error, making the algorithm particularly useful in the rare-event regime. Finally, Section 6.7 presents the output of simulation experiments that provide an indication of the achievable speedup, as well as of the accuracy of the approximation $p(u) \approx \bar{p}(u)$.

6.2 Model and preliminaries

In this section we first introduce the model that we consider in this chapter. This is followed by a description of our specific objective regarding the ruin probability. We then discuss the Wiener-Hopf decomposition for spectrally-positive Lévy processes and the result (the Number of Zeroes proposition) concerning the number of singularities of a spectrally one-sided MAP. Finally, we briefly outline the approach that we will follow in later sections to establish the exact asymptotics.

The model and preliminaries strongly resemble those featuring in Chapter 5, save for a few specific aspects (e.g. the restriction to spectrally-positive processes, the background process being irreducible, no killing, and slightly different versions of the Wiener-Hopf decomposition and Number of Zeroes Proposition). Accordingly, much of the notation of Chapter 5 carries over to the present chapter.

Model

We start by defining our model. Let the background process $J(\cdot) \equiv (J(t))_{t\geq 0}$ be an irreducible continuous-time Markov chain with state space $\{1, ..., d\}$ for $d > 1$. The corresponding generator matrix is given by $Q := (q_{ij})_{i,j=1}^{d}$, with $q_i := -q_{ii} > 0$, having invariant probability distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d)$. Associated with every state i, let $X_i(\cdot) \equiv (X_i(t))_{t\geq 0}$ be a spectrally-positive Lévy process, evolving independently of $J(\cdot)$. Our model is a MAP $Y(\cdot)$ with background chain $J(\cdot)$, underlying Lévy processes $X_1(\cdot), ..., X_d(\cdot)$ and jumps distributed as L_{ij} when a transition from state i to state j occurs. We assume that $Y(0) = 0$.

Objective

The aim of this chapter is to identify the tail asymptotics of the maximum of the MAP conditional on the initial background state being $i \in \{1, \dots, d\}$. That is, with

$$p_i(u) := \mathbb{P}\Big(\max_{s \geqslant 0} Y(s) \geqslant u \mid J(0) = i \Big),$$

we wish to find an explicit function $\bar{p}_i(u)$ such that $p_i(u)/\bar{p}_i(u) \to 1$ as $u \to \infty$. We say that $\bar{p}_i(u)$ are the *exact asymptotics* of $p_i(u)$. We throughout assume that $Y(\cdot)$ has a negative drift, i.e.,

$$\mu := \sum_{i=1}^{d} \pi_i \, \mathbb{E}\left(X_i(1)\right) + \sum_{i=1}^{d} \sum_{k \neq i} \pi_i q_{ik} \, \mathbb{E}\left(L_{ij}\right) < 0, \tag{6.1}$$

such that exceeding u is increasingly rare as $u \to \infty$; in the actuarial literature, this condition is known as the *net profit condition*. The focus is on a *light-tailed* spectrally-positive MAP, under which $p_i(u)$ decays effectively exponentially. We provide a more precise description of what we mean by the MAP being light-tailed in Section 6.3.

Preliminaries

As in Chapter 5, the Wiener-Hopf decomposition and the Number of Zeroes proposition play a crucial role in the analysis. Since we focus on the spectrally-positive case, we only use the corresponding part of the Wiener-Hopf decomposition.

Proposition 6.2.1 (Wiener-Hopf decomposition)**.** *Let $(X(t))_{t \geqslant 0}$ be a spectrally-positive Lévy process. Then $X(T_\nu)$ can be decomposed as the sum of the two independent quantities $\overline{X}(T_\nu)$ and $X(T_\nu) - \overline{X}(T_\nu)$. Moreover, $X(T_\nu) - \overline{X}(T_\nu)$ has the same distribution as $\underline{X}(T_\nu)$ which is distributed as $-T_{\psi(\nu)}$. In addition,*

$$\mathbb{E}\left(e^{-\gamma \overline{X}(T_\nu)}\right) = \frac{\nu}{\nu - \varphi(\gamma)} \left(1 - \frac{\gamma}{\psi(\nu)}\right).$$

For a detailed account, see e.g. Kyprianou [73, Chapter 6].

Likewise, we require the Number of Zeroes proposition [61, Theorem 2] for the spectrally-positive case only.

Proposition 6.2.2 (Number of Zeroes)**.** *Let $(Y(t))_{t \geqslant 0}$ be a spectrally-positive MAP, and let the underlying background process $(J(t))_{t \geqslant 0}$ be irreducible. Then the equation* $\det M(\gamma) = 0$ *has $d - |S^{\uparrow}| - \mathbb{1}_{\mu \leqslant 0}$ solutions in $\mathbb{C}$ with positive real part.*

Approach

We conclude this section with a short account of the approach followed in the derivation of the exact asymptotics, as covered by Sections 6.3–6.5. The MAP itself being a rather involved object, we work with a more manageable embedded version. This embedded process

is constructed in such a way that the event of the MAP exceeding u is equivalent to the event of the embedded process exceeding u. More specifically, throughout this chapter we restrict ourselves, similar to the approach followed in Chapter 5, to the value of $Y(\cdot)$ at only three types of time points. In the interval between two transitions of the background process we record the value of $Y(\cdot)$

(i) at the start of the interval (right after the jump at transition epoch, that is),

(ii) at the epoch that the maximum value (within the interval) is achieved, and

(iii) at the end of the interval (right before the jump at transition epoch, that is).

It can be seen that in order to verify whether $Y(\cdot)$ exceeds u we can restrict ourselves to the values of the MAP at epochs of type (ii). The increments of the MAP between the embedded time points are relatively straightforward to deal with, as a consequence of Proposition 6.2.1.

In Section 6.3 we work with the above embedding to find an alternative expression for the ruin probability under an alternative measure $\mathbb{Q}$. Concretely, we succeed in expressing $p_i(u)$ in terms of the overshoot of $Y(\cdot)$ over level u under $\mathbb{Q}$. This overshoot is then analyzed in Section 6.4, which eventually leads to the exact asymptotics of $p_i(u)$ in Section 6.5.

6.3 Change of measure

In this section we derive an alternative expression for the ruin probability $p_i(u)$ by applying a change of measure. The derivation consists of the following four steps:

1. constructing a system of equations for finding the positive solution to the so-called *Lundberg equation*, providing the candidate for the exponential decay rate $\theta^\star$ appearing in the exact asymptotics of $p_i(u)$ (we show later that $\theta^\star$ is independent of the initial state i);

2. defining an *embedding* of the MAP $Y(\cdot)$, based on the time points of type (i), (ii) and (iii) that were introduced in Section 6.2;

3. defining an alternative probability measure $\mathbb{Q}$, also corresponding to a MAP, but with different driving Lévy processes, a different Markovian background process, and jumps at transition epochs having different distributions;

4. calculating the likelihood ratio of the actual probability measure $\mathbb{P}$ with respect to the new probability measure $\mathbb{Q}$, through which we find a compact expression for our target probability $p_i(u)$.

The compact expression for $p_i(u)$ provides us with an explicit (exponentially decaying) bound on $p_i(u)$, uniformly in $u \geqslant 0$. More importantly, however, it also eventually allows us to evaluate the exact asymptotics of $p_i(u)$, as shown in Sections 6.4 and 6.5.

Step 1: the Lundberg equation

The Lundberg equation, which yields the candidate for the exponential decay rate $\theta^\star$ appearing in the exact asymptotics of $p_i(u)$, involves the increment of the MAP between two consecutive epochs that the background state changes to an (arbitrarily chosen) reference state $i_0 \in \{1, \ldots, d\}$. The goal of Step 1 is to quantify the moment generating function of this increment, providing the solution to the Lundberg equation.

Denote, for $i = 1, \ldots, d$, by $(V_i^n)_{n\in\mathbb{N}}$ a sequence of independent random variables being distributed as the generic random variable $V_i := \overline{X_i}(T_{q_i})$ (in words: the maximum of the spectrally-positive Lévy process $X_i(\cdot)$ over an exponentially distributed time with mean q_i^{-1}). Also, denote by $(W_i^n)_{n\in\mathbb{N}}$ a sequence of independent random variables distributed as the generic random variable $W_i := -\underline{X_i}(T_{q_i})$ (in words: minus the minimum of the spectrally-positive Lévy process $X_i(\cdot)$ over an exponentially distributed time with mean q_i^{-1}); recall from Proposition 6.2.1 that W_i is exponentially distributed with parameter $\delta_i := \psi_i(q_i)$. Let the resulting $2d$ sequences be independent. Due to Proposition 6.2.1, the Laplace-Stieltjes transforms of these random variables are given by

$$v_i(\alpha) := \mathbb{E}\left(e^{-\alpha V_i}\right) = \int_0^\infty e^{-\alpha x} f_i^{(\mathbb{P})}(x)\,\mathrm{d}x = \frac{q_i}{q_i - \varphi_i(\alpha)}\left(1 - \frac{\alpha}{\delta_i}\right),$$
$$w_i(\alpha) := \mathbb{E}\left(e^{-\alpha W_i}\right) = \int_0^\infty e^{-\alpha x} g_i^{(\mathbb{P})}(x)\,\mathrm{d}x = \frac{\delta_i}{\delta_i + \alpha},$$

with $f_i^{(\mathbb{P})}(\cdot)$ the density of V_i and $g_i^{(\mathbb{P})}(\cdot)$ the density of W_i (for ease assumed to exist), both under the original measure $\mathbb{P}$. Below, we will use the notation $h_{ij}^{(\mathbb{P})}(\cdot)$ for the density of L_{ij} under $\mathbb{P}$.

Now fix a reference state $i_0 \in \{1, \ldots, d\}$, and define by U the increment of $Y(\cdot)$ between two subsequent visits of the background process to this state i_0. That is, if t_n and t_m are two subsequent times that the background chain enters state i_0, then U is distributed as $Y(t_m) - Y(t_n)$. We assume that we are in the light-tailed regime, in the sense that the *Lundberg equation*

$$\tilde{u}(\theta) := \mathbb{E}\left(e^{\theta U}\right) = 1$$

has a positive solution, say $\theta^\star$; in the literature, $\tilde{u}(\theta)$ is referred to as the *moment generating function* (mgf) of the random variable U.

Two remarks are in place now. In the first place, the fact that the *Lundberg equation* has a positive solution implies that the jumps L_{ij} have a finite mgf in an open interval around the origin, and that the same holds for the mgfs of the possible upward jumps of the Lévy processes $X_i(\cdot)$. It is for this reason that we call the driving MAP $Y(\cdot)$ light-tailed. In the second place, we note that it may seem that $\theta^\star$ depends on the reference state i_0 chosen. Below, however, we argue that the choice of i_0 does not affect the value of $\theta^\star$.

Next, we show that we can express the root $\theta^\star$ in terms of a generalized eigensystem. Let $y_j \equiv y_j(\theta)$, for $j \neq i_0$, be the moment generating function of the increment of $Y(\cdot)$, with the background process starting in state j, before the background process reaches the reference state i_0. To derive a system of equations, consider an increment of $Y(\cdot)$ in an interval of the

type $(t_n, t_{n+1}]$, that is, an interval between two successive transition epochs of the background process $J(\cdot)$. By Proposition 6.2.1, we know that such an increment can be written as the sum of three independent variables: the maximum of the corresponding Lévy process in this interval, the decrease after this maximum, and the jump at the transition epoch. In terms of moment generating functions this leads to the equations

$$\tilde{u}(\theta) = y_{i_0} = v_{i_0}(-\theta) w_{i_0}(\theta) \sum_{j \neq i_0} \frac{q_{i_0 j}}{q_{i_0}} \lambda_{i_0 j}(-\theta)\, y_j,$$

and, for $j \neq i_0$,

$$y_j = v_j(-\theta) w_j(\theta) \left(\frac{q_{j i_0}}{q_j} \lambda_{j i_0}(-\theta) + \sum_{k \neq j, i_0} \frac{q_{jk}}{q_j} \lambda_{jk}(-\theta)\, y_k \right). \tag{6.2}$$

Then we equate $\tilde{u}(\theta)$ to 1, in order to find $\theta^\star$. We thus end up with the following system of equations: for $i = 1, \ldots, d$,

$$q_i y_i = v_i(-\theta) w_i(\theta) \sum_{j \neq i} q_{ij} \lambda_{ij}(-\theta)\, y_j, \tag{6.3}$$

with $y_{i_0} = 1$. Note that if $\boldsymbol{y}$ solves this system of equations, then so does $c \cdot \boldsymbol{y}$ for any c, implying that the solution $\theta^\star$ does not depend on the reference state i_0. It also means that $\boldsymbol{y}$ is unique up to a multiplication by a factor, so that ratios of the type y_i / y_j are uniquely determined. Due to the fact that the components of the vector $\boldsymbol{y} = \boldsymbol{y}(\theta^\star)$ can be interpreted as moment generating functions, we conclude that $\boldsymbol{y}$ is necessarily componentwise positive. Below we will use the eigenvalue/eigenvector pair $(\theta^\star, \boldsymbol{y})$ when defining the alternative measure $\mathbb{Q}$.

