Extreme-Value Theory for Large Fork-Join Queues

Extreme-Value Theory for Large Fork-Join Queues

This thesis is part of the PhD thesis series of the Beta Research School for Operations Management and Logistics (onderzoeksschool-beta.nl) in which the following universities cooperate: Eindhoven University of Technology, Ghent University, Maastricht University, Tilburg University, University of Twente, VU Amsterdam, Wageningen University & Research, Katholieke Universiteit Leuven, Universiteit Hasselt, VU Brussels, University of Antwerp and CWI, the Dutch national research institute for mathematics and computer science.

This work is part of the research program *Complexity in high-tech manufacturing* with project number 438.16.121, which is (partly) financed by the Dutch Research Council (NWO).

Extreme-Value Theory for Large Fork-Join Queues

A catalogue record and digital copies are available from the Eindhoven University of Technology Library.
ISBN: 978-90-386-5726-4

Printed by ProefschriftMaken || `www.proefschriftmaken.nl`
Cover design by Mercedes Benjaminse || `www.proefschriftmaken.nl`
Painting by Jessica Schol

The fork-join queue with three servers is visualized as leaf veins in a leaf. The cover is inspired by the story *Leaf by Niggle*, written by J.R.R. Tolkien. The story describes the life of Niggle, an artist who wants to paint a tree in full detail. At some point, Niggle is forced to make a journey to a country far away while having completed only one leaf. He is brought to a place where he gets the task to maintain a garden. In the middle of this garden, he finds the tree he wanted to paint.

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Beta
Research School for Operations Management and Logistics

Extreme-Value Theory for Large Fork-Join Queues

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus, prof.dr. S.K. Lenaerts, voor een
commissie aangewezen door het College voor
Promoties, in het openbaar te verdedigen
op dinsdag 16 mei 2023 om 16:00 uur

door

Cornelis Schol

geboren te Dirksland

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

voorzitter:	prof.dr. M.G.J. van den Brand
1^{e} promotor:	prof.dr. M. Vlasiou
2^{e} promotor:	prof.dr. A.P. Zwart
leden:	prof.dr.ir. S.C. Borst
	prof.dr.-ing. habil. K. Dębicki (Uniwersytet Wrocławski)
	prof.dr. B. Van Houdt (Universiteit Antwerpen)
	dr. W.L. van Jaarsveld
	prof.dr. N.V. Litvak

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

Acknowledgments

I am about to get to the end of a road. In the more than four years that are behind me, a lot has happened in my life, and I am happy and grateful to the Lord that I have reached this point.

When I applied for the position, Maria asked whether I did team sports. At that moment, it felt a bit silly to me to talk about handball during a job interview, and we talked about it for quite a while, but now I understand the question. Working as a scientist implies working as a team. Maria, I am very grateful for your guidance in so many ways. You helped me every time to transform a bulk of equations into an engaging paper. You encouraged me to 'mingle in the crowds'. You helped me to think about future perspectives. Moreover, even about parenting you could teach me a lot. I enjoyed our conversations. Bert, thank you so much for all the discussions we had about math. You showed me and infected me with the eagerness of the researcher to explore unknown areas. You taught me the patience to look further if a problem seems unsolvable. You helped me to become a more and more independent researcher.

I am grateful to the members of the doctorate committee for reading my thesis and providing valuable feedback.

I would like to thank my fellow PhD colleagues for having lunch together in the lounge, attending LNMB courses on Mondays, and going to conferences. I want to convey special thanks to Albert, Diego, Ivo, and Rowel. For me, our office was, thanks to you guys, a cozy place. I will miss our office vibe, the inside jokes (e.g., about statistics, Uruguay, weddings), and going to Gamestate. Mark and Tom, I appreciated hanging out with you, and watching football, especially during lockdown times. Tom, I liked your cookies. Jaap and Ellen, I very much enjoyed the trip we made to the Stochastic Networks Conference in Ithaca. I want to thank Peter for the nice days in Helsinki. I am grateful to Alessandro, Bart, Lorenzo, and Murtuza for welcoming me into a warm and friendly office. Wouter, I enjoyed the coffee breaks we had, and I loved the dinners that we had together with Esmée, Oedipus, Janine, and Yannick.

Mirjam and Willem, I appreciated collaborating with you on two papers. I am glad we shared our expertise, which resulted in Chapter 6 of this thesis. I am grateful that Willem agreed to be part of my committee.

I want to thank my family for their support during these more than four years. I especially want to thank Jessica for making paintings for the cover.

Janine, I owe you a lot. I apologize for coming home late when a young family should have had dinner long ago. I have to admit that while you were talking to me about raising our son and running a household, I was only thinking about queues, Brownian motions, newsvendor problems, and more. I thank you for your love and support throughout these years, and I am looking forward to traveling the road that lies ahead of us.

Finally, I want to thank my son Yannick for the gift that you are to me, for giving me joy, love, and friendship. And for keeping me awake at night.

Dennis Schol

Contents

Chapter 1

Introduction

1.1. Scope of the thesis

Queues are part of our everyday life. We need to wait behind each other in the supermarket, on the highway, and at the airport. In our modern society, queues appear in less visible places as well, such as data centers, telecommunication networks, and the Internet. Although these situations are all different, queueing systems typically consist of two types of actors: the *customers* and the *servers*. *Queueing theory* is the branch of applied mathematics that studies the behavior of these systems. Often, we cannot fully predict how many people will drive to their work or go to the supermarket, and we cannot fully predict how long we are in service. Hence, in the study of queues, stochasticity plays a key role.

The most basic queueing system is the *single-server* queue: incoming customers go to a single server to get service. If others are already there, the customer waits, and after getting service the customer leaves. Obviously, many real-life queueing systems are much more involved, such as the queueing system in a supermarket. When more than one checkout is open, customers typically join the queue which (they think) minimizes their waiting time. When personnel observes long queues, usually an extra checkout is opened. Hence, the structure of the queueing network is important.

Apart from the structure, what determines the performance of a queueing system is the behavior of the arriving customers and the service stations. A car driver spends more time on the same highway during rush hour than during midnight. An experienced worker will service customers faster than an inexperienced worker. The general behavior of arriving customers is governed by the *interarrival distribution*, while the service times are governed by the *service-time distribution*. Examples of other, more complicated types of behaviors or policies are: customers arrive in groups; certain types of customers get priority; customers abandon after having waited a certain amount of time, or queues are served in reverse order of arrival. Two comprehensive books on queueing theory are [13] and [26]. In summary, many real-life service systems can be modeled by a queueing network with a certain structure, and with certain interarrival and service-time distributions.

In this thesis, we focus on a specific type of queueing network that has been studied in the

past and is called the *fork-join queue*. The fork-join queue is a model of a parallel-processing system, where incoming *jobs* consist of *subtasks*, which are *forked* among different service stations. After completion of the subtasks, these are again *joined* and the job is finished. This queueing system is a suitable modeling tool for supply chains and communication networks. In the situation of the supply chain, the arriving jobs represent orders from a manufacturer, the service stations represent suppliers, and joining the completed subtasks represents assembling the components into a final product. In Figure 1.1, a schematic representation of such a supply chain is given; we see an arriving stream of orders from the manufacturer; each order is forked in N subtasks; each supplier operates as a single-server queue, and each supplier has to produce a component.

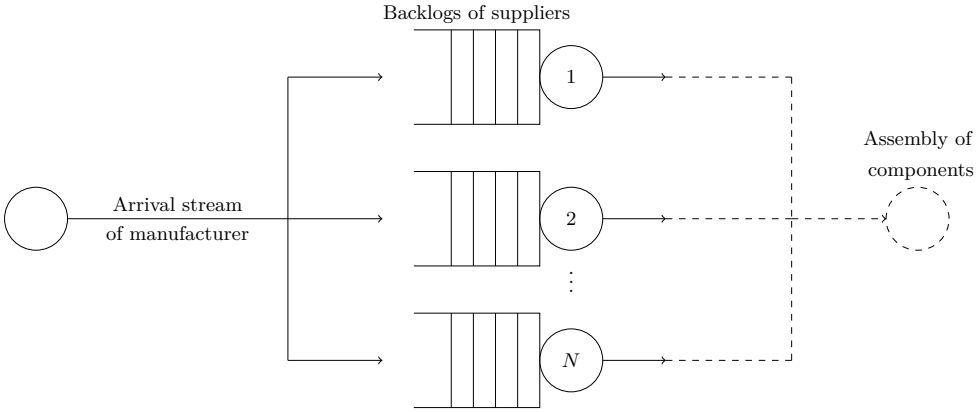


Figure 1.1 Fork-join queue with N servers

Obviously, when the demand exceeds the number of completed components, queues start to form in front of the service stations and this results in supply chain delays. The manufacturer in this simple model has a particular interest in the delay of the slowest supplier, as the slowest supplier determines the delay for the manufacturer. Hence the most important performance measures for the manufacturer are the length of the longest queue and the longest waiting time among the N subtasks before getting service.

This fork-join queue has been broadly studied, and we give an overview of this research in Section 1.3.2. However, the focus of this thesis lies on analyzing this type of queueing system where the number of service stations N is *large*. Because a large finite system is close to intractable, we aim to find convergence results for the maximum queue length and the longest waiting time as the number of service stations N goes to infinity. As we are interested in maximal quantities, we derive results that belong to the area of *extreme-value theory*, on which we provide a partial literature overview in Section 1.4. In this thesis, we give steady-state results in Chapters 2, 3, and 5, process convergence results in Chapters 2 and 5, we derive a large deviations principle in Chapter 4, and we apply these results to inventory problems in assembly systems in Chapter 6.

1.2. Motivation

Our inspiration to investigate large fork-join queues stems from supply chains in the high-tech industry. High-tech companies, such as ASML, Philips, and Boeing, are types of original equipment manufacturers (OEMs) that assemble thousands of components, each produced using specialized equipment, into complicated systems. As a result, these supply chains have typical characteristics.

The most important characteristic is that high-tech manufacturers outsource the production of components to many suppliers worldwide. High-tech companies typically have thousands of direct suppliers. For instance, the Dutch lithography equipment manufacturer ASML had 4750 direct suppliers in 2020; see [12, p. 53].

These supply chains often also contain multiple tiers, as suppliers make products involving parts delivered by other suppliers; see [12, p. 57]. Due to this structure, the state supply chain is less observable [108]. These supply chains are vulnerable as supply chains can be easily disrupted; see [12, p. 95].

Another important characteristic is that in the high-tech industry, many suppliers deliver products that have a very large technological value. For instance, of the 4750 suppliers of ASML, 188 are called *critical suppliers*. One example is the German company Carl Zeiss. This company is the sole supplier of lenses, mirrors, illuminators, collectors, and optics for ASML; see [12, p. 95]. For this reason, ASML has no back-up supplier available in case a disruption occurs at Carl Zeiss.

A further characteristic of high-tech manufacturing is that revenue is derived from a relatively small number of products. ASML sold 229 and 258 machines in 2019 and 2020, respectively; see [12, p. 92]. At the same time, the price of each of these machines is very high: one EUV lithography system costs around 150 million USD [144]. This underlines the fact that delays in these types of supply chains are very costly.

This all means that the size of the supply chain and the nature of the high-tech suppliers make the supply chains very exposed to adverse events, resulting in a significant loss of revenue. If one component is missing, the final product cannot be assembled, giving rise to costly delays. Hence, a straightforward measure for the performance of this high-tech supply chain is the delay of the slowest supplier. We can conclude that the fork-join queue functions as a stylized model for a high-tech supply chain, where the arrival stream of jobs represents the orders from the manufacturer to the suppliers, and where each service station represents a single supplier. Furthermore, the maximum queue length, and the longest waiting time, are the two performance measures that are of particular interest to the manufacturer in the high-tech supply chain.

Another area of application of the fork-join queue is parallel computing in data centers. Companies such as Google, Microsoft, and Alibaba have data centers with thousands of servers, that are available for cloud computing, where there is often a form of parallel scheduling [5, 101]. Jobs in these systems have large sizes and are often heavy-tailed, see for example [147], in which the Google Borg Scheduler is analyzed. However, most theoretical literature on parallel queues assumes service times to be light-tailed; see [70, p. 20]. This

motivates the analysis of parallel queueing networks with heavy-tailed job sizes.

1.3. Queueing theory

In this thesis, we study the fork-join queueing system with N service stations, in which each service station processes one subtask. The queueing system restricted to one service station in isolation can be seen as a single-server queue. Therefore, performance measures of fork-join queues are related to performance measures of single-server queues. In Section 1.3.1, we give an overview of the literature on the analysis of single-server queues, which we also use in this thesis. Although the dynamics of each service station in isolation are well understood, the fact that each service station has the same arrival stream complicates the analysis of the joint performance measures of the fork-join queue. Therefore, while a lot of research has been done on the topic of fork-join queues, only a few analytic results are known, e.g. [57, 58]. In Section 1.3.2, we give an outline of the results in the field of fork-join queues.

1.3.1 Single-server queue

One of the fundamental contributions to the analysis of queueing systems is due to Lindley [96]. In this study, it is shown that under the *first-come-first-served* (FCFS) policy, the waiting time of an arbitrary customer in a $GI/GI/1$ queue can be written as a relation with the waiting time of the previous customer, as $W(n+1) = \max(0, W(n) + S(n) - A(n))$, where $W(n)$ and $S(n)$ are the waiting time and service time of the n-th customer, respectively, and $A(n)$ is the interarrival time between the n-th and $(n+1)$-st customer. This expression is known as *Lindley's recursion*. One can inductively show that $W(n) := \sup_{1\leq k\leq n} \sum_{j=k}^{n-1}(S(j) - A(j))$ solves this equation, when $W(1) = 0$. Furthermore, when the queueing system is a stable $GI/GI/1$ queue, the steady-state waiting time satisfies $W \stackrel{d}{=} \max(0, W + S - A)^+$, in which case it is easy to show that

$$W \stackrel{d}{=} \sup_{k\geq 0} \sum_{j=1}^{k} (S(j) - A(j)).$$

As is intuitively clear, the average waiting time of a customer is connected to the average queue length. In [97], this is formalized: *Little's law* says that for a stable and non-preemptive queueing system, $\lambda\mathbb{E}[W] = \mathbb{E}[Q]$, where $\mathbb{E}[Q]$ is the expected steady-state queue length and λ is the arrival rate. In [68], an extension is given to the distributions of W and Q. When the arrival process is stationary, the queue discipline is FCFS, and the waiting time of a customer is independent of the number of arrivals during any time interval after its arrival, then the steady-state queue length has the same distribution as $\mathbf{N}_A(W)$, where W is the steady-state waiting time and $\mathbf{N}_A(t)$ is the number of arriving customers until time t.

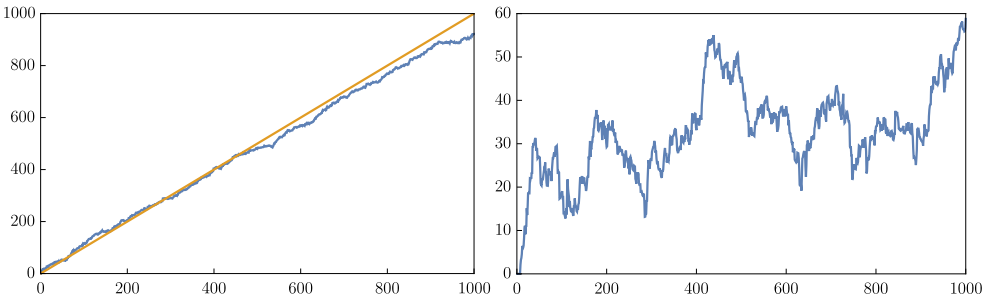
In the analysis of queues, the behavior of queueing systems is also investigated through the derivation of *fluid limits*. The idea is to scale the system in such a way that deterministic limits are obtained. This is an extension of the law of large numbers to time-dependent

processes. For instance, in [37, Thm. 6.5, p. 136], a standard fluid limit for the queue length in the $GI/GI/1$ queue is given. By the law of large numbers for renewal processes, it is known that for fixed t, $(\mathbf{N}_A(tn) - \mathbf{N}_S(tn))/n \xrightarrow{\mathbb{P}} (1/\mathbb{E}[A] - 1/\mathbb{E}[S])t$ as $n \to \infty$, with $\mathbf{N}_A(t)$ the number of arrivals until time t, and $\mathbf{N}_S(t)$ the number of services until time t. Now, we consider a sequence of queue lengths $(Q^{(n)}, n \geq 1)$. Furthermore, the number of items in queue at time 0 equals $Q^{(n)}(0)$ that satisfies $Q^{(n)}(0)/n \xrightarrow{\mathbb{P}} \bar{Q}(0) \geq 0$, as $n \to \infty$. Then the sequence of queue length processes satisfies $(Q^{(n)}(nt)/n, t \in [0,T]) \xrightarrow{\mathbb{P}} ((\bar{Q}(0) + (1/\mathbb{E}[A] - 1/\mathbb{E}[S])t)^+, t \in [0,T])$ in $C[0,T]$ as $n \to \infty$. The space $C[0,T]$ is the space of continuous functions on $[0,T]$, which we equip with the supremum norm.

As we can extend the law of large numbers, we can also extend the central limit theorem; this is called a *diffusion approximation* of a queue. By the functional central limit theorem [37, Thm. 5.11, p. 110], we know that $((\mathbf{N}_A(tn) - nt/\mathbb{E}[A])/\sqrt{n}, t \in [0,T]) \xrightarrow{d} (B_A(t), t \in [0,T])$ in $C[0,T]$ as $n \to \infty$, with $(B_A(t), t \geq 0)$ a Brownian motion with drift 0. There is an analogous limit $(B_S(t), t \in [0,T])$ for the number of services. Then, following [37, Thm. 6.8, p. 139], when $Q(0) = 0$, the queue length process $Q(t)$ satisfies $((Q(nt) - ((1/\mathbb{E}[A] - 1/\mathbb{E}[S])nt)^+)/\sqrt{n}, t \in [0,T]) \xrightarrow{d}$

- $(0, t \in [0,T])$ in $C[0,T]$ as $n \to \infty$, if $\mathbb{E}[A] > \mathbb{E}[S]$,
- $(\sup_{s \in [0,t]}((B_A(t) - B_S(t)) - (B_A(s) - B_S(s))), t \in [0,T])$ in $C[0,T]$ as $n \to \infty$, if $\mathbb{E}[A] = \mathbb{E}[S]$,
- $(B_A(t) - B_S(t), t \in [0,T])$ in $C[0,T]$ as $n \to \infty$, if $\mathbb{E}[A] < \mathbb{E}[S]$.

Obviously, the second case is the most intriguing, the limiting process is called a *reflected Brownian motion* [1, 71]. Simulation results of consecutive waiting times in the second and third cases are given in Figure 1.2.



(a) $S \sim \text{Exp}(1/2)$, $A \sim \text{Exp}(1)$, and fluid limit

(b) $S \sim \text{Exp}(1)$, $A \sim \text{Exp}(1)$

Figure 1.2 Waiting time of n-th arriving customer in the unstable $M/M/1$ queue as a function of n

Now, one might also be interested in what happens when $\mathbb{E}[A] - \mathbb{E}[S] > 0$, but is very close to 0. In other words, the queueing system is in heavy traffic. A lot of general results

can be derived. Early results on this topic are given in [78, 82, 83]. In [82], it is shown that in a $GI/GI/1$ queue with an FCFS policy, when the traffic intensity ρ converges to 1, $(1-\rho)W$ converges in distribution to an exponentially distributed random variable. In [83], this result is extended to $G/GI/1$ queues. Furthermore, it is shown that the queue length and waiting time processes converge to a reflected Brownian motion [127].

This analysis is also extended to queues with more servers [74, 75]. In [69], the focus lies on the performance in the $G/M/s$ queue in heavy traffic, where $s \to \infty$ and the traffic intensity depends on s. Special attention is given to the case that $(1-\rho)\sqrt{s} \to \beta$ as $s \to \infty$; non-trivial limit convergence results are obtained when the queue length is normalized with $\sqrt{s}$. This is called the *Halfin-Whitt* or the *Quality-and-Efficiency-Driven* regime; see also [95].

1.3.2 Fork-join queue

Fork-join queues have been extensively studied. The first papers on this topic are focused on two service stations. In [57, 58], the authors consider a Poisson arrival stream and two independently working servers with exponential service times. The authors describe the asymptotic behavior of the joint probability distribution where the number of tasks in one queue goes to infinity [58, Thm. 7.2]. Furthermore, asymptotic expressions are given for the expectation and distribution of the length of one queue conditioned on the number of tasks in the other queue, where the number of tasks in this other queue goes to infinity [57, Thm. 1 & 2]. These results are later extended and generalized [17, 84, 138, 155]. The fact that both service stations have the same arrival stream complicates the exact analysis of fork-join queues; already for fork-join queues with two servers, there are no exact expressions for the joint probability distribution of the number of tasks outside the asymptotic regime of one of the two queues. Furthermore, there is no extension of the asymptotic results known to fork-join queues with more than two servers.

For the fork-join queue with more than two servers, the main focus lies on finding bounds for performance measures, using inequalities on the maximum of associated random variables and stochastic orderings [18, 117, 118]. For instance, in [18, Cor. 3.4], the authors mention that the fork-join queue with i.i.d. single servers and deterministic arrivals minimizes the expected maximum response time among all possible interarrival distributions with fixed mean. This is a computable lower bound, as the fork-join queue with deterministic arrivals can be seen as a queueing system with N independent parallel queues. Thus, the cumulative distribution function of the longest waiting time of a fork-join queue with deterministic arrivals is known. When the service times are exponentially distributed, it is known that the steady-state waiting time is exponentially distributed, and the expectation of the maximum of N i.i.d. exponentials with parameter λ equals $(\sum_{j=1}^{N} 1/j)/\lambda$. In Chapter 2, we see that the maximum queue length of a fork-join queue with light-tailed services is relatively stable in probability, and possesses the same first-order behavior as the fork-join queue with deterministic arrivals. Thus, the first-order convergence result is the same as the first-order convergence result of the lower bound in [18] as $N \to \infty$. In [87], this analysis is extended to the fork-join queue with exponential interarrival times and service stations that all consist of

s memoryless servers. In [80], an algorithm is described to obtain the average response time when service times are Erlang distributed. Furthermore, in [152] the authors interpolate between light-traffic and heavy-traffic results and obtain approximations for symmetric fork-join queues. In [88], the authors present a closed-form approximation of the sojourn time of a job in a $G/M/1$ fork-join queue. In [128], approximations of the response time distribution are provided. In [153], the fork-join queue is studied under the assumption that the number of subtasks is less than or equal to the number of service stations. It is proven that when the number of subtasks k_N is $o(N^{1/4})$, the queues at any k_N servers are asymptotically independent as $N \to \infty$ [128, Thm. 4.1]. The authors also prove that for each $k_N \leq N$ the longest steady-state waiting time is stochastically dominated by the longest steady-state waiting time of an identical system, but with independent arrivals [128, Thm. 4.3].

Heavy-traffic approximations were discussed in [86, 98, 99, 100, 119, 120, 145, 151]. In [151], the author gives a heavy-traffic analysis for fork-join queues and shows weak convergence of several processes, such as the joint queue lengths in front of each server. Furthermore, in [119] it is proven that various emerging limiting processes are in fact multi-dimensional reflected Brownian motions. In [120], Nguyen extends this result to a fork-join queue with multiple job types. Lu and Pang study fork-join networks in [98, 99, 100]. In [98], they investigate a fork-join network where each service station has multiple servers under non-exchangeable synchronization and operates in the quality-driven regime. They derive functional central limit theorems for the number of tasks waiting in the waiting buffers for synchronization and for the number of synchronized jobs. In [99], they extend this analysis to a fork-join network with a fixed number of service stations, each having many servers, where the system operates in the Halfin-Whitt regime. In [100], the authors investigate these heavy-traffic limits for a fixed number of infinite-server stations, where services are dependent and could be disrupted.

Research has also been done on controlling performance measures in the fork-join queue, see for instance, Atar, Mandelbaum, and Zviran [16], who investigate the control of a fork-join queue in heavy traffic by using feedback procedures. Other studies are carried out in [103, 104].

Specific results on the interplay between fork-join queues and heavy-tailed services can be found in [129, 156, 157]. In [129, Thm. 2], asymptotic lower and upper bounds for the tail probability of the longest waiting time in steady state are given, however, these bounds are not sharp when N is large.

In [156] and [157] the authors investigate the fork-join queue with blocking. More work on fork-join queues with blocking is presented in [42, 43]. In [50], the fork-join queue, under different execution programs is studied. In [149], the mean-value approach is used to approximate performance measures of the fork-join queue. In [150], several bounds for performance measures of the fork-join queue with exponential interarrival and service times are given under a variable number of subtasks. In [146], approximation techniques are derived for the fork-join queue with exponential interarrival and general service times. In [105], a fork-join queueing model is studied where the available computational resources are allocated among the different servers, according to a certain algorithm, with the aim to minimize the

maximum queue length.

1.4. Extreme-value theory

When we are given a sequence of numerical data $(x_1, x_2, \ldots, x_n)$, the first characteristic that we study is the average of the data. The most elementary results on sequences of random variables $(X_1, X_2, \ldots, X_N)$, focus on the behavior of the average. For instance, when all the random variables are independent, identically distributed, and have a finite expectation, then the law of large numbers says that the average of the random variables converges to their expectation as $N \to \infty$. The central limit theorem is the natural extension of this law of large numbers and gives information on the fluctuation of the data around their average. We can use the Berry-Esséen inequality [54] to bound the distance between the cumulative distribution function of the scaled average and the cumulative distribution function of the Gaussian limit. When the moment-generating function exists, Cramér-Lundberg's large deviation principle [13, Thm. 5.2, p. 365] gives the probability that an average of random variables deviates a constant factor from the expectation.

Other than the average behavior, what is often of interest is the behavior of the extremes of the data. As mentioned in Section 1.2, high-tech manufacturers want to predict the delay of their slowest suppliers. Other examples of areas in which the extremes of samples emerge as an important random variable are insurance and water management, as insurance companies want to predict the size of the largest claim [53, Ch. 8.2 & 8.3], and the Dutch Deltawerken are constructed to stop the highest waves [39, p. 54]. As the law of large numbers, the central limit theorem, and the large deviation principle show, the easiest way to get insights on the behavior of sample extremes is by investigating limiting behavior as $N \to \infty$. The area of research called extreme-value theory is about studying limiting behavior of extremes like $\min(X_1, X_2, \ldots, X_N)$ and $\max(X_1, X_2, \ldots, X_N)$. Because a random variable is characterized by its cumulative distribution function, we typically focus on the limiting behavior of these extremes: $\mathbb{P}(\max_{i \leq N} X_i \leq x)$, where we write $\max_{i \leq N} X_i = \max(X_1, X_2, \ldots, X_N)$.

In this section, we give an overview of extreme-value results that we use in this thesis. We focus on one type of extreme, namely the maximum of a sample. In this thesis, we derive properties for the longest waiting time and the maximum queue length in a fork-join queue as N grows large. As Figure 1.1 shows, all service stations observe exactly the same orders. This means that when orders arrive in a deterministic fashion, queue lengths and waiting times corresponding to independently working single servers are independent, but when orders arrive in a stochastic fashion, queue lengths and waiting times corresponding to independently working single servers are dependent. For this reason, we distinguish between results on extremes of independent samples and extremes of dependent samples.

1.4.1 Basic extreme-value theory for independent random variables

When random variables are independent, we can simplify the cumulative distribution function of the sample extreme: because the cumulative distribution function of a maximum

of random variables is a joint probability, we easily see that

$$\mathbb{P}\left(\max_{i\leq N} X_i \leq x\right) = \mathbb{P}(X_1 \leq x, X_2 \leq x, \ldots, X_N \leq x) = \prod_{i=1}^{N} \mathbb{P}(X_i \leq x).$$

When random variables are also identically distributed, we can simplify this further and get

$$\mathbb{P}\left(\max_{i\leq N} X_i \leq x\right) = \prod_{i=1}^{N} \mathbb{P}(X_i \leq x) = \mathbb{P}(X_1 \leq x)^N.$$

Due to this convenient relation, elaborate convergence results for the maximum of N independent and identically distributed (i.i.d.) random variables have been found. For instance, if there exists a sequence $(a_N, N \geq 1)$ such that

$$\max_{i\leq N} X_i - a_N \xrightarrow{\mathbb{P}} 0,$$

as $N \to \infty$, then $\max_{i\leq N} X_i$ is *stable in probability*. A somewhat weaker result is that if,

$$\frac{\max_{i\leq N} X_i}{a_N} \xrightarrow{\mathbb{P}} 1,$$

as $N \to \infty$, then $\max_{i\leq N} X_i$ is *relatively stable in probability*. Necessary and sufficient conditions for these two types of convergence have been given in [64] and in [23, 62] for their almost sure equivalents. Furthermore, in [123], these necessary and sufficient conditions are extended to the convergence of the m-th absolute mean.

In [56], the most fundamental result on the convergence of sample extremes of i.i.d. random variables, often called the *Extremal Types Theorem*, was discussed for the first time, and in [64] a complete proof was given. When we assume that for sequences $(a_N, N \geq 1)$ and $(b_N, N \geq 1)$ the cumulative distribution function of $(\max_{i\leq N} X_i - a_N)/b_N$ has a nondegenerate limit, then, depending on the tail probability of X_1, $(\max_{i\leq N} X_i - a_N)/b_N$ converges to a Fréchet random variable, a Weibull random variable, or a Gumbel random variable; see [67, Thm. 1.2.1, p. 19].

These results have been applied in a variety of settings including the performance of queueing systems. For instance, under some conditions, the tail probability of the steady-state waiting time W of a $GI/GI/1$ queue satisfies $\mathbb{P}(W > u) \sim C\exp(-\gamma u)$ as $u \to \infty$, with $C > 0$, where γ solves the *Lundberg equation*. This is known as the *Cramér-Lundberg approximation* [13, Thm. 5.2, p. 365]. Now, when we consider N parallel and i.i.d. queues, the longest steady-state waiting time among those N queues is relatively stable in probability with $a_N = \log N/\gamma$, and is in the domain of attraction of the Gumbel random variable; see [13, Cor. 5.10, p. 369].

1.4.2 Basic extreme-value theory for dependent random variables

The results presented so far heavily rely on the fact that the random variables involved are independent. Obviously, when this is not the case, the cumulative distribution function of

a sample extreme cannot be written in such a convenient way, and convergence results for dependent extreme values are more specialized.

Several techniques can be used to analyze the limiting behavior of the maximum of N dependent random variables as $N \to \infty$. Though we can in general not express the cumulative distribution function of the sample extreme as a product of the N cumulative distribution functions, we can still find suitable lower and upper bounds. Since we can write the tail probability of $\max_{i \leq N} X_i$ as

$$\mathbb{P}(\max_{i \leq N} X_i \geq x) = \mathbb{P}(\cup_{i=1}^{N} X_i \geq x),$$

we obtain by the union bound and by basic set theory that

$$\sum_{i=1}^{N} \mathbb{P}(X_i \geq x) - \sum_{j=1}^{N} \sum_{k=1, j \neq k}^{N} \mathbb{P}(X_j \geq x, X_k \geq x) \leq \mathbb{P}(\max_{i \leq N} X_i \geq x) \leq \sum_{i=1}^{N} \mathbb{P}(X_i \geq x). \quad (1.1)$$

Both these inequalities are known as *Bonferroni's inequalities* [31]. Clearly, when $\sum_{j=1}^{N} \sum_{k=1, j \neq k}^{N} \mathbb{P}(X_j \geq x, X_k \geq x)$ is small compared to $\sum_{i=1}^{N} \mathbb{P}(X_i \geq x)$, these lower and upper bounds are sharp. In [92], convergence results are derived for the maximum of identically distributed random variables for which the tail probability equals the upper bound whenever the upper bound is smaller than 1; thus $\mathbb{P}(\max_{i \leq N} X_i \geq x) = \min(1, N\,\mathbb{P}(X_i \geq x))$. Other techniques are available to analyze the limiting behavior of specific dependent sample extremes; for instance, *Slepian's lemma* [142] and *Berman's inequality* [126, Thm. C.2, p. 6] give bounds for the cumulative distribution function of the sample extreme of dependent Gaussian random variables.

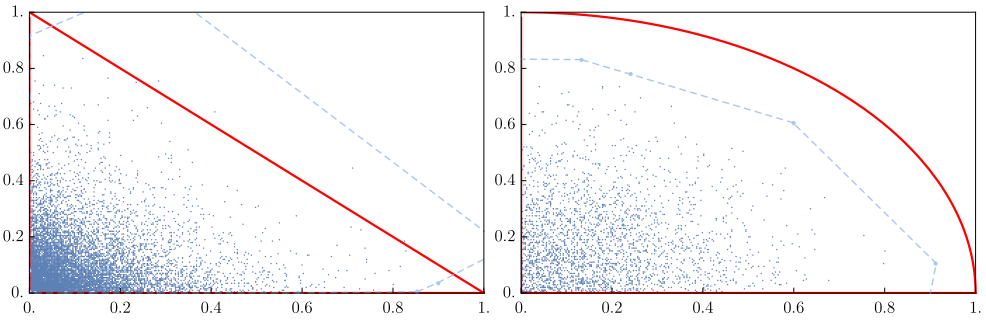
Extensive contributions to the development of independent and dependent extreme-value theory, applications, and statistics, are given in several books; see [24, 53, 67, 94]. We now focus on several specific results that are relevant to this thesis.

1.4.3 Extremes of multidimensional sets

In [20, 44, 55], the focus is broadened from univariate extreme values to multivariate extreme values. In these studies, the convergence of the convex hull of

$$\left\{ \left(X_1^{(1)}/a_N, X_1^{(2)}/a_N, \ldots, X_1^{(d)}/a_N \right), \ldots, \left(X_N^{(1)}/a_N, X_N^{(2)}/a_N, \ldots, X_N^{(d)}/a_N \right) \right\}$$

to a limit as $N \to \infty$ is proven, under several assumptions. In Figure 1.3, we show two examples together with their convex hull; first, the rescaled sample extremes of two i.i.d. exponentially distributed random variables with as limit the triangle between coordinates $(0,0)$, $(1,0)$, and $(0,1)$, and second, the rescaled sample extremes of two i.i.d. normally distributed random variables with as limit the circle with radius 1 and center $(0,0)$.



(a) $a_N = \log N, X^{(1,2)} \sim \text{Exp}(1)$ (b) $a_N = \sqrt{2 \log N}, X^{(1,2)} \sim \mathcal{N}(0, 1)$

Figure 1.3 Two-dimensional sample extremes of i.i.d. random variables in first quadrant, $N = 10^4$ and the convex hull. The dots indicate the two-dimensional samples, the dashed lines indicate the convex hulls, and the straight line and curve indicate the limits of these convex hulls.

In Chapter 2, we show that we can use these results to prove the convergence of sample extremes of finite sums of random variables.

1.4.4 Tail probabilities of growing sums

In [10], the analysis of univariate extreme-value theory is expanded in another direction. While in standard extreme-value theory the objective is the limiting behavior of $\max_{i \leq N} X_i$, one might also be interested in what happens when the distribution of the random variables $(X_i, i \leq N)$ changes with N. Thus, the target is to analyze the sample extremes of sequences $(X_i^{(N)}, i \leq N, N \geq 1)$, which are also known as *triangular arrays*. In this setting, the focus is particularly on triangular arrays of the form $X_i^{(N)} = \sum_{j=1}^{k_N} U_{i,j}$, with $k_N \overset{N \to \infty}{\longrightarrow} \infty$, where $U_{i,j}$ is i.i.d. for all i and j, is centered, and $X_i^{(N)}$ has unit variance. The authors investigate, among others, the case when $U_{i,j}$ has a moment-generating function, and $\log N = o(k_N^{(r+1)/(r+3)})$ for some integer $r \geq 0$. It turns out that the Extremal Types Theorem applies [10, Prop. 2]. Furthermore, the accompanying sequences $(a_N, N \geq 1)$ and $(b_N, N \geq 1)$ are very similar to the sequences belonging to the sample extremes of N i.i.d. Gaussians. This means that the sequence $(k_N, N \geq 1)$ grows fast enough that due to the central limit theorem, the sequence $(X_i^{(N)}, N \geq 1)$ can be replaced with a sequence of standard normally distributed random variables.

Obviously, when this sequence $(k_N, N \geq 1)$ does not grow fast enough, one cannot draw these conclusions. However, for general i.i.d. triangular arrays, the relation $\mathbb{P}(\max_{i \leq N} X_i^{(N)} \leq x) = \mathbb{P}(X_i^{(N)} \leq x)^N = (1 - \mathbb{P}(X_i^{(N)} > x))^N$ still holds. Therefore, the limiting behavior of the sample extremes of i.i.d. triangular arrays is determined by the tail probability of a single random variable $X_i^{(N)}$. A lot of research has been done on the tail probabilities of growing sums of i.i.d. random variables, which we discuss in the remaining part of this section.

One of the fundamental results says that if *Cramér's condition* is satisfied, i.e., the moment-generating function of $U_{i,j}$ is finite on an interval around the origin, then if $x_N/\sqrt{k_N} \overset{N\to\infty}{\longrightarrow} 0$,

$$\mathbb{P}\left(\frac{1}{\sqrt{k_N \mathrm{Var}(U_{i,j})}}\sum_{j=1}^{k_N}(U_{i,j}-\mathbb{E}[U_{i,j}]) > x_N\right) = (1-\Phi(x_N))\exp\left(\frac{x_N^3}{\sqrt{k_N}}\lambda(x_N/\sqrt{k_N})\right)(1+O(x_N/\sqrt{k_N})),$$

with the *Cramér series* $\lambda(t) = \sum_{k=0}^{\infty} c_k t^k$, and $(c_k, k \geq 1)$ depending on the moments of $U_{i,j}$ [40] and [122, p. 218]. In [59, 113, 114], bounds on the tail probability of sums of independent random variables are given, where those random variables satisfy Cramér's condition, which are applicable when $x_N/\sqrt{k_N}$ converges to ∞ as $N \to \infty$. For the case where $x_N/\sqrt{k_N} \overset{N\to\infty}{\longrightarrow} c$ with $c > 0$, *Cramér's theorem* [40] states that

$$\frac{-\log(\mathbb{P}(\sum_{j=1}^{k_N}(U_{i,j}-\mathbb{E}[U_{i,j}] > ck_N))}{k_N} \overset{N\to\infty}{\longrightarrow} \sup_{s>0}(cs - \log \mathbb{E}[\exp(s(U_{i,j}-\mathbb{E}[U_{i,j}])]).$$

Another line of research focuses on the tail behavior of sums of random variables for which Cramér's condition is violated. In [35, 111, 112], logarithmic and exact asymptotics of tail probabilities $\mathbb{P}(X_i^{(N)} \geq x_N)$ of sums of independent Weibull-like distributed random variables are analyzed, with $k_N, x_N \overset{N\to\infty}{\longrightarrow} \infty$ for particular choices of the sequences $(k_N, N \geq 1)$ and $(x_N, N \geq 1)$. Typical results are of the form: when the sequence $(k_N, N \geq 1)$ grows fast enough compared to $(x_N, N \geq 1)$, then, as in [10, Prop. 2], the asymptotics are the same as the asymptotics of the Gaussian tail probability [35, Thm. 1]. When the sequence $(k_N, N \geq 1)$ grows at a slower pace compared to $(x_N, N \geq 1)$, the asymptotics are the same as the asymptotics of the largest of the Weibull-like distributed random variables [35, Thm. 2] with some non-trivial behavior at the transition between these two regimes [35, Thm. 3]. In [110, Prop. 3.1], these results are extended to lognormal-type and regularly varying random variables.

These insights can be used for the analysis of the tail probabilities of steady-state waiting times of $GI/GI/1$ queues. As mentioned in Section 1.3.1, we know that the steady-state waiting time of a $GI/GI/1$ queue can be written as $W \overset{d}{=} \sup_{k\geq 0}\sum_{j=1}^{k}(S(j)-A(j))$, with $(S(j), j \geq 1)$ the service times and $(A(j), j \geq 1)$ the interarrival times. As mentioned earlier, when the moment-generating function exists and γ solves the Lundberg equation, then, if the system is stable, the Cramér-Lundberg approximation states that $\mathbb{P}(W > u) \sim C\exp(-\gamma u)$ as $u \to \infty$, with $C > 0$. This is connected to the analysis of tail probabilities of growing sums through the notion that the tail probability of the all-time supremum of a random walk

can be approximated by the tail probability of a growing sum;

$$\mathbb{P}\left(\sup_{k\geq 0}\sum_{j=1}^{k}(S(j)-A(j))>x_N\right)\approx\sup_{c\geq 0}\mathbb{P}\left(\sum_{j=1}^{cx_N}(S(j)-A(j))>x_N\right),$$

when $(S(j), j \geq 1)$ is light-tailed, and after applying Cramér's theorem the latter probability can be maximized over c.

In the case that service times are not light-tailed, Cramér's theorem cannot be applied. However, in [38, Thm. 1], it is shown that in the asymptotic regime, the survival probability of the steady-state waiting time can be expressed as a function of the survival probability of the service times in a $GI/GI/1$ queue, where the service times are regularly varying. In [89, Thm. 1], a similar result is given for the $GI/GI/1$ queue with subexponential service times. An overview of results on heavy-tailed phenomena can be found in [116].

1.4.5 Suprema of continuous-time stochastic processes

The analysis of suprema of stochastic processes is both connected with the extreme-value theory of sequences of random variables, and with the analysis of performance measures of queueing systems. As mentioned in Section 1.3.1, it is shown [1, 71] that when the $GI/GI/1$ queue is in heavy traffic, the waiting time and queue length can be approximated by a reflected Brownian motion. The reflected Brownian motion can be written as a supremum of a Brownian motion; $\sup_{s\in[0,t]}((B(t)-\mu t)-(B(s)-\mu s))$. As an extension of the reflected Brownian motion, which appears as the process limit of queue lengths and waiting times in heavy traffic, other continuous-time Gaussian processes and their suprema are studied in the literature. In [126], several methods are described which can be used in the analysis of these suprema. For instance, assume that for $k \in \mathbb{N}$, $t_0 = 0$, $t_k = t$, and for $i \in \{0, 1, \ldots, k-1\}$, $t_{i+1} > t_i$, then, by using Bonferroni's inequality given in (1.1), we obtain that

$$\sum_{i=1}^{k}\mathbb{P}\left(\sup_{s\in[t_{i-1},t_i]}X(s)>x\right)-\sum_{i\neq j}\mathbb{P}\left(\sup_{s\in[t_{i-1},t_i]}X(s)>x,\ \sup_{s\in[t_{j-1},t_j]}X(s)>x\right)$$
$$\leq\mathbb{P}\left(\sup_{s\in[0,t]}X(s)>x\right)\leq\sum_{i=1}^{k}\mathbb{P}\left(\sup_{s\in[t_{i-1},t_i]}X(s)>x\right).$$

Now, when the term $\sum_{i\neq j}\mathbb{P}(\sup_{s\in[t_{i-1},t_i]}X(s)>x,\sup_{s\in[t_{j-1},t_j]}X(s)>x)$ is small compared to $\sum_{i=1}^{k}\mathbb{P}(\sup_{s\in[t_{i-1},t_i]}X(s)>x)$, we obtain a sharp estimate of $\mathbb{P}(\sup_{s\in[0,t]}X(s)>x)$. Usually, this method is applied to find convergence results when $x\to\infty$. This method is called the *Double Sum Method* [126, p. 97–135]. Another result [4, Thm. 2.1.1, p. 50] gives an upper bound on the tail probability of a centered Gaussian process that is a.s. bounded on $[0, T]$:

$$\mathbb{P}\left(\sup_{s\in[0,t]}X(s)-\mathbb{E}\left[\sup_{s\in[0,t]}X(s)\right]>u\right)\leq\exp\left(-u^2/(2\sigma_T^2)\right),$$

with $\sigma_T^2 = \mathbb{E}[\sup_{s\in[0,t]} X(s)^2]$; this result is called the *Borell-TIS inequality*. Exact asymptotics are known as well: in [124, 125] it is shown that for centered stationary Gaussian processes with covariance $R(s) = 1 - |s|^\alpha + o(|s|^\alpha)$ as $s \to 0$, $\alpha \in (0, 2]$ and $R(s) < 1$ for all $s > 0$, it holds that

$$\mathbb{P}\left(\sup_{s\in[0,t]} X(s) > u\right) \sim \mathcal{H} t u^{2/\alpha}(1 - \Phi(u)),$$

as $u \to \infty$, with $\mathcal{H}$ the *Pickands constant*. In [4, 94, 126], a general overview of results on suprema of continuous-time processes is given. A specific result that is relevant to this thesis is found in [47]. The authors investigate the joint tail probability of the all-time suprema of two dependent Brownian motions. In the case that $(B_i(t), t \geq 0)$, $(B_j(t), t \geq 0)$, and $(B_A(t), t \geq 0)$ are independent Brownian motions with standard deviations σ, σ, and σ_A respectively, the result in [47, Thm. 2.3] simplifies to

$$\mathbb{P}\left(\sup_{s>0}(B_i(s) + B_A(s) - \beta s) > u, \sup_{s>0}(B_j(s) + B_A(s) - \beta s) > u\right) \sim \frac{\tilde{\mathcal{H}}}{2\sqrt{\pi\beta\sigma^2/(\sigma^2 + \sigma_A^2)}} u^{-1/2} \exp\left(-\frac{2\beta}{\sigma^2/2 + \sigma_A^2} u\right),$$

as $u \to \infty$, with $\tilde{\mathcal{H}}$ a constant similar to the Pickands constant. Together with Bonferroni's inequality, these asymptotics can be used to prove convergence of the sample extremes of N dependent Brownian motions as $N \to \infty$. We use this technique in Chapter 4.

1.5. Main contributions and outline

In this thesis, we study the fork-join queueing system. We focus on the fork-join queue with the following properties; first, we consider the fork-join queue with N service stations; second, each service station operates as a single server; third, we consider incoming jobs that are forked in N subtasks. Thus, the number of subtasks equals the number of servers. Finally, we do not consider blocking mechanisms, preemption, abandonment, etc. Now, we focus on the performance of the slowest server in the fork-join queue, as the slowest server represents the bottleneck in complex systems such as high-tech supply chains and parallel computing. Furthermore, we investigate this performance as the number of servers N becomes large.

In the literature, attention is given to performance measures for fork-join queueing systems with more than two servers, by developing bounds, approximations, and other techniques, which we discussed in Section 1.3.2. However, the existing literature lacks exact asymptotic results for the performance of the slowest server as $N \to \infty$. Below we give an overview of the contribution of this thesis to the existing literature.

1.5.1 Extreme-value results

As shown in Section 1.3.1, we can express the waiting time of a customer in a $GI/GI/1$ queue with the FCFS policy as the supremum of a random walk. Our main contribution is that

we obtain convergence results for the maximum queue length and the longest waiting time under the FCFS policy in the fork-join queue with N service stations, where each service station is an independently working single server. We therefore mainly investigate random variables that are of the form:

$$\max_{i\leq N} \sup_{k\geq 0} \sum_{j=1}^{k} (S_i(j) - A(j)).$$

This random variable has the same distribution as the longest steady-state waiting time in a fork-join queue with N service stations, where each service station works as a single-server queue and the sequences $(S_i(j), i \geq 1, j \geq 1)$ and $(A(j), j \geq 1)$ indicate the service times of subtask j at service station i and the interarrival times, respectively. Expressions for the maximum queue length and transient performance measures look similar.

In Chapter 2, we obtain relative stability or *first-order convergence* results for the maximum queue length as $N \to \infty$. Thus we show that $\max_{i\leq N} Q_i/a_N$ has a non-trivial limit as $N \to \infty$ with $(a_N, N \geq 1)$ appropriately chosen. In this chapter, we look at a particular setting in discrete time: arrivals and services are nearly deterministic, and the initial number of jobs in the queue satisfies the relative stability requirement; see Theorem 2.1. Because the system operates in heavy traffic, we see that we can approximate the separate queue lengths with reflected Brownian motions; see Section 2.2.2 for more details. One major observation that we make is that the first-order convergence result for the fork-join queue is the same as when we would replace the interarrival times with deterministic interarrival times equal to the first moment. Thus, we can conclude that the lower bound for the performance measures of the fork-join queue proven in [18, Cor. 3.4] is asymptotically sharp as $N \to \infty$, when we normalize the maximum queue lengths.

In Chapter 3, we prove a *second-order convergence* result for the longest waiting time and the maximum queue length. We find non-trivial limits for $(\max_{i\leq N} W_i - a_N^{(w)})/b_N^{(w)}$ and $(\max_{i\leq N} Q_i - a_N^{(q)})/b_N^{(q)}$ as $N \to \infty$. Thus, $a_N^{(w)}$ and $a_N^{(q)}$ indicate the typical size of the longest waiting time and maximum queue length, respectively, and $b_N^{(w)}$ and $b_N^{(q)}$ indicate the second-order terms. We obtain that these sequences of random variables converge in distribution to a normally distributed random variable; see Theorems 3.1 and 3.2. The standard deviation of the limiting distribution depends on the standard deviation of the arrival process and the most likely hitting time of a random walk to its supremum, which depends on the cumulant-generating function of the service process, and the solution of the Lundberg equation, which we explain in more detail in Section 3.3. We can conclude that in order to obtain a second-order convergence result of the longest waiting time and the maximum queue length, only the first and second moments of the arrival distribution need to be known. We subsequently show in Corollary 3.1 that the second-order convergence result of these quantities can be extended to fork-join queues where service times are not identically distributed but there is still some regulation among the different servers.

As we see in Chapters 2 and 3, in order to obtain first- and second-order convergence results, it suffices to know the first and second moments of the arrival process. In Chapter 4, we investigate the large deviations of the maximum queue length in the N-server fork-join

queue. Thus, having focused in Chapters 2 and 3 on the case $\max_{i\leq N} Q_i/a_N \xrightarrow{\mathbb{P}} c > 0$ as $N \to \infty$, we now look at $\mathbb{P}(\max_{i\leq N} Q_i > (c+x)a_N)$ with $x > 0$ as $N \to \infty$. It becomes clear that the distribution of the arrival process has a significant influence on these large deviations; see Theorem 4.1. In Chapter 4, we restrict ourselves to reflected Brownian motions. Our analysis relies heavily on the convergence of asymptotics of joint suprema of Brownian motions in [47]. We believe, however, that these results can be extended to the analysis of maximum queue lengths and longest waiting times as defined in Chapter 3.

For the results in Chapters 2–4, we need to use that the service times have light tails so that we can use the Cramér-Lundberg approximation. In Chapter 5, we drop this assumption and focus on heavy-tailed service times. We use results from extreme-value theory as well as results from the analysis of heavy tails, e.g. [35] and [67]. We analyze the random variable $\max_{i\leq N} W_i/a_N$ as $N \to \infty$. However, in contrast with Chapter 2, we now obtain a non-deterministic limit; see Theorems 5.2 and 5.3. Its distribution can be written as a supremum of a stochastic process with Fréchet-distributed marginals.

1.5.2 Process convergence

Besides looking at steady-state convergence of the maximum queue length and the longest waiting time, we also obtain process convergence results. In Chapter 2, we use techniques from diffusion approximations and induce a temporal and a spatial scaling; see Section 2.2.2. We then obtain convergence results for the process $(\max_{i\leq N} Q_i(tc_N)/a_N, t \in [0,T])$ to a limiting process $(q(t), t \in [0,T])$ in Proposition 2.1 and Theorem 2.1. We prove convergence of the finite-dimensional distributions in Theorem 2.3 and we prove tightness of $(\max_{i\leq N} Q_i(tc_N)/a_N, t \in [0,T])$ in order to prove convergence in $C[0,T]$. To achieve this, we use results from [28, Ch. 2] on convergence of stochastic processes in $C[0,T]$.

We also extend this result to queueing systems with non-empty queues at time 0. In order to achieve this, we prove in Lemma 2.13 a result of independent interest, which is the convergence of $\max_{i\leq N}(X_i^{(1)}/a_N^{(1)} + X_i^{(2)}/a_N^{(2)})$, where $X^{(1)}$ is relatively stable with sequence $(a_N^{(1)}, N \geq 1)$ and $X^{(2)}$ is relatively stable with sequence $(a_N^{(2)}, N \geq 1)$. Because deriving the cumulative distribution function of the sum of two random variables is usually difficult, deriving extreme-value results of N of these sums by using properties of the resulting tail probability, is also difficult. However, in Lemma 2.13, we prove that we only need to know the cumulative distribution functions of the two random variables to obtain an extreme-value limit.

The limiting process $(q(t), t \in [0,T])$ of the transient maximum queue length is deterministic. In Chapter 5, we prove process convergence of the longest waiting time, but now the service times are heavy-tailed. The resulting limiting process is still stochastic, is an *extremal process* minus a drift term [133], and is a function in $D[0,T]$. We use techniques described in [28, Ch. 3] to prove this result.

1.5.3 Applications to assembly systems

In Chapter 6, we use the obtained convergence results to analyze large-scale assembly systems. We focus on an *assemble-to-order* system, where incoming orders from a manufacturer are sent to each of the N suppliers at the same time. Finished components are stored in a warehouse, and are assembled when all suppliers have finished their components. This is visualized in Figure 1.4.

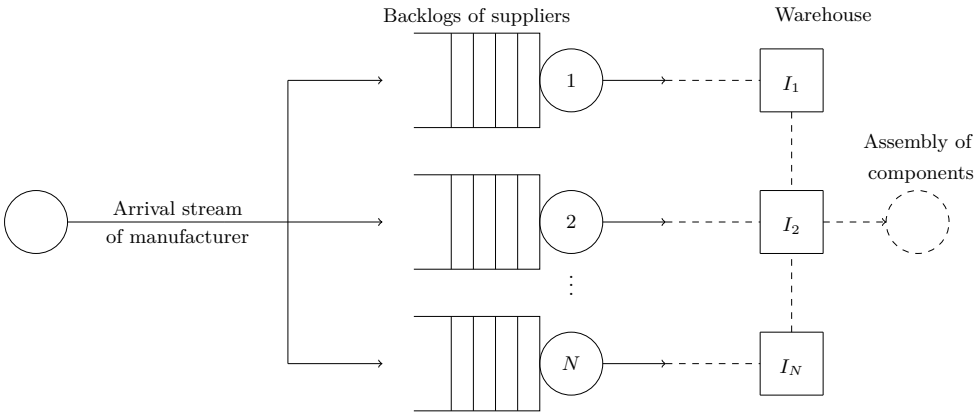


Figure 1.4 Fork-join queue with inventories

There are three types of costs made in this system: backlog costs; inventory costs, and capacity costs. The backlog costs in this system depend on the maximum delay among the N suppliers. The suppliers set up a base stock of components and choose a capacity. The aim is to minimize the resulting costs in the system over these base stocks and capacities. We do so by solving a newsvendor problem over $2N$ variables. We use the first- and second-order convergence results derived in Chapters 2 and 3 to derive an approximation of the cost function and give solutions that are asymptotically optimal as $N \to \infty$.

1.5.4 Summary

In this thesis, we focus on the maximum queue length and the longest waiting time in the N-server fork-join queue. All service stations operate as independent single servers but with coupled arrivals. We prove first- and second-order convergence results, steady-state results, and results for the system in a transient setting. We explore different interarrival and service distributions: we study a nearly deterministic setting, the general light-tailed setting, and a heavy-tailed setting. We derive tail asymptotics for the Brownian steady-state maximum queue length. Finally, we use the first- and second-order convergence result to provide a decision rule for a newsvendor problem in a large-scale assembly system, which is close to optimal and easy to compute.

Chapter 2

Nearly deterministic arrivals and service times

2.1. Introduction

In this chapter, we explore a discrete-time fork-join queue in which the arrival and service times are nearly deterministic. We consider a heavily loaded system. That is, we assume that the arrival rate to a queue times the expected service time of that queue, i.e., the traffic intensity per queue ρ_N, depends on the number of servers N and satisfies $(1-\rho_N)N^2 \overset{N\to\infty}{\longrightarrow} \beta$, with $\beta > 0$. Our main result is a fluid limit of the maximum queue length process as N goes to infinity, which holds under very mild conditions on the distribution of the number of jobs at time 0.

Our work adds to the literature on queueing systems with nearly deterministic arrivals and services. The only research line on queueing systems with nearly deterministic service times that we are aware of is Sigman and Whitt [139, 140], who investigate the $G/G/1$ and $G/D/N$ queues and establish heavy-traffic results for waiting times, queue lengths and other performance measures in stationarity as well as functional central limit theorems for the waiting time and for other performance measures. In these papers, they distinguish two cases, one in which $(1-\rho_N)\sqrt{N} \overset{N\to\infty}{\longrightarrow} \beta$ and one in which $(1-\rho_N)N \overset{N\to\infty}{\longrightarrow} \beta$, with ρ_N the traffic intensity and β some constant.

Our work also contributes to the literature on the process-level analysis of fork-join networks surveyed in Section 1.3.2 of this thesis. In particular, we derive a fluid limit of the stochastic process that keeps track of the largest queue length. This study seems to be the first explicit process-level approximation of a large fork-join queue.

We now turn to an overview of the techniques that we use in this chapter. Because we aim to obtain a fluid limit of a maximum of N queue lengths, we mainly use techniques from extreme-value theory in our proofs. This is, however, quite a challenge: while the queue lengths of the servers are mutually dependent, most results on extreme values hinge heavily on the assumption of mutual independence. Furthermore, we consider a fork-join queue in

This chapter is based on [134].

which the arrival and service probabilities depend on N, which makes the queue lengths triangular arrays with respect to N. This makes the analysis also rather uncommon as studies on triangular arrays are rare. One paper on this subject, relevant to us, is Anderson et al. [10], in which they study the maximum of a sum of i.i.d. triangular arrays; see also the discussion on [10] in Section 1.4.4.

In addition, in order to get fluid limits for the maximum queue lengths, we actually need to study diffusion limits for the individual queue lengths. We, thus, combine ideas from the literature on extreme-value theory with literature on diffusion approximations, which we show in detail in Section 2.2.2. Because we impose a heavy-traffic assumption, we obtain for each separate queue length a reflected Brownian motion as diffusion approximation. By using the well-known formula for the cumulative distribution function of a reflected Brownian motion (see Harrison [71, p. 49]), we investigate the maximum of N independent reflected Brownian motions to get an idea of the scaling of the maximum queue length.

Now, we give a brief sketch of how we apply these ideas to prove the fluid limit of this quantity. We start by considering the slightly simpler scenario that each queue is empty at time 0. Because we want to prove a fluid limit that holds uniformly on compact intervals (u.o.c.), we need to prove pointwise convergence and tightness of the collection of processes. Our first step in proving this is by showing that each queue length is in distribution the same as a supremum of an arrival process minus a service process. We then show in Section 2.2.2 that, under the temporal scaling of $tN^3 \log N$ and the spatial scaling of $N \log N$, the arrival process minus a drift term converges to $-\beta t$ as $N \to \infty$. Furthermore, we derive, under that same temporal scaling but under a spatial scaling of $N\sqrt{\log N}$, that the centralized service process satisfies the central limit theorem. This scaled centralized service process is given in Equation (2.3.4). We use the non-uniform Berry-Esséen inequality, which is described by Michel [109], to deduce the convergence rate of the cumulative distribution function of this scaled centralized service process to the cumulative distribution function of a normally distributed random variable, which is given in Equation (2.3.9). It turns out that this convergence rate is fast enough so that we can replace the scaled centralized service process with a normally distributed random variable in the expression for the maximum queue length in order to get the same limit. By Pickands' result [123] on the convergence of moments of the maximum of N scaled random variables, we know that the expectation of the maximum of standard normally distributed random variables divided by $\sqrt{\log N}$ converges to $\sqrt{2}$ as $N \to \infty$. This gives the convergence of the maximum of N scaled centralized service processes. After we have obtained these limiting results for the scaled arrival and service process, we use these, together with Doob's maximal submartingale inequality, to prove convergence in probability of the maximum queue length; we show this in Section 2.3.3. Finally, in Section 2.3.4, we use Doob's maximal submartingale inequality to bound the probability that the process makes large jumps and prove that this probability is small so that the maximum queue length is a tight process.

After we have considered the maximum queue length for the process with empty queues at time 0, we then turn to the scenario that the length of each queue at time 0 is identically distributed. In this case, we can use Lindley's recursion to express the maximum queue

length as the pairwise maximum of the maximum queue length with empty queues at time 0, which we now call term 1, and a part depending on the number of jobs at time 0, which we now call term 2; this formula is given in Equation (2.2.3) below. How to prove the fluid limit for term 1 is already sketched. In order to derive a fluid limit for term 2, we first observe that term 2 equals the maximum of N times the sum of the number of jobs at time 0 at each server plus the number of arrivals minus the number of services at each server. Following a similar path as for term 1, we can prove that the scaled centralized service process at a server behaves like a normally distributed random variable. Thus, we have to analyze a maximum of N pairwise sums of normally distributed random variables and random variables describing the number of jobs at time 0, which is stated in more detail in Lemma 2.9.

In Lemma 2.4, we prove a convergence result of this maximum. This is quite a challenge because we need to apply extreme-value theory to pairwise sums. In order to do this, we further develop the results from Davis, Mulrow, and Resnick [44] and Fisher [55] on the convergence of samples of random variables to limiting sets. The authors prove convergence results of the convex hull of $\{(Z_i^{(1)}/b_N, \ldots, Z_i^{(k)}/b_N)_{i\leq N}\}$ to a limiting set as $N \to \infty$, with $(Z_i^{(j)}, i \leq N)$ i.i.d., $Z_i^{(j)}$ and $Z_m^{(l)}$ independent and $(b_N, N \geq 1)$ a proper scaling sequence. We show in the proof of Lemma 2.13 that these results can be extended by establishing convergence of extreme values of $\max_{i\leq N} \sum_{j=1}^{k} Z_i^{(j)}/a_N^{(j)}$, where $a_N^{(l)}$ and $a_N^{(m)}$ are not necessarily the same, which is a stand-alone result of independent interest. We did not find this extension in other literature. The result in Lemma 2.4 follows from Lemma 2.13.

The rest of the chapter is organized as follows. In Section 2.2, we describe the fork-join system in more detail; we give a definition of the arrival and service processes, and we present a scaled version of the queueing model. In Section 2.2.1, we introduce the fluid limit and explain it heuristically. We elaborate on the scaling and the shape of the fluid limit in Sections 2.2.2 and 2.2.3. We give some examples and numerical results in Section 2.2.4. The proof of the fluid limit is given in Section 2.3. In Section 2.4, we elaborate on the convergence of the upper bound that was given in Lemma 2.7. In Section 2.5, we prove the lemmas stated in Section 2.3.2. We prove general convergence results of $\max_{i\leq N} \sum_{j=1}^{k} Z_i^{(j)}/a_N^{(j)}$ in Section 2.6, where we also give some numerical examples and provide a short discussion on the sufficient conditions for which these convergence results hold. Finally, we briefly discuss the nearly deterministic fork-join queue with a different parameter setting in Section 2.8.

2.2. Model description and main results

We now turn to a formal definition of the fork-join queue that we study. We consider a fork-join queue with integer-valued arrivals and services. In this queueing system, there is one arrival process. The arriving tasks are divided into N subtasks, which are completed by N servers. We assume that both the number of arrivals and services per time step are Bernoulli distributed. The parameters of the Bernoulli random variables depend on the number of servers. This is formalized in Definitions 2.1 and 2.2.

Definition 2.1 (Arrival process)**.** *The random variable* $\mathbf{N}_A^{(N)}(n)$ *indicates the number of arrivals up to time* n *and equals*

$$\mathbf{N}_A^{(N)}(n) := \sum_{j=1}^{\lfloor n \rfloor} X^{(N)}(j)$$

with $X^{(N)}(j)$ *indicating whether or not there is an arrival at time* j*. The random variable* $X^{(N)}(j)$ *is Bernoulli distributed with parameter* $p^{(N)}$*. So,*

$$X^{(N)}(j) := \begin{cases} 1 & \text{w.p. } p^{(N)}, \\ 0 & \text{w.p. } 1 - p^{(N)}. \end{cases}$$

Definition 2.2 (Service process i-th server)**.** *The random variable* $\mathbf{N}_{S,i}^{(N)}(n)$ *describes the number of potentially completed tasks of the* i*-th server in the fork-join queue at time* n *with*

$$\mathbf{N}_{S,i}^{(N)}(n) := \sum_{j=1}^{\lfloor n \rfloor} Y_i^{(N)}(j),$$

where $Y_i^{(N)}(j)$ *is a Bernoulli random variable with parameter* $q^{(N)}$ *indicating whether the* i*-th server completed a service at time* j*.*

$$Y_i^{(N)}(j) := \begin{cases} 1 & \text{w.p. } q^{(N)}, \\ 0 & \text{w.p. } 1 - q^{(N)}. \end{cases}$$

Both $p^{(N)}$ *and* $q^{(N)}$ *are taken as functions of* N*, which we specify in Definition 2.3 below.*

We assume that for all $N \geq 1$ the random variables $(X^{(N)}(j), j \geq 1)$ are mutually independent for all j and $(Y_i^{(N)}(j), i \geq 1, j \geq 1)$ are mutually independent for all j and i. We also assume that an incoming task can be completed in the same time slot as in which the task arrived. Finally, we assume that $X^{(N)}(j)$ and $Y_i^{(N)}(j)$ are independent; in other words, $Y_i^{(N)}(j)$ could still be 1 while there are no tasks to be served at server i at time j. Due to this assumption, we have on the one hand the beneficial situation that $(\mathbf{N}_A^{(N)}(n), n \geq 0)$ and $(\mathbf{N}_{S,i}^{(N)}(n), n \geq 0)$ are independent processes, but on the other hand, we should be careful with defining the queue length. However, it is a well-known result that we can use Lindley's recursion [96], and write the queue length of the i-th server at time n as

$$\sup_{0 \leq k \leq n} \left[\left(\mathbf{N}_A^{(N)}(n) - \mathbf{N}_A^{(N)}(k) \right) - \left(\mathbf{N}_{S,i}^{(N)}(n) - \mathbf{N}_{S,i}^{(N)}(k) \right) \right],$$

provided that the queue length is 0 at time 0. This is in distribution equal to

$$\sup_{0 \leq k \leq n} \left(\mathbf{N}_A^{(N)}(k) - \mathbf{N}_{S,i}^{(N)}(k) \right).$$

As can be seen in this expression, the queue lengths of different servers are mutually dependent, since the arrival process is the same. When at time 0 there are already jobs in the queue, then we can, after again applying Lindley's recursion, write the queue length of the i-th server at time n as

$$\max\left(\sup_{0\le k\le n}\left[\left(\mathbf{N}_A^{(N)}(n)-\mathbf{N}_A^{(N)}(k)\right)-\left(\mathbf{N}_{S,i}^{(N)}(n)-\mathbf{N}_{S,i}^{(N)}(k)\right)\right],\right.$$
$$\left.Q_i^{(N)}(0)+\mathbf{N}_A^{(N)}(n)-\mathbf{N}_{S,i}^{(N)}(n)\right),$$

with $Q_i^{(N)}(0)$ the number of jobs in front of the i-th server at time 0. Observe that the queue length of the i-th server equals the maximum of the queue length when the number of jobs at time 0 would be 0, and a random variable that depends on the initial number of jobs.

The aim of this work is to investigate the behavior of the fork-join queue when the number of servers N is very large. The main objective is deriving the distribution of the largest queue, as this represents the slowest server, which is thus the bottleneck of the system. Therefore, we define in Definition 2.3 a random variable indicating the maximum queue length at time n. Furthermore, we explore this model in the heavy-traffic regime. To this end, we let $p^{(N)}$ and $q^{(N)}$ go to 1 at similar rates, so that the arrivals and services are nearly deterministic processes.

Definition 2.3 (Maximum queue length at time n)**.** *Let $p^{(N)}=1-\alpha/N-\beta/N^2$ and $q^{(N)}=1-\alpha/N$, with $\alpha,\beta>0$. Let $Q_{(\alpha,\beta)}^{(N)}(n)$ be the maximum queue length of N parallel servers at time n, with $Q_{(\alpha,\beta)}^{(N)}(0)=0$. Then*

$$Q_{(\alpha,\beta)}^{(N)}(n):=\max_{i\le N}\sup_{0\le k\le n}\left[\left(\mathbf{N}_A^{(N)}(n)-\mathbf{N}_A^{(N)}(k)\right)-\left(\mathbf{N}_{S,i}^{(N)}(n)-\mathbf{N}_{S,i}^{(N)}(k)\right)\right]. \tag{2.2.1}$$

So,

$$Q_{(\alpha,\beta)}^{(N)}(n)\overset{d}{=}\max_{i\le N}\sup_{0\le k\le n}\left(\mathbf{N}_A^{(N)}(k)-\mathbf{N}_{S,i}^{(N)}(k)\right), \tag{2.2.2}$$

under the assumption that $Q_{(\alpha,\beta)}^{(N)}(0)=0$. From these choices of $p^{(N)}$ and $q^{(N)}$, it follows that the traffic intensity ρ_N of a single queue satisfies $(1-\rho_N)N^2\to\beta$ as $N\to\infty$. Furthermore, if $Q_i^{(N)}(0)>0$, the maximum queue length at time n can be written as

$$Q_{(\alpha,\beta)}^{(N)}(n):=\max_{i\le N}\max\left(\sup_{0\le k\le n}\left[\left(\mathbf{N}_A^{(N)}(n)-\mathbf{N}_A^{(N)}(k)\right)-\left(\mathbf{N}_{S,i}^{(N)}(n)-\mathbf{N}_{S,i}^{(N)}(k)\right)\right],\right.$$
$$\left.Q_i^{(N)}(0)+\mathbf{N}_A^{(N)}(n)-\mathbf{N}_{S,i}^{(N)}(n)\right). \tag{2.2.3}$$

Observe that we can interchange the order of the $\max_{i\le N}$ term and the max term, and rewrite the expression in (2.2.3) as the pairwise maximum of two random variables: one random variable is the maximum of N queue lengths with initial condition 0, as given in

Equation (2.2.1), and the other is the maximum of N sums of the queue length at time 0 plus the number of arrivals minus the number of services.

2.2.1 Fluid limit

As we just have formally defined the fork-join queue that we study, with the particular nearly deterministic setting, we now state and explain the main result of this chapter. Our central result is a fluid approximation for the rescaled maximum queue length process, which is given in Theorem 2.1. We prove that under a certain spatial and temporal scaling the maximum queue length converges to a continuous function, which depends on time t.

There is, however, no straightforward procedure for choosing the temporal and spatial scaling. There are namely more possibilities that lead to a non-trivial limit. For instance, when we choose a temporal scaling of N^3 and a spatial scaling of $N\sqrt{\log N}$, we get the fluid limit that is given in Proposition 2.1. For the main result given in Theorem 2.1, we choose a temporal scaling of $N^3 \log N$ and a spatial scaling of $N \log N$.

We now mention and discuss some assumptions under which the results hold. First, we assume that we have nearly deterministic arrivals and services.

Assumption 2.1. *$p^{(N)} = 1 - \alpha/N - \beta/N^2$ and $q^{(N)} = 1 - \alpha/N$, with $\alpha, \beta > 0$.*

Second, we have a basic assumption on the initial condition.

Assumption 2.2. *$(Q_i^{(N)}(0), i \leq N)$ are i.i.d. and non-negative for all N.*

Furthermore, we want to prove a fluid limit with a spatial scaling of $N \log N$. Therefore, we need to assume that the maximum number of jobs at time 0 also scales with $N \log N$. In order to do so, we allow $(Q_i^{(N)}(0), i \leq N, N \geq 1)$ to be a triangular array, i.e., a doubly indexed sequence with $i \leq N$. This is a necessity because otherwise we would be limited to distributions in which the maximum scales like $N \log N$, which would lead us to the family of the heavy-tailed distributions for which we do not have convergence in probability of its maximum. Thus in our setting, $Q_i^{(N)}(0)$ and $Q_i^{(N+1)}(0)$ do not need to be the same. Consequently, we need to have some regularity on $Q_i^{(N)}(0)$ as N increases to be able to prove a limit theorem. Therefore, we introduce a sequence of random variables $(U_i, i \leq N)$, such that $Q_i^{(N)}(0) = \lfloor r_N U_i \rfloor$, with $(r_N, N \geq 1)$. For Theorem 2.1 to hold, the sequence of random variables $(U_i, i \leq N)$ needs to satisfy Assumption 2.3 and either Assumption 2.4 or 2.5.

Assumption 2.3. *$Q^{(N)}_{(\alpha,\beta)}(0)/(N \log N) \xrightarrow{\mathbb{P}} q(0)$, with $q(0) \geq 0$, as $N \to \infty$, with $Q_i^{(N)}(0) = \lfloor r_N U_i \rfloor$, where $(r_N, N \geq 1)$ is a scaling sequence.*

Assumption 2.4. *U_i has a finite right endpoint.*

Assumption 2.5. *U_i is a continuous random variable and for all $v \in [0, 1]$,*

$$\lim_{t \to \infty} \frac{-\log\big(\mathbb{P}(U_i > vt)\big)}{-\log\big(\mathbb{P}(U_i > t)\big)} = h(v).$$

Before stating the results, we would like to give two remarks on Assumption 2.5. First, the function h has the property that for all $u, v \in [0,1]$, $h(uv) = h(u)h(v)$. Thus, if h is continuous, $h(v) = v^a$, with $a > 0$. When h is discontinuous, there are two possibilities: $h(v) = \mathbb{1}(v > 0)$ or $h(v) = \mathbb{1}(v = 1)$; this corresponds to $h(v) = v^a$ with $a = 0$ and $a = \infty$, respectively. Second, the assumption of continuity of U_i may be removed, which would lead to more cumbersome proofs.

In order to allow dependence between the initial number of jobs at different servers, we can also replace Assumptions 2.2 and 2.3 with the following assumption.

Assumption 2.6. *Let $Q_i^{(N)}(0) = U_i^{(N)} + V_i^{(N)}$, with $U_i^{(N)} = \lfloor r_N U_i \rfloor$, where $(U_i, i \leq N)$ are i.i.d. and non-negative, and satisfy either Assumption 2.4 or 2.5. Furthermore, the triangular array of random variables $(V_i^{(N)}, i \leq N, N \geq 1)$ is non-negative, and $\max_{i \leq N} V_i^{(N)}/(N \log N) \xrightarrow{\mathbb{P}} 0$ as $N \to \infty$.*

When Assumption 2.6 is satisfied, there may be mutual dependence between $Q_i^{(N)}(0)$ and $Q_j^{(N)}(0)$, because $V_i^{(N)}$ and $V_j^{(N)}$ may be mutually dependent.

Now, we state three results on the convergence of the maximum queue length as N grows large. First, we give a fluid limit for the maximum queue length with a temporal scaling of N^3 and a spatial scaling of $N\sqrt{\log N}$ in Proposition 2.1. The system is empty at time 0. Second, we give a steady-state result with a temporal scaling of $N \log N$ in Proposition 2.2. Finally, we give a fluid limit for the maximum queue length with a temporal scaling of $N^3 \log N$ and a spatial scaling of $N \log N$ in Theorem 2.1. This system is non-empty at time 0 and satisfies the Assumptions as described above.

Proposition 2.1 (Temporal scaling of N^3 and spatial scaling of $N\sqrt{\log N}$)**.** *Given Assumption 2.1 and $Q_{(\alpha,\beta)}^{(N)}(0) = 0$, we have for all $T > 0$, that*

$$\mathbb{P}\left(\sup_{t \in [0,T]} \left| \frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)}{N\sqrt{\log N}} - \sqrt{2\alpha t} \right| > \epsilon \right) \overset{N \to \infty}{\longrightarrow} 0 \ \forall\ \epsilon > 0.$$

The steady-state limit is given in Proposition 2.2.

Proposition 2.2 (Steady-state convergence)**.** *Given Assumption 2.1, we have*

$$\frac{Q_{(\alpha,\beta)}^{(N)}(\infty)}{N \log N} \xrightarrow{\mathbb{P}} \frac{\alpha}{2\beta},$$

as $N \to \infty$.

As we can see in Proposition 2.2, to obtain a non-trivial steady-state limit, we need a spatial scaling of $N \log N$. Since this is the only choice that leads to a non-trivial limit, it is a natural choice to look for a fluid limit that also has this spatial scaling. Our main result, stated in Theorem 2.1, is such a fluid limit, and it turns out that for establishing this limit, we need a temporal scaling of $N^3 \log N$. In Section 2.2.2, we explain why these scalings are natural. We give the proof of Proposition 2.1 in Section 2.7, and we explain how Proposition

2.1 is connected to Theorem 2.1 at the end of this section. Furthermore, we give a proof of Proposition 2.2 in Section 2.3.

Theorem 2.1 (Fluid limit with a non-zero initial condition)**.** *Given Assumptions 2.1–2.3, and either Assumption 2.4 or Assumption 2.5, then we have for all $T > 0$, that*

$$\mathbb{P}\left(\sup_{t\in[0,T]} \left| \frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3 \log N\right)}{N \log N} - q(t) \right| > \epsilon \right) \overset{N\to\infty}{\longrightarrow} 0 \ \forall\, \epsilon > 0, \tag{2.2.4}$$

with

$$q(t) = \max\left(\left(\sqrt{2\alpha t} - \beta t \right) \mathbb{1}\left(t < \frac{\alpha}{2\beta^2} \right) + \frac{\alpha}{2\beta} \mathbb{1}\left(t \geq \frac{\alpha}{2\beta^2} \right), g(t, q(0)) - \beta t \right). \tag{2.2.5}$$

The function $g(t, q(0))$ has the following properties:

1. *If Assumption 2.4 holds, then*

$$g(t, q(0)) = q(0) + \sqrt{2\alpha t}. \tag{2.2.6}$$

2. *If Assumption 2.5 holds, then*

$$g(t, q(0)) = \sup_{(u,v)} \{ \sqrt{2\alpha t} u + q(0) v : u^2 + h(v) \leq 1, 0 \leq u \leq 1, 0 \leq v \leq 1 \}. \tag{2.2.7}$$

As can be seen in Theorem 2.1, the fluid limit has an unusual form, $q(t)$ is namely a maximum of two functions. The first part of this maximum is the fluid limit when the initial number of jobs equals 0 and the second part is caused by the initial number of jobs. We elaborate on this more in Section 2.2.3. The $\log N$ term in the spatial and temporal scaling of the process is also unusual. We show in Section 2.2.2 that this is due to the fact that we take a maximum of N random variables, with N large. Scaling terms like $(\log N)^c$ are in this context very natural.