Step 2: the embedding

The idea is to embed the process $Y(\cdot)$ into a simpler process that still contains all information based on which it can be determined whether it exceeds level u. To this end, observe that it suffices to consider only the values of $Y(\cdot)$ that correspond to maxima between transitions of the background process. This observation is visualized in Figure 6.1; the maxima between transition epochs (i.e., the dots) give all information that is needed to verify whether level u is ever crossed.

Define the embedded process $(S_n)_n$ by

$$\begin{aligned} S_0 &:= V^0_{K_0}, \\ S_n &:= S_{n-1} - W^n_{K_{n-1}} + L^n_{K_{n-1}, K_n} + V^n_{K_n}, \quad n = 1, 2, \ldots, \end{aligned}$$

with $K_0 := J(0)$ and $K_n := J(t_n)$ denoting the state of the background process right after the n-th transition of the background process.

As discussed above, and appealing to Proposition 6.2.1 to justify the independence between the sequences $(V^n_i)_{n \in \mathbb{N}}$ and $(W^n_i)_{n \in \mathbb{N}}$, we can rewrite the ruin probability $p_i(u)$ in terms of the embedded process $(S_n)_n$, as follows:

$$p_i(u) = \mathbb{P}\big(\exists n \in \mathbb{N}_0 : S_n \geqslant u \mid K_0 = i\big).$$

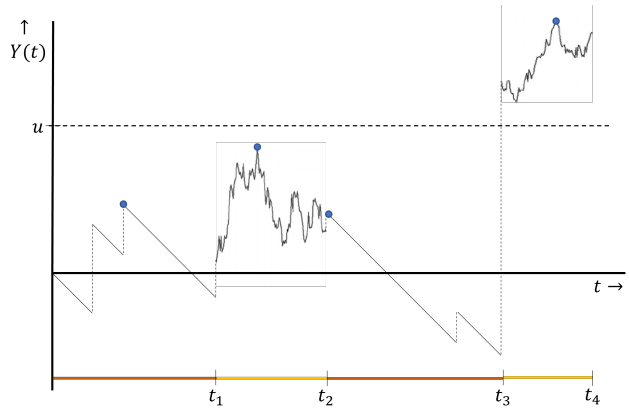

Figure 6.1: Example of a MAP with two background states, with dots representing the embedded process $(S_n)_n$. That is, the dots represent the maxima of the intervals $[t_n, t_{n+1})$ for $n \in \mathbb{N}$.

Step 3: construction of the alternative probability measure

To evaluate the ruin probability in the regime that $u \to \infty$, we work with a change of measure. Particularly, we set up the alternative probability measure $\mathbb{Q}$ (to be defined below), under which the process is still a MAP, but now has positive drift. The goal is to evaluate $p_i(u)$ where the path of $(S_n)_n$ is sampled under $\mathbb{Q}$. The measure $\mathbb{Q}$ is such that

$$\mathbb{E}_{\mathbb{Q}}\left(e^{\theta U}\right) = \mathbb{E}\left(e^{(\theta+\theta^\star)U}\right) = \tilde{u}(\theta + \theta^\star) \tag{6.4}$$

(where a more descriptive definition will be given below). From $\tilde{u}(\cdot)$ being convex, $\tilde{u}'(0) < 0$ (due to (6.1)) and $\tilde{u}(0) = \tilde{u}(\theta^\star) = 1$, it follows that $\tilde{u}'(\theta^\star) > 0$, which means that under the new measure $\mathbb{Q}$ the drift of $(S_n)_n$ has become positive, so that u is eventually exceeded almost surely under $\mathbb{Q}$. As an aside, we also note that if state i corresponds to a subordinator process $X_i(\cdot)$ under $\mathbb{P}$, then the same applies under $\mathbb{Q}$ (and vice versa).

Let $\ell \equiv \ell(S)$ be the appropriate likelihood ratio (or Radon-Nikodym derivative), which records the likelihood of the path of $(S_n)_n$ under $\mathbb{P}$ relative to $\mathbb{Q}$ (until $(S_n)_n$ exceeds u, that is). Then we have the following standard equality translating the likelihood of outcomes under $\mathbb{Q}$ into those under $\mathbb{P}$:

$$p_i(u) = \mathbb{E}\big(\mathbb{1}\{\exists n \in \mathbb{N}_0 : S_n \geqslant u\} \mid K_0 = i\big) = \mathbb{E}_{\mathbb{Q}}\big(\ell(S)\,\mathbb{1}\{\exists n \in \mathbb{N}_0 : S_n \geqslant u\} \mid K_0 = i\big).$$

An important benefit of working with $\mathbb{Q}$ is that under this measure, the event in the indicator function has probability 1 due to the positive drift, whereas under $\mathbb{P}$ the same probability vanishes as u becomes large; as a consequence,

$$p_i(u) = \mathbb{E}_{\mathbb{Q}}\big(\ell(S) \mid K_0 = i\big). \tag{6.5}$$

It is noted that by many other definitions of $\mathbb{Q}$ we could have achieved a positive drift, and thus the validity of (6.5). The crucial feature of our specific alternative measure, as given in (6.4), however, is that for this $\mathbb{Q}$ the likelihood ratio ℓ takes a simple form, as we will show below.

The alternative measure $\mathbb{Q}$, in an abstract sense defined via (6.4), is concretely described as follows. Under $\mathbb{Q}$, the distributions of $(V_i^n)_n$ and $(W_i^n)_n$ are characterized through the densities

$$f_i^{(\mathbb{Q})}(x) = f_i^{(\mathbb{P})}(x)\frac{e^{\theta^\star x}}{v_i(-\theta^\star)}, \quad g_i^{(\mathbb{Q})}(x) = g_i^{(\mathbb{P})}(x)\frac{e^{-\theta^\star x}}{w_i(\theta^\star)},$$

respectively, for $i = 1, \ldots, n$ and $x > 0$. Note that this means that the Laplace-Stieltjes transform of the $(V_i^n)_n$ under $\mathbb{Q}$ becomes

$$v_i^{(\mathbb{Q})}(\alpha) = \frac{v_i(\alpha - \theta^\star)}{v_i(-\theta^\star)}$$

and that under $\mathbb{Q}$ the $(W_i^n)_n$ are exponentially distributed with parameter

$$\delta_i^{(\mathbb{Q})} := \delta_i + \theta^\star.$$

In addition, the $(L_{ij}^n)_n$ under $\mathbb{Q}$ have density

$$h_{ij}^{(\mathbb{Q})}(x) = h_{ij}^{(\mathbb{P})}(x)\frac{e^{\theta^\star} x}{\lambda_{ij}(-\theta^\star)},$$

where, in the same way as above, the corresponding Laplace-Stieltjes transform can be verified to be $\lambda_{ij}^{(\mathbb{Q})}(\alpha) = \lambda_{ij}(\alpha - \theta^\star)/\lambda_{ij}(-\theta^\star)$.

Now that we have defined the driving Lévy processes and the jumps at transition epochs under $\mathbb{Q}$, we conclude by considering the transition rates of the background process under $\mathbb{Q}$. For $i \neq j$ these become

$$q_{ij}^{(\mathbb{Q})} = q_{ij}\lambda_{ij}(-\theta^\star)\frac{y_j}{y_i}.$$

It is readily verified that this choice implies that the diagonal elements of the transition rate matrix under $\mathbb{Q}$ are

$$q_{ii}^{(\mathbb{Q})} = -\sum_{j\neq i} q_{ij}^{(\mathbb{Q})} = -\sum_{j\neq i} q_{ij}\lambda_{ij}(-\theta^\star)\frac{y_j}{y_i} = -q_i\frac{1}{v_i(-\theta^\star)w_i(\theta^\star)},$$

where the last equality is by virtue of (6.3).

Step 4: the likelihood ratio

The next step is to evaluate the likelihood ratio $\ell(S)$ on a path such that $(S_n)_n$ reaches u, which is equal to $p_i(u)$ by (6.5). Let $N \equiv N(u)$ denote the first n for which $S_n \geqslant u$ (which is under the new measure $\mathbb{Q}$ finite almost surely). Then $\ell \equiv \ell(S)$ can be written as the product of four factors:

- In the first place there is the contribution of the 'in between maxima' $V_{K_0}^0, \ldots, V_{K_N}^N$:

$$\prod_{m=0}^{N} \frac{f_{K_m}^{(\mathbb{P})}(V_{K_m}^m)}{f_{K_m}^{(\mathbb{Q})}(V_{K_m}^m)} = \prod_{m=0}^{N} v_{K_m}(-\theta^\star)\, e^{-\theta^\star V_{K_m}^m}.$$

- Secondly, there is the contribution of the 'in between minima' $W^0_{K_0}, \dots, W^{N-1}_{K_{N-1}}$:

$$\prod_{m=0}^{N-1} \frac{g^{(\mathbb{P})}_{K_m}(W^m_{K_m})}{g^{(\mathbb{Q})}_{K_m}(W^m_{K_m})} = \prod_{m=0}^{N-1} w_{K_m}(\theta^\star)\, e^{\theta^\star W^m_{K_m}}.$$

- In the third place, there is the contribution of the jumps at transition epochs of the background process $L^1_{K_0,K_1}, \dots, L^N_{K_{N-1},K_N}$:

$$\prod_{m=1}^{N} \frac{h^{(\mathbb{P})}_{K_{m-1},K_m}(L^m_{K_{m-1},K_m})}{h^{(\mathbb{Q})}_{K_{m-1},K_m}(L^m_{K_{m-1},K_m})} = \prod_{m=1}^{N} \lambda_{K_{m-1},K_m}(-\theta^\star)\, e^{-\theta^\star L^m_{K_{m-1},K_m}}.$$

- And finally there is the contribution due to the jumps of the background process:

$$\prod_{m=1}^{N} \frac{q_{K_{m-1},K_m}/q_{K_{m-1}}}{q^{(\mathbb{Q})}_{K_{m-1},K_m}/q^{(\mathbb{Q})}_{K_{m-1}}}.$$

  It is readily verified (recognizing a 'telescopic product') that this contribution simplifies to

$$\frac{y_i}{y_{K_N}} \left(\prod_{m=0}^{N-1} v_{K_m}(-\theta^\star) \cdot \prod_{m=0}^{N-1} w_{K_m}(\theta^\star) \cdot \prod_{m=1}^{N} \lambda_{K_{m-1},K_m}(-\theta^\star) \right)^{-1}.$$

It is also noted that, by the definition of the process $(S_n)_n$,

$$\sum_{m=0}^{N} V^m_{K_m} - \sum_{m=0}^{N-1} W^m_{K_m} + \sum_{m=1}^{N} L^m_{K_{m-1},K_m} = S_N.$$

Multiplying the above four components of the likelihood ratio, the resulting expression greatly simplifies. It means that, upon combining the above, and recalling the identity (6.5), we have arrived at the following result.

Theorem 6.3.1. *For all $u \geqslant 0$ and $i \in \{1, \dots, d\}$,*

$$p_i(u) = \mathbb{E}_{\mathbb{Q}} \left(\frac{y_i}{y_{K_N}}\, v_{K_N}(-\theta^\star)\, e^{-\theta^\star S_N} \right).$$

Remark 6.3.1. To get insight into the expression found in Theorem 6.3.1, compare it to its counterpart for the maximum of a random walk with independent and identically distributed increments $(X_n)_{n\in\mathbb{N}}$ (distributed as the generic random variable X). With $p(u)$ the probability of the random walk exceeding level u, and S_N denoting the value of this random walk at the moment N that u is crossed, a similar change of measure yields $p(u) = \mathbb{E}_{\mathbb{Q}}(e^{-\theta^\star S_N})$, with $\theta^\star$ solving $\mathbb{E}[e^{\theta^\star X}] = 1$ and the X under $\mathbb{Q}$ having Laplace-Stieltjes transform

$$\mathbb{E}_{\mathbb{Q}}\left(e^{-\alpha X}\right) = \mathbb{E}\left(e^{-(\alpha-\theta^\star)X}\right).$$

This principle underlies the celebrated Siegmund algorithm for efficiently estimating $p(u)$; see for a detailed account e.g. [9, Equation (2.5)]. The additional factors for the MAP case, as appearing in Theorem 6.3.1, have two reasons. First, the factor y_i/y_{K_N} reflects the impact of

the initial and eventual background states of the MAP. Secondly, regarding $v_{K_N}(-\theta^\star)$, one could say that (by definition) the number of 'steps' of the embedded process $(S_n)_n$ is odd: in step n there have been $n+1$ contributions 'of the V-type', and just n contributions 'of the $(-W+L)$-type', the consequence being that at step N the contribution of the moment generating function of one of the V_i (namely the last one) has not been neutralized. This results in the additional factor $v_{K_N}(-\theta^\star)$ in the expression for $p_i(u)$. ◇

While the focus of the section lies on deriving an alternative expression for $p_i(u)$, as a by-product we obtain a uniform upper bound on $p_i(u)$. To this end, we define

$$y^+ := \max_{j=1,\dots,d} \frac{y_i}{y_j}, \quad v^+ := \max_{j=1,\dots,d} v_j(-\theta^\star).$$

Realizing that by definition $S_N \geqslant u$, the following result, which can be seen as the MAP-counterpart of the conventional Lundberg inequality, is an immediate consequence of Theorem 6.3.1.