We mentioned earlier that different choices for temporal and spatial scalings lead to a non-trivial fluid limit. We gave Proposition 2.1 as an example. Since we analyze one and only one system, the two fluid limits that we presented should be connected to each other. An easy way to see this is by observing that from Theorem 2.1 it follows that when $Q^{(N)}_{(\alpha,\beta)}(0) = 0$,

$$\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3 \log N\right)}{N \log N} \overset{\mathbb{P}}{\longrightarrow} \sqrt{2\alpha t} - \beta t,$$

as $N \to \infty$, for $t < \alpha/(2\beta^2)$. Thus, for all $t > 0$ and for N large, we expect that $Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)/(N\sqrt{\log N}) \approx \sqrt{2\alpha t} - \beta t/\sqrt{\log N} \overset{N\to\infty}{\longrightarrow} \sqrt{2\alpha t}$. This shows heuristically how Proposition 2.1 is connected with Theorem 2.1. The proof of Proposition 2.1 is given in Section 2.7.

2.2.2 Scaling

In Section 2.2.1, we presented the fluid limit under the rather unusual temporal scaling of $N^3 \log N$ and spatial scaling of $N \log N$. A heuristic justification for these scalings can be given by using extreme-value theory and ideas from literature on diffusion approximations. In particular, for the spatial scaling, we argue as follows: as we are interested in the convergence of the maximum queue length, we can use a central limit result to replace each separate queue length with a reflected Brownian motion and use extreme-value theory to get a heuristic idea of the convergence of the scaled maximum queue length. To argue this, first observe that the arrival and service processes are binomially distributed random variables, and we can compute the expectation and variance of the random walk $\left(\mathbf{N}_A^{(N)}\left(tN^3 \log N\right) - \mathbf{N}_{S,i}^{(N)}\left(tN^3 \log N\right)\right) \Big/ (N\sqrt{\log N})$ given in (2.2.2) as

$$\mathbb{E}\left[\frac{1}{N\sqrt{\log N}}\left(\mathbf{N}_A^{(N)}\left(tN^3 \log N\right) - \mathbf{N}_{S,i}^{(N)}\left(tN^3 \log N\right)\right)\right] = -\beta t\sqrt{\log N} + o(1), \tag{2.2.8}$$

and

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{N\sqrt{\log N}}\left(\mathbf{N}_A^{(N)}\left(tN^3 \log N\right) - \mathbf{N}_{S,i}^{(N)}\left(tN^3 \log N\right)\right)\right) \\
&\quad = \frac{1}{N^2 \log N}\lfloor tN^3 \log N \rfloor \left(\left(\frac{\alpha}{N} + \frac{\beta}{N^2}\right)\left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2}\right) + \frac{\alpha}{N}\left(1 - \frac{\alpha}{N}\right)\right) \\
&\quad = 2\alpha t + o(1).
\end{aligned} \tag{2.2.9}$$

From this, a non-trivial scaling limit can be easily deduced: observe that $\mathbf{N}_A^{(N)}\left(tN^3 \log N\right) - \mathbf{N}_{S,i}^{(N)}\left(tN^3 \log N\right)$ is a sum of independent and identically distributed random variables, so this implies that

$$\frac{1}{N\sqrt{\log N}}\left(\mathbf{N}_A^{(N)}\left(tN^3 \log N\right) - \mathbf{N}_{S,i}^{(N)}\left(tN^3 \log N\right)\right) \overset{d}{\approx} Z_i,$$

as N is large, with $Z_i \sim \mathcal{N}\left(-\beta t\sqrt{\log N}, 2\alpha t\right)$. Furthermore, because $\mathbf{N}_A^{(N)}\left(tN^3 \log N\right) - \mathbf{N}_{S,i}^{(N)}\left(tN^3 \log N\right)$ is, in fact, the difference of two random walks, we also have

$$\sup_{0 \le n \le tN^3 \log N} \frac{1}{N\sqrt{\log N}}\left(\mathbf{N}_A^{(N)}(n) - \mathbf{N}_{S,i}^{(N)}(n)\right) \overset{d}{\approx} R_i(t),$$

as N is large, with $R_i(t)$ a reflected Brownian motion for t fixed. We can apply extreme-value theory to show that $\max_{i \le N} R_i(t)$ scales with $\sqrt{\log N}$. This can be deduced from the cumulative distribution function of the reflected Brownian motion which is given in [71, p. 49]. Concluding, the proper spatial scaling of the fluid limit in Theorem 2.1 is $N \log N$.

2.2.3 Shape of the fluid limit

In Section 2.2.2, we gave a heuristic explanation of the temporal and spatial scaling of the process. Here, we do the same for the shape of the fluid limit. First, we rewrite the expression in (2.2.3) and get that the scaled maximum queue length satisfies

$$\begin{aligned}\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\log N\right)}{N\log N} = \\ \max\Bigg(\max_{i\le N}\sup_{s\in[0,t]}\Bigg(&\frac{\left(\mathbf{N}^{(N)}_A\left(tN^3\log N\right)-\mathbf{N}^{(N)}_A\left(sN^3\log N\right)\right)}{N\log N} \\ &-\frac{\left(\mathbf{N}^{(N)}_{S,i}\left(tN^3\log N\right)-\mathbf{N}^{(N)}_{S,i}\left(sN^3\log N\right)\right)}{N\log N}\Bigg), \\ &\max_{i\le N}\frac{\mathbf{N}^{(N)}_A\left(tN^3\log N\right)+\mathbf{N}^{(N)}_{S,i}\left(tN^3\log N\right)+Q^{(N)}_i(0)}{N\log N}\Bigg).\end{aligned} \tag{2.2.10}$$

Now, observe that when $Q^{(N)}_i(0)=0$ for all i, the pairwise maximum in (2.2.10) simplifies to the first part of the maximum. It turns out that the first and the second part of this maximum converge to the first and second part of the maximum in (2.2.5), respectively. To see the first limit heuristically, observe that due to the central limit theorem,

$$\frac{1}{N\sqrt{\log N}}\left(\mathbf{N}^{(N)}_A\left(tN^3\log N\right)-\mathbf{N}^{(N)}_{S,i}\left(tN^3\log N\right)\right)\overset{d}{\approx}\vartheta_i+\zeta,$$

with $\vartheta_i\sim\mathcal{N}(0,\alpha t)$, independently for all i, and $\zeta\sim\mathcal{N}(-\beta t\sqrt{\log N},\alpha t)$. We can write $\max_{i\le N}(\vartheta_i+\zeta)=\max_{i\le N}(\vartheta_i)+\zeta$. Then, by the basic convergence result that the maximum of N i.i.d. standard normal random variables scales like $\sqrt{2\log N}$, it is easy to see that $\max_{i\le N}(\vartheta_i+\zeta)/\sqrt{\log N}\xrightarrow{\mathbb{P}}\sqrt{2\alpha t}-\beta t$ as $N\to\infty$. Because of the fact that a queue length which is 0 at time 0, can be written as the supremum of the arrival process minus the service process up to time t, the fluid limit yields $\sup_{s\in[0,t]}(\sqrt{2\alpha s}-\beta s)$, which equals the first part of the maximum in (2.2.5).

Similarly, for the second part of (2.2.10), we observe that

$$\begin{aligned}&\max_{i\le N}\frac{\mathbf{N}^{(N)}_A\left(tN^3\log N\right)-\mathbf{N}^{(N)}_{S,i}\left(tN^3\log N\right)+Q^{(N)}_i(0)}{N\log N} \\ &\quad=\frac{\mathbf{N}^{(N)}_A\left(tN^3\log N\right)-\left(1-\alpha/N\right)tN^3\log N}{N\log N} \\ &\qquad+\max_{i\le N}\frac{\left(1-\alpha/N\right)tN^3\log N-\mathbf{N}^{(N)}_{S,i}\left(tN^3\log N\right)+Q^{(N)}_i(0)}{N\log N}.\end{aligned} \tag{2.2.11}$$

It is easy to see that the first term on the right-hand side of Equation (2.2.11) converges to $-\beta t$ as $N \to \infty$, and we prove later on that the second term converges to $g(t, q(0))$. This explains the second part of the fluid limit in (2.2.5).

Specific properties of the function g can be deduced. First, Assumption 2.4 considers the case where U_i has a finite right endpoint. In this scenario, we have that $Q_i^{(N)}(0)/(N \log N) = \lfloor r_N U_i \rfloor / (N \log N) = \lfloor N \log N U_i \rfloor / (N \log N) \approx U_i$. Now, the theorem says that $g(t, q(0)) = q(0) + \sqrt{2\alpha t}$. This actually means that for large N,

$$\max_{i \leq N} \left(U_i + \frac{\left(1 - \alpha/N\right) tN^3 \log N - \mathbf{N}_{S,i}^{(N)} \left(tN^3 \log N\right)}{N \log N} \right) \approx \max_{i \leq N} U_i + \max_{i \leq N} \frac{\left(1 - \alpha/N\right) tN^3 \log N - \mathbf{N}_{S,i}^{(N)} \left(tN^3 \log N\right)}{N \log N}.$$

This behavior can be very well explained, because due to the assumption that U_i has a finite right endpoint, there will be many observations of U_i that are close to the right endpoint, as N becomes large, and thus it will be more and more likely that there is a large observation $\Big(\left(1 - \alpha/N\right) (tN^3 \log N) - \mathbf{N}_{S,i^\star}^{(N)} \left(tN^3 \log N\right) \Big) \Big/ \Big(N \log N \Big)$, for which the observation $U_{i^\star}$ will also be large.

Furthermore, when Assumption 2.5 holds, $g(t, q(0))$ can be written as a supremum over a set. To give an idea of why this is the case, we first observe that we can write the last term in (2.2.11) as

$$\max_{i \leq N} \left(\frac{\left(1 - \alpha/N\right) (tN^3 \log N) - \mathbf{N}_{S,i}^{(N)} \left(tN^3 \log N\right)}{N \log N} + \frac{Q_i^{(N)}(0)}{N \log N} \right). \tag{2.2.12}$$

Thus, this maximum can be viewed as a maximum of N pairwise sums of random variables. For any $N > 0$, we can write down all the N pairs of random variables as

$$\left\{ \left(\frac{1}{\sqrt{2\alpha t}} \frac{\left(1 - \alpha/N\right) (tN^3 \log N) - \mathbf{N}_{S,i}^{(N)} \left(tN^3 \log N\right)}{N \log N}, \frac{1}{q(0)} \frac{Q_i^{(N)}(0)}{N \log N} \right)_{i \leq N} \right\}. \tag{2.2.13}$$

Now, the expression in Equation (2.2.12) can be written as $\sqrt{2\alpha t} u + q(0) v$ with (u, v) in the set in (2.2.13), such that $\sqrt{2\alpha t} u + q(0) v$ is maximized. Due to the central limit theorem, the first term in (2.2.13) can be approximated by $\vartheta_i / \sqrt{2\alpha t}$ with $\vartheta_i \sim \mathcal{N}(0, \alpha t)$ when N is large. Therefore, the convex hull of the set in (2.2.13) looks like the convex hull of the set

$$\left\{ \left(\frac{1}{\sqrt{2\alpha t}} \frac{\vartheta_i}{\sqrt{\log N}}, \frac{1}{q(0)} \frac{Q_i^{(N)}(0)}{N \log N} \right)_{i \leq N} \right\}.$$

The convex hull of this set can be seen as a random variable in the space of non-empty compact subsets of $\mathbb{R}^2$, and converges in probability, under an appropriate metric, to the

convex hull of the limiting set

$$\{(u,v) : u^2 + h(v) \leq 1, -1 \leq u \leq 1, 0 \leq v \leq 1\}, \tag{2.2.14}$$

in $\mathbb{R}^2$, as $N \to \infty$; see [44] and [55] for details on this. Our intuition says that the limit of the expression in (2.2.12) is attained at the coordinate (u,v) in the closure of the limiting set given in (2.2.14), such that $\sqrt{2\alpha t}u + q(0)v$ is maximized. We show that this is indeed correct. In fact, we prove this in Lemma 2.4 in a more general context than in [44] and [55]. In [44] and [55], the authors make the assumption that the scaling sequences are the same, so the analysis is restricted to samples of the type $\{(X_i/a_N, Y_i/a_N)_{i\leq N}\}$. However, we show that for proving convergence of the maximum of the pairwise sum, the scaling sequences do not need to be the same.

2.2.4 Examples and numerics

In Section 2.2.3, we showed that the shape of the fluid limit depends on the distribution of the number of jobs at time 0. Here, we give some basic examples of how the fluid limit is influenced by the distribution of the number of jobs at time 0. We also present and discuss some numerical results.

As a first example, we consider $U_i = X_i^+$, with $X_i \sim \mathcal{N}(0,1)$. We can write for $v > 0$, $\mathbb{P}(U_i > v) = \exp(-v^2\ell(v))$, such that ℓ is slowly varying. A function ℓ being slowly varying means that for all $x > 0$, we have that $\lim_{t\to\infty} \ell(tx)/\ell(x) = 1$; see [132, Def. 2.1, p. 20]. Thus for $v \in [0,1]$,

$$h(v) = \lim_{t\to\infty} \frac{-\log\left(\mathbb{P}(U_i > vt)\right)}{-\log\left(\mathbb{P}(U_i > t)\right)} = \lim_{t\to\infty} \frac{(vt)^2\ell(vt)}{t^2\ell(t)} = v^2.$$

Thus, the function g given in (2.2.7) equals

$$g(t, q(0)) = \sup_{(u,v)} \{\sqrt{2\alpha t}u + q(0)v : u^2 + v^2 \leq 1, -1 \leq u \leq 1, 0 \leq v \leq 1\} = \sqrt{q(0)^2 + 2\alpha t}.$$

Concluding, the limit of the second term of the pairwise maximum in (2.2.10) satisfies

$$\begin{aligned}
&\max_{i\leq N} \frac{\left(1-\alpha/N\right)\left(tN^3\log N\right) - \mathbf{N}_{S,i}^{(N)}\left(tN^3\log N\right) + Q_i^{(N)}(0)}{N\log N} \\
&\quad = \max_{i\leq N} \frac{\left(1-\alpha/N\right)\left(tN^3\log N\right) - \mathbf{N}_{S,i}^{(N)}\left(tN^3\log N\right) + \lfloor q(0)N\log NU_i/\sqrt{2\log N}\rfloor}{N\log N} \\
&\quad \xrightarrow{\mathbb{P}} \sqrt{q(0)^2 + 2\alpha t} - \beta t,
\end{aligned}$$

as $N \to \infty$, where $r_N = q(0)N\log N/\sqrt{2\log N}$, such that $Q_{(\alpha,\beta)}^{(N)}(0)/(N\log N) \xrightarrow{\mathbb{P}} q(0)$, as $N \to \infty$. Thus, we now have an expression for the second part of the pairwise maximum in (2.2.5). We see immediately that when $t = 0$, this function reduces to $q(0)$, as should be the case, since the rescaled number of jobs at time 0 converges to $q(0)$. We also see that

$\lim_{t\to\infty} \sqrt{q(0)^2 + 2\alpha t} - \beta t = -\infty$. The fluid limit in (2.2.5) is a pairwise maximum, of which the second part eventually converges to $-\infty$. This means that over time, the influence of the initial number of jobs on the fluid limit vanishes, and the system reaches the steady state.

Another example is when we assume that U_i is lognormally distributed. In this case, we know that $\mathbb{P}(U_i > v) = \mathbb{P}(X_i > \log v)$, with $X_i \sim \mathcal{N}(0,1)$. Thus, $\mathbb{P}(U_i > v) = \exp(-\mathbb{1}(v > 0)\log(v)^2 \ell(\log v))$ for a slowly varying function ℓ. Then, for $v \in [0,1]$,

$$h(v) = \lim_{t\to\infty} \frac{\mathbb{1}(v>0)\log(vt)^2\ell(\log(vt))}{\log(t)^2\ell(\log(t))} = \mathbb{1}(v>0).$$

In this case, we have for the function g given in (2.2.7) that

$$\begin{aligned} &g(t, q(0)) \\ &\quad = \sup_{(u,v)} \{\sqrt{2\alpha t}u + q(0)v : u^2 + \mathbb{1}(v>0) \le 1, -1 \le u \le 1, 0 \le v \le 1\} \\ &\quad = \max(q(0), \sqrt{2\alpha t}). \end{aligned}$$

We also consider the case $\mathbb{P}(U_i > v) = \exp(1 - \exp(v))$, hence; then for $v \in [0,1]$,

$$\lim_{t\to\infty} \frac{-\log\big(\mathbb{P}(U_i > vt)\big)}{-\log\big(\mathbb{P}(U_i > t)\big)} = \lim_{t\to\infty} \frac{\exp(vt) - 1}{\exp(t) - 1} = \mathbb{1}(v = 1).$$

Then, the function g given in (2.2.7) satisfies

$$\begin{aligned} &g(t, q(0)) \\ &\quad = \sup_{(u,v)} \{\sqrt{2\alpha t}u + q(0)v : u^2 + \mathbb{1}(v=1) \le 1, -1 \le u \le 1, 0 \le v \le 1\} = q(0) + \sqrt{2\alpha t}. \end{aligned}$$

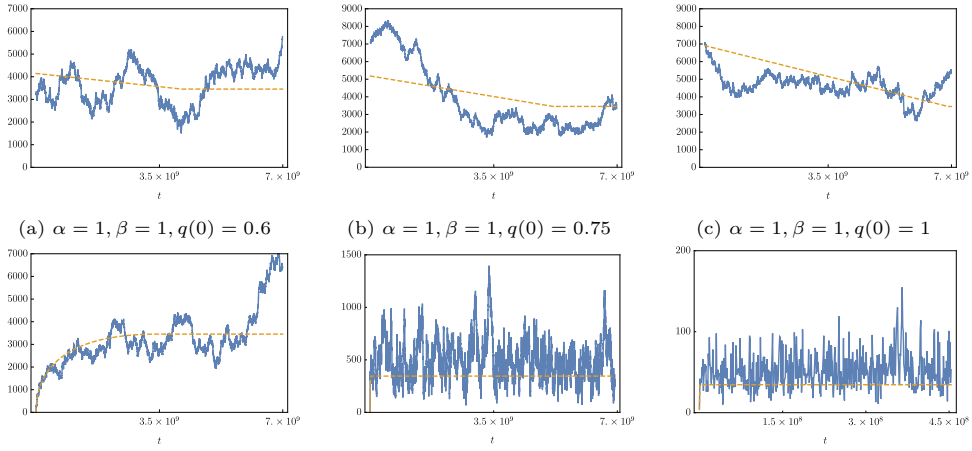
As a last example, we study the case where $\mathbb{P}(U_i > v) = \exp(-v\ell(v))$, with ℓ a slowly varying function; thus the function h, which is described in Assumption 2.5 and further, equals $h(v) = v$. Then,

$$\begin{aligned} g(t, q(0)) &= \sup_{(u,v)} \{\sqrt{2\alpha t}u + q(0)v : u^2 + v \le 1, 0 \le u \le 1, 0 \le v \le 1\} \\ &= \left(q(0) + \frac{\alpha t}{2q(0)}\right) \mathbb{1}\left(t < \frac{2q(0)^2}{\alpha}\right) + \sqrt{2\alpha t}\mathbb{1}\left(t \ge \frac{2q(0)^2}{\alpha}\right). \end{aligned}$$

We now give some extra attention to the case where $q(0) = \alpha/(2\beta)$. Then, it is not difficult to see that $q(t) \equiv \alpha/(2\beta)$. Thus, for these choices of $h(v)$ and $q(0)$, the system starts and stays in steady state. One can show that this limit is only obtained for $h(v) = v$, so this gives us some information on the joint steady-state distribution of *all* the queue lengths in the fork-join system.

Now, we turn to some numerical examples. In Figure 2.1, the simulated maximum queue length is plotted together with the scaled fluid limit $N \log N q(t/(N^3 \log N))$, with q given in Theorem 2.1 and $N = 1000$. The queue lengths at time 0 in Figures 2.1a, 2.1b, and 2.1c

are exponentially distributed. These figures show that for $N = 1000$, the maximum queue length is not close to its fluid limit.



(a) $\alpha = 1, \beta = 1, q(0) = 0.6$ (b) $\alpha = 1, \beta = 1, q(0) = 0.75$ (c) $\alpha = 1, \beta = 1, q(0) = 1$

(d) $\alpha = 1, \beta = 1, q(0) = 0$ (e) $\alpha = 1, \beta = 10, q(0) = 0$ (f) $\alpha = 1, \beta = 100, q(0) = 0$

Figure 2.1 Maximum queue length and fluid limit approximation (Thm. 2.1) for $N = 1000$

As these figures show, for $N = 1000$, the variance of the maximum queue length is still high. We give some heuristic arguments why these results are not very accurate. As mentioned before, we have that

$$\frac{\mathbf{N}_A^{(N)}\left(tN^3 \log N\right) - \left(1 - \alpha/N\right)\left(tN^3 \log N\right)}{N \log N} \xrightarrow{\mathbb{P}} -\beta t,$$

as $N \to \infty$, which is one building block of the fluid limit.
For $(\mathbf{N}_A^{(N)}\left(tN^3 \log N\right) - \left(1 - \alpha/N\right) tN^3 \log N)/(N \log N)$, we can compute the standard deviation. We have for $\alpha = \beta = t = 1$ and $N = 1000$ that

$$\begin{aligned}
&\sqrt{\mathrm{Var}\left(\mathbf{N}_A^{(N)}\left(tN^3 \log N\right) - \left(1 - \frac{\alpha}{N}\right)\left(tN^3 \log N\right)\right)} \\
&= \sqrt{\left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2}\right)\left(\frac{\alpha}{N} + \frac{\beta}{N^2}\right)\lfloor tN^3 \log N \rfloor} \\
&= 2628.26.
\end{aligned}$$

This is of the order of magnitude of the errors that we see in the figures.

Another way of seeing that there is a significant deviation is by looking at

$$\max_{i \le N}\left(\left(1 - \frac{\alpha}{N}\right) tN^3 \log N - \mathbf{N}_{S,i}^{(N)}\left(tN^3 \log N\right)\right).$$

As mentioned in Section 2.2.3, we have that

$$\frac{\left(1-\alpha/N\right)tN^3\log N - \mathbf{N}_{S,i}^{(N)}\left(tN^3\log N\right)}{N\sqrt{\log N}} \overset{d}{\approx} \vartheta_i,$$

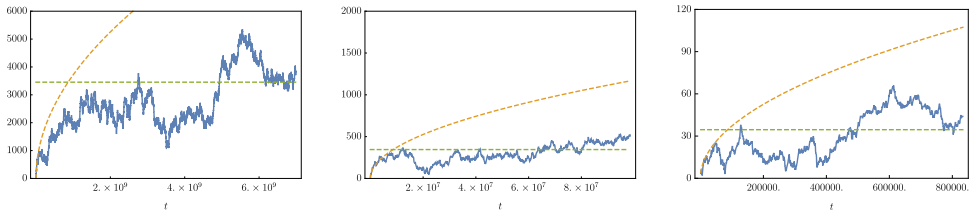
with $\vartheta_i \sim \mathcal{N}(0,\alpha t)$. Thus, this means that

$$\max_{i\leq N}\left(\left(1-\frac{\alpha}{N}\right)tN^3\log N - \mathbf{N}_{S,i}^{(N)}\left(tN^3\log N\right)\right) \overset{d}{\approx} \max_{i\leq N}\vartheta_i N\sqrt{\log N}.$$

When we choose $N = 1000$, $\alpha = t = 1$, and simulate enough samples of $\max_{i\leq N}\vartheta_i N\sqrt{\log N}$, we observe a standard deviation which is higher than 900.

In Figures 2.1a, 2.1b, and 2.1c, the high standard deviation is also caused by the distribution of the number of jobs at time 0. For example, for $E_i \sim \text{Exp}(1/N)$, i.i.d. for all i, and $N = 1000$, we have that $\sqrt{\text{Var}\left(\max_{i\leq N} E_i\right)} = 1,282.16$, so this is also of the order of magnitude of the errors that we see.

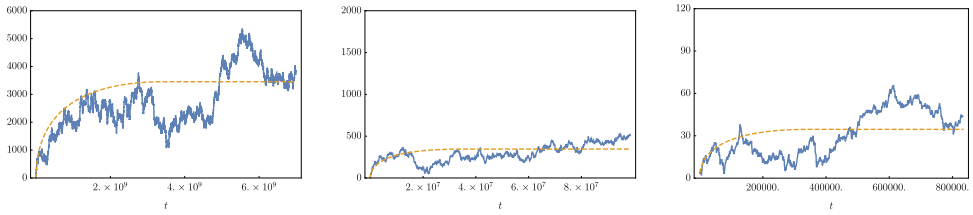
As mentioned, one can prove fluid limits under several temporal and spatial scalings. In Figure 2.2, the maximum queue length is plotted against the rescaled fluid limit given in Proposition 2.1, which is the curved dashed line, and the rescaled steady-state limit, which is the straight dashed line. In these plots, $N = 1000$. The rescaled fluid limit is $\sqrt{2\alpha t/N^3}N\sqrt{\log N}$, and the rescaled steady-state limit satisfies $\alpha/(2\beta)N\log N$.



(a) $\alpha = 1, \beta = 1$ (b) $\alpha = 1, \beta = 10$ (c) $\alpha = 1, \beta = 100$

Figure 2.2 Maximum queue length, fluid limit approximation (Prop. 2.1) and steady-state approximation for $N = 1000$

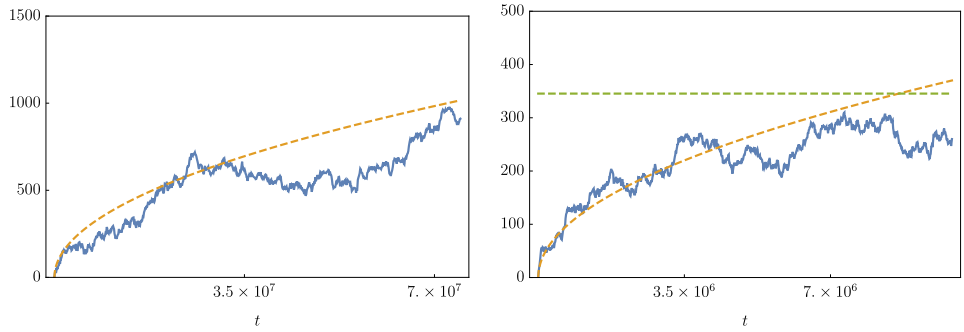
When we observe Figure 2.2, we see that for small time instances, the maximum queue length follows the fluid limit described in Proposition 2.1 with a negligible deviation, and we also see that, from the point that the fluid limit and steady state have intersected, the maximum queue length follows the steady state, though with a significant deviation. The latter behavior can be very well explained when we plot the same maximum queue lengths together with the fluid limit in Theorem 2.1. This is shown in Figure 2.3.



(a) $\alpha = 1, \beta = 1$ (b) $\alpha = 1, \beta = 10$ (c) $\alpha = 1, \beta = 100$

Figure 2.3 Maximum queue length and fluid limit approximation (Thm. 2.1) for $N = 1000$

In Figure 2.4, we zoom in on the graphs given in Figures 2.2a and 2.2b. As these figures show, for small time instances, the maximum queue length follows the fluid limit described in Proposition 2.1 quite well. Again, we can heuristically explain the deviations by approximating the maximum queue length with $\sqrt{1/N^3}N \max_{i \leq N} \vartheta_i$, with $\vartheta_i \sim \mathcal{N}(0, \alpha t)$, i.i.d. For $\alpha = 1$, and $t = 7 \cdot 10^7$, simulations show that this approximation has a standard deviation around 95, and for $t = 7 \cdot 10^6$, we get a standard deviation around 30. This is of the order of magnitude of the errors in Figures 2.4a and 2.4b, respectively.



(a) $\alpha = 1, \beta = 1$ (b) $\alpha = 1, \beta = 10$

Figure 2.4 Maximum queue length, fluid limit approximation (Prop. 2.1) and steady-state approximation for $N = 1000$

2.3. Proofs

In this section, we prove Theorem 2.1 and Proposition 2.2. Since each server has the same arrival process, the queue lengths are dependent. The general idea of proving Theorem 2.1 is to approximate the scaled centralized service process in (2.3.4) by a normally distributed random variable. We can use extreme-value theory to prove the convergence of the maximum of these normally distributed random variables in probability. By using the non-uniform version of the Berry-Esséen theorem [109], we show that the convergence result of the original process is the same as the convergence result with normally distributed random variables. Furthermore, we prove the convergence of the part involving non-zero starting points. This

gives us the pointwise convergence of the process, which we prove in Section 2.3.3. In this section, we also prove the convergence of the finite-dimensional distributions. Finally, we prove in Section 2.3.4 that the process $\left(Q^{(N)}_{(\alpha,\beta)}(tN^3\log N)\big/\left(N\log N\right), t\in[0,T]\right)$ is tight. These three results together prove the theorem. First, we define some auxiliary random variables in Section 2.3.1 and we state some lemmas in Section 2.3.2.

2.3.1 Definitions

We use the expressions given in Definition 2.4 to prove tightness.

Definition 2.4. *We define the random walk* $\tilde{R}^{(N)}_i(n)$ *as*

$$\tilde{R}^{(N)}_i(n) := \frac{\tilde{\mathbf{N}}^{(N)}_A(n) + \tilde{\mathbf{N}}^{(N)}_{S,i}(n)}{\log N}, \tag{2.3.1}$$

where

$$\tilde{\mathbf{N}}^{(N)}_A(n) := \frac{\mathbf{N}^{(N)}_A(n)}{N} - \left(1-\frac{\alpha}{N}\right)\frac{\lfloor n\rfloor}{N}, \tag{2.3.2}$$

and

$$\tilde{\mathbf{N}}^{(N)}_{S,i}(n) := -\frac{\mathbf{N}^{(N)}_{S,i}(n)}{N} + \left(1-\frac{\alpha}{N}\right)\frac{\lfloor n\rfloor}{N}. \tag{2.3.3}$$

Furthermore,

$$M^{(N)}_{i,1}(t) := \frac{\tilde{\mathbf{N}}^{(N)}_{S,i}\left(tN^3\log N\right)}{\sqrt{\alpha t(1-\alpha/N)\log N}}\frac{\sqrt{tN^3\log N}}{\sqrt{\lfloor tN^3\log N\rfloor}}, \tag{2.3.4}$$

and

$$M^{(N)}_{i,2}(t) := \frac{\tilde{\mathbf{N}}^{(N)}_{S,i}\left(tN^3\right)}{\sqrt{\alpha t(1-\alpha/N)}}\frac{\sqrt{tN^3}}{\sqrt{\lfloor tN^3\rfloor}}, \tag{2.3.5}$$

with $\mathbf{N}^{(N)}_A(n)$ *and* $\mathbf{N}^{(N)}_{S,i}(n)$ *given in Definitions 2.1 and Definition 2.2 respectively.*

As mentioned in Section 2.2.3, when $Q^{(N)}_{(\alpha,\beta)}(0)=0$, the quantity in (2.2.10) simplifies to

$$\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\log N\right)}{N\log N} = \max_{i\le N}\sup_{s\in[0,t]}\left(\frac{\left(\mathbf{N}^{(N)}_A\left(tN^3\log N\right)-\mathbf{N}^{(N)}_A\left(sN^3\log N\right)\right)}{N\log N} + \frac{\left(\mathbf{N}^{(N)}_{S,i}\left(tN^3\log N\right)-\mathbf{N}^{(N)}_{S,i}\left(sN^3\log N\right)\right)}{N\log N}\right).$$

Consequently, we can rewrite

$$\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\log N\right)}{N\log N} = \max_{i\leq N}\sup_{r\in[0,t]}\left(\tilde{R}^{(N)}_i(tN^3\log N) - \tilde{R}^{(N)}_i(rN^3\log N)\right). \tag{2.3.6}$$

2.3.2 Useful lemmas

In order to prove Theorem 2.1, a few preliminary results are needed. Observe that $\tilde{\mathbf{N}}^{(N)}_A(n)$ appearing in Definition 2.4 does not depend on i, while $\tilde{\mathbf{N}}^{(N)}_{S,i}(n)$ does. Hence, it is intuitively clear that $\tilde{\mathbf{N}}^{(N)}_A(n)$ pays no contribution to the maximum queue length. Therefore, in order to prove the pointwise convergence of the maximum queue length, we need to analyze $\tilde{\mathbf{N}}^{(N)}_{S,i}(n)/\log N$. Specifically, we use the fact that

$$M^{(N)}_{i,1}(t) \xrightarrow{d} Z,$$

as $N\to\infty$, with Z a standard normal random variable, which can be shown by the central limit theorem. We can use this result to approximate the maximum queue length because we know that the scaled maximum of N independent and normally distributed random variables converges to a Gumbel-distributed random variable. To prove the tightness of the maximum queue length, we have to prove that

$$\lim_{\delta\downarrow 0}\limsup_{N\to\infty}\frac{1}{\delta}\,\mathbb{P}\left(\sup_{s\in[t,t+\delta]}\left|\frac{Q^{(N)}_{(\alpha,\beta)}\left(sN^3\log N\right)}{N\log N} - \frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\log N\right)}{N\log N}\right| > \epsilon\right) = 0. \tag{2.3.7}$$

In Lemma 2.1, a useful upper bound for the absolute value in (2.3.7) is obtained, which we use to prove the tightness of the process $\left(Q^{(N)}_{(\alpha,\beta)}(tN^3\log N)/\left(N\log N\right), t\in[0,T]\right)$.

Lemma 2.1. *For fixed $t>0$, $\delta>0$ and $Q^{(N)}_{(\alpha,\beta)}(0)=0$, we have that*

$$\begin{aligned}
&\sup_{s\in[t,t+\delta]}\left|\frac{Q^{(N)}_{(\alpha,\beta)}\left(sN^3\log N\right)}{N\log N} - \frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\log N\right)}{N\log N}\right| \\
&\quad\leq \sup_{s\in[t,t+\delta]}\max_{i\leq N}\left(\tilde{R}^{(N)}_i(sN^3\log N) - \tilde{R}^{(N)}_i(tN^3\log N)\right) \\
&\qquad + 2\sup_{s\in[t,t+\delta]}\max_{i\leq N}\left(\tilde{R}^{(N)}_i(tN^3\log N) - \tilde{R}^{(N)}_i(sN^3\log N)\right).
\end{aligned} \tag{2.3.8}$$

In our proofs, we use the fact that the function $M^{(N)}_{i,1}(t)$, which is given in (2.3.4), converges in distribution to a normally distributed random variable. To be able to use this convergence result, we prove an upper bound of the convergence rate in Lemma 2.2.

Lemma 2.2. *For fixed $t>0$, we have that an upper bound of the rate of convergence of $\pm\tilde{\mathbf{N}}^{(N)}_{S,i}\left(tN^3\log N\right)\sqrt{tN^3\log N}/\sqrt{\alpha t(1-\alpha/N)\log N\lfloor tN^3\log N\rfloor}$ to a standard normal*

random variable is given by

$$\left|\mathbb{P}\Big(M_{i,1}^{(N)}(t) < y\Big) - \Phi(y)\right| \leq \frac{c_t}{N\sqrt{\log N}} \frac{1}{1+|y|^3}, \tag{2.3.9}$$

with $c_t > 0$. We have a similar result for the random variable $\pm\tilde{\mathbf{N}}_{S,i}^{(N)}\left(tN^3\right)\sqrt{tN^3}\big/\sqrt{\alpha t(1-\alpha/N)\lfloor tN^3\rfloor}$:

$$\left|\mathbb{P}\Big(M_{i,2}^{(N)}(t) < y\Big) - \Phi(y)\right| \leq \frac{c_t}{N} \frac{1}{1+|y|^3}, \tag{2.3.10}$$

Lemma 2.2 follows from the main result in [109], in which the author proves the non-uniform Berry-Esséen inequality. To prove tightness, we need the following lemma:

Lemma 2.3. *For fixed $t > 0$, we have that*

$$\limsup_{N\to\infty} \mathbb{E}\left[\max\left(\max_{i\leq N} \frac{\pm\tilde{\mathbf{N}}_{S,i}^{(N)}\left(tN^3\log N\right)}{\log N}, 0\right)^{5/2}\right] \leq (2\alpha t)^{5/4}, \tag{2.3.11}$$

and

$$\limsup_{N\to\infty} \mathbb{E}\left[\max\left(\max_{i\leq N} \frac{\pm\tilde{\mathbf{N}}_{S,i}^{(N)}\left(tN^3\right)}{\sqrt{\log N}}, 0\right)^{5/2}\right] \leq (2\alpha t)^{5/4}. \tag{2.3.12}$$

In order to prove pointwise convergence of the starting position, we show in Lemma 2.9 that

$$\max_{i\leq N}\left(\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(tN^3\log N\right)}{\log N} + \frac{Q_i^{(N)}(0)}{N\log N}\right) \approx \max_{i\leq N}\left(\frac{\sqrt{\alpha t}X_i}{\sqrt{\log N}} + \frac{Q_i^{(N)}(0)}{N\log N}\right),$$

with $X_i \sim \mathcal{N}(0,1)$, as N is large. In Lemma 2.4, we prove the convergence of $\max_{i\leq N}\left(\sqrt{\alpha t}X_i/\sqrt{\log N} + Q_i^{(N)}(0)/(N\log N)\right)$.

Lemma 2.4 (Pointwise convergence approximation starting position)**.**

$$\max_{i\leq N}\left(\frac{\sqrt{\alpha t}X_i}{\sqrt{\log N}} + \frac{Q_i^{(N)}(0)}{N\log N}\right) \xrightarrow{\mathbb{P}} g(t,q(0)),$$

as $N\to\infty$, with $X_i \sim \mathcal{N}(0,1)$ i.i.d. and the function g as given in Theorem 2.1.

The proofs of Lemmas 2.1, 2.2, 2.3, and 2.4 can be found in Section 2.5. Lemma 2.4 follows from Lemma 2.13, in which a more general result is proven for $\max_{i\leq N}\sum_{j=1}^{k} Z_i^{(j)}/a_N^{(j)}$.

2.3.3 Pointwise convergence

In this section, we prove pointwise convergence of the scaled maximum queue length appearing in Theorem 2.1.

Theorem 2.2 (Pointwise convergence)**.** *For fixed $t > 0$,*

$$\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3 \log N\right)}{N \log N} \xrightarrow{\mathbb{P}} q(t), \tag{2.3.13}$$

as $N \to \infty$, with $q(t)$ given in Equation (2.2.5)*.*

As Equation (2.2.10) shows, we can write the scaled maximum queue length as a maximum of two random variables, namely, one pertaining to a system starting empty and one pertaining to a system starting non-empty. We prove the pointwise convergence of the first part of this maximum in Lemma 2.5. In Lemma 2.9, we prove the pointwise convergence of the second part. In order to do so, we need some extra results, which are stated in Lemmas 2.4, 2.6, 2.7, and 2.8.

Lemma 2.5. *For fixed $t > 0$ and $Q^{(N)}_{(\alpha,\beta)}(0) = 0$,*

$$\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3 \log N\right)}{N \log N} \xrightarrow{\mathbb{P}} \left(\sqrt{2\alpha t} - \beta t\right)\mathbb{1}\left(t < \frac{\alpha}{2\beta^2}\right) + \frac{\alpha}{2\beta}\mathbb{1}\left(t \geq \frac{\alpha}{2\beta^2}\right),$$

as $N \to \infty$.

To prove the convergence of sequences of real-valued random variables to a constant, it suffices to show convergence in distribution. Therefore, we use Lemmas 2.6, 2.7 and 2.8 below to prove that the upper and lower bound of the cumulative distribution function converge to the same function.

Lemma 2.6. *For $\delta > 0$, $t < \alpha/(2\beta^2)$ and $Q^{(N)}_{(\alpha,\beta)}(0) = 0$,*

$$\limsup_{N\to\infty} \mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3 \log N\right)}{N \log N} > \sqrt{2\alpha t} - \beta t + \delta\right) = 0. \tag{2.3.14}$$

Proof. Let $\delta > 0$ be given. Let us assume that $t < \alpha/(2\beta^2)$. We then have that

$$\begin{aligned}
&\mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3 \log N\right)}{N \log N} > \sqrt{2\alpha t} - \beta t + \delta\right) \\
&\qquad = \mathbb{P}\left(\max_{i\leq N} \sup_{s\in[0,t]} \left(\frac{\tilde{\mathbf{N}}^{(N)}_A\left(sN^3 \log N\right) + \tilde{\mathbf{N}}^{(N)}_{S,i}\left(sN^3 \log N\right)}{\log N}\right) - \sqrt{2\alpha t} + \beta t > \delta\right).
\end{aligned}$$

For $t < \alpha/(2\beta^2)$, $\sqrt{2\alpha t} - \beta t$ is an increasing function. Therefore,

$$\mathbb{P}\left(\max_{i\leq N} \sup_{s\in[0,t]} \left(\frac{\tilde{\mathbf{N}}^{(N)}_A\left(sN^3 \log N\right) + \tilde{\mathbf{N}}^{(N)}_{S,i}\left(sN^3 \log N\right)}{\log N}\right) - \sqrt{2\alpha t} + \beta t > \delta\right)$$

$$\leq \mathbb{P}\left(\max_{i\leq N}\sup_{s\in[0,t]}\left(\frac{\tilde{\mathbf{N}}_A^{(N)}\left(sN^3\log N\right)+\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}-\sqrt{2\alpha s}+\beta s\right)>\delta\right)$$

$$=\mathbb{P}\left(\sup_{s\in[0,t]}\left(\max_{i\leq N}\frac{\tilde{\mathbf{N}}_A^{(N)}\left(sN^3\log N\right)+\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}-\sqrt{2\alpha s}+\beta s\right)>\delta\right).$$

Observe that

$$\mathbb{P}\left(\sup_{s\in[0,t]}\left(\max_{i\leq N}\frac{\tilde{\mathbf{N}}_A^{(N)}\left(sN^3\log N\right)+\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}-\sqrt{2\alpha s}+\beta s\right)>\delta\right)$$

$$\leq \mathbb{P}\left(\sup_{s\in[0,t]}\left|\max_{i\leq N}\frac{\tilde{\mathbf{N}}_A^{(N)}\left(sN^3\log N\right)+\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}-\sqrt{2\alpha s}+\beta s\right|>\delta\right)$$

$$\leq \mathbb{P}\left(\sup_{s\in[0,t]}\left|\frac{\tilde{\mathbf{N}}_A^{(N)}\left(sN^3\log N\right)}{\log N}+\beta s\right|>\frac{\delta}{2}\right)$$

$$+\mathbb{P}\left(\sup_{s\in[0,t]}\left|\frac{\max_{i\leq N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}-\sqrt{2\alpha s}\right|>\frac{\delta}{2}\right).$$

Moreover, $\tilde{\mathbf{N}}_A^{(N)}(n)/\log N+\beta n/(N^3\log N)$ is a martingale with mean 0. Therefore, by Doob's maximal submartingale inequality

$$\mathbb{P}\left(\sup_{s\in[0,t]}\left|\frac{\tilde{\mathbf{N}}_A^{(N)}\left(sN^3\log N\right)}{\log N}+\beta s\right|>\frac{\delta}{2}\right)$$

$$\leq \mathbb{P}\left(\sup_{s\in[0,t]}\left|\frac{\tilde{\mathbf{N}}_A^{(N)}\left(sN^3\log N\right)}{\log N}+\beta\frac{\lfloor sN^3\log N\rfloor}{N^3\log N}\right|>\frac{\delta}{4}\right) \tag{2.3.15}$$

$$+\mathbb{P}\left(\sup_{s\in[0,t]}\left|\beta\frac{\lfloor sN^3\log N\rfloor}{N^3\log N}-\beta s\right|>\frac{\delta}{4}\right)$$

$$\leq \frac{16}{\delta^2}\mathrm{Var}\left(\frac{\tilde{\mathbf{N}}_A^{(N)}\left(tN^3\log N\right)}{\log N}\right)+o(1)$$

$$=\frac{16}{\delta^2}\left(1-\frac{\alpha}{N}-\frac{\beta}{N^2}\right)\left(\frac{\alpha}{N}+\frac{\beta}{N^2}\right)\frac{\lfloor tN^3\log N\rfloor}{N^2(\log N)^2}+o(1)\overset{N\to\infty}{\longrightarrow}0. \tag{2.3.16}$$

Furthermore, in order to have

$$\mathbb{P}\left(\sup_{s\in[0,t]}\left|\frac{\max_{i\le N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}-\sqrt{2\alpha s}\right|>\frac{\delta}{2}\right)\overset{N\to\infty}{\longrightarrow}0, \tag{2.3.17}$$

we need to have that $\left(\max_{i\le N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)/\log N, s\in[0,t]\right)$ converges to $(\sqrt{2\alpha s}, s\in[0,t])$ u.o.c. Thus,

$$\lim_{N\to\infty}\mathbb{P}\left(\left|\frac{\max_{i\le N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}-\sqrt{2\alpha s}\right|>\epsilon\right)=0, \tag{2.3.18}$$

and for all $r\in[0,t]$,

$$\lim_{\eta\downarrow 0}\limsup_{N\to\infty}\frac{1}{\eta}\mathbb{P}\left(\sup_{s\in[r,r+\eta]}\left|\frac{\max_{i\le N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}-\frac{\max_{i\le N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(rN^3\log N\right)}{\log N}\right|>\epsilon\right)=0. \tag{2.3.19}$$

To prove the limit in (2.3.18), we use the result of Lemma 2.2 and observe that for all $\delta>0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\max_{i\le N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}>\sqrt{2\alpha s}+\delta\right)\\
&=1-\mathbb{P}\left(\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}<\sqrt{2\alpha s}+\delta\right)^N\\
&=1-\mathbb{P}\left(M_{i,1}^{(N)}(s)<\frac{\sqrt{2\alpha s}+\delta}{\sqrt{\alpha s(1-\alpha/N)}}\sqrt{\log N}\frac{\sqrt{sN^3\log N}}{\sqrt{\lfloor sN^3\log N\rfloor}}\right)^N\\
&\le 1-\left(\Phi\left(\frac{\sqrt{2\alpha s}+\delta}{\sqrt{\alpha s(1-\alpha/N)}}\sqrt{\log N}\frac{\sqrt{sN^3\log N}}{\sqrt{\lfloor sN^3\log N\rfloor}}\right)-\frac{c_s}{N\sqrt{\log N}}\right)^N\\
&\le 1-\Phi\left(\frac{\sqrt{2\alpha s}+\delta}{\sqrt{\alpha s(1-\alpha/N)}}\sqrt{\log N}\frac{\sqrt{sN^3\log N}}{\sqrt{\lfloor sN^3\log N\rfloor}}\right)^N+\left(1+\frac{c_s}{N\sqrt{\log N}}\right)^N-1\\
&\overset{N\to\infty}{\longrightarrow}0.
\end{aligned} \tag{2.3.20}$$

The proof that

$$\mathbb{P}\left(\frac{\max_{i\le N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}<\sqrt{2\alpha s}-\delta\right)\overset{N\to\infty}{\longrightarrow}0,$$

goes analogously. To prove the limit in (2.3.19), we observe that due to the facts that $\tilde{\mathbf{N}}_{S,i}^{(N)}(n)$ is a random walk that satisfies the duality principle, that $\max_{i\leq N} x_i - \max_{i\leq N} y_i \leq \max_{i\leq N}(x_i - y_i)$, and that $\mathbb{P}\big(|X| > \epsilon\big) \leq \mathbb{P}(X > \epsilon) + \mathbb{P}(-X > \epsilon)$, we have the upper bound

$$\begin{aligned}
&\frac{1}{\eta}\,\mathbb{P}\left(\sup_{s\in[r,r+\eta]}\left|\frac{\max_{i\leq N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N} - \frac{\max_{i\leq N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(rN^3\log N\right)}{\log N}\right| > \epsilon\right)\\
&\quad\leq \frac{1}{\eta}\,\mathbb{P}\left(\sup_{s\in[0,\eta]}\max_{i\leq N}\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N} > \epsilon\right)\\
&\qquad+\frac{1}{\eta}\,\mathbb{P}\left(\sup_{s\in[0,\eta]}\max_{i\leq N}\frac{-\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N} > \epsilon\right) + o(1).
\end{aligned}$$

The $o(1)$ term is due to the fact that

$$\lfloor (r+\eta)N^3\log N\rfloor - \lfloor rN^3\log N\rfloor \in \{\lfloor \eta N^3\log N\rfloor, \lfloor \eta N^3\log N\rfloor + 1\}.$$

Now, we have that $\pm\tilde{\mathbf{N}}_{S,i}^{(N)}(n)$ is a martingale with mean 0. The maximum of independent martingales is a submartingale; therefore,

$$\max\left(0, \max_{i\leq N}\frac{\pm\tilde{\mathbf{N}}_{S,i}^{(N)}\left(\eta N^3\log N\right)}{\log N}\right)^{5/2}$$

is a non-negative submartingale. Hence, by using Doob's maximal submartingale inequality, we can conclude that

$$\begin{aligned}
&\frac{1}{\eta}\,\mathbb{P}\left(\sup_{s\in[0,\eta]}\max_{i\leq N}\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N} > \epsilon\right)\\
&\qquad+\frac{1}{\eta}\,\mathbb{P}\left(\sup_{s\in[0,\eta]}\max_{i\leq N}\frac{-\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N} > \epsilon\right)\\
&\quad\leq \frac{1}{\eta\epsilon^{5/2}}\,\mathbb{E}\left[\max\left(\max_{i\leq N}\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(\eta N^3\log N\right)}{\log N}, 0\right)^{5/2}\right]\\
&\qquad+\frac{1}{\eta\epsilon^{5/2}}\,\mathbb{E}\left[\max\left(\max_{i\leq N}\frac{-\tilde{\mathbf{N}}_{S,i}^{(N)}\left(\eta N^3\log N\right)}{\log N}, 0\right)^{5/2}\right].
\end{aligned}$$

By taking the $\limsup_{N\to\infty}$ in this expression and applying Lemma 2.3, we see that this is upper bounded by $2\eta^{1/4}(2\alpha)^{5/4}/\epsilon^{5/2}$. This can be made as small as possible when η is chosen small enough. We also know that $\max_{i\leq N}\tilde{\mathbf{N}}_{S,i}^{(N)}(0)/\log N = 0$, and that the finite-dimensional distributions of $\left(\max_{i\leq N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)/\log N, s\in[0,t]\right)$ converge to the finite-dimensional distributions of $\left(\sqrt{2\alpha s}, s\in[0,t]\right)$, which follows from Theorem 2.3. The

lemma follows. □

Having examined $t \in [0, \alpha/(2\beta^2))$, we now turn to $t \in [\alpha/(2\beta^2), \infty]$.

Lemma 2.7. *For $\delta > 0$, $\alpha/(2\beta^2) \leq t \leq \infty$ and $Q^{(N)}_{(\alpha,\beta)}(0) = 0$,*

$$\limsup_{N\to\infty} \mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3 \log N\right)}{N \log N} > \frac{\alpha}{2\beta} + \delta \right) = 0.$$

Proof. We write

$$\mathbf{N}_A^{(u,N)}(n) = \sum_{j=1}^{n} X^{(u,N)}(j)$$

with

$$X^{(u,N)}(j) = \begin{cases} \alpha/N + \beta/N^2 - m/N^2 & \text{w.p. } 1 - \alpha/N - \beta/N^2, \\ -1 + \alpha/N + \beta/N^2 - m/N^2 & \text{w.p. } \alpha/N + \beta/N^2, \end{cases}$$

with $0 < m < \beta$. Furthermore, we write

$$\mathbf{N}_{S,i}^{(u,N)}(n) = \sum_{j=1}^{n} Y_i^{(u,N)}(j),$$

with

$$Y_i^{(u,N)}(j) = \begin{cases} -\alpha/N - \beta/N^2 + m/N^2 & \text{w.p. } 1 - \alpha/N, \\ 1 - \alpha/N - \beta/N^2 + m/N^2 & \text{w.p. } \alpha/N. \end{cases}$$

Thus,

$$\mathbf{N}_A^{(N)}(n) - \mathbf{N}_{S,i}^{(N)}(n) = \mathbf{N}_A^{(u,N)}(n) + \mathbf{N}_{S,i}^{(u,N)}(n),$$

and

$$\sup_{0\leq k\leq n} \left(\mathbf{N}_A^{(N)}(k) - \mathbf{N}_{S,i}^{(N)}(k) \right) \leq \sup_{0\leq k\leq n} \mathbf{N}_A^{(u,N)}(k) + \sup_{0\leq k\leq n} \mathbf{N}_{S,i}^{(u,N)}(k).$$

We obtain by using Doob's maximal submartingale inequality that

$$\mathbb{P}\left(\sup_{0\leq k\leq n} \mathbf{N}_A^{(u,N)}(k) \geq x \right) \leq \mathbb{E}\left[e^{\theta_A^{(u,N)} X^{(u,N)}(j)} \right] e^{-\theta_A^{(u,N)} x} = e^{-\theta_A^{(u,N)} x},$$

with $\theta_A^{(u,N)}$ the solution to the equation

$$\mathbb{E}\left[e^{\theta_A^{(u,N)} X^{(u,N)}(j)} \right] = \left(\frac{\alpha}{N} + \frac{\beta}{N^2} \right) \exp\left(\theta_A^{(u,N)} \left(-1 + \frac{\alpha}{N} + \frac{\beta}{N^2} - \frac{m}{N^2} \right) \right)$$

$$+\left(1-\frac{\alpha}{N}-\frac{\beta}{N^2}\right)\exp\left(\theta_A^{(u,N)}\left(\frac{\alpha}{N}+\frac{\beta}{N^2}-\frac{m}{N^2}\right)\right)=1.$$

When we consider the second-order Taylor approximation of this expression with $1/N$ around 0, we obtain

$$\theta_A^{(u,N)}=\frac{2mN^2}{-\alpha^2N^2+\alpha N^3-2\alpha\beta N-\beta^2+m^2+\beta N^2}+O\left(\frac{1}{N^2}\right).$$

Consequently, we have for N large $\theta_A^{(u,N)}\approx 2m/(\alpha N)$. By the monotone convergence theorem, we know that

$$\mathbb{P}\left(\sup_{k\geq 0}\mathbf{N}_A^{(u,N)}(k)\geq x\right)\leq e^{-\theta_A^{(u,N)}x}\approx e^{-2m/(\alpha N)x}.$$

In conclusion,

$$\frac{\sup_{k\geq 0}\mathbf{N}_A^{(u,N)}(k)}{N\log N}\xrightarrow{\mathbb{P}}0,$$

as $N\to\infty$. Similarly, by using Doob's maximal submartingale inequality, we obtain that

$$\mathbb{P}\left(\sup_{n\geq 0}\mathbf{N}_{S,i}^{(u,N)}(n)\geq x\right)\leq e^{-\theta_i^{(u,N)}x},$$

with $\theta_i^{(u,N)}$ the solution to the equation

$$\mathbb{E}\left[e^{\theta_i^{(u,N)}Y_i^{(u,N)}(j)}\right]=\frac{\alpha}{N}\exp\left(\theta_i^{(u,N)}\left(1-\frac{\alpha}{N}-\frac{\beta}{N^2}+\frac{m}{N^2}\right)\right)$$
$$+\left(1-\frac{\alpha}{N}\right)\exp\left(\theta_i^{(u,N)}\left(-\frac{\alpha}{N}-\frac{\beta}{N^2}+\frac{m}{N^2}\right)\right)=1.$$

The second-order Taylor approximation of $\mathbb{E}\left[e^{\theta_i^{(u,N)}Y_i^{(u,N)}(j)}\right]$ with $1/N$ around 0 gives

$$\theta_i^{(u,N)}=\frac{2N^2(\beta-m)}{-\alpha^2N^2+\alpha N^3+(\beta-m)^2}+O\left(\frac{1}{N^2}\right).$$

Thus, for N large, $\theta_i^{(u,N)}\approx 2(\beta-m)/(\alpha N)$. Concluding, $\sup_{n\geq 0}\mathbf{N}_{S,i}^{(u,N)}(n)$ is stochastically dominated by an exponentially distributed random variable $E_i^{(u,N)}$ with mean $\alpha N/(2(\beta-m))$. Because $\sup_{n\geq 0}\mathbf{N}_{S,i}^{(u,N)}(n)\perp\sup_{n\geq 0}\mathbf{N}_{S,j}^{(u,N)}(n)$ for $i\neq j$, we can conclude that also

$E_i^{(u,N)} \perp E_j^{(u,N)}$ for $i \neq j$. Therefore,

$$\mathbb{P}\left(\frac{\max_{i \leq N} E_i^{(u,N)}}{N} \leq \frac{\alpha}{2(\beta - m)} \left(x + \log N \right) \right) \overset{N \to \infty}{\longrightarrow} e^{-e^{-x}},$$

and

$$\frac{\max_{i \leq N} E_i^{(u,N)}}{N \log N} \xrightarrow{\mathbb{P}} \frac{\alpha}{2(\beta - m)},$$

as $N \to \infty$. Because

$$\frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3 \log N\right)}{N \log N} \leq_{st.} \frac{Q_{(\alpha,\beta)}^{(N)}(\infty)}{N \log N} \leq \frac{\sup_{k \geq 0} \mathbf{N}_A^{(u,N)}(k)}{N \log N} + \frac{\max_{i \leq N} \sup_{k \geq 0} \mathbf{N}_{S,i}^{(N)}(k)}{N \log N},$$

with $X \leq_{st.} Y$ meaning $\mathbb{P}(X \geq x) \leq \mathbb{P}(Y \geq x)$ for all x, the lemma follows. □

Lemma 2.8. *For $\delta > 0$ and $Q_{(\alpha,\beta)}^{(N)}(0) = 0$,*

$$\liminf_{N \to \infty} \mathbb{P}\left(\frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3 \log N\right)}{N \log N} \geq \left(\sqrt{2\alpha t} - \beta t \right) \mathbb{1}\left(t < \frac{\alpha}{2\beta^2} \right) + \frac{\alpha}{2\beta} \mathbb{1}\left(t \geq \frac{\alpha}{2\beta^2} \right) - \delta \right) = 1. \tag{2.3.21}$$

Proof. Let us first assume that $t \leq \alpha/(2\beta^2)$. We have the lower bound

$$\frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3 \log N\right)}{N \log N} \geq_{st.} \max_{i \leq N} \frac{\mathbf{N}_A^{(N)}\left(tN^3 \log N\right) - \mathbf{N}_{S,i}^{(N)}\left(tN^3 \log N\right)}{N \log N}.$$

By Equations (2.3.15) and (2.3.17), we know that

$$\max_{i \leq N} \frac{\mathbf{N}_A^{(N)}\left(tN^3 \log N\right) - \mathbf{N}_{S,i}^{(N)}\left(tN^3 \log N\right)}{N \log N} \xrightarrow{\mathbb{P}} \sqrt{2\alpha t} - \beta t,$$

as $N \to \infty$. Let us now assume that $t > \alpha/(2\beta^2)$. We have that

$$\frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3 \log N\right)}{N \log N} \geq_{st.} \max_{i \leq N} \frac{\mathbf{N}_A^{(N)}\left(\frac{\alpha}{2\beta^2} N^3 \log N\right) - \mathbf{N}_{S,i}^{(N)}\left(\frac{\alpha}{2\beta^2} N^3 \log N\right)}{N \log N} \xrightarrow{\mathbb{P}} \frac{\alpha}{2\beta},$$

as $N \to \infty$, by again using Lemma 2.6. This proves the lemma. □

Proof of Lemma 2.5. By combining the results of Lemmas 2.6, 2.7 and 2.8, Lemma 2.5 follows. □

In Lemma 2.9, we connect the convergence of

$$\max_{i \leq N} \frac{\mathbf{N}_A^{(N)}\left(tN^3 \log N\right) - \mathbf{N}_{S,i}^{(N)}\left(tN^3 \log N\right) + Q_i^{(N)}(0)}{N \log N}$$

to the convergence of

$$\max_{i\leq N}\left(\frac{\sqrt{\alpha t}X_i}{\sqrt{\log N}}+\frac{Q_i^{(N)}(0)}{N\log N}\right).$$

Lemma 2.9 (Convergence starting position)**.** *Assume that for X_i i.i.d. standard normally distributed,*

$$\max_{i\leq N}\left(\frac{\sqrt{\alpha t}X_i}{\sqrt{\log N}}+\frac{Q_i^{(N)}(0)}{N\log N}\right)\xrightarrow{\mathbb{P}} g(t,q(0)), \tag{2.3.22}$$

as $N\to\infty$, for a certain function g. Then

$$\max_{i\leq N}\frac{\mathbf{N}_A^{(N)}\left(tN^3\log N\right)-\mathbf{N}_{S,i}^{(N)}\left(tN^3\log N\right)+Q_i^{(N)}(0)}{N\log N}\xrightarrow{\mathbb{P}} g(t,q(0))-\beta t,$$

as $N\to\infty$.

Proof. We have

$$\begin{aligned}
&\max_{i\leq N}\frac{\mathbf{N}_A^{(N)}\left(tN^3\log N\right)-\mathbf{N}_{S,i}^{(N)}\left(tN^3\log N\right)+Q_i^{(N)}(0)}{N\log N}\\
&\quad=\frac{\mathbf{N}_A^{(N)}\left(tN^3\log N\right)-\left(1-\alpha/N\right)tN^3\log N}{N\log N} &(2.3.23)\\
&\qquad+\max_{i\leq N}\frac{\left(1-\alpha/N\right)tN^3\log N-\mathbf{N}_{S,i}^{(N)}\left(tN^3\log N\right)+Q_i^{(N)}(0)}{N\log N}. &(2.3.24)
\end{aligned}$$

We already proved in Equation (2.3.15) that the term in (2.3.23) converges to $-\beta t$. Furthermore, we can rewrite the term in (2.3.24) as

$$\max_{i\leq N}\left(\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(tN^3\log N\right)}{\log N}+\frac{Q_i^{(N)}(0)}{N\log N}+O\left(\frac{1}{N\log N}\right)\right).$$

We can easily deduce from Lemma 2.2 that

$$\left|\mathbb{P}\left(\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(tN^3\log N\right)}{\log N}<y\right)-\mathbb{P}\left(\frac{\sqrt{\alpha t(1-\alpha/N)}}{\sqrt{\log N}}\frac{\sqrt{\lfloor tN^3\log N\rfloor}}{\sqrt{tN^3\log N}}X_i<y\right)\right|\leq\frac{c_t}{N\sqrt{\log N}},$$

with $X_i\sim\mathcal{N}(0,1)$, and c_t given in Lemma 2.2. Then, it is easy to see that

$$\left|\mathbb{P}\left(\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(tN^3\log N\right)}{\log N}+\frac{Q_i^{(N)}(0)}{N\log N}<y\right)-\mathbb{P}\left(\frac{\sqrt{\alpha t(1-\alpha/N)}}{\sqrt{\log N}}\frac{\sqrt{\lfloor tN^3\log N\rfloor}}{\sqrt{tN^3\log N}}X_i+\frac{Q_i^{(N)}(0)}{N\log N}<y\right)\right|\leq\frac{c_t}{N\sqrt{\log N}}. \tag{2.3.25}$$

Now, because we assume the convergence result in (2.3.22), and

$$\frac{\sqrt{\alpha t(1-\alpha/N)}}{\sqrt{\log N}}\frac{\sqrt{\lfloor tN^3\log N\rfloor}}{\sqrt{tN^3\log N}}X_i=\frac{\sqrt{\alpha t}X_i}{\sqrt{\log N}}+o\left(\frac{1}{\sqrt{\log N}}\right)X_i,$$

it is easy to see that

$$\max_{i\leq N}\left(\frac{\sqrt{\alpha t(1-\alpha/N)}}{\sqrt{\log N}}\frac{\sqrt{\lfloor tN^3\log N\rfloor}}{\sqrt{tN^3\log N}}X_i+\frac{Q_i^{(N)}(0)}{N\log N}\right)\xrightarrow{\mathbb{P}}g(t,q(0)),$$

as $N\to\infty$. Let $\epsilon>0$; then, because of the bound given in (2.3.25), and the convergence result in (2.3.22),

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\leq N}\left(\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(tN^3\log N\right)}{\log N}+\frac{Q_i^{(N)}(0)}{N\log N}\right)<g(t,q(0))-\epsilon\right)\\
&=\mathbb{P}\left(\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(tN^3\log N\right)}{\log N}+\frac{Q_i^{(N)}(0)}{N\log N}<g(t,q(0))-\epsilon\right)^N\\
&\leq\mathbb{P}\left(\frac{\sqrt{\alpha t(1-\alpha/N)}}{\sqrt{\log N}}\frac{\sqrt{\lfloor tN^3\log N\rfloor}}{\sqrt{tN^3\log N}}X_i+\frac{Q_i^{(N)}(0)}{N\log N}<g(t,q(0))-\epsilon\right)^N\\
&\quad+\left(\frac{c_t}{N\sqrt{\log N}}+1\right)^N-1\\
&\xrightarrow{N\to\infty}0.
\end{aligned}$$

The proof that

$$\mathbb{P}\left(\max_{i\leq N}\left(\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(tN^3\log N\right)}{\log N}+\frac{Q_i^{(N)}(0)}{N\log N}\right)>g(t,q(0))+\epsilon\right)\xrightarrow{N\to\infty}0,$$

goes analogously. Hence, the lemma follows. □

Proof of Theorem 2.2. In Lemmas 2.5 and 2.9, we have proven that both parts in the maximum in (2.2.10) converge to a limit. The theorem follows. □

We can easily extend this result to finite-dimensional distributions.

Theorem 2.3 (The finite-dimensional distributions converge)**.** *If*

$$X^{(N)}(t)\xrightarrow{\mathbb{P}}f(t),$$

as $N\to\infty$, for all $t>0$, then for $(t_1,t_2,\ldots,t_k)$

$$\left(X^{(N)}(t_1),X^{(N)}(t_2),\ldots,X^{(N)}(t_k)\right)\xrightarrow{\mathbb{P}}\left(f(t_1),f(t_2),\ldots,f(t_k)\right),$$

as $N \to \infty$.

Proof.

$$\begin{aligned}
&\mathbb{P}\left(\left\|\left(X^{(N)}(t_1), X^{(N)}(t_2), \ldots, X^{(N)}(t_k)\right) - \left(f(t_1), f(t_2), \ldots, f(t_k)\right)\right\| > \epsilon\right) \\
&\quad \leq \mathbb{P}\left(\left|X^{(N)}(t_1) - f(t_1)\right| + \cdots + \left|X^{(N)}(t_k) - f(t_k)\right| > \epsilon\right) \\
&\quad \leq \mathbb{P}\left(\left|X^{(N)}(t_1) - f(t_1)\right| > \frac{\epsilon}{k}\right) + \cdots + \mathbb{P}\left(\left|X^{(N)}(t_k) - f(t_k)\right| > \frac{\epsilon}{k}\right) \overset{N\to\infty}{\longrightarrow} 0,
\end{aligned}$$

with $\|\cdot\|$ the Euclidean distance in $\mathbb{R}^k$. □

2.3.4 Tightness

It is known that when a sequence of random processes is tight and its finite-dimensional distributions converge to a continuous limit, then this sequence converges u.o.c.; see [28, Thm. 7.1, p. 80]. From [28, Thm. 7.3, p. 82], we know that a process $(X^{(N)}(t), t \in [0,T])$ is tight when for all positive η there exists an a and an integer N_0 such that

$$\mathbb{P}\left(\left|X^{(N)}(0)\right| > a\right) \leq \eta, \quad N \geq N_0, \tag{2.3.26}$$

and for all $\epsilon > 0$ and $\eta > 0$, there exists a $0 < \delta < 1$ and an integer N_0 such that

$$\frac{1}{\delta}\,\mathbb{P}\left(\sup_{s\in[t,t+\delta]}\left|X^{(N)}(s) - X^{(N)}(t)\right| > \epsilon\right) \leq \eta, \quad N \geq N_0. \tag{2.3.27}$$

The conditions given in Equations (2.3.26) and (2.3.27) hold for stochastic processes in the space of continuous functions. The process $\left(Q^{(N)}_{(\alpha,\beta)}(tN^3 \log N)\big/\left(N \log N\right), t \in [0,T]\right)$ does not lie in this space, because $Q^{(N)}_{(\alpha,\beta)}(n) = Q^{(N)}_{(\alpha,\beta)}\left(\lfloor n \rfloor\right)$. However, since the candidate limit $(q(t), t \in [0,T])$ is a continuous function, the conditions in (2.3.26) and (2.3.27) do also apply on $\left(Q^{(N)}_{(\alpha,\beta)}(tN^3 \log N)\big/\left(N \log N\right), t \in [0,T]\right)$; see [28, Cor. 13.4, p. 142].

In order to prove tightness for the process given in Theorem 2.1, we need to prove tightness of the maximum of two processes, as Equation (2.2.10) shows. In Lemma 2.10, we show that it suffices to prove tightness of the two processes separately. Then, in Lemmas 2.11 and 2.12, we prove the tightness of the two parts.

Lemma 2.10. *Assume that* $(X^{(N)}(s), s \in [0,t])$ *and* $(Y^{(N)}(s), s \in [0,t])$ *converge to functions* $(k(s), s \in [0,t])$ *and* $(l(s), s \in [0,t])$ *u.o.c., respectively. Then, the process* $(\max(X^{(N)}(s), Y^{(N)}(s)), s \in [0,t])$ *converges to* $(\max(k(s), l(s)), s \in [0,t])$ *u.o.c.*

Proof. The lemma holds because of the fact that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[0,t]}\left|\max(X^{(N)}(s),Y^{(N)}(s))-\max(k(s),l(s))\right|>\epsilon\right)\\
&\quad\le\mathbb{P}\left(\sup_{s\in[0,t]}\left(\max(X^{(N)}(s),Y^{(N)}(s))-\max(k(s),l(s))\right)>\epsilon\right)\\
&\qquad+\mathbb{P}\left(\sup_{s\in[0,t]}\left(\max(k(s),l(s))-\max(X^{(N)}(s),Y^{(N)}(s))\right)>\epsilon\right)\\
&\quad\le\mathbb{P}\left(\sup_{s\in[0,t]}\max(X^{(N)}(s)-k(s),Y^{(N)}(s)-l(s))>\epsilon\right)\\
&\qquad+\mathbb{P}\left(\sup_{s\in[0,t]}\max(k(s)-X^{(N)}(s),l(s)-Y^{(N)}(s))>\epsilon\right)\\
&\quad\le 2\,\mathbb{P}\left(\sup_{s\in[0,t]}\left|X^{(N)}(s)-k(s)\right|>\epsilon\right)+2\,\mathbb{P}\left(\sup_{s\in[0,t]}\left|Y^{(N)}(s)-l(s)\right|>\epsilon\right)\overset{N\to\infty}{\longrightarrow}0.
\end{aligned}$$

□

Lemma 2.11 (Tightness of the first part)**.** *For $\epsilon>0$, $\eta>0$, $T>0$ and $Q^{(N)}_{(\alpha,\beta)}(0)=0$, $\exists\,0<\delta<1$ and an integer N_0 such that for all $t\in[0,T]$,*

$$\frac{1}{\delta}\,\mathbb{P}\left(\sup_{s\in[t,t+\delta]}\left|\frac{Q^{(N)}_{(\alpha,\beta)}\left(sN^3\log N\right)}{N\log N}-\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\log N\right)}{N\log N}\right|\ge\epsilon\right)\le\eta,\quad N\ge N_0. \tag{2.3.28}$$

Proof. We take $t>0$. From Lemma 2.1, and the fact that $(\tilde{R}^{(N)}_i, i\le N)$ given in (2.3.1) is a sequence of random walks that satisfy the duality principle, we know that for N large enough,

$$\frac{1}{\delta}\,\mathbb{P}\left(\sup_{s\in[t,t+\delta]}\left|\frac{Q^{(N)}_{(\alpha,\beta)}\left(sN^3\log N\right)}{N\log N}-\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\log N\right)}{N\log N}\right|\ge\epsilon\right) \tag{2.3.29}$$

$$\le\frac{1}{\delta}\,\mathbb{P}\left(\sup_{0\le s\le\delta}\max_{i\le N}\tilde{R}^{(N)}_i(sN^3\log N)+2\sup_{0\le s\le\delta}\max_{i\le N}-\tilde{R}^{(N)}_i(sN^3\log N)\ge\epsilon\right)+o(1) \tag{2.3.30}$$

$$\le\frac{1}{\delta}\,\mathbb{P}\left(\sup_{0\le s\le\delta}\max_{i\le N}\tilde{R}^{(N)}_i(sN^3\log N)\ge\frac{\epsilon}{2}\right) \tag{2.3.31}$$

$$+\frac{1}{\delta}\,\mathbb{P}\left(2\sup_{0\le s\le\delta}\max_{i\le N}-\tilde{R}_i^{(N)}(sN^3\log N)\ge\frac{\epsilon}{2}\right)+o(1). \tag{2.3.32}$$

Now we focus on the term in (2.3.31). The analysis of the main term in (2.3.32) goes analogously:

$$\frac{1}{\delta}\,\mathbb{P}\left(\sup_{0\le s\le\delta}\max_{i\le N}\tilde{R}_i^{(N)}(sN^3\log N)\ge\frac{\epsilon}{2}\right) \tag{2.3.33}$$

$$=\frac{1}{\delta}\,\mathbb{P}\left(\sup_{0\le s\le\delta}\max_{i\le N}\frac{\tilde{\mathbf{N}}_A^{(N)}\left(sN^3\log N\right)+\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}\ge\frac{\epsilon}{2}\right) \tag{2.3.34}$$

$$\le\frac{1}{\delta}\,\mathbb{P}\left(\sup_{0\le s\le\delta}\frac{\tilde{\mathbf{N}}_A^{(N)}\left(sN^3\log N\right)}{\log N}\ge\frac{\epsilon}{4}\right)+\frac{1}{\delta}\,\mathbb{P}\left(\sup_{0\le s\le\delta}\max_{i\le N}\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\log N\right)}{\log N}\ge\frac{\epsilon}{4}\right). \tag{2.3.35}$$

In the proof of Lemma 2.6, we already showed that the second term in (2.3.35) is small. With a similar proof as in Lemma 2.6, one can also prove that the first term is small. Concluding, the process $\left(Q_{(\alpha,\beta)}^{(N)}(tN^3\log N)\big/\left(N\log N\right),t\in[0,T]\right)$ is tight, when $Q_{(\alpha,\beta)}^{(N)}(0)=0$. □

Lemma 2.12 (Tightness of the second part)**.** *For $\epsilon>0$, $\eta>0$ and $T>0$, $\exists\,0<\delta<1$ and an integer N_0 such that for all $t\in[0,T]$*

$$\frac{1}{\delta}\mathbb{P}\Bigg(\sup_{s\in[t,t+\delta]}\Bigg|\max_{i\le N}\frac{\mathbf{N}_A^{(N)}\left(sN^3\log N\right)-\mathbf{N}_{S,i}^{(N)}\left(sN^3\log N\right)+Q_i^{(N)}(0)}{N\log N} - \max_{i\le N}\frac{\mathbf{N}_A^{(N)}\left(tN^3\log N\right)-\mathbf{N}_{S,i}^{(N)}\left(tN^3\log N\right)+Q_i^{(N)}(0)}{N\log N}\Bigg|>\epsilon\Bigg)<\eta,\quad N\ge N_0. \tag{2.3.36}$$

Furthermore, for all η there exists an $a>0$ such that

$$\mathbb{P}\left(\frac{Q_{(\alpha,\beta)}^{(N)}(0)}{N\log N}>a\right)<\eta. \tag{2.3.37}$$

Proof. First, we observe that for a random variable X, $\mathbb{P}\big(|X|>\epsilon\big)\le\mathbb{P}(X>\epsilon)+\mathbb{P}(-X>\epsilon)$. Thus, we can remove the absolute values in (2.3.36) and examine both cases. Since both cases have similar proofs, we only write down the proof for the first case.

$$\frac{1}{\delta}\mathbb{P}\Bigg(\sup_{s\in[t,t+\delta]}\Bigg(\max_{i\le N}\frac{\mathbf{N}_A^{(N)}\left(sN^3\log N\right)-\mathbf{N}_{S,i}^{(N)}\left(sN^3\log N\right)+Q_i^{(N)}(0)}{N\log N} - \max_{i\le N}\frac{\mathbf{N}_A^{(N)}\left(tN^3\log N\right)-\mathbf{N}_{S,i}^{(N)}\left(tN^3\log N\right)+Q_i^{(N)}(0)}{N\log N}\Bigg)>\epsilon\Bigg)$$

$$\leq \frac{1}{\delta}\mathbb{P}\left(\sup_{s\in[t,t+\delta]}\left(\max_{i\leq N}\left(\frac{\mathbf{N}_A^{(N)}\left(sN^3\log N\right)-\mathbf{N}_{S,i}^{(N)}\left(sN^3\log N\right)}{N\log N}\right.\right.\right.$$
$$\left.\left.\left.-\frac{\mathbf{N}_A^{(N)}\left(tN^3\log N\right)-\mathbf{N}_{S,i}^{(N)}\left(tN^3\log N\right)}{N\log N}\right)>\epsilon\right)\right)$$
$$=\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\left(\max_{i\leq N}\frac{\mathbf{N}_A^{(N)}\left(sN^3\log N\right)-\mathbf{N}_{S,i}^{(N)}\left(sN^3\log N\right)}{N\log N}\right)>\epsilon\right)+o(1).$$

This is the same expression as Equation (2.3.34). In Lemma 2.11, it is proven that this expression will be small. At $t=0$, we should choose $a>0$ such that (2.3.37) holds for $N\geq N_0$. This is the case because we know that $Q_{(\alpha,\beta)}^{(N)}(0)/(N\log N)\xrightarrow{\mathbb{P}}q(0)$, as $N\to\infty$. The lemma follows. □

Corollary 2.1 (Tightness)**.** *The process* $\left(Q_{(\alpha,\beta)}^{(N)}(tN^3\log N)\big/\left(N\log N\right),t\in[0,T]\right)$ *is tight.*

Proof. The process $\left(Q_{(\alpha,\beta)}^{(N)}(tN^3\log N)\big/\left(N\log N\right),t\in[0,T]\right)$ can be written as a maximum of two processes. In Lemmas 2.11 and 2.12, it is proven that these processes are tight. Then from Lemma 2.10, it follows that $\left(Q_{(\alpha,\beta)}^{(N)}(tN^3\log N)\big/\left(N\log N\right),t\in[0,T]\right)$ is tight. □

Proof of Theorem 2.1. In Theorem 2.2, we proved that for fixed t, the stochastic process converges in probability to a constant, in Theorem 2.3, we proved that the finite-dimensional distributions converge and in Corollary 2.1, we showed that the process is tight. Thus the convergence holds u.o.c. □

We now prove that the scaled process in steady state converges to the constant $\alpha/(2\beta)$.