Corollary 6.3.1. *For all $u \geqslant 0$ and $i \in \{1,\dots,d\}$,*

$$p_i(u) \leqslant y^+ v^+ e^{-\theta^\star u}.$$

Observe that S_N can be decomposed into $u + R(u)$, with $R(u)$ denoting the 'overshoot' of the embedded process $(S_n)_n$ over level u (i.e., $S_{N(u)} - u \geqslant 0$); we also write $N(u)$ rather than just N to stress the dependence on u. This means that if we manage to compute, for $\theta > 0$, $i = 1, \dots, d$,

$$r_{ij}(\theta) := \lim_{u\to\infty} \mathbb{E}_{\mathbb{Q}}\left(e^{-\theta R(u)} \mathbb{1}\{K_{N(u)} = j\} \mid K_0 = i\right),$$

then Theorem 6.3.1 would entail

$$\lim_{u\to\infty} p_i(u)\, e^{\theta^\star u} = \sum_{j=1}^{d} r_{ij}(\theta^\star)\, \frac{y_i}{y_j}\, v_j(-\theta^\star), \tag{6.6}$$

which is the MAP-counterpart of the classical Cramér-Lundberg asymptotics that we were aiming at. Therefore, we would be done if we would be able to devise a way to compute the overshoot transform $r_{ik}(\theta)$. As it turns out, this can be done by taking a second transform (with respect to u): note that

$$r_{ij}(\theta) = \lim_{\alpha\downarrow 0} \alpha\, s_{ij}(\alpha,\theta),$$

where

$$s_{ij}(\alpha,\theta) := \int_0^\infty e^{-\alpha u}\, \mathbb{E}_{\mathbb{Q}}\left(e^{-\theta R(u)} \mathbb{1}\{K_{N(u)} = j\} \mid K_0 = i\right) \mathrm{d}u. \tag{6.7}$$

Informally, the object $\alpha\, s_{ij}(\alpha,\theta))$ corresponds to the overshoot over an exponentially distributed threshold with mean α^{-1}, which grows to ∞ when sending $\alpha \downarrow 0$. In other words: what is left is (i) computing $s_{ij}(\alpha,\theta)$, and (ii) let $\alpha \downarrow 0$ in $\alpha\, s_{ij}(\alpha,\theta)$. These are the topics of the next two sections.

6.4 Computing the overshoot transform

In this section we are interested in evaluating the double transform $s_{ij}(\alpha,\theta)$ corresponding to the target level u and the overshoot $R(u)$, as defined in (6.7). Recall that i and j respectively represent the initial background state and the state at the time level u is crossed. To find an expression for $s_{ij}(\alpha,\theta)$, we distinguish three scenarios for the MAP during the first background state i:

1. Level u has been crossed before the first transition of the background process. This only leads to a contribution if $i = j$.
2. Level u has not been crossed before the first transition of the background process, but due to the jump at the transition epoch it crosses level u. This only leads to a contribution if $i \neq j$.
3. Level u is not crossed before or at the first transition epoch of the background process (to state k, say), but later it is.

We now split $s_{ij}(\alpha,\theta)$ into the components $s_i^{(1)}(\alpha,\theta)$, $s_{ij}^{(2)}(\alpha,\theta)$, and $s_{ijk}^{(3)}(\alpha,\theta)$ corresponding to the above three scenarios. In the first place, for $i \neq j$ we have

$$s_{ij}(\alpha,\theta) = \frac{q_{ij}^{(\mathbb{Q})}}{q_i^{(\mathbb{Q})}} s_{ij}^{(2)}(\alpha,\theta)\, v_j^{(\mathbb{Q})}(\theta) + \sum_{k\neq i} \frac{q_{ik}^{(\mathbb{Q})}}{q_i^{(\mathbb{Q})}} s_{ijk}^{(3)}(\alpha,\theta), \tag{6.8}$$

where

$$v_i^{(\mathbb{Q})}(\theta) := \mathbb{E}_{\mathbb{Q}}\left(e^{-\theta V_i}\right) = \frac{v_i(\theta-\theta^\star)}{v_i(-\theta^\star)},$$

with $s_{ij}^{(2)}(\alpha,\theta)$ and $s_{ijk}^{(3)}(\alpha,\theta)$ evaluated below. The first term in the right-hand side of (6.8) is interpreted as the contribution due to the scenario in which V_i remains below u, a transition from background state i to background state j is made, and the corresponding jump brings $Y(\cdot)$ above u; in this scenario the overshoot includes an extra V_j (because we consider the embedded process, see Step 2 in Section 3). The second term in the right-hand side of (6.8) reflects the scenario in which V_i remains below u, and a transition from background state i to background state $k \neq i$ is made such that after the corresponding jump the process $Y(\cdot)$ is still below u. The transform $s_{ij}^{(2)}(\alpha,\theta)$ is formally defined by

$$\int_0^\infty e^{-\alpha u}\, \mathbb{E}_{\mathbb{Q}}\left(e^{-\theta R(u)} \mathbb{1}\{\bar{Y}(t_1) < u, Y(t_1) + L_{ij} \geqslant u, K_1 = j\} \mid K_0 = i\right) \mathrm{d}u,$$

in which case $R(u) = Y(t_1) + L_{ij} + V_j - u$, and the transform $s_{ijk}^{(3)}(\alpha,\theta)$ by

$$\int_0^\infty e^{-\alpha u}\, \mathbb{E}_{\mathbb{Q}}\left(e^{-\theta R(u)} \mathbb{1}\{\bar{Y}(t_1) < u, Y(t_1) + L_{ij} < u, K_1 = k, K_{N(u)} = j\} \mid K_0 = i\right) \mathrm{d}u.$$

In the case that $i = j$ we obtain along similar lines

$$s_{ii}(\alpha,\theta) = s_i^{(1)}(\alpha,\theta) + \sum_{k\neq j} \frac{q_{ik}^{(\mathbb{Q})}}{q_i^{(\mathbb{Q})}} s_{iik}^{(3)}(\alpha,\theta), \tag{6.9}$$

with $s_i^{(1)}(\alpha,\theta)$ evaluated below. The first term in the right-hand side of (6.9) is the contribution due to the scenario in which V_i exceeds u. In the second term in the right-hand side of (6.8) we have that V_i remains below u, and a transition from background state i to background state $k \neq j$ is made such that after the corresponding jump $Y(\cdot)$ is still below u. The transform $s_i^{(1)}(\alpha,\theta)$ is formally defined by

$$\int_0^\infty e^{-\alpha u}\,\mathbb{E}_{\mathbb{Q}}\left(e^{-\theta R(u)}\mathbb{1}\{\bar{Y}(t_1) \geqslant u\} \mid K_0 = i\right)\mathrm{d}u,$$

in which case $R(u) = \bar{Y}(t_1) - u$.

We now further evaluate the expressions of $s_i^{(1)}(\alpha,\theta)$, $s_{ij}^{(2)}(\alpha,\theta)$, and $s_{ijk}^{(3)}(\alpha,\theta)$. First, by conditioning on the value of $V_i \in [u,\infty)$, we directly find that

$$s_i^{(1)}(\alpha,\theta) = \int_{u=0}^\infty e^{-\alpha u}\int_{v=u}^\infty f_i^{(\mathbb{Q})}(v)\,e^{-\theta(v-u)}\,\mathrm{d}v\,\mathrm{d}u.$$

By conditioning on $V_i \in [0,u)$, W_i and $L_{ij} \in [u-(V_i-W_i),\infty)$, we also have that

$$s_{ij}^{(2)}(\alpha,\theta) = \int_{u=0}^\infty e^{-\alpha u}\int_{v=0}^u f_i^{(\mathbb{Q})}(v)\int_{w=0}^\infty g_i^{(\mathbb{Q})}(w)\int_{z=u-v+w}^\infty h_{ij}^{(\mathbb{Q})}(z)\,e^{-\theta(v+z-u-w)}\,\mathrm{d}z\,\mathrm{d}w\,\mathrm{d}v\,\mathrm{d}u.$$

Analogously, conditioning on $V_i \in [0,u)$, W_i and $L_{ij} \in [0,u-(V_i-W_i))$,

$$\begin{aligned} s_{ijk}^{(3)}(\alpha,\theta) = &\int_{u=0}^\infty e^{-\alpha u}\int_{v=0}^u f_i^{(\mathbb{Q})}(v)\\ &\int_{w=0}^\infty g_i^{(\mathbb{Q})}(w)\int_{z=0}^{u-v+w} h_{ik}^{(\mathbb{Q})}(z)\,r_{kj}(\theta,u-v+w-z)\,\mathrm{d}z\,\mathrm{d}w\,\mathrm{d}v\,\mathrm{d}u,\end{aligned}$$

where

$$r_{kj}(\theta,u) := \mathbb{E}_{\mathbb{Q}}(e^{-\theta R(u)}\mathbb{1}\{K_{N(u)} = j\} \mid K_0 = k).$$

Next, for each of the quantities, we swap the order of the integrals so that the most elementary integration (i.e., the one over u), becomes the innermost integral. Then this integral is computed, and after that the other integrals can be evaluated successively.

Along the lines that we sketched above, we obtain

$$\begin{aligned} s_i^{(1)}(\alpha,\theta) &= \int_{v=0}^\infty f_i^{(\mathbb{Q})}(v)\,e^{-\theta v}\int_{u=0}^v e^{(\theta-\alpha)u}\,\mathrm{d}u\,\mathrm{d}v\\ &= \frac{1}{\theta-\alpha}\int_{v=0}^\infty f_i^{(\mathbb{Q})}(v)\left(e^{-\alpha v}-e^{-\theta v}\right)\mathrm{d}v = \frac{v_i^{(\mathbb{Q})}(\alpha)-v_i^{(\mathbb{Q})}(\theta)}{\theta-\alpha}.\end{aligned}$$

For $s_{ij}^{(2)}(\alpha,\theta)$, a similar strategy can be followed. This yields

$$\begin{aligned} s_{ij}^{(2)}(\alpha,\theta) &= \int_{v=0}^\infty f_i^{(\mathbb{Q})}(v)\int_{z=0}^\infty h_{ij}^{(\mathbb{Q})}(z)\int_{w=0}^\infty \delta_i^{(\mathbb{Q})}e^{-\delta_i^{(\mathbb{Q})}w}\int_{u=v}^{z+v-w} e^{(\theta-\alpha)u}e^{-\theta(z+v-w)}\,\mathrm{d}u\,\mathrm{d}w\,\mathrm{d}z\,\mathrm{d}v\\ &= \frac{\delta_i^{(\mathbb{Q})}}{\alpha-\theta}\int_{v=0}^\infty f_i^{(\mathbb{Q})}(v)\int_{z=0}^\infty h_{ij}^{(\mathbb{Q})}(z)\int_{w=0}^\infty e^{-\delta_i^{(\mathbb{Q})}w}\left(e^{-\alpha v-\theta(z-w)}-e^{-\alpha(z+v-w)}\right)\mathrm{d}w\,\mathrm{d}z\,\mathrm{d}v\\ &= \frac{\delta_i^{(\mathbb{Q})}}{\alpha-\theta}\int_{v=0}^\infty e^{-\alpha v}f_i^{(\mathbb{Q})}(v)\,\mathrm{d}v\int_{z=0}^\infty h_{ij}^{(\mathbb{Q})}(z)\left(\frac{e^{-\theta z}}{\delta_i^{(\mathbb{Q})}-\theta}-\frac{e^{-\alpha z}}{\delta_i^{(\mathbb{Q})}-\alpha}\right)\mathrm{d}z\\ &= \frac{\delta_i^{(\mathbb{Q})}}{\alpha-\theta}\,v_i^{(\mathbb{Q})}(\alpha)\left(\frac{\lambda_{ij}^{(\mathbb{Q})}(\theta)}{\delta_i^{(\mathbb{Q})}-\theta}-\frac{\lambda_{ij}^{(\mathbb{Q})}(\alpha)}{\delta_i^{(\mathbb{Q})}-\alpha}\right),\end{aligned}$$

where

$$\lambda_{ij}^{(\mathbb{Q})}(\theta) := \mathbb{E}_{\mathbb{Q}}\left(e^{-\theta L_{ij}}\right) = \frac{\lambda_{ij}(\theta-\theta^\star)}{\lambda_{ij}(-\theta^\star)}.$$