Proof of Proposition 2.2. Since we look at the system in steady state, we can assume w.l.o.g. that $Q_{(\alpha,\beta)}^{(N)}(0)=0$. Then, we have

$$\frac{Q_{(\alpha,\beta)}^{(N)}(\infty)}{N\log N}\geq_{st.}\frac{Q_{(\alpha,\beta)}^{(N)}\left(\alpha/(2\beta^2)N^3\log N\right)}{N\log N},$$

because $Q_{(\alpha,\beta)}^{(N)}(n)\stackrel{d}{=}\max_{i\leq N}\sup_{0\leq k\leq n}(\mathbf{N}_A^{(N)}(k)-\mathbf{N}_{S,i}^{(N)}(k))$. We know by Lemma 2.5 that

$$\frac{Q_{(\alpha,\beta)}^{(N)}\left(\alpha/(2\beta^2)N^3\log N\right)}{N\log N}\xrightarrow{\mathbb{P}}\frac{\alpha}{2\beta},$$

as $N\to\infty$. Furthermore, we know by Lemma 2.7 that for all $\delta>0$,

$$\limsup_{N\to\infty}\mathbb{P}\left(\frac{Q_{(\alpha,\beta)}^{(N)}(\infty)}{N\log N}>\frac{\alpha}{2\beta}+\delta\right)=0.$$

The proposition follows. □

2.4. Taylor expansion of $\theta_A^{(u,N)}$

In Lemma 2.7, we used a second-order Taylor approximation to find an asymptotic solution of a moment-generating function. In this section, we elaborate on the techniques how to find this asymptotic solution.

The parameter $\theta_A^{(u,N)}$ is the strictly positive solution to the equation

$$\mathbb{E}\left[e^{\theta_A^{(u,N)}X^{(u,N)}(j)}\right] = \left(\frac{\alpha}{N}+\frac{\beta}{N^2}\right)\exp\left(\theta_A^{(u,N)}\left(-1+\frac{\alpha}{N}+\frac{\beta}{N^2}-\epsilon(N)\right)\right) + \left(1-\frac{\alpha}{N}-\frac{\beta}{N^2}\right)\exp\left(\theta_A^{(u,N)}\left(\frac{\alpha}{N}+\frac{\beta}{N^2}-\epsilon(N)\right)\right) = 1,$$

with $\epsilon(N) = m/N^2$. We found an approximation of $\theta_A^{(u,N)}$, of $2m/(\alpha N)$. To investigate the behavior of $\theta_A^{(u,N)}$ more carefully, we look at the function $\theta(x)$ such that

$$f(x,\theta(x)) = \left(\alpha x+\beta x^2\right)\exp\left(\theta(x)\left(-1+\alpha x+\beta x^2-mx^2\right)\right) + \left(1-\alpha x-\beta x^2\right)\exp\left(\theta(x)\left(\alpha x+\beta x^2-mx^2\right)\right) = 1.$$

When we set $x_N = 1/N$, we get $f(x_N,\theta(x_N)) = \mathbb{E}\left[e^{\theta_A^{(u,N)}X^{(u,N)}(j)}\right] = 1$, thus $\theta(x_N) = \theta_A^{(u,N)}$. We are interested in the case where N is large, therefore we have to investigate f for x around 0. Since $f(x,\theta(x)) = 1$, we know that $f^{(n)}(0,\theta(0)) = 0$ for all $n \geq 1$. When we solve these equations for θ iteratively, we can find $\theta^{(i)}(0)$ for all $i \geq 0$ and we get a Taylor expansion of $\theta(x)$ around 0. Since $f(x,\theta(x)) = 1$, we know that

$$\left.\frac{d}{dx}f(x,\theta(x))\right|_{x=0} = -\alpha+\alpha e^{-\theta(0)}+\alpha\theta(0) = 0.$$

Hence, $\theta(0) = 0$. When we look at the second and the third derivative of $f(x,\theta(x))$ around 0, while using that $\theta(0) = 0$, we see

$$\left.\frac{d^2}{dx^2}f(x,\theta(x))\right|_{x=0} = 0,$$

and

$$\left.\frac{d^3}{dx^3}f(x,\theta(x))\right|_{x=0} = 3\theta'(0)\left(\alpha\theta'(0)-2m\right).$$

Because we know that $f(x,\theta(x)) = 1$, we solve $3\theta'(0)\left(\alpha\theta'(0)-2m\right) = 0$. This gives $\theta'(0) = 0$ or $\theta'(0) = 2m/\alpha$. $\theta'(0) = 0$ indicates the situation that $\theta \equiv 0$. If we now use the information

that $\theta'(0) = 2m/\alpha$ and look at the fourth derivative of f, we see that

$$\left.\frac{d^4}{dx^4} f(x,\theta(x))\right|_{x=0} = 4m\left(3\theta''(0) - \frac{4m\left(3\alpha^2 - 3\beta + 2m\right)}{\alpha^2}\right) = 0.$$

This gives that $\theta''(0) = 4m\left(3\alpha^2 - 3\beta + 2m\right)/3\alpha^2$. In general, we can compute each derivative of $\theta(0)$ iteratively. This gives

$$\theta(x) = \frac{2m}{\alpha}x + \frac{4m\left(3\alpha^2 - 3\beta + 2m\right)}{3\alpha^2}\frac{x^2}{2} + O(x^3).$$

Since the function $f(x,\theta) - 1$ is analytic, we know by the implicit function theorem that the solution $\theta(x)$ is also analytic. So for $x = 1/N$ and N is large enough we know that $\theta_A^{(u,N)} = 2m/(\alpha N) + O\left(1/N^2\right)$.

2.5. Proofs of Lemmas 2.1, 2.2, 2.3, and 2.4

Proof of Lemma 2.1. We take $s > t > 0$. We write $t_N = tN^3 \log N$, $s_N = sN^3 \log N$, etc. We first prove that for $s_N > t_N$, the following upper bound holds:

$$\begin{aligned}\frac{Q^{(N)}_{(\alpha,\beta)}(s_N)}{N\log N} - \frac{Q^{(N)}_{(\alpha,\beta)}(t_N)}{N\log N} \leq & \max_{i\leq N}\left|\tilde{R}_i^{(N)}(s_N) - \tilde{R}_i^{(N)}(t_N)\right| \\ & + \max_{i\leq N}\sup_{r\in[t_N,s_N]}\left(\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(r)\right). \qquad (2.5.1)\end{aligned}$$

Due to the defined auxiliary processes in Definition 2.4, we can write the maximum queue length in terms of $\tilde{R}_i^{(N)}$ as in Equation (2.3.6). Similarly, we can rewrite $Q^{(N)}_{(\alpha,\beta)}(s_N)/(N\log N) - Q^{(N)}_{(\alpha,\beta)}(t_N)/(N\log N)$ as

$$\begin{aligned}&\max_{i\leq N}\sup_{r\in[0,s_N]}\left(\tilde{R}_i^{(N)}(s_N) - \tilde{R}_i^{(N)}(r)\right) - \max_{i\leq N}\sup_{u\in[0,t_N]}\left(\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(u)\right) \\ &\quad = \max_{i\leq N}\left[\tilde{R}_i^{(N)}(s_N) - \tilde{R}_i^{(N)}(t_N) + \sup_{r\in[0,s_N]}\left(\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(r)\right)\right] \\ &\qquad - \max_{i\leq N}\sup_{u\in[0,t_N]}\left(\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(u)\right).\end{aligned}$$

Therefore, the following upper bounds hold:

$$\begin{aligned}&\frac{Q^{(N)}_{(\alpha,\beta)}(s_N)}{N\log N} - \frac{Q^{(N)}_{(\alpha,\beta)}(t_N)}{N\log N} \\ &\quad \leq \max_{i\leq N}\left(\tilde{R}_i^{(N)}(s_N) - \tilde{R}_i^{(N)}(t_N)\right) + \max_{i\leq N}\sup_{r\in[0,s_N]}\left(\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(r)\right) \\ &\qquad - \max_{i\leq N}\sup_{u\in[0,t_N]}\left(\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(u)\right)\end{aligned}$$

$$\leq \max_{i\leq N}\left(\tilde{R}_i^{(N)}(s_N)-\tilde{R}_i^{(N)}(t_N)\right)$$
$$+\max_{i\leq N}\left[\sup_{r\in[0,s_N]}\left(\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(r)\right)-\sup_{u\in[0,t_N]}\left(\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(u)\right)\right].$$

Observe that both random variables $\sup_{r\in[0,s_N]}\left(\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(r)\right)$ and $\sup_{u\in[0,t_N]}\left(\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(u)\right)$ are non-negative. Furthermore,

$$\sup_{r\in[0,s_N]}\left(\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(r)\right)-\sup_{u\in[0,t_N]}\left(\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(u)\right)$$
$$\leq \sup_{r\in[t_N,s_N]}\left(\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(r)\right).$$

Now, we can conclude that

$$\frac{Q_{(\alpha,\beta)}^{(N)}(s_N)}{N\log N}-\frac{Q_{(\alpha,\beta)}^{(N)}(t_N)}{N\log N}$$
$$\leq \max_{i\leq N}\left(\tilde{R}_i^{(N)}(s_N)-\tilde{R}_i^{(N)}(t_N)\right)$$
$$+\max_{i\leq N}\left[\sup_{r\in[0,s_N]}\left(\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(r)\right)-\sup_{u\in[0,t_N]}\left(\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(u)\right)\right]$$
$$\leq \max_{i\leq N}\left|\tilde{R}_i^{(N)}(s_N)-\tilde{R}_i^{(N)}(t_N)\right|+\max_{i\leq N}\sup_{r\in[t_N,s_N]}\left(\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(r)\right),$$

and hence the inequality in Equation (2.5.1) is satisfied. We can similarly deduce the lower bound

$$\frac{Q_{(\alpha,\beta)}^{(N)}(s_N)}{N\log N}-\frac{Q_{(\alpha,\beta)}^{(N)}(t_N)}{N\log N}\geq -\max_{i\leq N}\left|\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(s_N)\right|. \tag{2.5.2}$$

To show this, we write

$$\frac{Q_{(\alpha,\beta)}^{(N)}(s_N)}{N\log N}-\frac{Q_{(\alpha,\beta)}^{(N)}(t_N)}{N\log N}$$
$$=\max_{i\leq N}\sup_{r\in[0,s_N]}\left(\tilde{R}_i^{(N)}(s_N)-\tilde{R}_i^{(N)}(r)\right)$$
$$-\max_{i\leq N}\sup_{u\in[0,t_N]}\left(\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(u)\right)$$
$$=\max_{i\leq N}\sup_{r\in[0,s_N]}\left(\tilde{R}_i^{(N)}(s_N)-\tilde{R}_i^{(N)}(r)\right)$$
$$-\max_{i\leq N}\left[\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(s_N)+\sup_{u\in[0,t_N]}\left(\tilde{R}_i^{(N)}(s_N)-\tilde{R}_i^{(N)}(u)\right)\right]$$

$$\geq \max_{i\leq N} \sup_{r\in[0,s_N]} \left(\tilde{R}_i^{(N)}(s_N) - \tilde{R}_i^{(N)}(r)\right)$$
$$- \max_{i\leq N} \left(\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(s_N)\right) - \max_{i\leq N} \sup_{u\in[0,t_N]} \left(\tilde{R}_i^{(N)}(s_N) - \tilde{R}_i^{(N)}(u)\right).$$

Observe that

$$\sup_{r\in[0,s_N]} \left(\tilde{R}_i^{(N)}(s_N) - \tilde{R}_i^{(N)}(r)\right) \geq \sup_{u\in[0,t_N]} \left(\tilde{R}_i^{(N)}(s_N) - \tilde{R}_i^{(N)}(u)\right),$$

because $s_N > t_N$, so on the left side of the inequality, the supremum is taken over a larger interval than on the right side of the inequality. From this, we can conclude that

$$\frac{Q_{(\alpha,\beta)}^{(N)}(s_N)}{N\log N} - \frac{Q_{(\alpha,\beta)}^{(N)}(t_N)}{N\log N}$$
$$\geq \max_{i\leq N} \sup_{r\in[0,s_N]} \left(\tilde{R}_i^{(N)}(s_N) - \tilde{R}_i^{(N)}(r)\right)$$
$$- \max_{i\leq N} \left(\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(s_N)\right) - \max_{i\leq N} \sup_{u\in[0,t_N]} \left(\tilde{R}_i^{(N)}(s_N) - \tilde{R}_i^{(N)}(u)\right)$$
$$\geq -\max_{i\leq N} \left(\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(s_N)\right) \geq -\max_{i\leq N} \left|\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(s_N)\right|,$$

and indeed (2.5.2) holds. Combining (2.5.1) and (2.5.2) gives

$$\left|\frac{Q_{(\alpha,\beta)}^{(N)}(s_N)}{N\log N} - \frac{Q_{(\alpha,\beta)}^{(N)}(t_N)}{N\log N}\right|$$
$$\leq \max_{i\leq N} \left|\tilde{R}_i^{(N)}(s_N) - \tilde{R}_i^{(N)}(t_N)\right| + \max_{i\leq N} \sup_{r\in[t_N,s_N]} \left(\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(r)\right).$$

Thus,

$$\sup_{s\in[t_N,t_N+\delta_N]} \left|\frac{Q_{(\alpha,\beta)}^{(N)}(s)}{N\log N} - \frac{Q_{(\alpha,\beta)}^{(N)}(t_N)}{N\log N}\right|$$
$$\leq \sup_{s\in[t_N,t_N+\delta_N]} \max_{i\leq N} \left|\tilde{R}_i^{(N)}(s) - \tilde{R}_i^{(N)}(t_N)\right| + \sup_{s\in[t_N,t_N+\delta_N]} \max_{i\leq N} \left(\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(s)\right). \tag{2.5.3}$$

Since both random variables $\sup_{s\in[t_N,t_N+\delta_N]} \left(\tilde{R}_i^{(N)}(t_N) - \tilde{R}_i^{(N)}(s)\right)$ and $\sup_{s\in[t_N,t_N+\delta_N]} \left(\tilde{R}_i^{(N)}(s) - \tilde{R}_i^{(N)}(t_N)\right)$ are non-negative, we have that

$$\sup_{s\in[t_N,t_N+\delta_N]} \max_{i\leq N} \left|\tilde{R}_i^{(N)}(s) - \tilde{R}_i^{(N)}(t_N)\right|$$

$$\leq \sup_{s\in[t_N,t_N+\delta_N]} \max_{i\leq N}\left(\tilde{R}_i^{(N)}(s)-\tilde{R}_i^{(N)}(t_N)\right)+\sup_{s\in[t_N,t_N+\delta_N]} \max_{i\leq N}\left(\tilde{R}_i^{(N)}(t_N)-\tilde{R}_i^{(N)}(s)\right). \tag{2.5.4}$$

Combining the inequalities in (2.5.3) and (2.5.4) gives us the desired result. □

Proof of Lemma 2.2. We first prove the bound in (2.3.9). The random variable $\tilde{\mathbf{N}}_{S,i}^{(N)}(n)$ is a sum of independent and identically distributed random variables with $\mathbb{E}\left[\pm\tilde{\mathbf{N}}_{S,i}^{(N)}(1)\right]=0$, and $\mathrm{Var}\left(\pm\tilde{\mathbf{N}}_{S,i}^{(N)}(1)\right)=\left(1-\alpha/N\right)\alpha/N^3$. So, the random variable

$$\pm M_{i,1}^{(N)}(t)=\frac{\pm\tilde{\mathbf{N}}_{S,i}^{(N)}\left(tN^3\log N\right)\sqrt{tN^3\log N}}{\sqrt{\alpha t(1-\alpha/N)\log N\lfloor tN^3\log N\rfloor}}$$

has mean 0 and variance 1, and satisfies the central limit theorem. From [109] it follows that for all y,

$$\left|\mathbb{P}\left(\pm M_{i,1}^{(N)}(t)<y\right)-\Phi(y)\right| \leq C\frac{1}{\sqrt{\lfloor tN^3\log N\rfloor}}\,\mathbb{E}\left[\left|\frac{\pm\tilde{\mathbf{N}}_{S,i}^{(N)}(1)}{\sqrt{\alpha t(1-\alpha/N)\log N}}\sqrt{tN^3\log N}\right|^3\right]\frac{1}{1+|y|^3}.$$

Observe that for N large enough and $0<\epsilon<t$, $\left\lfloor tN^3\log N\right\rfloor>(t-\epsilon)N^3\log N$. We also have that

$$\mathbb{E}\left[\left|\frac{\pm\tilde{\mathbf{N}}_{S,i}^{(N)}(1)}{\sqrt{\alpha t(1-\alpha/N)\log N}}\sqrt{tN^3\log N}\right|^3\right] = \frac{N^4\sqrt{N}}{\alpha(1-\alpha/N)\sqrt{\alpha(1-\alpha/N)}N^3}\left(\left(1-\frac{\alpha}{N}\right)^3\frac{\alpha}{N}+\frac{\alpha^3}{N^3}\left(1-\frac{\alpha}{N}\right)\right)\leq 2\sqrt{N}\frac{(1+\alpha^2)}{\sqrt{\alpha}},$$

which holds for $N>\max(1,2\alpha)$. Thus, the bound in (2.3.9) follows for N large enough, with $c_t=2C(1+\alpha^2)/\sqrt{\alpha(t-\epsilon)}$.

The proof of the bound given in (2.3.10) follows along the same lines. The random variable $\tilde{\mathbf{N}}_{S,i}^{(N)}(n)$ is a sum of independent and identically distributed random variables with $\mathbb{E}\left[\pm\tilde{\mathbf{N}}_{S,i}^{(N)}(1)\right]=0$, and $\mathrm{Var}\left(\pm\tilde{\mathbf{N}}_{S,i}^{(N)}(1)\right)=\left(1-\alpha/N\right)\alpha/N^3$. So, $\pm M_{i,2}^{(N)}(t)=\pm\tilde{\mathbf{N}}_{S,i}^{(N)}\left(tN^3\right)\sqrt{tN^3}/\sqrt{\alpha t(1-\alpha/N)\lfloor tN^3\rfloor}$ has mean 0 and variance 1, and satisfies the central limit theorem. From [109] it follows that for all y,

$$\left|\mathbb{P}\left(\pm M_{i,2}^{(N)}(t)<y\right)-\Phi(y)\right|\leq C\frac{1}{\sqrt{\lfloor tN^3\rfloor}}\,\mathbb{E}\left[\left|\frac{\pm\tilde{\mathbf{N}}_{S,i}^{(N)}(1)}{\sqrt{\alpha t(1-\alpha/N)}}\sqrt{tN^3}\right|^3\right]\frac{1}{1+|y|^3}.$$

Similar to before, we have

$$\mathbb{E}\left[\left|\frac{\pm\tilde{\mathbf{N}}_{S,i}^{(N)}(1)}{\sqrt{\alpha t(1-\alpha/N)}}\sqrt{tN^3}\right|^3\right] \leq 2\sqrt{N}\frac{(1+\alpha^2)}{\sqrt{\alpha}},$$

for N large enough, thus the bound in (2.3.10) follows for N large enough. □

Proof of Lemma 2.3. We first prove the limsup in (2.3.11). We write

$$K_1(x) = 1 - \left(\Phi\left(\frac{x^{2/5}\sqrt{N^3}\log N}{\sqrt{\alpha\left(1-\alpha/N\right)\lfloor tN^3\log N\rfloor}}\right) - \frac{c_t}{N\sqrt{\log N}}\frac{1}{1+\left(\frac{x^{2/5}\sqrt{N^3}\log N}{\sqrt{\alpha\left(1-\alpha/N\right)\lfloor tN^3\log N\rfloor}}\right)^3}\right)^N$$

and

$$K_2(x) = -\Phi\left(\frac{x^{2/5}\sqrt{N^3}\log N}{\sqrt{\alpha\left(1-\alpha/N\right)\lfloor tN^3\log N\rfloor}}\right)^N + \left(1+\frac{c_t}{N\sqrt{\log N}}\frac{1}{1+\left(\frac{x^{2/5}\sqrt{N^3}\log N}{\sqrt{\alpha\left(1-\alpha/N\right)\lfloor tN^3\log N\rfloor}}\right)^3}\right)^N.$$

We argue that for all $x > 0$, $K_1(x) \leq K_2(x)$. This inequality has the form $1-(u-v)^N \leq -u^N + (1+v)^N$ for $0 \leq u \leq 1$, $0 \leq v \leq 1$, and $N > 1$. To see why this is true, we first observe for $v = 0$, we get that $1-(u-v)^N = -u^N + (1+v)^N = 1-u^N$. Furthermore, when we take the derivatives of $1-(u-v)^N$ and $-u^N + (1+v)^N$ with respect to v, we get

$$\frac{\partial}{\partial v}(1-(u-v)^N) = N(u-v)^{N-1},$$

and

$$\frac{\partial}{\partial v}(-u^N + (1+v)^N) = N(1+v)^{N-1}.$$

Because both u and v are in $[0,1]$, we see that $\frac{\partial}{\partial v}(1-(u-v)^N) = N(u-v)^{N-1} \leq \frac{\partial}{\partial v}(-u^N + (1+v)^N) = N(1+v)^{N-1}$. Thus, the inequality follows.

Now, we can write

$$\begin{aligned}
&\mathbb{E}\left[\max\left(0, \frac{\max_{i\le N} \pm\tilde{\mathbf{N}}^{(N)}_{S,i}\left(tN^3\log N\right)}{\log N}\right)^{5/2}\right] \\
&= \int_0^\infty \mathbb{P}\left(\frac{\max_{i\le N} \pm\tilde{\mathbf{N}}^{(N)}_{S,i}\left(tN^3\log N\right)}{\log N} > x^{2/5}\right)dx \\
&= \int_0^\infty \mathbb{P}\left(\max_{i\le N} \pm M^{(N)}_{i,1}(t) > x^{2/5}\frac{\log N}{\sqrt{\alpha t\left(1-\alpha/N\right)\log N}}\frac{\sqrt{tN^3\log N}}{\sqrt{\lfloor tN^3\log N\rfloor}}\right)dx \\
&= \int_0^\infty 1-\mathbb{P}\left(\pm M^{(N)}_{i,1}(t) < \frac{x^{2/5}\sqrt{N^3}\log N}{\sqrt{\alpha\left(1-\alpha/N\right)\lfloor tN^3\log N\rfloor}}\right)^N dx \\
&\le \int_0^\infty K_1(x)dx \le \int_0^\infty K_2(x)dx \\
&= \mathbb{E}\left[\max\left(0, \frac{\sqrt{\alpha t(1-\alpha/N)}\max_{i\le N} X_i}{\sqrt{\log N}}\frac{\sqrt{\lfloor tN^3\log N\rfloor}}{\sqrt{tN^3\log N}}\right)^{5/2}\right] \qquad (2.5.5) \\
&\quad + \int_0^\infty -1 + \left(1 + \frac{c_t}{N\sqrt{\log N}}\frac{1}{1+\left(\frac{x^{2/5}\sqrt{N^3}\log N}{\sqrt{\alpha\left(1-\alpha/N\right)\lfloor tN^3\log N\rfloor}}\right)^3}\right)^N dx, \qquad (2.5.6)
\end{aligned}$$

with X_i standard normally distributed. By Pickands' theorem [123, Thm. 3.2, p. 888], we know that the expectation in (2.5.5) converges to $(2\alpha t)^{5/4}$. Furthermore, the term in (2.5.6) is upper bounded by

$$\int_0^\infty -1 + \exp\left(c_t \Big/ \left(\sqrt{\log N}\left(1+\left(\frac{x^{2/5}\sqrt{N^3}\log N}{\sqrt{\alpha\left(1-\alpha/N\right)\lfloor tN^3\log N\rfloor}}\right)^3\right)\right)\right)dx. \qquad (2.5.7)$$

We let $y = 1\Big/\left(1+\left(\frac{x^{2/5}\sqrt{N^3}\log N}{\sqrt{\alpha\left(1-\alpha/N\right)\lfloor tN^3\log N\rfloor}}\right)^3\right)$. Then the term in (2.5.7) can be rewritten as

$$\left(\frac{\lfloor tN^3\log N\rfloor}{N^3(\log N)^2}\right)^{5/4}\int_0^1 \frac{5\left(\sqrt{\alpha\left(1-\alpha/N\right)}\right)^{5/2}}{6(1-y)^{1/6}y^{11/6}}\left(-1+e^{\frac{c_t}{\sqrt{\log N}}y}\right)dy \overset{N\to\infty}{\longrightarrow} 0. \qquad (2.5.8)$$

The limsup in (2.3.11) follows.

The analysis to prove the limsup in (2.3.12) is similar. The term $\exp(c_t/\sqrt{\log N}y)$ on the left-hand side in (2.5.8) should be replaced with $\exp(c_t y)$, which is due to the differences between (2.3.9) and (2.3.10). Furthermore, as we have a different temporal and spatial scaling, the term

$$\left(\frac{\lfloor tN^3 \log N \rfloor}{N^3(\log N)^2}\right)^{5/4}$$

should be replaced with

$$\left(\frac{\lfloor tN^3 \rfloor}{N^3(\log N)}\right)^{5/4}.$$

Still, we get that the resulting quantity

$$\left(\frac{\lfloor tN^3 \rfloor}{N^3(\log N)}\right)^{5/4} \int_0^1 \frac{5\left(\sqrt{\alpha\left(1-\alpha/N\right)}\right)^{5/2}}{6(1-y)^{1/6}y^{11/6}}\left(-1+e^{c_t y}\right)dy$$

converges to 0 as $N \to \infty$. □

Proof of Lemma 2.4. In order to prove that

$$\max_{i\le N}\left(\frac{\sqrt{\alpha t}X_i}{\sqrt{\log N}} + \frac{Q_i^{(N)}(0)}{N\log N}\right) \xrightarrow{\mathbb{P}} g(t,q(0)),$$

as $N \to \infty$, we first observe that, from the definition of $Q_i^{(N)}(0)$ in Theorem 2.1, it is easy to see that

$$\left|\max_{i\le N}\left(\frac{\sqrt{\alpha t}X_i}{\sqrt{\log N}} + \frac{Q_i^{(N)}(0)}{N\log N}\right) - \max_{i\le N}\left(\frac{\sqrt{\alpha t}X_i}{\sqrt{\log N}} + \frac{r_N U_i}{N\log N}\right)\right| \le \max_{i\le N}\frac{V_i^{(N)}}{N\log N} + \frac{1}{N\log N} \xrightarrow{\mathbb{P}} 0,$$

as $N \to \infty$. Thus, from this, it follows that

$$\max_{i\le N}\left(\frac{\sqrt{\alpha t}X_i}{\sqrt{\log N}} + \frac{Q_i^{(N)}(0)}{N\log N}\right) \xrightarrow{\mathbb{P}} g(t,q(0)) \iff \max_{i\le N}\left(\frac{\sqrt{\alpha t}X_i + \frac{r_N U_i}{N\sqrt{\log N}}}{\sqrt{\log N}}\right) \xrightarrow{\mathbb{P}} g(t,q(0)),$$

as $N \to \infty$. Let us first consider that U_i satisfies Assumption 2.4, thus U_i has a finite right endpoint. Theorem 2.1 says that when U_i has a finite right endpoint, that $g(t,q(0)) = \sqrt{2\alpha t} + q(0)$. To prove this, first observe that $g(t,q(0)) \le \sqrt{2\alpha t} + q(0)$ because $\max_{i\le N}\sqrt{\alpha t}X_i/\sqrt{\log N} \xrightarrow{\mathbb{P}} \sqrt{2\alpha t}$ and $Q_{(\alpha,\beta)}^{(N)}(0)/(N\log N) \xrightarrow{\mathbb{P}} q(0)$ as $N \to \infty$. Hence,

the only thing we need to establish is that for all $\gamma < \sqrt{2\alpha t} + q(0)$,

$$N\,\mathbb{P}\left(\sqrt{\alpha t}X_i + \frac{r_N U_i}{N\sqrt{\log N}} \geq \gamma\sqrt{\log N}\right) \overset{N\to\infty}{\longrightarrow} \infty.$$

When $\gamma < \sqrt{2\alpha t}$, this is obvious, because $U_i > 0$, and $\max_{i\leq N}\sqrt{\alpha t}X_i/\sqrt{\log N} \overset{\mathbb{P}}{\longrightarrow} \sqrt{2\alpha t}$ as $N \to \infty$. So, let us assume that $\sqrt{2\alpha t} \leq \gamma < \sqrt{2\alpha t} + q(0)$. Because U_i has a finite right endpoint, $r_N/(N\sqrt{\log N}) = \sqrt{\log N}$. By convolution, we have that

$$\begin{aligned}
&N\,\mathbb{P}\left(\sqrt{\alpha t}X_i + \sqrt{\log N}U_i \geq \gamma\sqrt{\log N}\right)\\
&\quad = N\,\mathbb{P}\left(\sqrt{\alpha t}X_i \geq \gamma\sqrt{\log N}\right)\\
&\qquad + N\int_{-\infty}^{\gamma\sqrt{\log N}} \mathbb{P}\left(\sqrt{\log N}U_i > \gamma\sqrt{\log N} - z\right)\frac{e^{-z^2/(2\alpha t)}}{\sqrt{2\alpha t\pi}}dz\\
&\quad \geq N\int_{-\infty}^{\gamma} \mathbb{P}(U_i > \gamma - v)\frac{N^{-v^2/(2\alpha t)}}{\sqrt{2\alpha t\pi}}\sqrt{\log N}dv\\
&\quad = \int_{\gamma - q(0)}^{\gamma} \mathbb{P}(U_i > \gamma - v)\frac{N^{1-v^2/(2\alpha t)}}{\sqrt{2\alpha t\pi}}\sqrt{\log N}dv.
\end{aligned}$$

From this, it follows, that when $1 - v^2/(2\alpha t) > 0$, this integral converges to ∞. We chose $\sqrt{2\alpha t} \leq \gamma < \sqrt{2\alpha t} + q(0)$; thus the lower bound $\gamma - q(0)$ in the integral is smaller than $\sqrt{2\alpha t}$ and hence this integral converges to ∞. Thus $g(t, q(0)) = \sqrt{2\alpha t} + q(0)$.

Let us now consider the scenario described in Assumption 2.5. Then $g(t, q(0))$ satisfies the limit given in (2.2.7). We have the straightforward limit result that for standard normally distributed X_i, $\lim_{t\to\infty} -\log(\mathbb{P}(X_i \geq ut))/-\log(\mathbb{P}(X_i \geq t)) = u^2$. Furthermore, following the assumptions on U_i in Theorem 2.1, we know that $\lim_{t\to\infty} -\log(\mathbb{P}(U_i \geq vt))/-\log(\mathbb{P}(U_i \geq t)) = h(v)$. Thus from Lemma 2.13, we know that for sequences $(a_N, N \geq 1), (b_N, N \geq 1)$ with $\mathbb{P}(X_i \geq a_N) = \mathbb{P}(U_i \geq b_N) = 1/N$, that

$$\max_{i\leq N}\left(\frac{X_i}{a_N} + \frac{U_i}{b_N}\right) \overset{\mathbb{P}}{\longrightarrow} \sup_{(u,v)}\{u + v : u^2 + h(v) \leq 1, 0 \leq u \leq 1, 0 \leq v \leq 1\},$$

as $N \to \infty$. Now, we can use this result to prove that
$\max_{i\leq N}\left(\sqrt{\alpha t}X_i/\sqrt{\log N} + r_N U_i/(N\log N)\right)$ converges to the limit in (2.2.7). We first observe that

$$\max_{i\leq N}\left(\frac{\sqrt{\alpha t}X_i}{\sqrt{\log N}} + \frac{r_N U_i}{N\log N}\right) = \max_{i\leq N}\left(\sqrt{2\alpha t}\frac{X_i}{\sqrt{2\log N}} + q(0)\frac{r_N U_i}{q(0)N\log N}\right).$$

We have that $a_N/\sqrt{2\log N} \overset{N\to\infty}{\longrightarrow} 1$, because $\max_{i\leq N} X_i/a_N \overset{\mathbb{P}}{\longrightarrow} 1$,

and $\max_{i\le N} X_i/\sqrt{2\log N} \xrightarrow{\mathbb{P}} 1$, as $N\to\infty$. Analogously, $b_N q(0) N\log N/r_N \overset{N\to\infty}{\longrightarrow} 1$. Thus,

$$\left|\max_{i\le N}\left(\sqrt{2\alpha t}\frac{X_i}{a_N}+q(0)\frac{U_i}{b_N}\right)-\max_{i\le N}\left(\frac{\sqrt{\alpha t}X_i}{\sqrt{\log N}}+\frac{r_N U_i}{N\log N}\right)\right|\xrightarrow{\mathbb{P}} 0,$$

as $N\to\infty$. With an analogous proof as before, $\max_{i\le N}\left(\sqrt{2\alpha t}X_i/a_N+q(0)U_i/b_N\right)$ converges to the limit in (2.2.7). □

2.6. Extreme values of sums of random variables

In this section, we prove convergence results of the maximum of N sums of k random variables. In order to do so, we use and extend results from [44] and [55].

Lemma 2.13. *Consider sequences of continuous random variables $(Y_i^{(1)}, i\ge 1)$, $(Y_i^{(2)}, i\ge 1)$, …, $(Y_i^{(k)}, i\ge 1)$, where all random variables in the sequence $(Y_i^{(j)}, i\ge 1)$ are identically and independently distributed and have infinite right endpoints. Furthermore, $Y_i^{(j)}$ and $Y_m^{(l)}$ are independent for all $j,l\in\{1,2,\dots,k\}$ and $i,m\ge 1$, and $Y_i^{(j)}$ satisfies Assumption 2.5 with function $h^{(j)}(u^{(j)})$. Finally, we have sequences $(a_N^{(j)}, N\ge 1)$ such that $\mathbb{P}(Y_i^{(j)}\ge a_N^{(j)})=1/N$. We assume that the random variables $Y_i^{(j)}$ are relatively stable, thus $\max_{i\le N} Y_i^{(j)}/a_N^{(j)} \xrightarrow{\mathbb{P}} 1$, as $N\to\infty$. Then*

$$\max_{i\le N}\left(\sum_{j=1}^{k}\frac{Y_i^{(j)}}{a_N^{(j)}}\right)\xrightarrow{\mathbb{P}}\sup_{(u^{(j)},j\le k)}\left\{\sum_{j=1}^{k}u^{(j)}:\sum_{j=1}^{k}h^{(j)}(u^{(j)})\le 1, u^{(j)}\le 1\ \forall\, j\le k\right\},$$

as $N\to\infty$.