We calculate $s_{ijk}^{(3)}(\alpha,\theta)$ in a similar fashion. In addition to the usual rearrangement of the integrals, we perform the change-of-variable $y = x - u + v$, so as to obtain

$$\begin{aligned}
&s_{ijk}^{(3)}(\alpha,\theta)\\
&= \int_{u=0}^{\infty} e^{-\alpha u} \int_{v=0}^{u} f_i^{(\mathbb{Q})}(v) \int_{x=u-v}^{\infty} \delta_i^{(\mathbb{Q})} e^{-\delta_i^{(\mathbb{Q})}(x-u+v)} \int_{z=0}^{x} h_{ik}^{(\mathbb{Q})}(z)\, r_{kj}(\theta, x-z)\,\mathrm{d}z\,\mathrm{d}x\,\mathrm{d}v\,\mathrm{d}u\\
&= \delta_i^{(\mathbb{Q})} \int_{z=0}^{\infty} h_{ik}^{(\mathbb{Q})}(z) \int_{x=z}^{\infty} e^{-\delta_i^{(\mathbb{Q})}x} r_{jk}(\theta, x-z) \int_{v=0}^{\infty} e^{-\delta_i^{(\mathbb{Q})}v} f_i^{(\mathbb{Q})}(v) \int_{u=v}^{v+x} e^{(\delta_i^{(\mathbb{Q})}-\alpha)u}\,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}x\,\mathrm{d}z\\
&= \frac{\delta_i^{(\mathbb{Q})}}{\alpha-\delta_i^{(\mathbb{Q})}} \int_{z=0}^{\infty} h_{ik}^{(\mathbb{Q})}(z) \int_{x=z}^{\infty} \left(e^{-\delta_i^{(\mathbb{Q})}x} - e^{-\alpha x}\right) r_{jk}(\theta, x-z) \int_{v=0}^{\infty} e^{-\alpha v} f_i^{(\mathbb{Q})}(v)\,\mathrm{d}v\,\mathrm{d}x\,\mathrm{d}z\\
&= \frac{\delta_i^{(\mathbb{Q})}}{\alpha-\delta_i^{(\mathbb{Q})}}\, v_i^{(\mathbb{Q})}(\alpha) \int_{z=0}^{\infty} h_{ik}^{(\mathbb{Q})}(z) \int_{x=z}^{\infty} \left(e^{-\delta_i^{(\mathbb{Q})}x} - e^{-\alpha x}\right) r_{jk}(\theta, x-z)\,\mathrm{d}x\,\mathrm{d}z\\
&= \frac{\delta_i^{(\mathbb{Q})}}{\alpha-\delta_i^{(\mathbb{Q})}}\, v_i^{(\mathbb{Q})}(\alpha) \int_{z=0}^{\infty} h_{ik}^{(\mathbb{Q})}(z) \int_{x=0}^{\infty} \left(e^{-\delta_i^{(\mathbb{Q})}(x+z)} - e^{-\alpha(x+z)}\right) r_{kj}(\theta, x)\,\mathrm{d}x\\
&= \frac{\delta_i^{(\mathbb{Q})}}{\alpha-\delta_i^{(\mathbb{Q})}}\, v_i^{(\mathbb{Q})}(\alpha) \left(\lambda_{ik}^{(\mathbb{Q})}(\delta_i^{(\mathbb{Q})})\, s_{kj}(\delta_i^{(\mathbb{Q})},\theta) - \lambda_{ik}^{(\mathbb{Q})}(\alpha)\, s_{kj}(\alpha,\theta)\right).
\end{aligned}$$

We have thus managed to express the entries of the vector

$$\boldsymbol{s}_j(\alpha,\theta) \equiv (s_{1j}(\alpha,\theta),\ldots,s_{dj}(\alpha,\theta))^\top$$

in themselves. We proceed by writing the resulting linear equations in vector/matrix-form. To this end, note that under $\mathbb{Q}$ the matrix $M(\alpha)$ has entries

$$m_{ij}(\alpha) := q_{ij}^{(\mathbb{Q})}\, \lambda_{ij}^{(\mathbb{Q})}(\alpha) + \varphi_i^{(\mathbb{Q})}(\alpha)\, \mathbb{1}_{\{i=j\}}.$$

For the remainder of this section, the notation $M(\alpha)$ implies that we are working under the measure $\mathbb{Q}$. Also, with

$$\boldsymbol{b}_j(\alpha,\theta) \equiv (b_{1j}(\alpha,\theta),\ldots,b_{dj}(\alpha,\theta))^\top,$$

we define

$$\begin{aligned}
b_{ij}(\alpha,\theta) := &\sum_{k\neq i} q_{ik}^{(\mathbb{Q})}\lambda_{ik}^{(\mathbb{Q})}(\delta_i^{(\mathbb{Q})})\, s_{kj}(\delta_i^{(\mathbb{Q})},\theta) + \mathbb{1}_{\{i=j\}}\left(\varphi_i^{(\mathbb{Q})}(\alpha) - q_i^{(\mathbb{Q})}\right)\frac{v_j^{(\mathbb{Q})}(\alpha) - v_j^{(\mathbb{Q})}(\theta)}{\theta-\alpha} +\\
&\mathbb{1}_{\{i\neq j\}} q_{ij}^{(\mathbb{Q})}\, v_j^{(\mathbb{Q})}(\theta)\, \frac{\alpha-\delta_i^{(\mathbb{Q})}}{\alpha-\theta}\left(\frac{\lambda_{ij}^{(\mathbb{Q})}(\theta)}{\delta_i^{(\mathbb{Q})}-\theta} - \frac{\lambda_{ij}^{(\mathbb{Q})}(\alpha)}{\delta_i^{(\mathbb{Q})}-\alpha}\right).
\end{aligned}$$

It is useful to observe that if background state i corresponds to a non-decreasing subordinator, we have $\delta_i \to \infty$, so the expression simplifies to

$$b_{ij}(\alpha,\theta) := \mathbb{1}_{\{i=j\}}\left(\varphi_i^{(\mathbb{Q})}(\alpha) - q_i^{(\mathbb{Q})}\right)\frac{v_j^{(\mathbb{Q})}(\alpha) - v_j^{(\mathbb{Q})}(\theta)}{\theta-\alpha}.$$

To obtain a system of equations for $\boldsymbol{s}_j(\alpha,\theta)$, we combine (6.8) and (6.9) with the expressions for $s_i^{(1)}(\alpha,\theta)$, $s_{ij}^{(2)}(\alpha,\theta)$, and $s_{ijk}^{(3)}(\alpha,\theta)$. When multiplying (6.8) and (6.9) by

$$\frac{q_i^{(\mathbb{Q})}}{v_i^{(\mathbb{Q})}(\alpha)}\frac{\alpha-\delta_i^{(\mathbb{Q})}}{\delta_i^{(\mathbb{Q})}} = \varphi_i^{(\mathbb{Q})}(\alpha) - q_i^{(\mathbb{Q})}$$

for each $i = 1, ..., d$, it is seen that, for any given α and θ, we obtain the following system of linear equations.

Theorem 6.4.1. *For any $\alpha > 0$ and $\theta > 0$, and for $j = 1, \ldots, d$,*

$$M(\alpha)\,\boldsymbol{s}_j(\alpha,\theta) = \boldsymbol{b}_j(\alpha,\theta).$$

We would be able to determine the vector $\boldsymbol{s}_j(\alpha,\theta)$ from Theorem 6.4.1, were it not for the fact that for $i \notin S^{\uparrow}, k \in \{1, ..., d\}$, the quantities $s_{kj}(\delta_i,\theta)$ appearing in $b_{ij}(\alpha,\theta)$ are unknown. Defining

$$\omega_{ij} \equiv \omega_{ij}(\theta) := \sum_{k\neq i} q_{ik}^{(\mathbb{Q})}\lambda_{ik}^{(\mathbb{Q})}(\delta_i^{(\mathbb{Q})})\, s_{kj}(\delta_i^{(\mathbb{Q})},\theta),$$

we now turn our attention towards finding $(\omega_{ij})_{i\notin S^{\uparrow}}$.

Note that, using the linear equations given in Theorem 6.4.1, one may express the vector $\boldsymbol{s}_j(\alpha,\theta)$ by relying on Cramer's rule. More concretely, with the matrix $M_{\boldsymbol{b}_j,i}(\alpha,\theta)$ denoting the matrix $M(\alpha)$ in which the i-th column is replaced by the vector $\boldsymbol{b}_j(\alpha)$, we have that

$$s_{ij}(\alpha,\theta) = \frac{\det M_{\boldsymbol{b}_j,i}(\alpha,\theta)}{\det M(\alpha)}. \tag{6.10}$$

Since $s_{ij}(\alpha,\theta)$ is finite, any zero of the denominator should be a zero of the numerator. According to Proposition 6.2.2, $\det M(\alpha) = 0$ has $d^{\circ} := d - |S^{\uparrow}|$ zeroes in the right half of the complex plane (recalling that the asymptotic drift is positive under $\mathbb{Q}$). For ease of exposition, we let these zeroes have multiplicity 1 (and we call them, say, $\alpha_1, \ldots, \alpha_{d^{\circ}}$). When this multiplicity property does not hold, a reasoning similar to the one below still applies, but one needs to resort to the concept of Jordan chains; we do not discuss this procedure in detail, but instead refer to the in-depth treatment in [43].

Having distinct zeroes guarantees that we have d° equations to identify the ω_{ij}. That is, for $i = 1, \ldots, d$ and $j = 1, \ldots, d^{\circ}$,

$$\det M_{\boldsymbol{b}_j,i}(\alpha_j,\theta) = 0; \tag{6.11}$$

in other words, the zeroes of $\det M$ (in the right half of the complex plane, that is) are also zeroes of $\det M_{\boldsymbol{b}_j,i}$, for each $i = 1, \ldots, d$.

By precisely the same argument as the one given in Section 5.3.3, Equation (6.11) provides the same information for any $i = 1, ..., d$, so it suffices to consider just the equation $M_{\boldsymbol{b}_j,1}(\alpha_j,\theta) = 0$. With $\bar{M}_{ik}(\alpha)$ representing the $(d-1)\times(d-1)$ matrix which results after deleting the i-th column and the k-th row from $M(\alpha)$, this equation can be rewritten as

$$\sum_{i\notin S^{\uparrow}}(-1)^{i-1}\left(\omega_{ij} + \mathbb{1}_{\{i\neq j\}}q_{ij}^{(\mathbb{Q})}\,v_j^{(\mathbb{Q})}(\theta)\,\frac{\alpha_j-\delta_i^{(\mathbb{Q})}}{\alpha_j-\theta}\left(\frac{\lambda_{ij}^{(\mathbb{Q})}(\theta)}{\delta_i^{(\mathbb{Q})}-\theta} - \frac{\lambda_{ij}^{(\mathbb{Q})}(\alpha_j)}{\delta_i^{(\mathbb{Q})}-\alpha_j}\right)\right)\det\bar{M}_{i1}(\alpha_j)$$
$$+ (-1)^{j-1}\mathbb{1}_{\{i=j\}}\left(\varphi_i^{(\mathbb{Q})}(\alpha_j) - q_i^{(\mathbb{Q})}\right)\frac{v_j^{(\mathbb{Q})}(\alpha_j) - v_j^{(\mathbb{Q})}(\theta)}{\theta-\alpha_j}\det\bar{M}_{j1}(\alpha_j) = 0.$$

We thus obtain d° equations (one for each α_j) that are linear in the unknowns $\omega_{1j}, \ldots, \omega_{d^\circ j}$, which can be dealt with in the standard manner, thus yielding a solution for the ω_{ij}. This procedure can be repeated for each eventual state j. Now that the quantities ω_{ij} (for $i \notin S^\uparrow$) are known, Equation (6.10) expresses $s_{ij}(\alpha, \theta)$ in terms of known quantities.

6.5 Exact tail asymptotics

As pointed out at the end of Section 3, in order to evaluate $r_{ij}(\theta^\star)$, we are interested in $\lim_{\alpha \downarrow 0} \alpha s_{ij}(\alpha, \theta^\star)$. The purpose of this section is to find an explicit expression for this limit, and consequently determining the exact asymptotics of $p_i(u)$ by Equation (6.6). We rely on the fact that from Section 6.4 we know how $s_{ij}(\alpha, \theta^\star)$ can be evaluated.

It is first observed that as $\alpha \downarrow 0$, the denominator $\partial(\alpha) := \det M(\alpha)$ of (6.10) tends to 0, so that L'Hopital's rule gives

$$\lim_{\alpha \downarrow 0} \frac{\alpha}{\partial(\alpha)} = \frac{1}{\partial'(0)}.$$

This explains why we first study the behavior of $\partial(\alpha)$ as $\alpha \downarrow 0$. To this end, we write, taking entry-wise Taylor expansions at $\alpha = 0$,

$$M(\alpha) = Q + \alpha Z + O(\alpha^2),$$

where $Z = (z_{ij})_{i,j=1}^d$ is given by

$$z_{ij} := \varphi_i^{(\mathbb{Q})\prime}(0) \mathbb{1}_{\{i=j\}} + q_{ij}^{(\mathbb{Q})} \lambda_{ij}^{(\mathbb{Q})\prime}(0) \mathbb{1}_{\{i \neq j\}}.$$

Hence, we can express $\partial(\alpha)$ as the determinant of a sum. In order to work with determinants of sums, we have the following lemma. Let C_k^E be the matrix consisting of all columns of C, but with its k-th column replaced with the k-th column of E.