Proof. First, let us choose $u^{(1)},\dots,u^{(k)}$ such that $u^{(j)}\le 1$ for all j. It is a well-known result [67, Eq. (5.4.5), p. 188] that

$$\mathbb{P}\left(\cup_{i=1}^{N}\cap_{j=1}^{k}\left\{Y_i^{(j)}\ge u^{(j)}a_N^{(j)}\right\}\right)\overset{N\to\infty}{\longrightarrow}1$$

$$\iff N\,\mathbb{P}\left(\cap_{j=1}^{k}\left\{Y_i^{(j)}\ge u^{(j)}a_N^{(j)}\right\}\right)\overset{N\to\infty}{\longrightarrow}\infty.$$

From this, it follows that

$$\log N+\sum_{j=1}^{k}\log\left(\mathbb{P}\left(Y_i^{(j)}\ge u^{(j)}a_N^{(j)}\right)\right)\overset{N\to\infty}{\longrightarrow}\infty.$$

This is the case when

$$\limsup_{N\to\infty}\left(\sum_{j=1}^{k}\frac{-\log\left(\mathbb{P}\left(Y_i^{(j)}\geq u^{(j)}a_N^{(j)}\right)\right)}{\log N}\right)<1.$$

Similarly,

$$\liminf_{N\to\infty}\left(\sum_{j=1}^{k}\frac{-\log\left(\mathbb{P}\left(Y_i^{(j)}\geq u^{(j)}a_N^{(j)}\right)\right)}{\log N}\right)>1$$

$$\Rightarrow\mathbb{P}\left(\cup_{i=1}^{N}\cap_{j=1}^{k}\left\{Y_i^{(j)}\geq u^{(j)}a_N^{(j)}\right\}\right)\overset{N\to\infty}{\longrightarrow}0.$$

Because of the fact that we have $\mathbb{P}(Y_i^{(j)}\geq a_N^{(j)})=1/N$, we can conclude that

$$\lim_{N\to\infty}\left(\sum_{j=1}^{k}\frac{-\log\left(\mathbb{P}\left(Y_i^{(j)}\geq u^{(j)}a_N^{(j)}\right)\right)}{\log N}\right)$$

$$=\lim_{N\to\infty}\left(\sum_{j=1}^{k}\frac{-\log\left(\mathbb{P}\left(Y_i^{(j)}\geq u^{(j)}a_N^{(j)}\right)\right)}{-\log\left(\mathbb{P}\left(Y_i^{(j)}\geq a_N^{(j)}\right)\right)}\right)=\sum_{j=1}^{k}h^{(j)}(u^{(j)}).$$

Let us now call

$$c^\star=\sup_{(u^{(j)},j\leq k)}\left\{\sum_{j=1}^{k}u^{(j)}:\sum_{j=1}^{k}h^{(j)}(u^{(j)})\leq 1,u^{(j)}\leq 1\;\forall\;j\leq k\right\},$$

and let $\epsilon>0$ be small. Then, we distinguish two scenarios. First, we consider the case where $\#\{j\leq k:h^{(j)}(u^{(j)})=\mathbb{1}(u^{(j)}>0)\}\leq k-2$. Then, there exists a sequence $(u_\epsilon^{(1)},\ldots,u_\epsilon^{(k)})$ such that $\sum_{j=1}^{k}u_\epsilon^{(j)}=c^\star-\epsilon$, and $\sum_{j=1}^{k}h^{(j)}(u_\epsilon^{(j)})<1$. Therefore,

$$\mathbb{P}\left(\max_{i\leq N}\left(\sum_{j=1}^{k}\frac{Y_i^{(j)}}{a_N^{(j)}}\right)>c^\star-\epsilon\right)>\mathbb{P}\left(\cup_{i=1}^{N}\cap_{j=1}^{k}\left\{Y_i^{(j)}\geq u_\epsilon^{(j)}a_N^{(j)}\right\}\right)\overset{N\to\infty}{\longrightarrow}1.$$

If $\#\{j \le k : h^{(j)}(u^{(j)}) = \mathbb{1}(u^{(j)} > 0)\} \ge k-1$, we know that $c^\star = 1$, because

$$\sup_{(u^{(j)}, j \le k)} \left\{ \sum_{j=1}^{k} u^{(j)} : \sum_{j=1}^{k-1} \mathbb{1}(u^{(j)} > 0) + h^{(k)}(u^{(k)}) \le 1, u^{(j)} \le 1 \; \forall \; j \le k \right\} = 1,$$

and

$$\sup_{(u^{(j)}, j \le k)} \left\{ \sum_{j=1}^{k} u^{(j)} : \sum_{j=1}^{k} \mathbb{1}(u^{(j)} > 0) \le 1, u^{(j)} \le 1 \; \forall \; j \le k \right\} = 1.$$

Furthermore, we know that

$$\max_{i \le N} \left(\sum_{j=1}^{k} \frac{Y_i^{(j)}}{a_N^{(j)}} \right) \ge_{st.} \max_{i \le N} \left(\frac{Y_i^{(1)}}{a_N^{(1)}} \right) + \sum_{j=2}^{k} \frac{Y_i^{(j)}}{a_N^{(j)}} \xrightarrow{\mathbb{P}} 1,$$

as $N \to \infty$, because we have for sequences $(a_i, i \ge 1)$ and $(b_i, i \ge 1)$, that $\max_{i \le N}(a_i + b_i) \ge \max_{i \le N}(a_i) + b_{i^\star}$, with $i^\star$ satisfying $a_{i^\star} = \max_{i \le N}(a_i)$. Thus, at this moment we can conclude that the limit cannot be smaller than $c^\star$. To prove that

$$\mathbb{P}\left(\max_{i \le N} \left(\sum_{j=1}^{k} \frac{Y_i^{(j)}}{a_N^{(j)}} \right) > c^\star + \epsilon \right) \overset{N \to \infty}{\longrightarrow} 0, \tag{2.6.1}$$

we first observe that the boundary is given by $\{(u^{(j)}, j \le k) : \sum_{j=1}^{k} u^{(j)} = c^\star + \epsilon\}$. We already know that $N\,\mathbb{P}\Big(Y_i^{(j)} > u^{(j)} a_N\Big) \overset{N \to \infty}{\longrightarrow} 0$, for $u^{(j)} > 1$. Hence,

$$\limsup_{N \to \infty} \mathbb{P}\left(\max_{i \le N} \left(\sum_{j=1}^{k} \frac{Y_i^{(j)}}{a_N^{(j)}} \right) > c^\star + \epsilon \right) > 0,$$

would mean that there are limiting points in the set $\{(u^{(j)}, j \le k) : \sum_{j=1}^{k} u^{(j)} \ge c^\star + \epsilon, u^{(j)} \le 1 \; \forall \; j\}$. However, we know that for all $(u^{(j)}, j \le k)$ with $c^\star < \sum_{j=1}^{k} u^{(j)} < c^\star + \epsilon$ that

$$\begin{aligned}
&N\,\mathbb{P}(\cap_{j=1}^{k} \{Y_i^{(j)} \ge u^{(j)} a_N^{(j)}\}) \\
&\quad = \exp\left(\log N + \sum_{j=1}^{k} \log(\mathbb{P}(Y_i^{(j)} \ge u^{(j)} a_N^{(j)})) \right) \\
&\quad = \exp\left(\log N \left(1 - \sum_{j=1}^{k} h^{(j)}(u^{(j)}) \right) (1 + o(1)) \right) \\
&\quad \overset{N \to \infty}{\longrightarrow} 0.
\end{aligned}$$

Thus, we know that there are no limiting points in the positive quadrants with starting points $(u^{(1)}, \ldots, u^{(k)})$ with $c^\star < \sum_{j=1}^k u^{(j)} < c^\star + \epsilon$. The union of a finite number of these quadrants covers the set $\{(u^{(j)}, j \leq k) : \sum_{j=1}^k u^{(j)} \geq c^\star + \epsilon, u^{(j)} \leq 1 \; \forall \; j\}$, as the set $\{(u^{(j)}, j \leq k) : \sum_{j=1}^k u^{(j)} \geq c^\star + \epsilon, u^{(j)} \leq 1 \; \forall \; j\}$ is compact. For example, in the case where $k = 2$,

$$\{(u,v) : u + v \geq c^\star + \epsilon, u \in [0,1], v \in [0,1]\}$$
$$\subset \cup_{m=1}^{\lceil 2/\epsilon - 1/2 \rceil + 1} \left\{(u,v) : u \geq c^\star - 1 + \frac{m\epsilon}{2}, v \geq 1 + \frac{\epsilon}{4} - \frac{m\epsilon}{2}\right\}. \quad (2.6.2)$$

For $k > 2$, we can show in an inductive manner that this property also holds. For instance, for $k = 3$, we have that

$$\{(u,v,w) : u + v + w \geq c^\star + \epsilon, u \in [0,1], v \in [0,1], w \in [0,1]\}$$
$$\subset \cup_{m=1}^{\lceil 2/\epsilon - 1/2 \rceil + 1} \left\{(u,v,w) : u + v \geq c^\star - 1 + \frac{m\epsilon}{2}, w \geq 1 + \frac{\epsilon}{4} - \frac{m\epsilon}{2}, u \in [0,1], v \in [0,1]\right\}. \quad (2.6.3)$$

Now, we can repeat the procedure in (2.6.2) for the union in (2.6.3), and we get that the union of a finite number of three-dimensional quadrants with starting points $c^* < u+v+w < c^*+\epsilon$ covers the set $\{(u,v,w) : u + v + w \geq c^\star + \epsilon, u \in [0,1], v \in [0,1], w \in [0,1]\}$. Hence, the limit in (2.6.1) and the lemma follows. □

Remark 2.1. *We believe that this result can be extended to extremes of sums of dependent random variables, when the joint tail probability is given by a copula function, and satisfies some conditions. First, we consider sequences of continuous random variables* $(Y_i^{(1)}, i \geq 1)$, $(Y_i^{(2)}, i \geq 1)$, …, $(Y_i^{(k)}, i \geq 1)$, *where all random variables in the sequence* $(Y_i^{(j)}, i \geq 1)$ *are i.i.d. and have infinite right endpoints, and we have a function* $C : \mathbb{R}^k \to \mathbb{R}$ *such that for all* $(u^{(1)}, \ldots, u^{(k)})$,

$$\mathbb{P}\left(\cap_{j=1}^k \left\{Y_i^{(j)} \geq u^{(j)}\right\}\right) = C(\mathbb{P}(Y_i^{(1)} \geq u^{(1)}), \ldots, \mathbb{P}(Y_i^{(k)} \geq u^{(k)})).$$

Second, we have a function $H : \mathbb{R}^k \to \mathbb{R}$ *such that for* $(u^{(1)}, \ldots, u^{(k)}) \in [0,1]^k$,

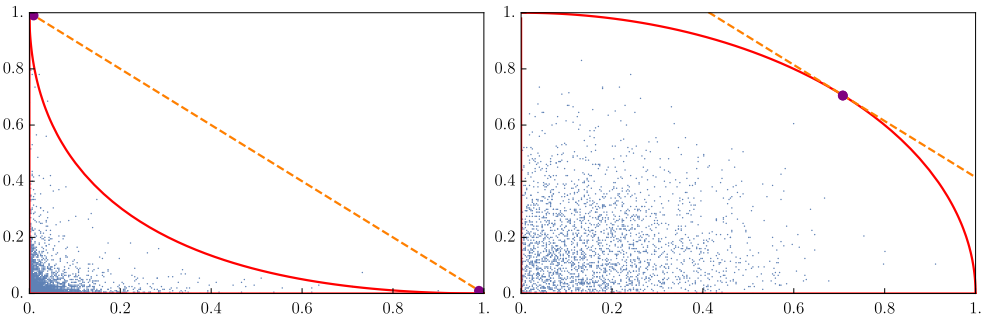
$$\lim_{t \to \infty} \frac{-\log\left(C(\mathbb{P}(Y_i^{(1)} \geq u^{(1)}t), \ldots, \mathbb{P}(Y_i^{(k)} \geq u^{(k)}t))\right)}{-\log\left(C(\mathbb{P}(Y_i^{(1)} \geq t), 1, \ldots, 1)\right)} = H(u^{(1)}, \ldots, u^{(k)}).$$

Finally, we have a sequence $(a_N, N \geq 1)$ *such that for all* j, $\mathbb{P}(Y_i^{(j)} \geq a_N) \sim 1/N$ *as* $N \to \infty$, *with* $Y_i^{(j)}$ *relatively stable. Here, the function* $H(u^{(1)}, \ldots, u^{(k)})$ *plays the same role as* $\sum_{j=1}^k h^{(j)}(u^{(j)})$ *in Lemma 2.13. In the proof of Lemma 2.13, we exploit the property of mutual independence and we distinguish between different cases of the function* $h^{(j)}(u^{(j)})$. *Thus, this result does not follow trivially and requires another proof.*

In Figure 2.5, we plot the part of the realization of the set

$$\left\{\left(X_1^{(1)}/a_N, X_1^{(2)}/a_N\right), \ldots, \left(X_N^{(1)}/a_N, X_N^{(2)}/a_N\right)\right\},$$

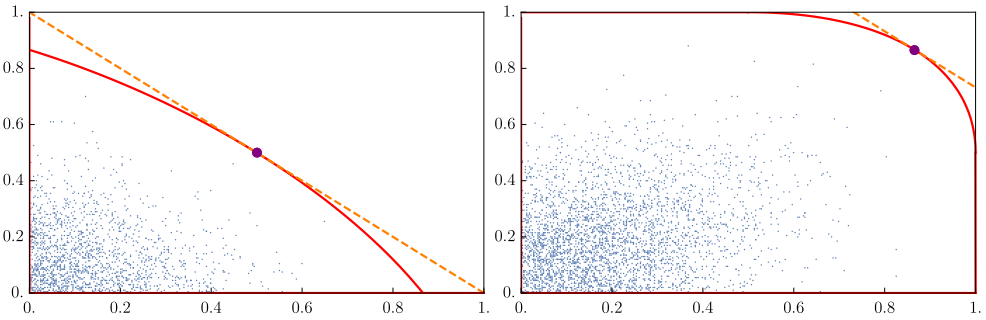
for which both coordinates have a positive value, together with the boundaries of the limiting sets $\{x^{1/2} + y^{1/2} \leq 1, 0 \leq x \leq 1, 0 \leq y \leq 1\}$ and $\{x^2 + y^2 \leq 1, 0 \leq x \leq 1, 0 \leq y \leq 1\}$, respectively. The dashed lines have slope -1, and the points where these lines touch the boundaries indicate the supremum of $x + y$ over the limiting sets. As we see in Figure 2.5a, the dashed line touches the curve in $(0, 1)$ and in $(1, 0)$. From this, it follows that when $\mathbb{P}(X^{(1,2)} \geq x) = \exp(-\sqrt{x})$ for all $x \geq 0$, thus $X^{(1,2)} \sim \text{Weibull}(1/2)$, we have that $\max_{i \leq N}(X_1^{(1)}/a_N + X_1^{(2)}/a_N) \xrightarrow{\mathbb{P}} 1$, as $N \to \infty$. In the second case, the dashed line touches the limiting curve in the point $(\sqrt{2}/2, \sqrt{2}/2)$, thus, when $X^{(1,2)} \sim \mathcal{N}(0, 1)$, we have that $\max_{i \leq N}(X_1^{(1)}/a_N + X_1^{(2)}/a_N) \xrightarrow{\mathbb{P}} \sqrt{2}/2 + \sqrt{2}/2 = \sqrt{2}$, as $N \to \infty$. In general, we have that when the tail of $X^{(1,2)}$ is exponential or heavier, but still relatively stable, $\max_{i \leq N}(X_1^{(1)}/a_N + X_1^{(2)}/a_N) \xrightarrow{\mathbb{P}} 1$, as $N \to \infty$. When the tail of $X^{(1,2)}$ is lighter than exponential, a non-trivial limit emerges.



(a) $a_N = (\log N)^2$, $X^{(1,2)} \sim \text{Weibull}(1/2)$ (b) $a_N = \sqrt{2 \log N}$, $X^{(1,2)} \sim \mathcal{N}(0, 1)$

Figure 2.5 Two-dimensional sample extremes of i.i.d. random variables in first quadrant with extreme of sum, $N = 10^4$. The blue dots indicate the two-dimensional samples, the curve indicates the border of the set $\{h^{(1)}(u^{(1)}) + h^{(2)}(u^{(2)}) \leq 1, u^{(j)} \leq 1\}$. The supremum of the set $\{\sum_{j=1}^{2} u^{(j)} : h^{(1)}(u^{(1)}) + h^{(2)}(u^{(2)}) \leq 1, u^{(j)} \leq 1\}$ is achieved at the points where the dashed line touches the red curve.

In Figure 2.6, we show the realization of the two-dimensional samples of binormally distributed random variables, with $\rho \neq 0$. When $\rho = -1/2$, the limiting extreme-value result of the sum equals 1, while for $\rho = 1/2$, the limit equals $\sqrt{3}$.



(a) $a_N = \sqrt{2\log N}, X^{(1,2)} \sim \mathcal{N}(0,1), \rho = -1/2$ (b) $a_N = \sqrt{2\log N}, X^{(1,2)} \sim \mathcal{N}(0,1), \rho = 1/2$

Figure 2.6 Two-dimensional sample extremes of binormally distributed random variables in first quadrant with extreme of sum, $N = 10^4$. The blue dots indicate the two-dimensional samples, the curve indicates the border of the set $\{H(u^{(1)}, u^{(2)}) \leq 1, u^{(j)} \leq 1\}$. The supremum of the set $\{\sum_{j=1}^{2} u^{(j)} : H(u^{(1)}, u^{(2)}) \leq 1, u^{(j)} \leq 1\}$ is achieved at the points where the dashed line touches the red curve.

2.6.1 How restrictive is Assumption 2.5?

In order to be able to prove the relative stability of sums of random variables, we impose Assumption 2.5 on each random variable: for $v \in [0,1]$,

$$\lim_{t\to\infty} \frac{-\log \mathbb{P}(U_i > vt)}{-\log \mathbb{P}(U_i > t)} = h(v).$$

We will now investigate whether this is the case for all random variables U_i in the domain of attraction of the Gumbel random variable.

Because U_i in the domain of attraction of the Gumbel random variable, we know that for t large enough

$$\mathbb{P}(U_i > t) = c(t)\exp\left(- \int_{t_0}^{t} \frac{1}{f_U(s)} ds \right),$$

with c, f_U positive, c converging to a positive constant, and $f_U'(t)$ converging to 0 as $t \to \infty$; see [67, Thm. 1.2.6, p. 22]. Thus, for t large enough,

$$\frac{-\log \mathbb{P}(U_i > vt)}{-\log \mathbb{P}(U_i > t)} = \frac{-\log c(vt) + \int_{t_0}^{vt} 1/f_U(s) ds}{-\log c(t) + \int_{t_0}^{t} 1/f_U(s) ds}.$$

Because $c(t)$ converges to a constant as $t \to \infty$, we can look at

$$\frac{\int_{t_0}^{vt} 1/f_U(s) ds}{\int_{t_0}^{t} 1/f_U(s) ds}.$$

By l'Hôspital we get that

$$\lim_{t\to\infty} \frac{\int_{t_0}^{vt} 1/f_U(s)ds}{\int_{t_0}^{t} 1/f_U(s)ds} = \lim_{t\to\infty} v\frac{f_U(t)}{f_U(vt)}.$$

Thus the somewhat restrictive assumption on f_U, will be that for all $v \in [0,1]$,

$$\lim_{t\to\infty} \frac{f_U(t)}{f_U(vt)}$$

exists. A choice of f_U where this is violated, is

$$f_U(t) = \frac{1}{(\log t)(2+\sin(\log t))}.$$

We take U_i a random variable with support on $[2,\infty)$, then for $t > 2$,

$$\mathbb{P}(U_i > t) = \exp\left(-\int_2^t (\log s)(2+\sin(\log s))ds\right).$$

We have

$$\begin{aligned}
&\int_2^t (\log s)(2+\sin(\log s))ds \\
&\quad = \frac{1}{2}(-4t + t\log(t)(\sin(\log(t)) + 4) \\
&\qquad + t(\log(t)-1)(-\cos(\log(t))) + 8 - \log(4)(4+\sin(\log(2))) + (\log(4)-2)\cos(\log(2))).
\end{aligned}$$

Then, for $v \in (0,1)$,

$$\frac{-\log \mathbb{P}(U_i > vt)}{-\log \mathbb{P}(U_i > t)}$$

does not converge as $t \to \infty$.

2.7. Proof of Proposition 2.1

The proof of Proposition 2.1 follows the same lines as the proof of Theorem 2.1: we need to prove pointwise convergence, the convergence of the finite-dimensional distributions, and tightness of the maximum queue length under the temporal scaling of N^3 and the spatial scaling of $N\sqrt{\log N}$. The convergence of the finite-dimensional distributions follows directly from Lemma 2.3. In Lemmas 2.14 and 2.15, we prove the pointwise convergence and tightness of the maximum queue length, respectively.

Lemma 2.14 (Pointwise convergence)**.** *Given Assumption 2.1 and* $Q^{(N)}_{(\alpha,\beta)}(0) = 0$*, then we*

have for all $T > 0$, that

$$\mathbb{P}\left(\left|\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)}{N\sqrt{\log N}} - \sqrt{2\alpha t}\right| > \epsilon\right) \overset{N\to\infty}{\longrightarrow} 0 \ \forall\, \epsilon > 0.$$

Proof. First, we observe that by using the union bound, we get that

$$\begin{aligned}\mathbb{P}&\left(\left|\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)}{N\sqrt{\log N}} - \sqrt{2\alpha t}\right| > \epsilon\right)\\ &\leq \mathbb{P}\left(\left|\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)}{N\sqrt{\log N}} - \left(\sqrt{2\alpha t} - \frac{\beta t}{\sqrt{\log N}}\right)\right| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(\frac{\beta t}{\sqrt{\log N}} > \frac{\epsilon}{2}\right).\end{aligned}$$

The second term on the right-hand side equals 0 when N is large enough. Thus, we only need to focus on the first term. Now, we need to prove that

$$\limsup_{N\to\infty} \mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)}{N\sqrt{\log N}} > \sqrt{2\alpha t} - \frac{\beta t}{\sqrt{\log N}} + \frac{\epsilon}{2}\right) = 0, \tag{2.7.1}$$

and

$$\liminf_{N\to\infty} \mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)}{N\sqrt{\log N}} > \sqrt{2\alpha t} - \frac{\beta t}{\sqrt{\log N}} - \frac{\epsilon}{2}\right) = 1. \tag{2.7.2}$$

In order to prove the first limit, we follow the proof of Lemma 2.6, where the same result is proven for the maximum queue length under the temporal scaling of $N^3 \log N$ and the spatial scaling of $N \log N$. Following the proof of Lemma 2.6, we see that

$$\begin{aligned}&\mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)}{N\sqrt{\log N}} > \sqrt{2\alpha t} - \frac{\beta t}{\sqrt{\log N}} + \frac{\epsilon}{2}\right)\\ &\quad\leq \mathbb{P}\left(\sup_{s\in[0,t]}\left|\frac{\tilde{\mathbf{N}}^{(N)}_A\left(sN^3\right)}{\sqrt{\log N}} + \beta\frac{s}{\sqrt{\log N}}\right| > \frac{\epsilon}{4}\right)\\ &\qquad+ \mathbb{P}\left(\sup_{s\in[0,t]}\left|\frac{\max_{i\leq N}\tilde{\mathbf{N}}^{(N)}_{S,i}\left(sN^3\right)}{\sqrt{\log N}} - \sqrt{2\alpha s}\right| > \frac{\epsilon}{4}\right).\end{aligned} \tag{2.7.3}$$

Then, for the first term on the right-hand side of (2.7.3), we use a similar bound as in (2.3.16) and get

$$\mathbb{P}\left(\sup_{s\in[0,t]}\left|\frac{\tilde{\mathbf{N}}_A^{(N)}\left(sN^3\right)}{\sqrt{\log N}}+\beta\frac{s}{\sqrt{\log N}}\right|>\frac{\epsilon}{4}\right)$$
$$\leq\frac{64}{\epsilon^2}\left(1-\frac{\alpha}{N}-\frac{\beta}{N^2}\right)\left(\frac{\alpha}{N}+\frac{\beta}{N^2}\right)\frac{\lfloor tN^3\rfloor}{N^2(\log N)}+o(1)\overset{N\to\infty}{\longrightarrow}0. \quad (2.7.4)$$

For the second term on the right-hand side of (2.7.3), we prove, along the same lines as in the proof of Lemma 2.6, that $\left(\max_{i\leq N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\right)/\sqrt{\log N},s\in[0,t]\right)$ converges to $(\sqrt{2\alpha s},s\in[0,t])$ u.o.c. Similarly to the limit given in (2.3.20), we use (2.3.10) from Lemma 2.2 to conclude that

$$\begin{aligned}&\mathbb{P}\left(\frac{\max_{i\leq N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\right)}{\sqrt{\log N}}>\sqrt{2\alpha s}+\frac{\epsilon}{4}\right)\\&\leq 1-\Phi\left(\frac{\sqrt{2\alpha s}+\epsilon/4}{\sqrt{\alpha s(1-\alpha/N)}}\sqrt{\log N}\frac{\sqrt{sN^3}}{\sqrt{\lfloor sN^3\rfloor}}\right)^N\\&\quad+\left(1+\frac{c_s}{N}\frac{1}{1+(\sqrt{2+\epsilon/(4\sqrt{\alpha s})}\sqrt{\log N}(1+o(1)))^3}\right)^N-1\\&\overset{N\to\infty}{\longrightarrow}0.\end{aligned} \quad (2.7.5)$$

The proof that

$$\mathbb{P}\left(\frac{\max_{i\leq N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\right)}{\sqrt{\log N}}>\sqrt{2\alpha s}-\frac{\epsilon}{4}\right)\overset{N\to\infty}{\longrightarrow}1$$

is analogous. As in the proof of Lemma 2.6, we use Doob's maximal submartingale inequality together with the upper bound in (2.3.12) in Lemma 2.3 to prove that for all $\eta>0$ and $\epsilon>0$,

$$\begin{aligned}&\limsup_{N\to\infty}\frac{1}{\eta}\mathbb{P}\left(\sup_{s\in[0,\eta]}\max_{i\leq N}\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\right)}{\sqrt{\log N}}>\epsilon\right)+\limsup_{N\to\infty}\frac{1}{\eta}\mathbb{P}\left(\sup_{s\in[0,\eta]}\max_{i\leq N}\frac{-\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\right)}{\sqrt{\log N}}>\epsilon\right)\\&\leq\limsup_{N\to\infty}\frac{1}{\eta\epsilon^{5/2}}\mathbb{E}\left[\max\left(\max_{i\leq N}\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(\eta N^3\right)}{\sqrt{\log N}},0\right)^{5/2}\right]\\&\quad+\limsup_{N\to\infty}\frac{1}{\eta\epsilon^{5/2}}\mathbb{E}\left[\max\left(\max_{i\leq N}\frac{-\tilde{\mathbf{N}}_{S,i}^{(N)}\left(\eta N^3\right)}{\sqrt{\log N}},0\right)^{5/2}\right]\\&\leq\frac{2\eta^{1/4}(2\alpha)^{5/4}}{\epsilon^{5/2}}.\end{aligned}$$

From this, it follows by the same arguments that $\left(\max_{i\leq N}\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\right)/\sqrt{\log N},s\in[0,t]\right)$ converges to $(\sqrt{2\alpha s},s\in[0,t])$ u.o.c. Thus, we can conclude that the limsup given in (2.7.1)

holds.

Now, in order to prove that the liminf in (2.7.2) also holds, we follow the proof of Lemma 2.8. We have the lower bound

$$\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)}{N\sqrt{\log N}} \geq_{st.} \max_{i\leq N} \frac{\mathbf{N}^{(N)}_A\left(tN^3\right) - \mathbf{N}^{(N)}_{S,i}\left(tN^3\right)}{N\sqrt{\log N}}.$$

From the limit in (2.7.4) and the fact that $\left(\max_{i\leq N} \tilde{\mathbf{N}}^{(N)}_{S,i}\left(sN^3\right)/\sqrt{\log N}, s\in[0,t]\right)$ converges to $(\sqrt{2\alpha s}, s\in[0,t])$ u.o.c., we can conclude that

$$\max_{i\leq N} \frac{\mathbf{N}^{(N)}_A\left(tN^3\right) - \mathbf{N}^{(N)}_{S,i}\left(tN^3\right)}{N\sqrt{\log N}} \xrightarrow{\mathbb{P}} \sqrt{2\alpha t},$$

as $N\to\infty$. The result in (2.7.2) follows.

Since we have a sharp upper and lower bound in (2.7.1) and (2.7.2), we have pointwise convergence of the rescaled maximum queue length $Q^{(N)}_{(\alpha,\beta)}(tN^3)/\left(N\sqrt{\log N}\right)$ to the limiting function $\sqrt{2\alpha t}$. □

Lemma 2.15 (Tightness)**.** *For $Q^{(N)}_{(\alpha,\beta)}(0)=0$, $\epsilon>0$, $\eta>0$, $T>0$, $\exists\ 0<\delta<1$ and an integer N_0 such that for all $t\in[0,T]$*

$$\frac{1}{\delta}\mathbb{P}\left(\sup_{s\in[t,t+\delta]}\left|\frac{Q^{(N)}_{(\alpha,\beta)}\left(sN^3\right)}{N\sqrt{\log N}} - \frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)}{N\sqrt{\log N}}\right|\geq\epsilon\right)\leq\eta,\ \ N\geq N_0. \tag{2.7.6}$$

Proof. We follow the proof of Lemma 2.11, which states the tightness of the process $\left(Q^{(N)}_{(\alpha,\beta)}(tN^3\log N)/\left(N\log N\right), t\in[0,T]\right)$. As in the proof of Lemma 2.11, we have for N large enough that

$$\begin{aligned}
&\frac{1}{\delta}\mathbb{P}\left(\sup_{s\in[t,t+\delta]}\left|\frac{Q^{(N)}_{(\alpha,\beta)}\left(sN^3\right)}{N\sqrt{\log N}} - \frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)}{N\sqrt{\log N}}\right|\geq\epsilon\right)\\
&\leq \frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\max_{i\leq N}\tilde{R}^{(N)}_i(sN^3)\sqrt{\log N}\geq\frac{\epsilon}{2}\right)\\
&\quad+\frac{1}{\delta}\mathbb{P}\left(2\sup_{0\leq s\leq\delta}\max_{i\leq N}-\tilde{R}^{(N)}_i(sN^3)\sqrt{\log N}\geq\frac{\epsilon}{2}\right)+o(1).
\end{aligned}$$

Now we focus on the first term on the right-hand side. The analysis of the second term on the right-hand side goes analogously:

$$\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\max_{i\leq N}\tilde{R}^{(N)}_i(sN^3)\sqrt{\log N}\geq\frac{\epsilon}{2}\right)$$

$$= \frac{1}{\delta}\,\mathbb{P}\left(\sup_{0\le s\le\delta}\max_{i\le N}\frac{\tilde{\mathbf{N}}_A^{(N)}\left(sN^3\right)+\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\right)}{\sqrt{\log N}}\ge\frac{\epsilon}{2}\right)$$
$$\le \frac{1}{\delta}\,\mathbb{P}\left(\sup_{0\le s\le\delta}\frac{\tilde{\mathbf{N}}_A^{(N)}\left(sN^3\right)}{\sqrt{\log N}}\ge\frac{\epsilon}{4}\right)+\frac{1}{\delta}\,\mathbb{P}\left(\sup_{0\le s\le\delta}\max_{i\le N}\frac{\tilde{\mathbf{N}}_{S,i}^{(N)}\left(sN^3\right)}{\sqrt{\log N}}\ge\frac{\epsilon}{4}\right).$$

In the proof of Lemma 2.14, we already showed that the second term in the last display is small. With a similar proof, one can also prove that the first term is small. Concluding, the process $\left(Q_{(\alpha,\beta)}^{(N)}(tN^3)/\left(N\sqrt{\log N}\right), t\in[0,T]\right)$ is tight. □

Proposition 2.1 follows from Lemmas 2.14 and 2.15.

2.8. Other model parameters

In this section, we again look at the fork-join queueing system with nearly deterministic arrivals and services, but we look at the setting in which the arrival and service probabilities satisfy $N(p^{(N)}-q^{(N)}) \overset{N\to\infty}{\longrightarrow} c<0$. The emerging convergence result is fundamentally different from what we saw before.

Proposition 2.3. *Let $p^{(N)}=1-a/N$ and $q^{(N)}=1-b/N$ with $0<b<a$. Then we get that the maximum queue length $\max_{i\le N}\sup_{0\le k\le n}[(\mathbf{N}_A^{(N)}(n)-\mathbf{N}_A^{(N)}(k))-(\mathbf{N}_{S,i}^{(N)}(n)-\mathbf{N}_{S,i}^{(N)}(k))]$ satisfies for all $t>0$,*

$$\frac{\log\log N}{\log N}\max_{i\le N}\sup_{s\in[0,t]}\left[\left(\mathbf{N}_A^{(N)}(tN)-\mathbf{N}_A^{(N)}(sN)\right)-\left(\mathbf{N}_{S,i}^{(N)}(tN)-\mathbf{N}_{S,i}^{(N)}(sN)\right)\right]\xrightarrow{\mathbb{P}}1,$$

as $N\to\infty$.

Proof. First, we observe that

$$\max_{i\le N}\sup_{s\in[0,t]}\left[\left(\mathbf{N}_A^{(N)}(tN)-\mathbf{N}_A^{(N)}(sN)\right)-\left(\mathbf{N}_{S,i}^{(N)}(tN)-\mathbf{N}_{S,i}^{(N)}(sN)\right)\right]$$
$$\overset{d}{=}\max_{i\le N}\sup_{s\in[0,t]}\left(\mathbf{N}_A^{(N)}(sN)-\mathbf{N}_{S,i}^{(N)}(sN)\right).$$

Then, we have by using subadditivity that

$$\max_{i\le N}\sup_{0\le s\le t}\left(\mathbf{N}_A^{(N)}(sN)-\mathbf{N}_{S,i}^{(N)}(sN)\right)\le\sup_{0\le s\le t}\left(\mathbf{N}_A^{(N)}(sN)-\mathbb{E}\left[\mathbf{N}_A^{(N)}(sN)\right]\right)$$
$$+\max_{i\le N}\sup_{0\le s\le t}\left(\mathbb{E}\left[\mathbf{N}_A^{(N)}(sN)\right]-\mathbf{N}_{S,i}^{(N)}(sN)\right).$$

Now, the process $(\mathbf{N}_A^{(N)}(j)-\mathbb{E}[\mathbf{N}_A^{(N)}(j)], j\ge1)$ is a martingale with mean 0. Furthermore, $\mathrm{Var}(\mathbf{N}_A^{(N)}(tN))=\lfloor tN\rfloor a/N\left(1-a/N\right)\overset{N\to\infty}{\longrightarrow}at$. Then, by Doob's maximal submartingale inequality, we get that for all $x>0$,

$$\mathbb{P}\left(\sup_{0\le s\le t}\left(\mathbf{N}_A^{(N)}(sN)-\mathbb{E}\left[\mathbf{N}_A^{(N)}(sN)\right]\right)>x\frac{\log N}{\log\log N}\right)$$
$$\le\frac{\sqrt{\mathrm{Var}(\mathbf{N}_A^{(N)}(tN))}\log\log N}{x\log N}\overset{N\to\infty}{\longrightarrow}0.$$

Thus,

$$\frac{\log\log N}{\log N}\sup_{0\le s\le t}\left(\mathbf{N}_A^{(N)}(sN)-\mathbb{E}\left[\mathbf{N}_A^{(N)}(sN)\right]\right)\overset{\mathbb{P}}{\longrightarrow}0,$$

as $N\to\infty$. With an analogous technique we can treat a lower bound.

Now, we are left with proving the convergence of $\max_{i\le N}\sup_{0\le s\le t}(\mathbb{E}[\mathbf{N}_A^{(N)}(sN)]-\mathbf{N}_{S,i}^{(N)}(sN))$. We have

$$\mathbb{E}\left[\mathbf{N}_A^{(N)}(sN)\right]-\mathbf{N}_{S,i}^{(N)}(sN)=\left(1-\frac{a}{N}\right)\lfloor sN\rfloor-\mathbf{N}_{S,i}^{(N)}(sN).$$

The random variable $\mathbf{N}_{S,i}^{(N)}(sN)$ is Bin $\left(\lfloor sN\rfloor,1-b/N\right)$ distributed. Therefore, we can write

$$\left(\mathbb{E}[\mathbf{N}_A^{(N)}(sN)]-\mathbf{N}_{S,i}^{(N)}(sN),s\in[0,t]\right)=\left(-\frac{a}{N}\lfloor sN\rfloor+B_i^{(N)}(sN),s\in[0,t]\right).$$

with $(B_i^{(N)}(j),j\ge1)$ a random walk with $B_i^{(N)}(j)\sim\mathrm{Bin}\left(j,b/N\right)$, $B_i^{(N)}(j)$ and $B_k^{(N)}(j)$ independent, and we write $B_i^{(N)}(j)=B_i^{(N)}(\lfloor j\rfloor)$. Then,

$$-\frac{a}{N}\lfloor tN\rfloor+B_i^{(N)}(tN)\le\sup_{0\le s\le t}\left(-\frac{a}{N}\lfloor sN\rfloor+B_i^{(N)}(sN)\right)\le B_i^{(N)}(tN).$$

This means that $\max_{i\le N}\sup_{0\le s\le t}(-a/N\lfloor sN\rfloor+B_i^{(N)}(sN))$ scales in the same way as $\max_{i\le N}B_i^{(N)}(tN)$. In order to analyze this latter random variable, we first note that it is a standard result that

$$B_i^{(N)}(tN)\overset{d}{\longrightarrow}P_i,$$

as $N\to\infty$, with $P_i\sim\mathrm{Poi}\,(bt)$; see [141, Eq. (1)]. In [8, 9], it is shown that

$$\mathbb{P}\left(\max_{i\le N}P_i\in\{I_N,I_N+1\}\right)\overset{N\to\infty}{\longrightarrow}1,$$

for a sequence of integers $(I_N,N\ge1)$ satisfying $I_N\sim\log N/\log\log N$ as $N\to\infty$; see [81]. Hence

$$\frac{\log\log N}{\log N}\max_{i\le N}P_i\overset{\mathbb{P}}{\longrightarrow}1,$$

as $N \to \infty$. Le Cam's theorem [93] states that

$$\limsup_{N\to\infty} N \sum_{x=0}^{\infty} \left| \mathbb{P}(P_i = x) - \mathbb{P}(B_i^{(N)}(tN) = x) \right| \leq 2b^2 t.$$

Now, for all sequences $(y_N, N \geq 1)$ with $y_N < I_N$,

$$\mathbb{P}\left(\max_{i\leq N} P_i \leq y_N \right) \overset{N\to\infty}{\longrightarrow} 0.$$

Hence for N large enough,

$$\mathbb{P}\left(\max_{i\leq N} B_i^{(N)}(tN) \leq y_N \right) = \mathbb{P}\left(B_i^{(N)}(tN) \leq y_N \right)^N \leq \left(\mathbb{P}(P_i \leq y_N) + \frac{2b^2 t}{N} \right)^N \overset{N\to\infty}{\longrightarrow} 0.$$

Similarly, for all sequences $(z_N, N \geq 1)$ with $z_N > I_N + 1$,

$$\mathbb{P}\left(\max_{i\leq N} P_i \leq z_N \right) \overset{N\to\infty}{\longrightarrow} 1.$$

In [11, Cor. 2.1], it is stated that $\mathbb{P}(P_i \leq z_N) \leq \mathbb{P}(B_i^{(N)}(tN) \leq z_N)$. Therefore,

$$\mathbb{P}\left(\max_{i\leq N} B_i^{(N)}(tN) \leq z_N \right) \overset{N\to\infty}{\longrightarrow} 1.$$

In conclusion,

$$\frac{\log\log N}{\log N} \max_{i\leq N} B_i^{(N)}(tN) \overset{\mathbb{P}}{\longrightarrow} 1,$$

as $N \to \infty$. Our result follows.

□

A remarkable fact to notice is that $\log\log N \max_{i\leq N} P_i / \log N \overset{\mathbb{P}}{\longrightarrow} 1$, as $N \to \infty$, with $P_i \sim \text{Poi}(\lambda)$, for all $\lambda > 0$. Thus the parameter λ does not appear in the limit. A well-known result is that the sum of independent Poisson-distributed random variables is again Poisson distributed. Thus, this means that the behavior of N extremes of n independent sums of Poisson-distributed random variables is the same as the behavior of the sample extremes of N Poisson-distributed random variables. Hence this result gives us information on the tail of the Poisson distribution. So, on the one hand, following Lemma 2.13, we know that the limit of $-\log(\mathbb{P}(P_i > ut))/-\log(\mathbb{P}(P_i > t))$ should be such that this property is satisfied. On the other hand, we know that the moment-generating function exists everywhere on the real line. Thus the tail distribution is not heavier than the exponential distribution. Combining these two facts, we get that

$$\lim_{t\to\infty} \frac{-\log(\mathbb{P}(P_i > ut))}{-\log(\mathbb{P}(P_i > t))} = u.$$

Chapter 3

Limiting behavior of the invariant distribution

3.1. Introduction

In Chapter 2, we derived a first-order convergence result for the maximum queue length process. This convergence result provides a prediction of the typical delay. The obtained limit is deterministic and can be viewed as a law of large numbers. In this chapter, we aim to derive a second-order convergence result, which resembles a central limit theorem. Furthermore, we look at a different setting. In particular, we investigate the longest steady-state waiting time among the N servers with a common arrival process; i.e., $\max_{i\leq N} W_i(\infty) \stackrel{d}{=} \max_{i\leq N} \sup_{k\geq 0} \sum_{j=1}^{k}(S_i(j) - A(j))$. This expression follows from Lindley's recursion. Furthermore, we have that both $(S_i(j), i \geq 1, j \geq 1)$ and $(A(j), j \geq 1)$ are i.i.d. and the service times and interarrival times are mutually independent. Thus, $S_i(j)$ indicates the service time of the j-th task in queue i, and $A(j)$ indicates the interarrival time between the $(j-1)$-st and the j-th task. We see that the longest steady-state waiting time is a maximum of N dependent random variables due to the common arrival process $(A(j), j \geq 1)$.

Our main finding of this chapter is that the longest steady-state waiting time in this queueing system scales around $\frac{1}{\gamma}\log N$, where γ is determined by the cumulant-generating function Λ of the service distribution and solves the Cramér-Lundberg equation with stochastic service times and deterministic interarrival times. We exploit the properties of the all-time suprema of random walks in order to derive this result. We show that, as N becomes large,

$$\max_{i\leq N} \sup_{k\geq 0} \sum_{j=1}^{k}(S_i(j) - A(j)) \approx \max_{i\leq N} \sum_{j=1}^{\tau^{(N)}}(S_i(j) - A(j)),$$

This chapter is based on [107] and [135].

with $\tau^{(N)} = \frac{1}{\Lambda'(\gamma)\gamma} \log N$. Then we can further rewrite

$$\max_{i \leq N} \sum_{j=1}^{\tau^{(N)}} (S_i(j) - A(j)) = \max_{i \leq N} \sum_{j=1}^{\tau^{(N)}} \left(S_i(j) - \frac{1}{\lambda} \right) + \sum_{j=1}^{\tau^{(N)}} \left(\frac{1}{\lambda} - A(j) \right).$$

The first term on the right-hand side reaches the value $\frac{1}{\gamma} \log N$, and the second term satisfies the central limit theorem, with standard deviation $\frac{\sigma_A}{\sqrt{\Lambda'(\gamma)\gamma}}$. By using distributional Little's law [68], we can prove the second-order convergence for the maximum queue length as well.

In order to prove the convergence of the longest steady-state waiting time, we use the results given in [13, Ch. XIII, Par. 5] on Cramér-Lundberg theory, as discussed in Sections 1.4.1 and 1.4.4. Further results in this area are given in [59, 113, 114]. In these studies, bounds on the tail probability of sums of independent random variables are given. Another important result in this area is the Bahadur-Rao theorem [19], which provides exact asymptotics for the large deviations of the sum of independent random variables. An overview of results on large deviations in queueing theory can be found in [45, 60]. We add to the existing literature by proving second-order convergence of the extremes of dependent random variables, using the aforementioned large deviations results.

This chapter is organized as follows. In Section 3.2, we present our main results; in Theorem 3.1 we state that the longest steady-state waiting time satisfies a central limit result; in Theorem 3.2 we show that a similar result holds for the maximum queue length, and in Corollary 3.1 we present a similar result when the service distributions can differ among the different queues. Finally, we extend the convergence result of the longest waiting time to the maximum of N all-time suprema of dependent Brownian motions in Corollary 3.2 and we prove L_1-convergence for the maximum of N all-time suprema of dependent Brownian motions in Lemma 3.3. In Section 3.3, we give an intuition of why the results hold and how we prove these. Section 3.4 is devoted to proofs.

3.2. Model

We investigate a fork-join queue with N servers. Each of the N servers has the same arrival stream of jobs and works independently from all other servers but with the same service distribution. In this section, we state the main result for the longest steady-state waiting time in Theorem 3.1. We also show that a similar result holds for the maximum queue length in Lemma 3.2 and Theorem 3.2. Furthermore, we extend the result in Theorem 3.2 to a heterogeneous model in Corollary 3.1 and consider Brownian motions in Corollary 3.2. Finally, we prove L_1-convergence of the maximum of N suprema of dependent Brownian motions in Lemma 3.3.

We now specify some properties of the service times and interarrival times in this fork-join queueing system. First, the sequence of non-negative random variables $(S_i(j), i \geq 1, j \geq 1)$ are i.i.d. with $S_i(j) \sim S$, and $S_i(j)$ indicating the service time of the j-th subtask in queue i. Furthermore, the sequence of non-negative random variables $(A(j), j \geq 1)$ are i.i.d. with $A(j) \sim A$, $\mathbb{E}[A(j)] = 1/\lambda$, $\mathrm{Var}(A(j)) = \sigma_A^2$, and $A(j)$ indicating the interarrival time between

the $(j-1)$-st and the j-th task. Finally, we have that $\mathbb{E}[S_i(j) - A(j)] = -\mu$, with $\mu > 0$, and $(A(j), j \geq 1)$ and $(S_i(j), i \geq 1, j \geq 1)$ are mutually independent.

We can now write the cumulative distribution function of the longest steady-state waiting time as the cumulative distribution function of the maximum of N all-time suprema of random walks involving the interarrival and service times.

Lemma 3.1. *For the model given in Section 3.2 with $W_i(1) = 0$ for all $i \leq N$, we have that the longest waiting time in steady state satisfies*

$$\max_{i \leq N} W_i(\infty) \stackrel{d}{=} \max_{i \leq N} \sup_{k \geq 0} \sum_{j=1}^{k} (S_i(j) - A(j)). \tag{3.2.1}$$

Proof. Using Lindley's recursion [96], we get a similar expression for the longest waiting time as we got for the maximum queue length in Equation (2.2.1):

$$\max_{i \leq N} W_i(n) = \max_{i \leq N} \sup_{0 \leq k \leq n} \sum_{j=k+1}^{n} (S_i(j) - A(j)).$$

We have that

$$\mathbb{P}(\max_{i \leq N} W_i(\infty) \geq x) = \lim_{n \to \infty} \mathbb{P}(\max_{i \leq N} W_i(n) \geq x).$$

Because, as in (2.2.2),

$$\max_{i \leq N} W_i(n) \stackrel{d}{=} \max_{i \leq N} \sup_{0 \leq k \leq n} \sum_{j=1}^{k} (S_i(j) - A(j)), \tag{3.2.2}$$

we obtain the lemma by using the monotone convergence theorem.

□

In order to be able to prove convergence of the longest steady-state waiting time, we need some additional structure for the service-time distribution. We define

$$\Lambda(\theta) := \log(\mathbb{E}[\exp(\theta(S - 1/\lambda)]). \tag{3.2.3}$$

Moreover, we write $\mathcal{D}(\Lambda) := \{\theta : \Lambda(\theta) < \infty\}$ and $\mathcal{D}^\circ(\Lambda)$ as the interior of $\mathcal{D}(\Lambda)$.

Assumption 3.1. *We assume there exists a $\gamma > 0$ such that*

1. $\Lambda(\gamma) = 0$,
2. $\gamma \in \mathcal{D}^\circ(\Lambda)$.

The first assumption indicates that the random variable $S - 1/\lambda$ has a tail that is bounded by an exponential. The second assumption is needed for our proofs. In [45, Ex. 2.2.24], it is namely stated that when $\gamma \in \mathcal{D}^\circ(\Lambda)$, Λ is infinitely differentiable at the point γ. For example, when $S - 1/\lambda$ has density function $f_{S-1/\lambda}(x) = c_1 \exp(-x)/(1+x^2)$ for $x > 0$, where c_1, λ are chosen such that $\mathbb{P}(S - 1/\lambda < x)$ is a cumulative distribution function and $\gamma = 1$, then the

first assumption is satisfied but the second is not, since $\Lambda(\theta)$ is not differentiable at $\theta = \gamma$. Our main result is given in Theorem 3.1.

Theorem 3.1. *For the model in Section 3.2 where the sequence of service times $(S_i(j), i \geq 1, j \geq 1)$ satisfies Assumption 3.1, we have that*

$$\frac{\max_{i\leq N} W_i(\infty) - \frac{1}{\gamma}\log N}{\sqrt{\log N}} \xrightarrow{d} \frac{\sigma_A}{\sqrt{\Lambda'(\gamma)\gamma}} X, \tag{3.2.4}$$

with $X \sim \mathcal{N}(0,1)$, as $N \to \infty$.

Lemma 3.2 (Distributional Little's Law)**.** *Let for $t \geq 0$, $\mathbf{N}_A(t)$ indicate the number of arrivals up to time t, where the interarrival times are i.i.d. with $A(j) \sim A$. Then*

$$\max_{i\leq N} Q_i(\infty) \stackrel{d}{=} \mathbf{N}_A\left(\max_{i\leq N} W_i(\infty)\right). \tag{3.2.5}$$

Proof. In [68], a short proof is given that for the $GI/GI/1$ queue under the FCFS policy, $Q \stackrel{d}{=} \mathbf{N}_A(W)$. We follow the same steps to prove that $\max_{i\leq N} Q_i(\infty) \stackrel{d}{=} \mathbf{N}_A\left(\max_{i\leq N} W_i(\infty)\right)$.

First, let $t > 0$ be given such that the system is in steady state. Furthermore, let $\tilde{W}_i(j)$ be the waiting time of the i-th subtask of the j-th task numbered backward in time, beginning at time t. Thus, $\tilde{W}_i(1)$ is the waiting time of the i-th subtask of the last task arriving before time t. Now, let the random variable $T(j)$ be such that $t - T(j)$ is the arrival time of the j-th task numbered backward in time. Then, observe that the event $\{\max_{i\leq N} Q_i(t) \geq j\}$ is equivalent to the event that at least one subtask of the j-th task numbered backward in time is still in the queue at time t. Thus,

$$\left\{\max_{i\leq N} Q_i(t) \geq j\right\} = \left\{\max_{i\leq N} \tilde{W}_i(j) \geq T(j)\right\},$$

for $j \geq 1$. The event $\{T(j) \leq x\}$ is equivalent to the event that the number of arrivals during the period $[t-x, t)$ is larger than or equal to j. The arrival process is a stationary process, thus the event $\{T(j) \leq x\}$ is equivalent to the event $\{\mathbf{N}_A(x) \geq j\}$. Additionally, the random variables $\max_{i\leq N} \tilde{W}_i(j)$ and $T(j)$ are independent. Therefore,

$$\left\{\max_{i\leq N} Q_i(t) \geq j\right\} = \left\{\mathbf{N}_A\big(\max_{i\leq N} \tilde{W}_i(j)\big) \geq j\right\}.$$

As the system is in steady state, we get that

$$\left\{\max_{i\leq N} Q_i(\infty) \geq j\right\} = \left\{\mathbf{N}_A\big(\max_{i\leq N} \tilde{W}_i(\infty)\big) \geq j\right\}.$$

□

Now, combining the result in Lemma 3.2 with the main result in Theorem 3.1, we can find a similar convergence result for the maximum queue length in steady state.

Theorem 3.2. *For the model in Section 3.2 where the sequence of service times* $(S_i(j), i \geq 1, j \geq 1)$ *satisfies Assumption 3.1, we have that*

$$\frac{\max_{i\leq N} Q_i(\infty) - \frac{\lambda}{\gamma}\log N}{\sqrt{\log N}} \xrightarrow{d} \sqrt{\frac{\lambda^2\sigma_A^2}{\Lambda'(\gamma)\gamma} + \frac{\lambda^3\sigma_A^2}{\gamma}}X, \tag{3.2.6}$$

with $X \sim \mathcal{N}(0,1)$*, as* $N \to \infty$.

Proof. Let $\hat{A}(j) \sim A$, let $(\hat{A}(j), j \geq 1)$ be mutually independent, and $\hat{A}(j)$ and $\max_{i\leq N} W_i(\infty)$ be mutually independent for all $j \geq 1$. Then, using Lemma 3.2 and Theorem 3.1, we get that

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\leq N} Q_i(\infty) \leq \frac{\lambda}{\gamma}\log N + x\sqrt{\log N}\right)\\
&= \mathbb{P}\left(\mathbf{N}_A\left(\max_{i\leq N} W_i(\infty)\right) \leq \left\lfloor \frac{\lambda}{\gamma}\log N + x\sqrt{\log N}\right\rfloor\right)\\
&= \mathbb{P}\left(\max_{i\leq N} W_i(\infty) \leq \sum_{j=1}^{\left\lfloor \frac{\lambda}{\gamma}\log N + x\sqrt{\log N}\right\rfloor} \hat{A}(j)\right)\\
&= \mathbb{P}\left(\frac{\max_{i\leq N} W_i(\infty) - \frac{1}{\gamma}\log N}{\sqrt{\log N}} \leq \frac{\sum_{j=1}^{\left\lfloor \frac{\lambda}{\gamma}\log N + x\sqrt{\log N}\right\rfloor} \hat{A}(j) - \frac{1}{\gamma}\log N}{\sqrt{\log N}}\right)\\
&\xrightarrow{N\to\infty} \mathbb{P}\left(\frac{\sigma_A}{\sqrt{\Lambda'(\gamma)\gamma}}X_1 \leq \frac{\sigma_A\sqrt{\lambda}}{\sqrt{\gamma}}X_2 + \frac{x}{\lambda}\right),
\end{aligned}$$

with X_1, X_2 independent and standard normally distributed, this convergence holds, as $(\hat{A}(j), j \geq 1)$ and $\max_{i\leq N} W_i(\infty)$ are independent. Thus, the theorem follows. □

Until now, we considered the fork-join queueing system where each server has the same service distribution. In Corollary 3.1, we show that we can extend the convergence of the longest steady-state waiting time to a more heterogeneous setting. We examine a fork-join queueing system with N servers, where each of these N servers belongs to one of K classes. Additionally, we assume that the size of class k with $k \in \{1, \ldots, K\}$ grows as $\alpha_k N$, as N becomes large, with $0 < \alpha_k < 1$.

Corollary 3.1. *Let* $K \in \mathbb{N}$*, let* $k = 1, \ldots, K$*, furthermore, take an increasing sequence of integers given by* $M_0^{(N)}, M_1^{(N)}, M_2^{(N)}, \ldots, M_K^{(N)} > 0$ *with* $M_0^{(N)} = 1$*,* $M_K^{(N)} = N$*, and* $M_k^{(N)} - M_{k-1}^{(N)} \in \mathbb{N}$*. Moreover,* $(M_k^{(N)} - M_{k-1}^{(N)})/N \xrightarrow{N\to\infty} \alpha_k \in (0,1]$ *with* $\sum_{k=1}^K \alpha_k = 1$*. Let* $(S_i(j), j \geq 1, M_{k-1}^{(N)} < i \leq M_k^{(N)})$ *be i.i.d. with* $S_i(j) \sim S_k$*,* $(A(j), j \geq 1)$ *be i.i.d. with* $A(j) \sim A$*,* $\mathbb{E}[A(j)] = 1/\lambda$*, Var*$(A(j)) = \sigma_A^2$*,* $\mathbb{E}[S_i(j) - A(j)] = -\mu_k$ *with* $\mu_k > 0$*,* $\Lambda_k(\theta) = \log(\mathbb{E}[\exp(\theta(S_k - 1/\lambda)])$*,* Λ_k *satisfies Assumption 3.1. Furthermore,* $S_{i_1}(j_1)$ *and*

$S_{i_2}(j_2)$ are mutually independent for all i_1, i_2, j_1, j_2. Let $K^* = \arg\min\{\gamma_k, k = 1, \ldots, K\}$. We assume that $|K^*| = 1$ and $k^* \in K^*$. Then,

$$\frac{\max_{i\leq N} W_i(\infty) - \frac{1}{\gamma_{k^*}} \log N}{\sqrt{\log N}} \xrightarrow{d} \frac{\sigma_A}{\sqrt{\Lambda'_{k^*}(\gamma_{k^*})\gamma_{k^*}}} X, \tag{3.2.7}$$

with $X \sim \mathcal{N}(0,1)$, as $N \to \infty$.

Proof. We prove this corollary by giving an asymptotically sharp lower and upper bound. First, observe that

$$\max_{i\leq N} W_i(\infty) \geq_{st.} \max_{M^{(N)}_{k^*-1} < i \leq M^{(N)}_{k^*}} \sup_{k\geq 0} \sum_{j=1}^{k} (S_i(j) - A(j)),$$

with $X \geq_{st.} Y$ meaning that $\mathbb{P}(X \geq x) \geq \mathbb{P}(Y \geq x)$ for all x. Applying the result from Theorem 3.1 on the lower bound results in (3.2.7). By using the union bound we get the following upper bound:

$$\begin{aligned}\mathbb{P}\left(\max_{i\leq N} W_i(\infty) \geq \frac{1}{\gamma_{k^*}} \log N + x\sqrt{\log N}\right) \\ = \sum_{l=1}^{K} \mathbb{P}\left(\max_{M^{(N)}_{l-1} < i \leq M^{(N)}_{l}} \sup_{k\geq 0} \sum_{j=1}^{k} (S_i(j) - A(j)) \geq \frac{1}{\gamma_{k^*}} \log N + x\sqrt{\log N}\right).\end{aligned}$$

When $l \neq k^*$, we get after applying the results from Theorem 3.1 that

$$\begin{aligned}\mathbb{P}\left(\max_{M^{(N)}_{l-1} < i \leq M^{(N)}_{l}} \sup_{k\geq 0} \sum_{j=1}^{k} (S_i(j) - A(j)) \geq \frac{1}{\gamma_{l}} \log N + x\sqrt{\log N}\right) \\ \xrightarrow{N\to\infty} 1 - \Phi\left(\frac{\sqrt{\Lambda'_l(\gamma_l)\gamma_l}}{\sigma_A} x\right),\end{aligned}$$

with Φ the cumulative distribution function of a standard normal random variable. Because $\gamma_{k^*} < \gamma_l$ we get that

$$\mathbb{P}\left(\max_{M^{(N)}_{l-1} < i \leq M^{(N)}_{l}} \sup_{k\geq 0} \sum_{j=1}^{k} (S_i(j) - A(j)) \geq \frac{1}{\gamma_{K^*}} \log N + x\sqrt{\log N}\right) \xrightarrow{N\to\infty} 0.$$

The corollary follows. □

Remark 3.1. *In Corollary 3.1 we assume that $|K^*| = 1$. The situation that $|K^*| > 1$ follows analogously. Assume for instance that $|K^*| = 2$, then we can introduce a new random variable $\tilde{S}$ such that $\tilde{S}_i(j) \sim S_1$ with probability α and $\tilde{S}_i(j) \sim S_2$ with probability $1-\alpha$, such*

that $\gamma_1 = \gamma_2 = \gamma_{K^*}$. As N is large enough this fork-join queueing system behaves analogous to the original fork-join queue, and for this system $|K^*| = 1$.

We can extend the results from Theorems 3.1 and 3.2 to the extremes of the all-time suprema of N dependent Brownian motions. We will use this convergence result in Chapter 6.

Corollary 3.2. *Let $(B_i(t), t \geq 0)$ and $(B_A(t), t \geq 0)$ be Brownian motions with mean 0 and standard deviations σ and σ_A, respectively, and $(B_i(t), t \geq 0)$ and $(B_j(t), t \geq 0)$ independent for $i \neq j$, then we get that*

$$\frac{\max_{i \leq N} \sup_{s>0} \left(B_i(s) + B_A(s) - \beta s\right) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \xrightarrow{d} \frac{\sigma\sigma_A}{\sqrt{2}\beta} X, \tag{3.2.8}$$

with $X \sim \mathcal{N}(0,1)$, as $N \to \infty$.

Remark 3.2. *In the proof of Corollary 3.2, we use the same ideas as in the proof of Theorem 3.1. However, we also exploit specific properties of Brownian motions, as Corollary 3.2 does not trivially follow from Theorem 3.1. The extension of this result to general Lévy processes requires another proof. A possible direction to a proof is by giving lower and upper bounds of the Lévy processes in terms of random walks; see for instance [48].*

We prove an even stronger statement, which we will also use in Chapter 6.

Lemma 3.3. *Let $(B_i(t), t \geq 0)$ and $(B_A(t), t \geq 0)$ be Brownian motions with mean 0 and standard deviations σ and σ_A, respectively, and $(B_i(t), t \geq 0)$ and $(B_j(t), t \geq 0)$ independent for $i \neq j$. We define $X_N := \frac{\sqrt{2}\beta}{\sigma\sigma_A} \frac{B_A\left(\frac{\sigma^2}{2\beta^2} \log N\right)}{\sqrt{\log N}}$. Then,*

$$\mathbb{E}\left[\left|\frac{\max_{i \leq N} \sup_{s>0} \left(B_i(s) + B_A(s) - \beta s\right) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} - \frac{\sigma\sigma_A}{\sqrt{2}\beta} X_N\right|\right] \xrightarrow{N \to \infty} 0.$$

We give the proofs of the convergence of the longest steady-state waiting time in Section 3.4. In this section, we also prove Corollary 3.2 and Lemma 3.3. First, we give a heuristic explanation of why the convergence result in Theorem 3.1 is true, and we illustrate the structure of the proof.

3.3. Heuristic analysis

To prove Theorem 3.1, we analyze lower and upper bounds of the tail probability of the longest steady-state waiting time among the N servers $\mathbb{P}(\max_{i \leq N} W_i(\infty) > \frac{1}{\gamma} \log N + x\sqrt{\log N})$ and we show that these lower and upper bounds converge to the same limit as $N \to \infty$. The longest steady-state waiting time has the form $\max_{i \leq N} W_i(\infty) \stackrel{d}{=} \sup_{k \geq 0} \max_{i \leq N} \sum_{j=1}^{k} (S_i(j) - A(j))$. Thus the longest steady-state waiting time is the all-time supremum of the maximum of N random walks. For all processes $(X(t), t \geq 0)$, we

have for all $t > 0$

$$\mathbb{P}\left(\sup_{s>0} X(s) > x\right) \geq \mathbb{P}\big(X(t) > x\big). \tag{3.3.1}$$

Furthermore, due to the union bound, we have for all $0 < t_1 < t_2$ that

$$\begin{aligned}&\mathbb{P}\left(\sup_{s>0} X(s) > x\right)\\ &\qquad \leq \mathbb{P}\left(\sup_{0<s<t_1} X(s) > x\right) + \mathbb{P}\left(\sup_{t_1\leq s<t_2} X(s) > x\right) + \mathbb{P}\left(\sup_{s\geq t_2} X(s) > x\right).\end{aligned} \tag{3.3.2}$$

We use these types of lower and upper bounds to prove Theorem 3.1. Obviously, not all choices of t, t_1, and t_2 give sharp bounds. We can however make an educated guess about which choices will give the sharpest bounds. Let us first replace the sequence of random variables $(A(j), j \geq 1)$ with their expectation $1/\lambda$. Thus, we look at a simplified fork-join queue with deterministic arrivals. Because the arrivals are deterministic, the waiting times are mutually independent, and we are able to use standard extreme-value theory. We know from the Cramér-Lundberg approximation [13, Ch. XIII, Thm. 5.2] that $\mathbb{P}(\sup_{k\geq 0}\sum_{j=1}^{k}(S_i(j) - 1/\lambda) > x) \sim C\exp(-\gamma x)$, as $x \to \infty$, with $0 < C < 1$. Thus, $\mathbb{P}(\sup_{k\geq 0}\sum_{j=1}^{k}(S_i(j) - 1/\lambda) > \frac{1}{\gamma}\log N) \sim C/N$, as $N \to \infty$. Now we can conclude by using basic extreme-value results; see [67, Thm. 5.4.1, p. 188], that

$$\frac{\max_{i\leq N}\sup_{k\geq 0}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right)}{\log N} \xrightarrow{\mathbb{P}} \frac{1}{\gamma},$$

as $N \to \infty$. The analysis is similar to the nearly deterministic case in Chapter 2. Thus, we know that $\max_{i\leq N}\sup_{k\geq 0}\sum_{j=1}^{k}(S_i(j) - 1/\lambda)$ centers around $\frac{1}{\gamma}\log N$. In order to find suitable lower and upper bounds of the form as given in (3.3.1) and (3.3.2), we need to estimate the hitting time

$$\tau^{(N)} := \inf\left\{k \geq 0 : \max_{i\leq N}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \geq \frac{1}{\gamma}\log N\right\}.$$

As mentioned before, we have that $\mathbb{P}(\sup_{k\geq 0}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) > \frac{1}{\gamma}\log N) \sim C/N$ as $N \to \infty$. Thus, a good estimate $\hat{\tau}^{(N)}$ for $\tau^{(N)}$ should also satisfy the property that

$$\liminf_{N\to\infty} N\,\mathbb{P}\left(\sum_{j=1}^{\hat{\tau}^{(N)}}\left(S_i(j) - \frac{1}{\lambda}\right) > \frac{1}{\gamma}\log N\right) > 0 \tag{3.3.3}$$

and

$$\limsup_{N\to\infty} N\,\mathbb{P}\left(\sum_{j=1}^{\hat{\tau}^{(N)}}\left(S_i(j)-\frac{1}{\lambda}\right) > \frac{1}{\gamma}\log N\right) < \infty. \tag{3.3.4}$$

Now, by using Cramér's theorem and by using the fact that Λ is at least twice differentiable at γ, we know that

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\mathbb{P}\left(\sum_{j=1}^{n}\left(S_i(j)-\frac{1}{\lambda}\right)\geq nx\right)\right) = -\Lambda^*(x), \tag{3.3.5}$$

for all $x > \mathbb{E}[S_i(j) - 1/\lambda]$ with $\Lambda^*(x) = \sup_{t\in\mathbb{R}}(tx - \Lambda(t))$; see [13, Ch. XIII, Thm. 2.1 (2.3)]. We write $\hat{\tau}^{(N)} = \hat{c}\log N$. Then we can conclude from Equation (3.3.5) that

$$\lim_{N\to\infty}\frac{1}{\log N}\log\left(\mathbb{P}\left(\sum_{j=1}^{\lfloor \hat{c}\log N\rfloor}\left(S_i(j)-\frac{1}{\lambda}\right)\geq x\hat{c}\log N\right)\right) = -\Lambda^*(x)\hat{c}. \tag{3.3.6}$$

Thus, in order to find a good estimate $\hat{\tau}^{(N)}$ for the hitting time $\tau^{(N)}$ we need to solve two equations. First, $x\hat{c} = 1/\gamma$, because we know that the longest steady-state waiting time under deterministic arrivals is approximately equal to $\frac{1}{\gamma}\log N$. Therefore the expression $x\hat{c}\log N$ in (3.3.6) should be the same as $\frac{1}{\gamma}\log N$. Second, $-\Lambda^*(x)\hat{c} = -1$, because we know from (3.3.3), (3.3.4), and (3.3.6) that for large N

$$\mathbb{P}\left(\sum_{j=1}^{\lfloor \hat{c}\log N\rfloor}\left(S_i(j)-\frac{1}{\lambda}\right)\geq x\hat{c}\log N\right) \approx \frac{1}{N} = \exp(-\Lambda^*(x)\hat{c}\log N).$$

Combining these two equations gives $\hat{c} = \frac{1}{\Lambda'(\gamma)\gamma}$ and $x = \Lambda'(\gamma)$. Clearly, $x\hat{c} = 1/\gamma$, and

$$\Lambda^*(x)\hat{c} = \frac{\Lambda^*(\Lambda'(\gamma))}{\gamma\Lambda'(\gamma)}.$$

From [45, Lem. 2.2.5(c)], we know that $\Lambda^*(\Lambda'(\gamma)) = \gamma\Lambda'(\gamma)$, thus indeed, $\Lambda^*(x)\hat{c} = 1$. Finally, we can conclude that $\hat{\tau}^{(N)} = \hat{c}\log N = \frac{1}{\gamma\Lambda'(\gamma)}\log N$. Obviously, in order to be a good estimation for a hitting time we need to have that $\Lambda'(\gamma) > 0$. This is the case because $\Lambda(\theta)$ is convex; see [13, Ch. XIII, Thm. 5.1].

Until this point, we know the first-order scaling of the largest of N steady-state waiting times with deterministic arrivals, and we can give an estimation of the hitting time of this value. Now, we can use these results to obtain a second-order convergence result for the longest steady-state waiting time with stochastic arrivals. Following the analysis above together with the lower bound in (3.3.1), we see that

$$\mathbb{P}\left(\frac{\max_{i\leq N} W_i(\infty) - \frac{1}{\gamma}\log N}{\sqrt{\log N}} \geq x\right)$$
$$\geq \mathbb{P}\left(\frac{\max_{i\leq N} \sup_{\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\log N < k < \frac{1}{\Lambda'(\gamma)\gamma}\log N} \sum_{j=1}^{k}(S_i(j) - A(j)) - \frac{1}{\gamma}\log N}{\sqrt{\log N}} \geq x\right), \tag{3.3.7}$$

with $\epsilon > 0$ and small. In Lemma 3.4, we prove that the right-hand side in (3.3.7) converges to a function that is close to the tail probability of a normally distributed random variable. Furthermore, we show in Lemmas 3.5, 3.6, and 3.7, that this lower bound is sharp. To achieve this, we first divide the supremum over all positive numbers in the random variable $\max_{i\leq N} W_i(\infty)$ in three parts. After that, we take the supremum over the intervals $\left[0, \left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right)\log N\right]$, $\left(\left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right)\log N, \left(\frac{1}{\Lambda'(\gamma)\gamma} + \epsilon\right)\log N\right]$, and $\left(\left(\frac{1}{\Lambda'(\gamma)\gamma} + \epsilon\right)\log N, \infty\right)$, with $\epsilon > 0$ and small. Consequently, we show that the tail probabilities of the first and third suprema of the maximum of N random walks asymptotically vanish, while

$$\mathbb{P}\left(\max_{i\leq N} \sup_{\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\log N < k < \left(\frac{1}{\Lambda'(\gamma)\gamma}+\epsilon\right)\log N} \sum_{j=1}^{k}(S_i(j) - A(j)) > \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)$$

converges to a limit close to the lower bound as $N \to \infty$.

Remark 3.3. *The lower bound presented in Equation* (3.3.1) *gives us information about the convergence rate of the result in Theorem 3.1. From the Berry-Esséen theorem [109], we know that when* $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \xrightarrow{d} X \sim \mathcal{N}(0,1)$*, the convergence rate is of order* $1/\sqrt{n}$*. Thus, the lower bound in* (3.3.1) *shows that the convergence rate is of order* $1/\sqrt{\log N}$*.*

3.4. Proofs

Lemma 3.4. *Given the model in Section 3.2 where the sequence of service times* $(S_i(j), i \geq 1, j \geq 1)$ *satisfies Assumption 3.1,* $0 < \epsilon < \frac{1}{\Lambda'(\gamma)\gamma}$*,* $t_1^{(N)} = \left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right)\log N$*, and* $t_2^{(N)} = \frac{1}{\Lambda'(\gamma)\gamma}\log N$*, then for all* $x \in \mathbb{R}$*, we have that*

$$\liminf_{N\to\infty} \mathbb{P}\left(\max_{i\leq N} \sup_{t_1^{(N)} < k < t_2^{(N)}} \sum_{j=1}^{k}(S_i(j) - A(j)) > \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)$$
$$\geq \mathbb{P}\left(\sigma_A\sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon}X_1 - \sigma_A\sqrt{\epsilon}\,|X_2| > x\right), \tag{3.4.1}$$

with $X_1, X_2 \sim \mathcal{N}(0,1)$ and independent.

Proof. In order to prove this convergence result, we first bound

$$\max_{i\leq N} \sup_{t_1^{(N)}<k<t_2^{(N)}} \sum_{j=1}^{k} (S_i(j) - A(j)) \\ \geq \max_{i\leq N} \sup_{t_1^{(N)}<k<t_2^{(N)}} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda}\right) + \inf_{t_1^{(N)}<k<t_2^{(N)}} \sum_{j=1}^{k} \left(\frac{1}{\lambda} - A(j)\right).$$

We treat the terms on the right-hand side separately. We first prove that

$$\frac{\inf_{t_1^{(N)}<k<t_2^{(N)}} \sum_{j=1}^{k} \left(\frac{1}{\lambda} - A(j)\right)}{\sqrt{\log N}} \xrightarrow{d} \sigma_A \sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon} X_1 - \sigma_A \sqrt{\epsilon}\, |X_2|\,, \tag{3.4.2}$$

as $N \to \infty$. Afterwards, we prove that

$$\frac{\max_{i\leq N} \sup_{t_1^{(N)}<k<t_2^{(N)}} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda}\right) - \frac{1}{\gamma}\log N}{\sqrt{\log N}} \xrightarrow{\mathbb{P}} 0, \tag{3.4.3}$$

as $N \to \infty$.

The first convergence result follows from Donsker's theorem. The left-hand side in (3.4.2) is an infimum of a random walk with drift 0. Then for $(B(t), t \geq 0)$ a Brownian motion with drift 0 and standard deviation 1, by using Donsker's theorem [49] and the fact that the infimum is a continuous functional, we obtain that

$$\mathbb{P}\left(\frac{\inf_{t_1^{(N)}<k<t_2^{(N)}} \sum_{j=1}^{k} (\frac{1}{\lambda} - A(j))}{\sqrt{\log N}} > x\right) \\ \xrightarrow{N\to\infty} \mathbb{P}\left(\inf_{\left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right) < s < \frac{1}{\Lambda'(\gamma)\gamma}} \sigma_A B(s) > x\right).$$

Furthermore, we can rewrite

$$\inf_{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon < s < \frac{1}{\Lambda'(\gamma)\gamma}} \sigma_A B(s) \stackrel{d}{=} \sigma_A B\left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right) - \inf_{0<s<\epsilon} \sigma_A \tilde{B}(s),$$

where $\tilde{B}$ is an independent copy of B. Obviously, we have that $\sigma_A B\left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right) \stackrel{d}{=} \sigma_A \sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon} X_1$ with $X_1 \sim \mathcal{N}(0,1)$. Because $\inf_{0<s<\epsilon} \sigma_A \tilde{B}(s) \stackrel{d}{=} \sigma_A \sqrt{\epsilon} |X_2|$, with $X_2 \sim \mathcal{N}(0,1)$, we have that the limit in (3.4.2) follows.

In order to prove the second convergence result, we define for $A \in \mathcal{F}_k$, with $\{\mathcal{F}_k, k \geq 1\}$

the natural filtration, the probability measure

$$\mathbb{P}_i(A) := \mathbb{E}\left[\exp\left(\gamma\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\right)\mathbb{1}(A)\right];$$

see [13, Ch. XIII, Par. 3]. Now, we know that

$$\mathbb{E}_i\left[S_i(j)-\frac{1}{\lambda}\right] = \mathbb{E}\left[\left(S_i(j)-\frac{1}{\lambda}\right)\exp\left(\gamma\left(S_i(j)-\frac{1}{\lambda}\right)\right)\right] = \Lambda'(\gamma).$$

Thus, by checking the conditions in [13, Ch. XIII, Thm. 5.6], we see that

$$\begin{aligned}\mathbb{P}\left(\sup_{0\le k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\ge\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\\ = C\exp\left(-\gamma\left(\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\right)\Phi\left(-x\frac{\sqrt{\gamma\Lambda'(\gamma)}}{\sqrt{\Lambda''(\gamma)}}\right)(1+o(1)).\end{aligned}\tag{3.4.4}$$

With the same approach, we get from [13, Ch. XIII, Thm. 5.6] that

$$\begin{aligned}\mathbb{P}\left(\sup_{0\le k<t_1^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\ge\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\\ = o\left(C\exp\left(-\gamma\left(\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\right)\right),\end{aligned}\tag{3.4.5}$$

as $N\to\infty$, for all $x\in\mathbb{R}$. By applying the union bound, we get that

$$\begin{aligned}&\mathbb{P}\left(\sup_{0\le k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\ge\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\\ &\le\mathbb{P}\left(\sup_{0\le k<t_1^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\ge\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\\ &\qquad+\mathbb{P}\left(\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\ge\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\\ &\le\mathbb{P}\left(\sup_{0\le k<t_1^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\ge\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\end{aligned}$$

$$+\mathbb{P}\left(\sup_{0\le k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\ge\frac{1}{\gamma}\log N+x\sqrt{\log N}\right).$$

We can conclude from these bounds, together with (3.4.4) and (3.4.5) that

$$\mathbb{P}\left(\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\ge\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)$$
$$=C\exp\left(-\gamma\left(\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\right)\Phi\left(-x\frac{\sqrt{\gamma\Lambda'(\gamma)}}{\sqrt{\Lambda''(\gamma)}}\right)(1+o(1)). \quad (3.4.6)$$

By using this expression it is easy to derive that for $x>0$

$$\mathbb{P}\left(\max_{i\le N}\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\le\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)$$
$$=\mathbb{P}\left(\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\le\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)^N\overset{N\to\infty}{\longrightarrow}1.$$

Similarly, for $x<0$,

$$\mathbb{P}\left(\max_{i\le N}\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\le\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)$$
$$=\mathbb{P}\left(\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\le\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)^N\overset{N\to\infty}{\longrightarrow}0.$$

Combining these two results gives us the limit in (3.4.3). Finally, the convergence result in (3.4.1) follows from the two limits in (3.4.2) and (3.4.3). □

Lemma 3.5. *Given the model in Section 3.2 where the sequence of service times $(S_i(j), i\ge 1, j\ge 1)$ satisfies Assumption 3.1, $t_1^{(N)}=\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\log N$, $\delta=\frac{\delta_1}{\Lambda'(\gamma)\gamma}+\delta_2$ with $\delta_{1,2}>0$ and small, and $\epsilon=\delta^{1/4}$, then for all $x\in\mathbb{R}$, we have that*

$$\mathbb{P}\left(\max_{i\le N}\sup_{0\le k<t_1^{(N)}}\sum_{j=1}^{k}(S_i(j)-A(j))>\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\overset{N\to\infty}{\longrightarrow}0. \quad (3.4.7)$$

Proof. We derive upper bounds for the left-hand side of (3.4.7) that converge to 0 as $N\to\infty$.