Lemma 6.5.1. *If C and E are $d \times d$ matrices, then, as $\varepsilon \downarrow 0$,*

$$\det(C + \varepsilon E) = \det(C) + \varepsilon \sum_{k=1}^d \det(C_k^E) + O(\varepsilon^2).$$

Proof. Recall that $\det(C + \varepsilon E)$ is the sum of 2^d determinants; one for each possible matrix in which each of the columns equals the corresponding column of C or εE. Using this rule, one can write $\det(C + \varepsilon E)$ as a polynomial in ε of degree d where the coefficients of the ε^ℓ are sums of $\binom{d}{\ell}$ determinants of the above type. The result follows by isolating the terms that do not depend on ε and those that are linear in ε, and by in addition aggregating all terms that correspond to $\varepsilon^2, \ldots, \varepsilon^d$. □

The idea is now to set $C = Q + \alpha Z$, $\varepsilon = \alpha^2$ and E any finite matrix in the above lemma. It immediately follows that $\partial(\alpha) = \det(Q + \alpha Z) + O(\alpha^2)$. We then use the lemma a second time, but now with $C = Q$, $\varepsilon = \alpha$ and $E = Z$, so as to obtain

$$\partial(\alpha) = \det Q + \alpha \sum_{k=1}^d \det(Q_k^Z) + O(\alpha^2) = \alpha \sum_{k=1}^d \det(Q_k^Z) + O(\alpha^2).$$

Hence, we obtain

$$\lim_{\alpha\downarrow 0} \frac{\alpha}{\partial(\alpha)} = \frac{1}{\partial'(0)} = \frac{1}{\sum_{k=1}^{d} \det(Q_k^Z)}.$$

We thus conclude that

$$r_{ij}(\theta) = \lim_{\alpha\downarrow 0} \alpha\, s_{ij}(\alpha,\theta) = \lim_{\alpha\downarrow 0} \alpha \frac{\det M_{\boldsymbol{b}_j,i}(\alpha,\theta)}{\partial(\alpha)} = \frac{\det M_{\boldsymbol{b}_j,i}(0,\theta)}{\sum_{k=1}^{d} \det(Q_k^Z)} = \frac{\det Q_{\boldsymbol{b}_j,i}(\theta)}{\sum_{k=1}^{d} \det(Q_k^Z)}.$$

Combining this with (6.6), we obtain the main result of the chapter: the exact asymptotics for $p_i(u)$. As mentioned, this result can be considered the MAP-counterpart of the classical Cramér-Lundberg result.

Theorem 6.5.1. *For any initial state $i \in \{1, ..., d\}$,*

$$\lim_{u\to\infty} p_i(u)\, e^{\theta^\star u} = \alpha_i := \frac{y_i}{\sum_{k=1}^{d} \det(Q_k^Z)} \sum_{j=1}^{d} \frac{\det Q_{\boldsymbol{b}_j,i}(\theta^\star)}{y_j}\, v_j(-\theta^\star).$$

The above theorem provides a possible approximation for the ruin probability $p_i(u)$ in the regime that u is large. Concretely, we propose the approximation

$$p_i(u) \approx \bar{p}_i(u) := \alpha_i\, e^{-\theta^\star u}. \tag{6.12}$$

In Section 6.7 the accuracy of (6.12) is assessed.

6.6 Efficient simulation

In this section we point out how to efficiently estimate $p_i(u)$ by simulation, with emphasis on the regime that u is large. The main idea is to rely on importance sampling, using a generalized version of the celebrated Siegmund algorithm. More specifically, we propose to run independent simulations of the MAP under $\mathbb{Q}$, and subsequently average the values of

$$\gamma := \frac{y_i}{y_{K_N}}\, v_{K_N}(-\theta^\star)\, e^{-\theta^\star S_N}$$

that are sampled in each of the runs. By Theorem 6.3.1 this results in an unbiased estimator. The main objective of this section is to prove that this estimator has bounded relative error. Its efficiency gain (relative to direct simulation, that is) is demonstrated in Section 6.7 through a series of numerical experiments.

A pseudo-code corresponding to a single run of this generalized Siegmund algorithm is given by Algorithm 1. Here s records the current value of the embedded MAP and j the current background state. Also, the function 'sample(X)' generates and returns a sample of the random variable X, independent of everything what has been sampled before. Finally, the function 'sampleNextState($j, Q^{(\mathbb{Q})}$)' returns a new state of the background chain sampled under $\mathbb{Q}$, when the current state is j.

The while loop in Algorithm 1 updates the value of the MAP at maxima between two successive background transition epochs. Lines 3 through 7 respectively correspond to sampling the

Data: Initial state i of background process and target level u;
distributions of V_j, W_j, and L_{jk} under $\mathbb{Q}$;
transition rate matrix $Q = (q_{ij})_{i,j=1}^{d}$ under $\mathbb{Q}$;
eigenvalue $\theta^\star$ and corresponding eigenvector $\boldsymbol{y}$.
Result: Unbiased sample of $p_i(u)$.

```
Initialization: s ← sample(V_i^(Q)) and j ← i.
while s < u do
    w ← sample(W_j^(Q));
    j ← sampleNextState(j, Q^(Q));
    ℓ ← sample(L_jk^(Q));
    v ← sample(V_k^(Q));
    s ← s − w + ℓ + v and j ← k.
end while
return (y_i/y_j) × v_j(−θ*) × e^(−θ* s).
```

Algorithm 1: A single run in the generalized Siegmund algorithm

decrease of the MAP before the next background transition, sampling the next background state, sampling the jump at the transition epoch and sampling the maximum between the current and next background transitions.

An important performance measure of algorithms estimating small probabilities is their relative error, defined by the standard deviation of the estimate divided by the estimated probability. Not only does $\gamma \equiv \gamma(u)$ yield an unbiased estimator of $p_i(u)$, the next theorem entails that the relative error is bounded in u. We refer to this property as *bounded relative error* [9, Chapter VI.1].

Theorem 6.6.1. *A sample of $p_i(u)$ as returned by Algorithm 1 has bounded relative error.*

Proof. The proof is closely related to its counterpart for the random walk case [9, Chapter VI.2]. Denote, in line with the notation that we have previously used, by $R(u)$ and $K_{N(u)}$ the overshoot and the background state, respectively, at the time that level u is crossed. Now consider the process $(R(u), K_{N(u)})_{u\geqslant 0}$, conditional on $K_0 = i$. Above, we have computed the transform $r_{ij}(\theta)$, by which we uniquely characterized the limiting distribution of $(R(u), K_{N(u)})$ as $u \to \infty$; we let (R, K) be distributed as the corresponding limiting random vector.

As a result of the above, the ruin probability (which equals the mean of $\gamma(u)$, as defined above) satisfies the following asymptotics, as $u \to \infty$:

$$e^{\theta^\star u}\,\mathbb{E}_{\mathbb{Q}}(\gamma(u)) = e^{\theta^\star u}\,p_i(u) = \mathbb{E}_{\mathbb{Q}}\left(\frac{y_i}{y_{K_{N(u)}}}\,v_{K_{N(u)}}(-\theta^\star)\,e^{-\theta^\star(S_{N(u)}-u)}\right)$$
$$\to \mathbb{E}_{\mathbb{Q}}\left(\frac{y_i}{y_K}\,v_K(-\theta^\star)\,e^{-\theta^\star R}\right) = \sum_{j=1}^{d} r_{ij}(\theta^\star)\,\frac{y_i}{y_j}\,v_j(-\theta^\star) =: C_1.$$

We have bounded relative error as the second moment of $\gamma(u)$ vanishes at essentially the same

rate as the square of $p_i(u)$; to see this, observe that

$$e^{2\theta^\star u}\,\mathbb{E}_{\mathbb{Q}}\left(\gamma^2(u)\right) = e^{2\theta^\star u}\,\mathbb{E}_{\mathbb{Q}}\left(\frac{y_i^2}{y_{K_{N(u)}}^2}\,v_{K_{N(u)}}^2(-\theta^\star)\,e^{-2\theta^\star S_{N(u)}}\right)$$
$$\to \mathbb{E}_{\mathbb{Q}}\left(\frac{y_i^2}{y_K^2}\,v_K^2(-\theta^\star)\,e^{-2\theta^\star R}\right) = \sum_{j=1}^{d} r_{ij}(2\theta^\star)\left(\frac{y_i}{y_j}\,v_j(-\theta^\star)\right)^2 =: C_2,$$

where it is noted that $C_2 > C_1^2$ due to Jensen's inequality. The variance of a single observation $\gamma(u)$ in our generalized Siegmund algorithm thus satisfies

$$e^{2\theta^\star u}\,\mathrm{Var}_{\mathbb{Q}}\left(\frac{y_i}{y_{K_{N(u)}}}\,v_{K_{N(u)}}(-\theta^\star)\,e^{-\theta^\star S_{N(u)}}\right) \to C_2 - C_1^2.$$

It now directly follows that for u large the relative error tends to

$$c := \frac{\sqrt{C_2 - C_1^2}}{C_1} \in (0,\infty). \tag{6.13}$$

We observe that apparently the relative error loses its dependence on u as u grows, and that it is bounded by a constant. □

6.7 Numerical experiments

In this section we numerically study the asymptotic behavior of $p(u)$, measuring in particular the efficiency of the generalized Siegmund algorithm compared to direct estimation. We in addition include an experiment that studies the impact of the background process on the ruin probability.

To obtain a direct estimation of the ruin probability $p(u)$, one may simulate the model at hand say n times under the original measure $\mathbb{P}$, and then determine the fraction of runs in which the process exceeds level u. This leads to an unbiased estimator, with the relative error equalling

$$\frac{\sqrt{\frac{1}{n}\,p(u)(1-p(u))}}{p(u)} = \sqrt{\frac{1-p(u)}{np(u)}}.$$

In order to achieve a relative error of at most ε, one thus requires roughly

$$n \approx \frac{1-p(u)}{\varepsilon^2\,p(u)}$$

runs. The particularly worrisome element in this quantity is the $p(u)$ in the denominator, being very small when u is large. Concretely, with $p(u)$ decaying roughly as $e^{-\theta^\star u}$ and $1-p(u)\approx 1$, we conclude that n blows up like $e^{\theta^\star u}$ as u grows.

The generalized Siegmund algorithm, in which the process is simulated under $\mathbb{Q}$, on the other hand has bounded relative error. With c given in (6.13), and again aiming for a relative error ε, this algorithm thus requires roughly $n \approx c/\varepsilon^2$ runs for large u. As a result, the generalized

Siegmund algorithm gives an accurate estimation for any u, in that for large u the number of runs required becomes independent of u. This in particular means that this number of runs does not blow up as $p(u) \to 0$.

In the remainder of this section we discuss experiments in which we apply our generalized Siegmund algorithm. It should be noted that executing this algorithm requires being able to sample random variables $(V_k)_{k\in\{1,\ldots,d\}}$, while only their respective Laplace-Stieltjes transforms are known (see Proposition 6.2.1). To this end we first apply numerical inversion to obtain a discretization of the distribution function of each of the V_k. In our numerics we have used the intensively tested and frequently cited algorithm that was developed in [5]; in the experiments reported in this section we use 10^3 mass points. With this approximate distribution function at our disposal, we can use the inverse distribution function method to sample a random variable distributed according to V_k. Three remarks are in place here.

- First observe that, for any k, the Laplace inversion has to be performed *just once* (say, in the pre-processing phase); once the approximate distribution function of V_k has been computed, the generalized Siegmund algorithm can be repeatedly executed until an estimate of sufficient precision has been produced.
- Secondly, the simulation of the embedded process under the original measure $\mathbb{P}$ has the same inherent issue that the V_k must be sampled. In other words, the need to perform Laplace inversion is not specific for the generalized Siegmund algorithm.
- The fact that we propose to use numerical Laplace inversion to run our generalized Siegmund algorithm, triggers the question why we do not simply numerically invert the transform of $p(u)$ (as can be obtained with the results from Chapter 5). However, the latter method is typically inferior to the generalized Siegmund algorithm, in particular in the current context where the ruin probability $p(u)$ is small, as a consequence of the fact that the Laplace inversion becomes increasingly inaccurate further along the tail.

We now describe the specific MAP we use in our experiments. It consists of two background states, the first and second corresponding respectively to a standard Brownian motion $X_1(\cdot)$ with drift $-\frac{1}{3}$, and a compound Poisson process $X_2(\cdot)$ with drift -1 and jumps of $\mathrm{Exp}(1)$ size arriving at rate $\frac{2}{3}$. Note that we constructed this instance such that $\mathbb{E}(X_1(1)) = \mathbb{E}(X_2(1)) = -\frac{1}{3}$, allowing for better comparison between the impacts of both processes on the ruin probability. To make our model as elementary as possible, our setup does not contain jumps at transition epochs; we stress however that adding those does not lead to any conceptual complications.

Experiment 1. In the first experiment we vary the value of the ruin level (or, in risk applications, the initial reserve) u, so as to provide empirical backing for the claims of Theorems 6.5.1 and 6.6.1. In addition, we obtain insight into the accuracy of the approximation (6.12).

We consider the instance $q_1 = q_2 = 1$, and we vary u (with steps of 10) from 10 to 80, see Table 6.1. Two approximations of the ruin probability are presented: the second column shows estimates of $p_1(u)$ that are generated by 10^4 runs of Algorithm 1, and the third column presents the approximation $\bar{p}_1(u) = \alpha_1\, e^{-\theta^\star u}$ of (6.12) (with $\alpha_1 \approx 0.6390$ and $\theta^\star \approx 0.4066$). The

u	$p_1(u)$ estimated by Algorithm 1	$\alpha_1 e^{-\theta^\star u}$	relative error per run under $\mathbb{Q}$	relative error per run under $\mathbb{P}$
10	$1.084 \cdot 10^{-2}$	$1.09537 \cdot 10^{-2}$	0.623	9.55
20	$1.862 \cdot 10^{-4}$	$1.87784 \cdot 10^{-4}$	0.630	73.3
30	$3.224 \cdot 10^{-6}$	$3.21924 \cdot 10^{-6}$	0.626	$5.57 \cdot 10^{2}$
40	$5.519 \cdot 10^{-8}$	$5.51886 \cdot 10^{-8}$	0.622	$4.26 \cdot 10^{3}$
50	$9.460 \cdot 10^{-10}$	$9.46117 \cdot 10^{-10}$	0.623	$3.25 \cdot 10^{4}$
60	$1.620 \cdot 10^{-11}$	$1.62196 \cdot 10^{-11}$	0.629	$2.48 \cdot 10^{5}$
70	$2.789 \cdot 10^{-13}$	$2.78058 \cdot 10^{-13}$	0.621	$1.89 \cdot 10^{6}$
80	$4.798 \cdot 10^{-15}$	$4.76685 \cdot 10^{-15}$	0.619	$1.44 \cdot 10^{7}$

Table 6.1: Ruin probabilities as a function of u, and the relative error per run under $\mathbb{Q}$ and $\mathbb{P}$.

approximations of both methods differ around 1%, even for small values of u. This indicates, for our specific MAP, fast convergence of the expression in Theorem 6.5.1.

In addition, Table 6.1 shows the average relative error of a single run under $\mathbb{Q}$ (Algorithm 1), based on the sample (fourth column). This is compared to the same error when one would estimate the ruin probability directly under $\mathbb{P}$ (fifth column). Where the relative error of Algorithm 1 is fairly constant, the same error under direct estimation shows exponential increase in u, as anticipated. If, say, one is interested in an estimate for $p_1(u)$ with relative error at most 5%, the instances $u = 10, 40, 80$ respectively require approximately $4 \cdot 10^4$, $7 \cdot 10^9$, and $8 \cdot 10^{16}$ runs. The number of runs required under $\mathbb{Q}$, on the other hand, is around 250 for any value of u.

Experiment 2. In the second experiment the background chain transition rates q_1 and q_2 are varied, and with them the proportion of time spent in each of the two background states. For each combination of these parameter values we run Algorithm 1 with $u = 40$ a total of 10^4 times. The output consists of the estimated ruin probability and the relative error per run based on the sample. The results are shown in Table 6.2.

As we can see, the ruin probability heavily depends on the transition rates of the background chain. The larger the proportion of time spent in the compound Poisson state (state 2), the larger the ruin probability, and the larger the proportion of time spent in the Brownian state (state 1), the smaller the ruin probability. It is also reassuring to see that, for this MAP, the (bounded) relative error per run is rather low. With 10^4 runs this results in a relative error in the order of $5 \cdot 10^{-3}$.

When $q_1 \gg q_2$, one expects that the net cumulative claim process effectively coincides with a compound Poisson process, which has a ruin probability that asymptotically decays as $\frac{2}{3}e^{-\frac{u}{3}}$. Substituting $u = 40$ gives $1.0797 \cdot 10^{-6}$, close to the values of $p_1(40)$ in the top rows. Conversely, when $q_1 \ll q_2$, the net cumulative claim process should be close to a Brownian motion, which has a ruin probability $e^{-\frac{2}{3}u}$. Substituting $u = 40$ gives $2.6231 \cdot 10^{-12}$, in line with the values of $p_1(40)$ in the bottom rows.

q_1	q_2	fraction of time in state 1	$p_1(40)$	relative error per run
10^4	1	1/10001	$1.078 \cdot 10^{-6}$	0.520
10^3	1	1/1001	$1.072 \cdot 10^{-6}$	0.519
10^2	1	1/101	$1.026 \cdot 10^{-6}$	0.527
10	1	1/11	$7.043 \cdot 10^{-7}$	0.536
5	1	1/6	$4.873 \cdot 10^{-7}$	0.543
2	1	1/3	$1.833 \cdot 10^{-7}$	0.577
1	1	1/2	$5.519 \cdot 10^{-8}$	0.622
1	2	2/3	$1.075 \cdot 10^{-8}$	0.594
1	5	5/6	$8.559 \cdot 10^{-10}$	0.584
1	10	10/11	$1.374 \cdot 10^{-10}$	0.575
1	10^2	100/101	$5.005 \cdot 10^{-12}$	0.465
1	10^3	1000/1001	$2.821 \cdot 10^{-12}$	0.408
1	10^4	10000/10001	$2.629 \cdot 10^{-12}$	0.408

Table 6.2: Ruin probabilities as a function of the fraction of time spent in each of the two background states.

6.8 Discussion and concluding remarks

In this chapter we have considered a ruin model driven by a light-tailed spectrally-positive Markov additive process. We first identified the exact asymptotics of the ruin probability, and then devised an efficient simulation algorithm. Since previous literature restricts to MAPs with specific underlying Lévy processes, our results narrow the gap in the understanding of these asymptotics for arbitrary MAPs. Nevertheless, various directions for further research are possible, of which we list a few.

When the net cumulative claim process is a spectrally-negative MAP (i.e., having jumps in the downward direction only), the all-time maximum has a phase-type distribution (see Section 5.4.1). This means that in that case the exact asymptotics of the ruin probability can be dealt with relatively easily. Considerably more challenging, however, is the case of a MAP with jumps in both directions; this would extend the result for the Lévy case that was established in [17]. A possible first step in this direction could be to assume that the jumps in one of the directions are of phase-type (cf. [64] for instance).

A second interesting extension would concern the inclusion of heavy-tailed jump size distributions, applying to jumps of the Lévy processes and/or jumps at transition epochs. Such distributions deny the existence of a positive solution $\theta^\star$ to the Lundberg equation, thus invalidating our change-of-measure approach. Instead, we suggest to build upon the results of [47], in which the 'principle of a single big jump' is exploited.

In this chapter we used an embedded process to simplify the analysis while maintaining all

information on the ruin probability over level u. For this embedded process, we in particular identified the transform $s_{ij}(\alpha, \theta)$ of the overshoot over level u. However, it is clear that the overshoot over level u of the embedded process can be substantially different from its counterpart for the original MAP. This leaves open the problem of characterizing the distribution of the overshoot for our MAP. Extensions in this direction would be valuable additions to the literature on first passage times.

Publications

The Chapters 2 up to 6 are based, in their respective order, on the papers [115, 116, 123, 124, 125].

Part I

O. Boxma and J. Dorsman proposed the idea for [115], and Z. Scully initiated the work on [116]. The mathematical analysis and writing tasks of both papers were mainly shared by Z. Scully and L. van Kreveld. For [116] specifically, J. Dorsman, O. Boxma and A. Wierman continuously provided helpful input in terms of analysis and editing.

Chapter 2: [115] Scully, Z., van Kreveld, L., Boxma, O., Dorsman, J., and Wierman, A. (2020c). Characterizing policies with optimal response time tails under heavy-tailed job sizes. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(2):1–32.

Chapter 3: [116] Scully, Z. and van Kreveld, L. (2021). When does the Gittins policy have asymptotically optimal response time tail? *arXiv preprint arXiv:2110.06326*.

Part II

L. van Kreveld took the lead in the mathematical analysis and writing of [122] and its more general successor [123]. Aside from that, all authors contributed equally to both papers.

[122] van Kreveld, L., Boxma, O., Dorsman, J., and Mandjes, M. (2020). Scaling analysis of an extended machine-repair model. In *Proceedings of the 13th EAI International Conference on Performance Evaluation Methodologies and Tools*, pages 172–179.

Chapter 4: [123] van Kreveld, L., Boxma, O., Dorsman, J., and Mandjes, M. (2021). Scaling limits for closed product-form queueing networks. *Performance Evaluation*, 151:102220.

Part III

The majority of the analysis and writing of [124, 125] was performed by L. van Kreveld and M. Mandjes. The role of J. Dorsman was mainly advisory and editorial.

Chapter 5: [124] van Kreveld, L., Mandjes, M., and Dorsman, J. (2022a). Extreme value analysis for a Markov additive process driven by a nonirreducible background chain. *Stochastic Systems*, 12(3):293–317.

Chapter 6: [125] van Kreveld, L., Mandjes, M., and Dorsman, J. (2022b). Cramér-Lundberg asymptotics for spectrally positive Markov additive processes. *Submitted*.

Bibliography

[1] Aalto, S., Ayesta, U., Borst, S., Misra, V., and Núñez-Queija, R. (2007). Beyond processor sharing. *ACM SIGMETRICS Performance Evaluation Review*, 34(4):36–43.

[2] Aalto, S., Ayesta, U., and Righter, R. (2009). On the Gittins index in the M/G/1 queue. *Queueing Systems*, 63:437–458.

[3] Aalto, S., Ayesta, U., and Righter, R. (2011). Properties of the Gittins index with application to optimal scheduling. *Probability in the Engineering and Informational Sciences*, 25(3):269–288.

[4] Abate, J. and Whitt, W. (1997). Asymptotics for M/G/1 low-priority waiting-time tail probabilities. *Queueing Systems*, 25:173–233.

[5] Abate, J. and Whitt, W. (2006). A unified framework for numerically inverting Laplace transforms. *INFORMS Journal on Computing*, 18(4):408–421.

[6] Altman, E., Avrachenkov, K., and Ayesta, U. (2006). A survey on discriminatory processor sharing. *Queueing Systems*, 53:53–63.

[7] Asmussen, S. (2003). *Applied Probability and Queues*, volume 2. Springer.

[8] Asmussen, S. and Albrecher, H. (2010). *Ruin Probabilities*, volume 14. World Scientific.

[9] Asmussen, S. and Glynn, P. (2007). *Stochastic Simulation: Algorithms and Analysis*, volume 57. Springer.

[10] Asmussen, S. and Ivanovs, J. (2018). A factorization of a Lévy process over a phase-type horizon. *Stochastic Models*, 34(4):397–408.

[11] Asmussen, S. and Kella, O. (2000). A multi-dimensional martingale for Markov additive processes and its applications. *Advances in Applied Probability*, 32(2):376–393.

[12] Balsamo, S. and de Nitto Personè, V. (1994). A survey of product form queueing networks with blocking and their equivalences. *Annals of Operations Research*, 48:31–61.

[13] Bansal, N., Kamphorst, B., and Zwart, B. (2018). Achievable performance of blind policies in heavy traffic. *Mathematics of Operations Research*, 43(3):949–964.

[14] Baskett, F., Chandy, K., Muntz, R., and Palacios, F. (1975). Open, closed, and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22:248–260.

[15] Becchetti, L. and Leonardi, S. (2004). Nonclairvoyant scheduling to minimize the total flow time on single and parallel machines. *Journal of the ACM*, 51(4):517–539.

[16] Bertoin, J. (1996). *Lévy Processes*. Cambridge University Press.

[17] Bertoin, J. and Doney, R. (1994). Cramér's estimate for Lévy processes. *Statistics & Probability Letters*, 21(5):363–365.

[18] Bingham, N., Goldie, C., and Teugels, J. (1987). *Regular Variation*. Cambridge University Press.

[19] Birman, A. and Kogan, Y. (1996). Error bounds for asymptotic approximations of the partition function. *Queueing Systems*, 23:217–234.

[20] Borst, S., Boxma, O., Núñez-Queija, R., and Zwart, B. (2003). The impact of the service discipline on delay asymptotics. *Performance Evaluation*, 54(2):175–206.

[21] Borst, S., Mandelbaum, A., and Reiman, M. (2004). Dimensioning large call centers. *Operations Research*, 52(1):17–34.

[22] Borst, S., Núñez-Queija, R., and Zwart, B. (2006). Sojourn-time asymptotics in processor-sharing queues. *Queueing Systems*, 53:31–51.

[23] Boucherie, R. (1998). Norton's equivalent for queueing networks comprised of quasire-versible components linked by state-dependent routing. *Performance Evaluation*, 32(2):83–99.

[24] Boucherie, R. and van Dijk, N. (1991). Product forms for queueing networks with state-dependent multiple job transitions. *Advances in Applied Probability*, 23(1):152–187.

[25] Boucherie, R. and van Dijk, N. (2010). *Queueing Networks: A Fundamental Approach*. Springer.

[26] Boxma, O. and Denisov, D. (2011). Sojourn time tails in the single server queue with heavy-tailed service times. *Queueing Systems*, 69(2):101–119.

[27] Boxma, O. and Mandjes, M. (2021). Affine storage and insurance risk models. *Mathematics of Operations Research*, 46(4):1282–1302.

[28] Boxma, O. and Mandjes, M. (2023). *The Cramér-Lundberg Model and Its Variants*. To appear.

[29] Boxma, O. and Zwart, B. (2007). Tails in scheduling. *ACM SIGMETRICS Performance Evaluation Review*, 34(4):13–20.

[30] Breuer, L. (2008). First passage times for Markov additive processes with positive jumps of phase type. *Journal of Applied Probability*, 45(3):779–799.

[31] Casale, G. (2011). A generalized method of moments for closed queueing networks. *Performance Evaluation*, 68(2):180–200.

[32] Chao, X., Miyazawa, M., Serfozo, R., and Takada, H. (1998). Markov network processes with product form stationary distributions. *Queueing Systems*, 28:377–401.

[33] Çinlar, E. (1972). Markov additive processes. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 24(2):85–93.

[34] Cline, D. (1994). Intermediate regular and Π variation. *Proceedings of the London Mathematical Society*, 3(3):594–616.

[35] Cohen, J. (1973). Some results on regular variation for distributions in queueing and fluctuation theory. *Journal of Applied Probability*, 10(2):343–353.

[36] Crovella, M. and Bestavros, A. (1997). Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846.

[37] Dębicki, K. and Mandjes, M. (2015). *Queues and Lévy Fluctuation Theory*. Springer.

[38] Delsing, G. and Mandjes, M. (2021). A transient Cramér–Lundberg model with applications to credit risk. *Journal of Applied Probability*, 58(3):721–745.

[39] den Hollander, F. (2000). *Large Deviations*, volume 14. American Mathematical Society.

[40] Dieker, A. and Mandjes, M. (2011). Extremes of Markov-additive processes with one-sided jumps, with queueing applications. *Methodology and Computing in Applied Probability*, 13(2):221–267.

[41] Dorsman, J. (2015). *Layered Queueing Networks: Performance Modelling, Analysis and Optimisation*. PhD thesis, Technische Universiteit Eindhoven.

[42] Dumitriu, I., Tetali, P., and Winkler, P. (2003). On playing golf with two balls. *SIAM Journal on Discrete Mathematics*, 16(4):604–615.

[43] D'Auria, B., Ivanovs, J., Kella, O., and Mandjes, M. (2010). First passage of a Markov additive process and generalized Jordan chains. *Journal of Applied Probability*, 47(4):1048–1057.

[44] Erlang, A. (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineer's Journal*, 10:189–197.

[45] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley.

[46] Fischer, H. (2011). *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer.

[47] Foss, S., Konstantopoulos, T., and Zachary, S. (2007). Discrete and continuous time modulated random walks with heavy-tailed increments. *Journal of Theoretical Probability*, 20(3):581–612.

[48] Foss, S., Korshunov, D., and Zachary, S. (2013). *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer.

[49] Gelenbe, E. (1991). Product-form queueing networks with negative and positive customers. *Journal of Applied Probability*, 28(3):656–663.

[50] George, D. and Xia, C. (2011). Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *European Journal of Operational Research*, 211(1):198–207.

[51] George, D., Xia, C., and Squillante, M. (2012). Exact-order asymptotic analysis for closed queueing networks. *Journal of Applied Probability*, 49(2):503–520.

[52] Gittins, J. (1989). *Multi-Armed Bandit Allocation Indices*. Wiley-Interscience Series in Systems and Optimization. Wiley.

[53] Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-armed Bandit Allocation Indices*. John Wiley & Sons.

[54] Grosof, I., Yang, K., Scully, Z., and Harchol-Balter, M. (2021). Nudge: stochastically improving upon FCFS. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 5(2):1–29.

[55] Guillemin, F., Robert, P., and Zwart, B. (2004). Tail asymptotics for processor-sharing queues. *Advances in Applied Probability*, 36(2):525–543.

[56] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–587.

[57] Harchol-Balter, M. (2013). *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press.

[58] Harvey, C. and Hills, C. (1979). Determining grades of service in a network. *Ninth International Teletraffic Conference, Torremolinos, Spain*.

[59] Ivanovs, J. (2011). *One-sided Markov Additive Processes and Related Exit Problems*. PhD thesis, University of Amsterdam.

[60] Ivanovs, J. (2017). Splitting and time reversal for Markov additive processes. *Stochastic Processes and their Applications*, 127(8):2699–2724.

[61] Ivanovs, J., Boxma, O., and Mandjes, M. (2010). Singularities of the matrix exponent of a Markov additive process with one-sided jumps. *Stochastic Processes and their Applications*, 120(9):1776–1794.

[62] Jackson, J. (1957). Networks of waiting lines. *Operations Research*, 5(4):518–521.

[63] Jackson, R. (1954). Queueing systems with phase type service. *Journal of the Operational Research Society*, 5(4):109–120.

[64] Jacobsen, M. (2005). The time to ruin for a class of Markov additive risk process with two-sided jumps. *Advances in Applied Probability*, 37(4):963–992.

[65] Jasiulewicz, H. (2001). Probability of ruin with variable premium rate in a Markovian environment. *Insurance: Mathematics and Economics*, 29(2):291–296.

[66] Jelenković, P., Kondev, J., Mohapatra, L., and Momčilović, P. (2021). A probabilistic approach to growth networks. *Operations Research.* To appear.

[67] Jelenković, P. and Momčilović, P. (2022). Scaling regimes of growth networks. *Queueing Systems*, 100(3):313–315.

[68] Johnson, W. (2002). The curious history of Faà di Bruno's formula. *The American Mathematical Monthly*, 109(3):217–234.

[69] Kalyanasundaram, B. and Pruhs, K. (1997). Minimizing flow time nonclairvoyantly. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*, pages 345–352. IEEE.

[70] Kelly, F. (1976). Networks of queues. *Advances in Applied Probability*, 8(2):416–432.

[71] Kelly, F. (2011). *Reversibility and Stochastic Networks.* Cambridge University Press.

[72] Kingman, J. (1962). On queues in heavy traffic. *Journal of the Royal Statistical Society: Series B (Methodological)*, 24(2):383–392.

[73] Kyprianou, A. (2006). *Introductory Lectures on Fluctuations of Lévy Processes with Applications.* Springer.

[74] Kyprianou, A. and Palmowski, Z. (2008). Fluctuations of spectrally negative Markov additive processes. In *Séminaire de probabilités XLI*, pages 121–135. Springer.

[75] Lam, S. (1982). Dynamic scaling and growth behavior of queuing network normalization constants. *Journal of the ACM*, 29(2):492–513.

[76] Lavenberg, S. and Reiser, M. (1980). Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers. *Journal of Applied Probability*, 17(4):1048–1061.

[77] Lévy, P. (1954). *Théorie de l'Addition des Variables Aléatoires.* Gauthier-Villars.

[78] Lewis, A. and Mordecki, E. (2008). Wiener-hopf factorization for Lévy processes having positive jumps with rational transforms. *Journal of Applied Probability*, 45(1):118–134.

[79] Mandelbaum, A., Massey, W., and Reiman, M. (1998). Strong approximations for Markovian service networks. *Queueing Systems*, 30:149–201.

[80] Mandjes, M. and Nuyens, M. (2005). Sojourn times in the M/G/1 FB queue with light-tailed service times. *Probability in the Engineering and Informational Sciences*, 19(3):351–361.

[81] Markov, A. (1906). Extension of the law of large numbers to dependent quantities. *Izvestiya Fiziko-Matematicheskogo Obschestva pri Kazanskom Universitete.(2nd Ser)*, 15(1):135–156.

[82] Mimica, A. (2016). Exponential decay of measures and Tauberian theorems. *Journal of Mathematical Analysis and Applications*, 440(1):266–285.

[83] Mitra, D. and McKenna, J. (1986). Asymptotic expansions for closed Markovian networks with state-dependent service rates. *Journal of the ACM*, 33(3):568–592.

[84] Miyazawa, M. (2002). A Markov renewal approach to the asymptotic decay of the tail probabilities in risk and queuing processes. *Probability in the Engineering and Informational Sciences*, 16(2):139–150.

[85] Miyazawa, M. (2004). Hitting probabilities in a Markov additive process with linear movements and upward jumps: applications to risk and queueing processes. *The Annals of Applied Probability*, 14(2):1029–1054.

[86] Morris, N., Stewart, C., Chen, L., Birke, R., and Kelley, J. (2018). Model-driven computational sprinting. In *Proceedings of the Thirteenth EuroSys Conference*, pages 1–13.

[87] Nair, J., Wierman, A., and Zwart, B. (2010). Tail-robust scheduling via limited processor sharing. *Performance Evaluation*, 67(11):978–995.

[88] Nair, J., Wierman, A., and Zwart, B. (2022). *The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation*. Cambridge University Press.

[89] Nakagawa, K. (2005). Tail probability of random variable and Laplace transform. *Applicable Analysis*, 84(5):499–522.

[90] Nakagawa, K. (2007). Application of Tauberian theorem to the exponential decay of the tail probability of a random variable. *IEEE Transactions on Information Theory*, 53(9):3239–3249.

[91] Nelson, E. (2020). *Dynamical Theories of Brownian Motion*. Princeton University Press.

[92] Neveu, J. (1961). Une généralisation des processus à accroissements positifs indépendants. In *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, volume 25, pages 36–61. Springer.

[93] NIST Digital Library of Mathematical Functions (2021). http://dlmf.nist.gov/.

[94] Norris, J. (1998). *Markov Chains*. Cambridge University Press.

[95] Núñez-Queija, R. (2002). Queues with equally heavy sojourn time and service requirement distributions. *Annals of Operations Research*, 113(1):101–117.

[96] Nuyens, M. and Wierman, A. (2008). The foreground-background queue: a survey. *Performance Evaluation*, 65(3–4):286–307.

[97] Nuyens, M., Wierman, A., and Zwart, B. (2008). Preventing large sojourn times using SMART scheduling. *Operations Research*, 56(1):88–101.

[98] Park, K. and Willinger, W. (2000). *Self-similar Network Traffic and Performance Evaluation*. Wiley.

[99] Peskir, G. and Shiryaev, A. (2006). *Optimal Stopping and Free-Boundary Problems*. Lectures in Mathematics. ETH Zürich. Birkhäuser Verlag.

[100] Peterson, D. (1996). Data center I/O patterns and power laws. *CMG Proceedings*.

[101] Raghavan, A., Luo, Y., Chandawalla, A., Papaefthymiou, M., Pipe, K., Wenisch, T., and Martin, M. (2012). Computational sprinting. In *Proceedings of the 18th Symposium on High Performance Computer Architecture*, pages 1–12. IEEE.

[102] Reiser, M. (1979). A queueing network analysis of computer communication networks with window flow control. *IEEE Transactions on Communications*, 27(8):1199–1209.

[103] Reiser, M. and Kobayashi, H. (1975). Queuing networks with multiple closed chains: theory and computational algorithms. *IBM Journal of Research and Development*, 19(3):283–294.

[104] Remerova, M., Foss, S., and Zwart, B. (2014). Random fluid limits of an overloaded polling model. *Advances in Applied Probability*, 46(1):76–101.

[105] Schrage, L. (1968). A proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 16(3):687–690.

[106] Schrage, L. and Miller, L. (1966). The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, 14(4):670–684.

[107] Scully, Z., Grosof, I., and Harchol-Balter, M. (2020a). The Gittins policy is nearly optimal in the M/G/k under extremely general conditions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(3):1–29.

[108] Scully, Z., Grosof, I., and Harchol-Balter, M. (2021). Optimal multiserver scheduling with unknown job sizes in heavy traffic. *Performance Evaluation*, 145:102150.

[109] Scully, Z., Grosof, I., and Mitzenmacher, M. (2022). Uniform bounds for scheduling with job size estimates. In *13th Innovations in Theoretical Computer Science Conference*, Leibniz International Proceedings in Informatics. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[110] Scully, Z. and Harchol-Balter, M. (2018). SOAP bubbles: robust scheduling under adversarial noise. In *56th Annual Allerton Conference on Communication, Control, and Computing*, pages 144–154.

[111] Scully, Z. and Harchol-Balter, M. (2021). The Gittins policy in the M/G/1 queue. In *19th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*. IEEE.

[112] Scully, Z., Harchol-Balter, M., and Scheller-Wolf, A. (2018a). Optimal scheduling and exact response time analysis for multistage jobs. *arXiv preprint arXiv:1805.06865*.

[113] Scully, Z., Harchol-Balter, M., and Scheller-Wolf, A. (2018b). SOAP: one clean analysis of all age-based scheduling policies. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1):1–30.

[114] Scully, Z., Harchol-Balter, M., and Scheller-Wolf, A. (2020b). Simple near-optimal scheduling for the M/G/1. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(1):1–29.

[115] Scully, Z., van Kreveld, L., Boxma, O., Dorsman, J., and Wierman, A. (2020c). Characterizing policies with optimal response time tails under heavy-tailed job sizes. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(2):1–32.

[116] Scully, Z. and van Kreveld, L. (2021). When does the Gittins policy have asymptotically optimal response time tail? *arXiv preprint arXiv:2110.06326*.

[117] Shiryaev, A. (2008). *Optimal Stopping Rules*. Springer.

[118] Siegl, T. and Tichy, R. (1999). A process with stochastic claim frequency and a linear dividend barrier. *Insurance: Mathematics and Economics*, 24(1-2):51–65.

[119] Starreveld, N., Bekker, R., and Mandjes, M. (2016). Transient analysis of one-sided Lévy-driven queues. *Stochastic Models*, 32(3):481–512.

[120] Stolyar, A. and Ramanan, K. (2001). Largest weighted delay first scheduling: large deviations and optimality. *Annals of Applied Probability*, 11(1):1–48.

[121] Stoyan, D. (1983). *Comparison Methods for Queues and Other Stochastic Models.* John Wiley & Sons.

[122] van Kreveld, L., Boxma, O., Dorsman, J., and Mandjes, M. (2020). Scaling analysis of an extended machine-repair model. In *Proceedings of the 13th EAI International Conference on Performance Evaluation Methodologies and Tools*, pages 172–179.

[123] van Kreveld, L., Boxma, O., Dorsman, J., and Mandjes, M. (2021). Scaling limits for closed product-form queueing networks. *Performance Evaluation*, 151:102220.

[124] van Kreveld, L., Mandjes, M., and Dorsman, J. (2022a). Extreme value analysis for a Markov additive process driven by a nonirreducible background chain. *Stochastic Systems*, 12(3):293–317.

[125] van Kreveld, L., Mandjes, M., and Dorsman, J. (2022b). Cramér-Lundberg asymptotics for spectrally positive Markov additive processes. *Submitted*.

[126] Whitt, W. (2002). *Stochastic-Process Limits: an Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer.

[127] Wierman, A. and Zwart, B. (2012). Is tail-optimal scheduling possible? *Operations Research*, 60(5):1249–1257.

[128] Yin, G., Liu, Y., and Yang, H. (2006). Bounds of ruin probability for regime-switching models using time scale separation. *Scandinavian Actuarial Journal*, 2006(2):111–127.

[129] Zhu, J. and Yang, H. (2009). On differentiability of ruin functions under Markov-modulated models. *Stochastic Processes and their Applications*, 119(5):1673–1695.

[130] Zwart, B. and Boxma, O. (2000). Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Systems*, 35(1-4):141–166.

Acknowledgements

My first thanks goes out to my excellent supervisors. Michel, it's incredible how you are always able to find time in your busy schedule (I remember a meeting starting at 12:05). Onno, I'm still amazed that you come up with the perfect suggestion for seemingly any arising mathematical problem. Jan-Pieter, I feel extremely lucky to have had a supervisor who's door is (literally) always open and is willing to think along in every situation.

I also want to express my gratitude to my American co-authors Ziv Scully and Adam Wierman. Working with you has been a true pleasure, and it has enriched me with the ability of writing in the engaging CMU style. Ziv, I enjoyed the open discussions during our numerous overseas video calls.

The Korteweg-de Vries Institute has facilitated all the work put in this thesis. Special thanks to Mariska, Nicos, Sven, Rens, Nikki, Ben and Bharti; having colleagues like you is a privilege. A big thank you to the entire support staff as well.

Another organization I should mention is the NETWORKS-consortium. Its regular events have broadened my knowledge and have kept me up to date on the developments of other interesting research areas. Being part of such a close research group has brought valuable variation to the many solitary PhD-hours.

Finally, thanks to my family, to my ever supporting parents and my brother Ivo. Kyra, you bring joy to my life in ways that keep on surprising me. Lastly, I feel indescribably lucky for having Kris in my life; during the last few months of writing you have been a much-needed source of distraction. May you be as lucky as your dad.

Summary

Asymptotic Analysis of Stochastic Systems

In this thesis, we discuss a number of problems in stochastic systems involving queues and Lévy processes. For the more advanced systems in these categories, it is commonly seen that exact analytical results are unachievable and/or that direct numerical computations are too slow. As an alternative, we aim our attention at asymptotic approaches. In some cases, we study the stochastic system as one parameter tends to a threshold value, whereas in other cases it proves useful to asymptotically scale the system as a whole.

Three general topics separate the thesis into equally many parts that can be read individually. Part I concerns the effect of the scheduling policy on the asymptotic response time tail in the $M/G/1$ queue. In Part II a scaling approach is adopted to approximate queue-length distributions in a closed product-form network. Lastly, the main content of Part III is the analysis of the (asymptotic) maximum of a Markov additive process (MAP).

In the scheduling context of Part I, it is known that the performance of a policy strongly depends on the job size distribution. We analyze the asymptotic response time tail for the class of SOAP policies. Specifically, the focus in Chapter 2 lies on heavy-tailed job sizes. We characterize which SOAP policies have a response time tail of the same order as the job size tail, and are thus optimal. Our characterization takes the form of an easily verifiable condition on the rank function. The set of new policies for which tail optimality is established includes Foreground-Background with limited preemption, Shortest Expected Remaining Processing Time and Randomized Multi-Level Feedback. In the course of our analysis, we derive new bounds on fractional moments of busy periods.

Chapter 3 extends the main result of Chapter 2 in two ways. Firstly, it relaxes the condition for tail optimality in the case of heavy-tailed job sizes. Secondly, a characterization for tail optimality of SOAP policies is derived for light-tailed job sizes. For this class of job size distributions, a policy is tail-optimal (tail-pessimal) if it maximizes (minimizes) the decay rate of the response time tail. The two extensions allow us to characterize the tail performance of the prominent Gittins policy. Particularly, for heavy-tailed job sizes it is always tail-optimal, and in the light-tailed case it may be optimal, pessimal or in between, depending on the nature of the job size distribution. For instances in which Gittins is pessimal, we derive a condition under which a slight modification of its rank function increases its response time decay rate, and thus avoids pessimality.

In Part II, closed product-form queueing networks with single-server and infinite-server stations are considered. The joint stationary queue-length distribution of such models is a set of geometric (single-server) and Poisson (infinite-server) distributions, together with a population size constraint. As this constraint makes the corresponding distribution challenging

to numerically analyze, we consider in Chapter 4 a scaling of Halfin-Whitt type. The limiting normalized distribution of the joint queue lengths is identified using asymptotic analysis of Laplace-Stieltjes transforms. We show that the normalized queue lengths are exponentially (single-server) and normally (infinite-server) distributed, and the population size constraint translates to a single constraint on only the most heavily loaded station(s).

Finally, the objective of Part III is to identify the distribution of the maximum of a MAP. Knowledge on this distribution is valuable in several respects: it contributes to the fundamental understanding of MAPs, it has a direct implication for first passage processes, and it is the key performance metric in risk applications. Chapter 5 characterizes the distribution of the maximum of a spectrally one-sided MAP over a phase-type horizon (which includes exponential horizons). In the spectrally-positive case, we provide a recursive procedure for computing the Laplace-Stieltjes transform of the maximum. Our characterization of the maximum in the spectrally-negative case is more explicit in the sense that it is neither recursive nor in terms of transforms. By letting the state of the background chain correspond to the phase of a phase-type distribution, our results also allow computation of the maximum of any spectrally one-sided Lévy process over a phase-type horizon.

In Chapter 6 we focus specifically on a spectrally-positive light-tailed MAP, and analyze the asymptotic tail probability of its maximum (or, equivalently, the asymptotic ruin probability for a Cramér-Lundberg risk model in the MAP setting). Our analysis relies on the introduction of an alternative measure, under which the ruin probability is one. An expression is derived for the exact asymptotics in terms of the solution to a system of equations. In addition, the change-of-measure approach allows for an algorithm that estimates the ruin probability with bounded relative error, yielding much shorter running times than direct simulation.

Samenvatting

Asymptotische Analyse van Stochastische Systemen

In dit proefschrift bespreken we een aantal vraagstukken voor stochastische systemen. We richten onze aandacht in het bijzonder op wachtrijen en Lévy processen. Bij geavanceerdere systemen uit deze categorie geldt doorgaans dat exacte analytische resultaten onhaalbaar zijn en/of directe numerieke berekeningen te langzaam. Als alternatief vestigen we onze aandacht op asymptotische aanpakken. Daarbij maken we gebruik van twee soorten asymptotiek: die waar een enkele parameter een drempelwaarde nadert, en die waar het systeem als geheel asymptotisch wordt geschaald.

Dit proefschrift is onderverdeeld in drie delen. Elk van deze delen kan los van de andere gelezen worden en bestrijkt de asymptotische analyse van bepaalde stochastische systemen. Deel I betreft het effect van de bedieningsdiscipline op de asymptotische staart van de verblijftijdverdeling in de $M/G/1$-wachtrij. In Deel II hanteren we een schalingsaanpak om wachtrijlengteverdelingen te benaderen van een gesloten wachtrijnetwerk met een stationaire product-vorm verdeling. De voornaamste inhoud van Deel III, ten slotte, is de analyse van het (asymptotische) maximum van een Markov additief proces (MAP).

In de bedieningsdiscipline-context van Deel I is algemeen bekend dat de prestatie van een discipline sterk afhangt van de taakgrootteverdeling. We analyseren de asymptotische staart van de verblijftijdverdeling voor de klasse van SOAP-disciplines. In het bijzonder gaat de aandacht in Hoofdstuk 2 uit naar zwaarstaartige taakgroottes. We karakteriseren welke SOAP-disciplines een verblijftijdstaart van dezelfde orde hebben als de staart van de taakgrootte, en daarmee optimaal zijn. Onze karakterisering heeft de vorm van een gemakkelijk verifieerbare conditie op de rangfunctie. De verzameling nieuwe disciplines waarvoor staartoptimaliteit wordt bevestigd, bevat Foreground-Background met beperkte onderbreking, Shortest Expected Remaining Processing Time en Randomized Multi-Level Feedback. In de loop van onze analyse leiden we nieuwe bovengrenzen af voor fractionele momenten van aaneengesloten bedieningsperiodes.

Hoofdstuk 3 breidt het hoofdresultaat van Hoofdstuk 2 op twee manieren uit. In de eerste plaats zwakt het de conditie voor staartoptimaliteit af in het geval van zwaarstaartige taakgroottes. Ten tweede wordt een karakterisering voor staartoptimaliteit van SOAP-disciplines afgeleid voor lichtstaartige taakgroottes. Bij deze klasse taakgrootteverdelingen is een discipline staartoptimaal (staartinferieur) als het de vervalsnelheid van de staart van de verblijftijd maximaliseert (minimaliseert). De twee uitbreidingen stellen ons in staat om het staartgedrag van de belangrijke Gittinsdiscipline te bepalen. In het bijzonder geldt dat deze discipline bij zwaarstaartige taakgroottes altijd optimaal is. In het lichtstaartige geval kan de Gittinsdiscipline optimaal zijn, inferieur of ergens daartussen, afhankelijk van de aard van de taakgrootteverdeling. Voor gevallen waarin Gittins inferieur is, leiden we een conditie af onder welke een kleine wijziging van de rangfunctie resulteert in een verhoogde vervalsnelheid van de

verblijftijdstaart, waardoor inferioriteit wordt vermeden.

In Deel II worden gesloten wachtrijnetwerken met stationaire product-vorm verdelingen beschouwd, bestaande uit rijen met één bediende of een oneindig aantal bedienden. De gezamenlijke stationaire rijlengteverdeling van zulke modellen is een verzameling van geometrische verdelingen (één bediende) en Poissonverdelingen (oneindig aantal bedienden), samen met een beperking op de populatiegrootte. Omdat deze beperking het numeriek analyseren van de corresponderende verdeling bemoeilijkt, beschouwen we in Hoofdstuk 4 een schaling van het type Halfin-Whitt. De limietverdeling van de genormaliseerde gezamenlijke wachtrijlengten wordt bepaald door middel van asymptotische analyse van Laplace-Stieltjes transformaties. We laten zien dat de genormaliseerde wachtrijlengten exponentieel (één bediende) en normaal (oneindig aantal bedienden) verdeeld zijn, en dat de beperking op de populatiegrootte zich vertaalt naar een enkele beperking op slechts de meest zwaarbeladen wachtrij(en).

Ten slotte is het doel van Deel III om de verdeling van het maximum van een MAP te bepalen. Bekendheid van deze verdeling is waardevol in meerdere opzichten: het draagt bij aan de fundamentele kennis over MAPs, het verloop van intreetijdprocessen kan hiermee worden bepaald, en het is de belangrijkste prestatiemaat in de verzekeringswiskunde. Hoofdstuk 5 karakteriseert de verdeling van het maximum van een spectraal eenzijdige MAP gedurende een fase-type interval (dus ook voor exponentiële intervallen). In het spectraal positieve geval geven we een recusieve procedure voor het berekenen van de Laplace-Stieltjes transformatie van het maximum. Onze karakterisering van het maximum in het spectraal negatieve geval is explicieter in die zin dat deze noch recursief, noch in termen van transformaties is gedefinieerd. Door de toestand van de achtergrondketen te laten samenhangen met de fase van een fase-type verdeling, staan onze resultaten een berekening toe van het maximum van een willekeurig spectraal eenzijdig Lévy proces gedurende een fase-type interval.

In Hoofdstuk 6 concentreren we ons op een spectraal positieve en lichtstaartige MAP, en analyseren we de asymptotische staartkans van het maximum (met andere woorden, de asymptotische ruïneringskans van een Cramér-Lundberg model in de MAP-context). Onze analyse bouwt op de introductie van een alternatieve kansmaat, onder welke de ruïneringskans één is. Een uitdrukking voor de exacte asymptotiek wordt afgeleid in termen van de oplossing van een stelsel vergelijkingen. Daarnaast maakt de verandering van maat een algoritme mogelijk dat de ruïneringskans schat met begrensde relatieve fout, wat veel kortere looptijden oplevert dan directe simulatie.