We get by using the subadditivity property of the sup operator and the union bound

that

$$\mathbb{P}\left(\max_{i\le N}\sup_{0\le k<t_1^{(N)}}\sum_{j=1}^{k}(S_i(j)-A(j))>\frac{1}{\gamma}\log N\right) \tag{3.4.8}$$

$$\le \mathbb{P}\left(\max_{i\le N}\sup_{0\le k<t_1^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}+\delta_1\right)>\left(\frac{1}{\gamma}-\delta_2\right)\log N\right) \tag{3.4.9}$$

$$+\mathbb{P}\left(\sup_{k\ge 0}\sum_{j=1}^{k}\left(\frac{1}{\lambda}-\delta_1-A(j)\right)>\delta_2\log N+x\sqrt{\log N}\right). \tag{3.4.10}$$

First, because $\mathbb{E}[\frac{1}{\lambda}-\delta_1-A(j)]<0$, we get that

$$\mathbb{P}\left(\sup_{k\ge 0}\sum_{j=1}^{k}\left(\frac{1}{\lambda}-\delta_1-A(j)\right)>\delta_2\log N+x\sqrt{\log N}\right)\overset{N\to\infty}{\longrightarrow}0.$$

Second, we can bound the term in (3.4.9) as follows;

$$\mathbb{P}\left(\max_{i\le N}\sup_{0\le k<t_1^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}+\delta_1\right)>\left(\frac{1}{\gamma}-\delta_2\right)\log N\right)$$

$$\le \mathbb{P}\left(\max_{i\le N}\sup_{0\le k<t_1^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)>\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)\log N\right).$$

Now, we can bound this further to

$$\mathbb{P}\left(\max_{i\le N}\sup_{0\le k<t_1^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)>\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)\log N\right)$$

$$\le \sum_{k=0}^{\lfloor t_1^{(N)}\rfloor}N\,\mathbb{P}\left(\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)>\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)\log N\right).$$

By using Chernoff's bound we obtain that for $\Lambda(\theta)<\infty$

$$\sum_{k=0}^{\lfloor t_1^{(N)}\rfloor}N\,\mathbb{P}\left(\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)>\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)\log N\right) \tag{3.4.11}$$

$$\le N\sum_{k=0}^{\lfloor t_1^{(N)}\rfloor}\exp(k\Lambda(\theta))\exp\left(-\theta\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)\log N\right)$$

$$= N\frac{-1+\exp\left((\lfloor t_1^{(N)}\rfloor+1)\Lambda(\theta)\right)}{\exp(\Lambda(\theta))-1}\exp\left(-\theta\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)\log N\right). \tag{3.4.12}$$

Now,

$$\frac{\log\left(N\frac{-1+\exp\left((\lfloor t_1^{(N)}\rfloor+1)\Lambda(\theta)\right)}{\exp(\Lambda(\theta))-1}\exp\left(-\theta\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)\log N\right)\right)}{\log N}$$
$$\overset{N\to\infty}{\longrightarrow} 1-\left(\theta\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\Lambda(\theta)\right).$$

In order to make the bound in (3.4.12) as sharp as possible, we need to choose a convenient θ. The choice of θ that gives the sharpest bound maximizes the function $\theta\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\Lambda(\theta)$. We have that $\delta=\frac{\delta_1}{\Lambda'(\gamma)\gamma}+\delta_2$ and $\epsilon=\delta^{1/4}$. Furthermore, we choose $\theta=\gamma+\sqrt{\delta}$. This gives us a sharp enough bound in (3.4.12). We obviously have that

$$\sup_{\eta\in\mathbb{R}}\left(\eta\left(\frac{1}{\gamma}-\delta\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}-\delta^{1/4}\right)\Lambda(\eta)\right)$$
$$\geq\left((\gamma+\sqrt{\delta})\left(\frac{1}{\gamma}-\delta\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}-\delta^{1/4}\right)\Lambda(\gamma+\sqrt{\delta})\right).$$

The first order Taylor series of $\Lambda(\gamma+\sqrt{\delta})$ around γ gives

$$\Lambda(\gamma+\sqrt{\delta})=\Lambda(\gamma)+\sqrt{\delta}\Lambda'(\gamma)+O(\delta)=\sqrt{\delta}\Lambda'(\gamma)+O(\delta).$$

Thus,

$$\left((\gamma+\sqrt{\delta})\left(\frac{1}{\gamma}-\delta\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}-\delta^{1/4}\right)\Lambda(\gamma+\sqrt{\delta})\right)=1+\delta^{3/4}\Lambda'(\gamma)+O(\delta)>1,$$

for δ small enough. Thus the expression in (3.4.12) is upper bounded by the term $N^{-\delta^{3/4}\Lambda'(\gamma)-O(\delta)}\overset{N\to\infty}{\longrightarrow}0$. □

Lemma 3.6. *Given the model in Section 3.2 where the sequence of service times $(S_i(j), i\geq 1, j\geq 1)$ satisfies Assumption 3.1, $0<\epsilon<\frac{1}{\Lambda'(\gamma)\gamma}$, $t_1^{(N)}=\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\log N$, and $t_3^{(N)}=\left(\frac{1}{\Lambda'(\gamma)\gamma}+\epsilon\right)\log N$, then for all $x\in\mathbb{R}$, we have that*

$$\limsup_{N\to\infty}\mathbb{P}\left(\max_{i\leq N}\sup_{t_1^{(N)}\leq k<t_3^{(N)}}\sum_{j=1}^{k}(S_i(j)-A(j))>\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)$$

$$\leq \mathbb{P}\left(\sigma_A\sqrt{\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon}X_1+\sigma_A\sqrt{2\epsilon}\,|X_2|>x\right), \quad (3.4.13)$$

with $X_1, X_2 \sim \mathcal{N}(0,1)$ *and independent.*

Proof. In order to prove this lemma, we first rewrite

$$\begin{aligned}
&\frac{\max_{i\leq N}\sup_{t_1^{(N)}\leq k<t_3^{(N)}}\sum_{j=1}^{k}(S_i(j)-A(j))-\frac{1}{\gamma}\log N}{\sqrt{\log N}}\\
&\quad\leq \frac{\max_{i\leq N}\sup_{t_1^{(N)}\leq k<t_3^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)-\frac{1}{\gamma}\log N}{\sqrt{\log N}}\\
&\qquad\qquad + \frac{\sup_{t_1^{(N)}\leq k<t_3^{(N)}}\sum_{j=1}^{k}\left(\frac{1}{\lambda}-A(j)\right)}{\sqrt{\log N}}\\
&\quad\leq \frac{\max_{i\leq N}\sup_{k\geq 0}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)-\frac{1}{\gamma}\log N}{\sqrt{\log N}} + \frac{\sup_{t_1^{(N)}\leq k<t_3^{(N)}}\sum_{j=1}^{k}\left(\frac{1}{\lambda}-A(j)\right)}{\sqrt{\log N}}.
\end{aligned} \quad (3.4.14)$$

We first look at the first term in (3.4.14). This term gives the rescaled longest steady-state waiting time of N i.i.d. $D/G/1$ queues. We know that

$$\mathbb{P}\left(\sup_{k\geq 0}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)>x\right)\sim C\exp(-\gamma x),$$

as $x\to\infty$, with $0<C<1$; see [13, Ch. XIII, Thm. 5.2]. Thus for $x>0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\max_{i\leq N}\sup_{k\geq 0}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)-\frac{1}{\gamma}\log N}{\sqrt{\log N}}>x\right)\\
&\qquad\qquad\sim 1-\left(1-C\exp(-\gamma(1/\gamma\log N+x\sqrt{\log N}))\right)^N \overset{N\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Similarly, for $x<0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\max_{i\leq N}\sup_{k\geq 0}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)-\frac{1}{\gamma}\log N}{\sqrt{\log N}}>x\right)\\
&\qquad\qquad\sim 1-\left(1-C\exp(-\gamma(1/\gamma\log N+x\sqrt{\log N}))\right)^N \overset{N\to\infty}{\longrightarrow} 1.
\end{aligned}$$

Thus, the first term in (3.4.14) converges in probability to 0.

Now, we prove convergence of the tail probability of the second term in (3.4.14). This term is a supremum of a random walk with drift 0. Then for $(B(t), t\geq 0)$ a Brownian motion with drift 0 and standard deviation 1, by using Donsker's theorem [49] and the fact that the

supremum is a continuous functional, we obtain with a similar analysis as in Lemma 3.4, that

$$\mathbb{P}\left(\frac{\sup_{t_1^{(N)}\le k<t_3^{(N)}}\sum_{j=1}^{k}(\frac{1}{\lambda}-A(j))}{\sqrt{\log N}}>x\right) \overset{N\to\infty}{\longrightarrow} \mathbb{P}\left(\sigma_A\sqrt{\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon}X_1+\sigma_A\sqrt{2\epsilon}\,|X_2|>x\right).$$

□

Lemma 3.7. *Given the model in Section 3.2 where the sequence of service times $(S_i(j), i \ge 1, j \ge 1)$ satisfies Assumption 3.1, $\delta = \frac{\delta_1}{\Lambda'(\gamma)\gamma} + \delta_2$ with $\delta_{1,2} > 0$ and small, $\epsilon = \delta^{1/4}$, and $t_3^{(N)} = \left(\frac{1}{\Lambda'(\gamma)\gamma}+\epsilon\right)\log N$, then for all $x \in \mathbb{R}$, we have that*

$$\mathbb{P}\left(\max_{i\le N}\sup_{k\ge t_3^{(N)}}\sum_{j=1}^{k}(S_i(j)-A(j))>\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\overset{N\to\infty}{\longrightarrow}0. \tag{3.4.15}$$

Proof. As in the proof of Lemma 3.5, we derive upper bounds for

$$\mathbb{P}\left(\max_{i\le N}\sup_{k\ge t_3^{(N)}}\sum_{j=1}^{k}(S_i(j)-A(j))>\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)$$

that converge to 0 as $N\to\infty$.

First, we see that by using subadditivity and the union bound, we obtain

$$\begin{aligned}&\mathbb{P}\left(\max_{i\le N}\sup_{k\ge t_3^{(N)}}\sum_{j=1}^{k}(S_i(j)-A(j))>\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\\ &\le \mathbb{P}\left(\max_{i\le N}\sup_{k\ge t_3^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}+\delta_1\right)>\left(\frac{1}{\gamma}-\delta_2\right)\log N\right)\\ &\quad+\mathbb{P}\left(\sup_{k\ge 0}\sum_{j=1}^{k}\left(\frac{1}{\lambda}-\delta_1-A(j)\right)>\delta_2\log N+x\sqrt{\log N}\right).\end{aligned}$$

As in the proof of Lemma 3.5, we have that

$$\mathbb{P}\left(\sup_{k\ge 0}\sum_{j=1}^{k}\left(\frac{1}{\lambda}-\delta_1-A(j)\right)>\delta_2\log N+x\sqrt{\log N}\right)\overset{N\to\infty}{\longrightarrow}0.$$

Furthermore, observe that $\log\mathbb{E}[\exp(\theta(S_i(j)-1/\lambda+\delta_1))]=\Lambda(\theta)+\theta\delta_1$. Now, as in the proof

of Lemma 3.5, we can bound

$$\mathbb{P}\left(\max_{i\leq N}\sup_{k\geq t_3^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}+\delta_1\right)>\left(\frac{1}{\gamma}-\delta_2\right)\log N\right) \tag{3.4.16}$$

$$\leq N\sum_{k=\lfloor t_3^{(N)}\rfloor}^{\infty}\mathbb{P}\left(\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}+\delta_1\right)>\left(\frac{1}{\gamma}-\delta_2\right)\log N\right) \tag{3.4.17}$$

$$\leq N\sum_{k=\lfloor t_3^{(N)}\rfloor}^{\infty}\exp(k(\Lambda(\theta)+\theta\delta_1))\exp\left(-\theta\left(\frac{1}{\gamma}-\delta_2\right)\log N\right)$$

$$=N\frac{\exp\left(\lfloor t_3^{(N)}\rfloor(\Lambda(\theta)+\theta\delta_1)\right)}{\exp(\Lambda(\theta)+\theta\delta_1)-1}\exp\left(-\theta\left(\frac{1}{\gamma}-\delta_2\right)\log N\right), \tag{3.4.18}$$

when $\Lambda(\theta)+\theta\delta_1<0$. When $\Lambda(\theta)+\theta\delta_1\geq 0$ the sum in the upper bound diverges to ∞. Now, for the case $\Lambda(\theta)+\theta\delta_1<0$, we have that

$$\frac{\log\left(N\frac{\exp\left(\lfloor t_3^{(N)}\rfloor(\Lambda(\theta)+\theta\delta_1)\right)}{\exp(\Lambda(\theta)+\theta\delta_1)-1}\exp\left(-\theta\left(\frac{1}{\gamma}-\delta_2\right)\log N\right)\right)}{\log N}$$

$$\overset{N\to\infty}{\longrightarrow}1+\left(\frac{1}{\Lambda'(\gamma)\gamma}+\epsilon\right)(\Lambda(\theta)+\theta\delta_1)-\theta\left(\frac{1}{\gamma}-\delta_2\right).$$

As in the proof of Lemma 3.5, we have $\delta=\frac{\delta_1}{\Lambda'(\gamma)\gamma}+\delta_2$ and $\epsilon=\delta^{1/4}$. We now get after a similar derivation as in the proof of Lemma 3.5 that $\theta=\gamma-\sqrt{\delta}$ gives a sharp bound. First, observe that $\Lambda(\gamma-\sqrt{\delta})=-\sqrt{\delta}\Lambda'(\gamma)+O(\delta)$, thus $\Lambda(\theta)+\theta\delta_1=-\sqrt{\delta}\Lambda'(\gamma)+(\gamma-\sqrt{\delta})\delta_1+O(\delta)=-\sqrt{\delta}\Lambda'(\gamma)+O(\delta)<0$ for δ small enough, thus the upper bound in (3.4.18) holds. Second, we see that

$$\sup_{\eta\in\mathbb{R}}\left(\eta\left(\frac{1}{\gamma}-\delta_2\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}+\epsilon\right)(\Lambda(\eta)+\eta\delta_1)\right)$$

$$\geq(\gamma-\sqrt{\delta})\left(\frac{1}{\gamma}-\delta_2\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}+\epsilon\right)(\Lambda(\gamma-\sqrt{\delta})+(\gamma-\sqrt{\delta})\delta_1).$$

So, we can conclude that

$$(\gamma-\sqrt{\delta})\left(\frac{1}{\gamma}-\delta_2\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}+\delta^{1/4}\right)(\Lambda(\gamma-\sqrt{\delta})+(\gamma-\sqrt{\delta})\delta_1)=1+\delta^{3/4}\Lambda'(\gamma)+O(\delta)>1$$

for δ small enough, thus the expression in (3.4.18) converges to 0 as $N\to\infty$. □

Proof of Theorem 3.1. First, to prove a lower bound, we see that

$$\max_{i\leq N} W_i(\infty) \geq_{st.} \max_{i\leq N} \sum_{j=1}^{\left\lfloor \frac{1}{(\Lambda'(\gamma)\gamma)} \log N \right\rfloor} (S_i(j) - A(j)).$$

Thus, combining this inequality with the result from Lemma 3.4, we see that

$$\liminf_{N\to\infty} \mathbb{P}\left(\max_{i\leq N} W_i(\infty) > \frac{1}{\gamma}\log N + x\sqrt{\log N} \right) \geq \mathbb{P}\left(\sigma_A \sqrt{\frac{1}{\Lambda'(\gamma)\gamma}} X > x \right).$$

Second, by using the union bound of the types as given in (3.3.2) and explained in Section 3.3, we get from Lemmas 3.5, 3.6, and 3.7, with $t_1^{(N)} = \left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right)\log N$ and $t_3^{(N)} = \left(\frac{1}{\Lambda'(\gamma)\gamma} + \epsilon\right)\log N$, that

$$\begin{aligned}
&\limsup_{N\to\infty} \mathbb{P}\left(\max_{i\leq N} W_i(\infty) > \frac{1}{\gamma}\log N + x\sqrt{\log N} \right) \\
&\quad \leq \limsup_{N\to\infty} \mathbb{P}\left(\max_{i\leq N} \sup_{t_1^{(N)} \leq k < t_3^{(N)}} \sum_{j=1}^{k} (S_i(j) - A(j)) > \frac{1}{\gamma}\log N + x\sqrt{\log N} \right) \\
&\quad \leq \mathbb{P}\left(\sigma_A \sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon} X_1 + \sigma_A \sqrt{2\epsilon}\, |X_2| > x \right).
\end{aligned}$$

Finally, we have that

$$\mathbb{P}\left(\sigma_A \sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon} X_1 + \sigma_A \sqrt{2\epsilon}\, |X_2| > x \right) \xrightarrow{\epsilon\downarrow 0} \mathbb{P}\left(\sigma_A \sqrt{\frac{1}{\Lambda'(\gamma)\gamma}} X > x \right).$$

□

Proof of Corollary 3.2. Since Brownian motions are continuous-time processes, the extension from random walks is not entirely trivial. We first observe that due to the fact that Brownian motions are infinitely divisible, we can write $B_i(\lfloor t \rfloor)$ as $\sum_{j=1}^{\lfloor t \rfloor} X_i(j)$, with $(X_i(j), j \geq 1)$ i.i.d. normally distributed random variables with mean 0 and standard deviation σ. Similarly, we can write $B_A(\lfloor t \rfloor)$ as $\sum_{j=1}^{\lfloor t \rfloor} X_A(j)$, with $(X_A(j), j \geq 1)$ i.i.d. normally distributed random variables with mean 0 and standard deviation σ_A. Thus, it follows that

$$\max_{i\leq N} \sup_{s>0} (B_i(s) + B_A(s) - \beta s) \geq \max_{i\leq N} \sup_{k\geq 0} \sum_{j=1}^{k} (X_i(j) + X_A(j) - \beta j). \tag{3.4.19}$$

Now, following Theorem 3.1, we see that for the lower bound in (3.4.19), we have that

$$\Lambda(\theta) = \log(\mathbb{E}[\exp(\theta(X_i(1) - \beta))]) = \frac{\theta^2\sigma^2}{2} - \beta\theta.$$

Therefore, $\gamma = 2\beta/\sigma^2$, and $\Lambda'(\gamma) = \beta$, which means that the lower bound satisfies the limit in (3.2.8). Thus, we have a sharp lower bound.

To prove a converging upper bound, we use similar techniques as in Lemmas 3.5, 3.6, and 3.7, but now for continuous-time processes. We write, as before, that $t_1^{(N)} = \left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right)\log N = \left(\frac{\sigma^2}{2\beta^2} - \epsilon\right)\log N$ and $t_3^{(N)} = \left(\frac{1}{\Lambda'(\gamma)\gamma} + \epsilon\right)\log N = \left(\frac{\sigma^2}{2\beta^2} + \epsilon\right)\log N$, with $0 < \epsilon < \frac{\sigma^2}{2\beta^2}$. We have that

$$\mathbb{P}\left(\max_{i\le N}\sup_{0<s<t_1^{(N)}}(B_i(s) + B_A(s) - \beta s) > \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right) \overset{N\to\infty}{\longrightarrow} 0. \tag{3.4.20}$$

To prove this limit, we first give the following upper bound:

$$\max_{i\le N}\sup_{0<s<t_1^{(N)}}(B_i(s) + B_A(s) - \beta s) \le \max_{i\le N}\sup_{0<s<t_1^{(N)}}(B_i(s) - \beta s) + \sup_{0<s<t_1^{(N)}} B_A(s). \tag{3.4.21}$$

The first term on the right-hand side of (3.4.21) is a maximum of N i.i.d. random variables. The tail probability of these random variables is known: we have that

$$\mathbb{P}\left(\sup_{0<s<t_1^{(N)}}(B_i(s) - \beta s) > y\right) = \Phi\left(\frac{-y - \beta t_1^{(N)}}{\sigma\sqrt{t_1^{(N)}}}\right) + \exp\left(-\frac{2\beta}{\sigma^2}y\right)\Phi\left(\frac{-y + \beta t_1^{(N)}}{\sigma\sqrt{t_1^{(N)}}}\right); \tag{3.4.22}$$

see [1, Eq. (1.1)]. Then, by the union bound, we have that

$$\mathbb{P}\left(\max_{i\le N}\sup_{0<s<t_1^{(N)}}(B_i(s) - \beta s) > \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right) \le N\,\mathbb{P}\left(\sup_{0<s<t_1^{(N)}}(B_i(s) - \beta s) > \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right).$$

We now use the expression in (3.4.22) to prove that this upper bound converges to 0 as $N \to \infty$. We see that by letting $y = \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}$, the first term on the right-hand

side of (3.4.22) becomes

$$\Phi\left(\frac{-\frac{\sigma^2}{2\beta}\log N - x\sqrt{\log N} - \beta\left(\frac{\sigma^2}{2\beta^2}-\epsilon\right)\log N}{\sigma\sqrt{\left(\frac{\sigma^2}{2\beta^2}-\epsilon\right)\log N}}\right) = \Phi\left(\frac{-\frac{\sigma^2}{\beta}+\beta\epsilon}{\sigma\sqrt{\frac{\sigma^2}{2\beta^2}-\epsilon}}\sqrt{\log N}(1+o(1))\right).$$

The cumulative distribution function of the normal distribution Φ satisfies $\Phi(-x) = 1-\Phi(x)$. Furthermore, we have that $1-\Phi(x) \sim \exp(-x^2/2)/(\sqrt{2\pi}x)$ as $x \to \infty$; see [4, Eq. (2.1.1), p. 49], with $f(x) \sim g(x)$ as $x \to \infty$ meaning that $\lim_{x\to\infty} f(x)/g(x) = 1$. Therefore, because

$$\frac{-\frac{\sigma^2}{\beta}+\beta\epsilon}{\sigma\sqrt{\frac{\sigma^2}{2\beta^2}-\epsilon}} < -\sqrt{2},$$

we get that

$$N\Phi\left(\frac{-\frac{\sigma^2}{2\beta}\log N - x\sqrt{\log N} - \beta\left(\frac{\sigma^2}{2\beta^2}-\epsilon\right)\log N}{\sigma\sqrt{\left(\frac{\sigma^2}{2\beta^2}-\epsilon\right)\log N}}\right) \overset{N\to\infty}{\longrightarrow} 0. \tag{3.4.23}$$

By letting $y = \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}$, the second term on the right-hand side of (3.4.22) becomes

$$\begin{aligned}
&\exp\left(-\frac{2\beta}{\sigma^2}\left(\frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right)\right) \\
&\quad\cdot\Phi\left(\frac{-\left(\frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right) + \beta\left(\left(\frac{\sigma^2}{2\beta^2}-\epsilon\right)\log N + x\sqrt{\log N}\right)}{\sigma\sqrt{\left(\frac{\sigma^2}{2\beta^2}-\epsilon\right)\log N}}\right) \\
&= \frac{1}{N}\exp\left(-\frac{2\beta}{\sigma^2}x\sqrt{\log N}\right)\Phi\left(\frac{-\beta\epsilon}{\sqrt{\frac{\sigma^2}{2\beta^2}-\epsilon}}\sqrt{\log N}(1+o(1))\right).
\end{aligned}$$

By again using the fact that $\Phi(-x) \sim \exp(-x^2/2)/(\sqrt{2\pi}x)$ as $x \to \infty$, we obtain that

$$N\exp\left(-\frac{2\beta}{\sigma^2}\left(\frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right)\right)$$

$$\cdot\,\Phi\left(\frac{-\left(\frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right) + \beta\left(\left(\frac{\sigma^2}{2\beta^2}-\epsilon\right)\log N + x\sqrt{\log N}\right)}{\sigma\sqrt{\left(\frac{\sigma^2}{2\beta^2}-\epsilon\right)\log N}}\right)$$

$$= N\frac{1}{N}\exp\left(-\frac{2\beta}{\sigma^2}x\sqrt{\log N}\right)\Phi\left(\frac{-\beta\epsilon}{\sqrt{\frac{\sigma^2}{2\beta^2}-\epsilon}}\sqrt{\log N}(1+o(1))\right) \overset{N\to\infty}{\longrightarrow} 0. \tag{3.4.24}$$

From these two limits in (3.4.23) and (3.4.24) it follows that

$$N\,\mathbb{P}\left(\sup_{0<s<t_1^{(N)}}(B_i(s)-\beta s) > \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right) \overset{N\to\infty}{\longrightarrow} 0.$$

For the second term on the right-hand side of (3.4.21), we have, because $(B_A(t), t \geq 0)$ is a Brownian motion with drift 0, that

$$\sup_{0<s<t_1^{(N)}} B_A(s) \overset{d}{=} |B_A(t_1^{(N)})| \overset{d}{=} \sqrt{\left(\frac{\sigma^2}{2\beta^2}-\epsilon\right)\log N}|X|,$$

with $X \sim \mathcal{N}(0,1)$. One can easily see this by filling in $\beta = 0$ and by replacing σ with σ_A in (3.4.22).

Therefore, we can use the upper bound in (3.4.21), and get by the union bound that

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\leq N}\sup_{0<s<t_1^{(N)}}(B_i(s)+B_A(s)-\beta s) > \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right)\\
&\quad\leq N\,\mathbb{P}\left(\sup_{0<s<t_1^{(N)}}(B_i(s)-\beta s) > \frac{\sigma^2}{2\beta}\log N + (x-y)\sqrt{\log N}\right)\\
&\qquad+\mathbb{P}\left(\sup_{0<s<t_1^{(N)}} B_A(s) > y\sqrt{\log N}\right)\\
&\overset{N\to\infty}{\longrightarrow}\mathbb{P}\left(|X| > \frac{y}{\sqrt{\frac{\sigma^2}{2\beta^2}-\epsilon}}\right).
\end{aligned}$$

This last expression converges to 0 as $y \to \infty$. Thus, the limit in (3.4.20) follows.

Similarly to the limit in (3.4.20), we also have that

$$\mathbb{P}\left(\max_{i\leq N}\sup_{s>t_3^{(N)}}(B_i(s)+B_A(s)-\beta s) > \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right) \overset{N\to\infty}{\longrightarrow} 0. \tag{3.4.25}$$

To prove this limit, we first take

$$0 < \delta < \min\left(\beta + \frac{\beta\sigma^2}{2\beta^2\epsilon + \sigma^2} - 2\frac{\beta\sigma}{\sqrt{2\beta^2\epsilon + \sigma^2}}, \frac{2\beta^3\epsilon}{2\beta^2\epsilon + \sigma^2}, \beta\right).$$

Then, we bound

$$\max_{i\leq N} \sup_{s>t_3^{(N)}} (B_i(s) + B_A(s) - \beta s) \leq \max_{i\leq N} \sup_{s>t_3^{(N)}} (B_i(s) - (\beta-\delta)s) + \sup_{s>0}(B_A(s) - \delta s). \tag{3.4.26}$$

The second term on the right-hand side is exponentially distributed with mean $\sigma_A^2/(2\delta)$. We can, due to the fact that Brownian motions have independent increments, rewrite the first term as follows:

$$\max_{i\leq N} \sup_{s>t_3^{(N)}} (B_i(s) - (\beta-\delta)s)$$
$$= \max_{i\leq N}\left(B_i(t_3^{(N)}) - (\beta-\delta)t_3^{(N)} + \sup_{s>0}(\hat{B}_i(s) - (\beta-\delta)s)\right),$$

with $(\hat{B}_i(t), t \geq 0)$ an independent copy of $(B_i(t), t \geq 0)$. By using the union bound, we have that

$$\mathbb{P}\left(\max_{i\leq N}\left(B_i(t_3^{(N)}) - (\beta-\delta)t_3^{(N)} + \sup_{s>0}(\hat{B}_i(s) - (\beta-\delta)s)\right) > \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right)$$
$$\leq N\,\mathbb{P}\left(B_i(t_3^{(N)}) - (\beta-\delta)t_3^{(N)} + \sup_{s>0}(\hat{B}_i(s) - (\beta-\delta)s) > \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right).$$

Now, by Chernoff's bound, we obtain for all $x \in \mathbb{R}$ and $0 < \theta < 2(\beta-\delta)/\sigma^2$ that

$$N\,\mathbb{P}\left(B_i(t_3^{(N)}) - (\beta-\delta)t_3^{(N)} + \sup_{s>0}(\hat{B}_i(s) - (\beta-\delta)s) > \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right)$$
$$\leq N\exp\left(t_3^{(N)}\frac{\theta^2\sigma^2}{2} - \theta(\beta-\delta)t_3^{(N)} - \theta\frac{\sigma^2}{2\beta}\log N - \theta x\sqrt{\log N}\right)\frac{2(\beta-\delta)/\sigma^2}{2(\beta-\delta)/\sigma^2 - \theta}. \tag{3.4.27}$$

We choose $\theta = (\beta-\delta)/\sigma^2 + \beta/(\sigma^2 + 2\beta^2\epsilon)$. Because

$$\delta < \frac{2\beta^3\epsilon}{2\beta^2\epsilon + \sigma^2},$$

we have that $\theta < 2(\beta-\delta)/\sigma^2$. Thus, in order to derive the asymptotics of the term on the right-hand side of (3.4.27), we can ignore the term

$$\frac{2(\beta-\delta)/\sigma^2}{2(\beta-\delta)/\sigma^2 - \theta}.$$

Moreover, the term $-\theta x\sqrt{\log N} = o(\log N)$. The main remaining terms are

$$\begin{aligned}
&N\exp\left(t_3^{(N)}\frac{\theta^2\sigma^2}{2} - \theta(\beta-\delta)t_3^{(N)} - \theta\frac{\sigma^2}{2\beta}\log N + o(\log N)\right)\\
&\quad= \exp\left(\left(1+\left(\frac{\sigma^2}{2\beta^2}+\epsilon\right)\frac{\theta^2\sigma^2}{2} - \theta(\beta-\delta)\left(\frac{\sigma^2}{2\beta^2}+\epsilon\right) - \theta\frac{\sigma^2}{2\beta}\right)\log N(1+o(1))\right)\\
&\quad= \exp\left(\frac{1}{4}\left(\frac{\delta(4\beta-\delta)}{\beta^2} + \frac{2\beta^2\epsilon}{2\beta^2\epsilon+\sigma^2} - \frac{2\epsilon(\beta-\delta)^2}{\sigma^2}\right)\log N(1+o(1))\right).
\end{aligned}$$

Because

$$0 < \delta < \beta + \frac{\beta\sigma^2}{2\beta^2\epsilon+\sigma^2} - 2\frac{\beta\sigma}{\sqrt{2\beta^2\epsilon+\sigma^2}},$$

we have that

$$\frac{1}{4}\left(\frac{\delta(4\beta-\delta)}{\beta^2} + \frac{2\beta^2\epsilon}{2\beta^2\epsilon+\sigma^2} - \frac{2\epsilon(\beta-\delta)^2}{\sigma^2}\right) < 0.$$

Therefore, we have that

$$N\exp\left(t_3^{(N)}\frac{\theta^2\sigma^2}{2} - \theta(\beta-\delta)t_3^{(N)} - \theta\frac{\sigma^2}{2\beta}\log N - \theta x\sqrt{\log N}\right)\frac{2(\beta-\delta)/\sigma^2}{2(\beta-\delta)/\sigma^2-\theta} \overset{N\to\infty}{\longrightarrow} 0.$$

Thus,

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\le N}\sup_{s>t_3^{(N)}}(B_i(s)+B_A(s)-\beta s) > \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right)\\
&\quad\le N\,\mathbb{P}\left(\sup_{s>t_3^{(N)}}(B_i(s)-(\beta-\delta)s) > \frac{\sigma^2}{2\beta}\log N + (x-y)\sqrt{\log N}\right)\\
&\qquad+\mathbb{P}\left(\sup_{s>0}(B_A(s)-\delta s) > y\sqrt{\log N}\right)\\
&\quad\overset{N\to\infty}{\longrightarrow} 0,
\end{aligned}$$

for $y > 0$. Therefore, the limit in (3.4.25) follows.

Finally, we show that

$$\begin{aligned}
&\limsup_{N\to\infty}\mathbb{P}\left(\frac{\max_{i\le N}\sup_{t_1^{(N)}\le s\le t_3^{(N)}}(B_i(s)+B_A(s)-\beta s) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}} \ge x\right)\\
&\qquad\le \limsup_{N\to\infty}\mathbb{P}\left(\sigma_A\sqrt{\frac{\sigma^2}{2\beta^2}-\epsilon}X_1 + \sqrt{2\epsilon}\sigma_A|X_2| \ge x\right), \quad (3.4.28)
\end{aligned}$$

as $N \to \infty$, with $X_1, X_2 \sim \mathcal{N}(0,1)$ and independent.

To prove this, we bound

$$\begin{aligned}
&\frac{\max_{i\le N} \sup_{t_1^{(N)}\le s\le t_3^{(N)}} \left(B_i(s) + B_A(s) - \beta s\right) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}} \\
&\quad\le \sup_{t_1^{(N)}\le s\le t_3^{(N)}} \frac{B_A(s)}{\sqrt{\log N}} + \frac{\max_{i\le N} \sup_{t_1^{(N)}\le s\le t_3^{(N)}} (B_i(s) - \beta s) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}} \\
&\quad\le \sup_{t_1^{(N)}\le s\le t_3^{(N)}} \frac{B_A(s)}{\sqrt{\log N}} + \frac{\max_{i\le N} \sup_{s>0} (B_i(s) - \beta s) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}}.
\end{aligned}$$

We can write

$$\begin{aligned}
\sup_{t_1^{(N)}\le s\le t_3^{(N)}} \frac{B_A(s)}{\sqrt{\log N}} &= \frac{B_A(t_1^{(N)})}{\sqrt{\log N}} + \sup_{0\le s<2\epsilon\log N} \frac{\hat{B}_A(s)}{\sqrt{\log N}} \\
&\stackrel{d}{=} \sigma_A \sqrt{\frac{\sigma^2}{2\beta^2} - \epsilon} X_1 + \sqrt{2\epsilon}\sigma_A |X_2|,
\end{aligned}$$

with $X_1, X_2 \sim \mathcal{N}(0,1)$ and independent, and $(\hat{B}_A(t), t \ge 0)$ an independent copy of $(B_A(t), t \ge 0)$. Furthermore, because the random variable $\max_{i\le N} \sup_{s>0}(B_i(s) - \beta s)$ is a maximum of N i.i.d. exponentially distributed random variables, we have that

$$\frac{2\beta}{\sigma^2}\left(\max_{i\le N} \sup_{s>0}(B_i(s) - \beta s) - \frac{\sigma^2}{2\beta}\log N\right) \xrightarrow{d} G,$$

as $N \to \infty$, with $G \sim$ Gumbel; see [67, Thm. 1.2.1, p. 19]. Therefore,

$$\frac{\max_{i\le N} \sup_{s>0}(B_i(s) - \beta s) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}} \xrightarrow{\mathbb{P}} 0,$$

as $N \to \infty$. The limit in (3.4.28) follows. Combining the limits in (3.4.20), (3.4.25), and (3.4.28) together gives that

$$\begin{aligned}
&\limsup_{N\to\infty} \mathbb{P}\left(\frac{\max_{i\le N} \sup_{s>0} \left(B_i(s) + B_A(s) - \beta s\right) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}} \ge x\right) \\
&\qquad\le \mathbb{P}\left(\sigma_A \sqrt{\frac{\sigma^2}{2\beta^2} - \epsilon} X_1 + \sqrt{2\epsilon}\sigma_A |X_2| \ge x\right) \xrightarrow{\epsilon\downarrow 0} \mathbb{P}\left(\frac{\sigma\sigma_A}{\sqrt{2}\beta} X \ge x\right).
\end{aligned}$$

The corollary follows.

□

Proof of Lemma 3.3. Without loss of generality, we assume that $\beta = 1$. We write $Y_i = \sup_{s>0}(B_i(s) + B_A(s) - s)$. Let $d = \frac{\sigma^2}{2}$, and $X_N = \frac{\sqrt{2}}{\sigma\sigma_A} \frac{B_A(d\log N)}{\sqrt{\log N}}$. It is easy to see that

$X_N \sim \mathcal{N}(0,1)$. We want to prove that

$$\mathbb{E}\left[\left|\frac{\max_{i\leq N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\sigma\sigma_A}{\sqrt{2}}X_N\right|\right] \overset{N\to\infty}{\longrightarrow} 0. \tag{3.4.29}$$

First observe that

$$\mathbb{E}\left[\left|\frac{\max_{i\leq N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\sigma\sigma_A}{\sqrt{2}}X_N\right|\right] \tag{3.4.30}$$

$$\leq \mathbb{E}\left[\left|\frac{\max_{i\leq N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i\leq N} B_i(d\log N) + B_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right|\right] \tag{3.4.31}$$

$$+ \mathbb{E}\left[\left|\frac{\max_{i\leq N} B_i(d\log N) + B_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}} - \frac{\sigma\sigma_A}{\sqrt{2}}X_N\right|\right]. \tag{3.4.32}$$

Because $Y_i > B_i(d\log N) + B_A(d\log N) - d\log N$, we can rewrite (3.4.31):

$$\mathbb{E}\left[\left|\frac{\max_{i\leq N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i\leq N} B_i(d\log N) + B_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right|\right]$$

$$= \mathbb{E}\left[\frac{\max_{i\leq N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i\leq N} B_i(d\log N) + B_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right]. \tag{3.4.33}$$

Moreover, due to [123, Th. 3.1]:

$$\mathbb{E}\left[\left|\frac{\max_{i\leq N} B_i(d\log N) + B_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}} - \frac{\sigma\sigma_A}{\sqrt{2}}X_N\right|\right]$$

$$= \mathbb{E}\left[\left|\frac{\max_{i\leq N} B_i(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right|\right] \overset{N\to\infty}{\longrightarrow} 0. \tag{3.4.34}$$

From this, it also follows that

$$\mathbb{E}\left[\frac{\max_{i\leq N} B_i(d\log N) + B_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right] \overset{N\to\infty}{\longrightarrow} \mathbb{E}\left[\frac{\sigma\sigma_A}{\sqrt{2}}X\right] = 0. \tag{3.4.35}$$

Thus, from the convergence results in (3.4.34) and (3.4.35) together with the bounds in (3.4.31) and (3.4.32), we can conclude that we only need to show that

$$\mathbb{E}\left[\frac{\max_{i\leq N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right] \overset{N\to\infty}{\longrightarrow} 0,$$

in order to prove Lemma 3.3. Because $Y_i > B_i(d\log N) + B_A(d\log N) - d\log N$ and due to the convergence result in (3.4.35), we see that

$$\liminf_{N\to\infty} \mathbb{E}\left[\frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right] \ge 0.$$

In order to prove a converging upper bound, we write

$$\begin{aligned}
&\mathbb{E}\left[\frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right] \\
&\quad\le \mathbb{E}\left[\frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \mathbb{1}\left(-M \le \frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} < M\right)\right] \\
&\qquad + \mathbb{E}\left[\frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \mathbb{1}\left(\frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \ge M\right)\right].
\end{aligned}$$

By Corollary 3.2 and the dominated convergence theorem we have that

$$\begin{aligned}
&\mathbb{E}\left[\frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \mathbb{1}\left(-M \le \frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} < M\right)\right] \\
&\qquad\qquad \overset{N\to\infty}{\longrightarrow} \mathbb{E}\left[\frac{\sigma\sigma_A}{\sqrt{2}} X \mathbb{1}\left(-M \le \frac{\sigma\sigma_A}{\sqrt{2}} X < M\right)\right] = 0.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\limsup_{N\to\infty} \mathbb{E}\left[\frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right] \\
&\qquad\qquad \le \limsup_{N\to\infty} \mathbb{E}\left[\frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \mathbb{1}\left(\frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \ge M\right)\right]
\end{aligned}$$

for all $M > 0$. Now, we bound

$$\max_{i\le N} Y_i \le \max_{i\le N} \sup_{s>0}(B_i(s) - (1 - 1/\sqrt{\log N})s) + \sup_{s>0}(B_A(s) - s/\sqrt{\log N}) =: Z_N.$$

Then, we have the bound

$$\mathbb{E}\left[\frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \mathbb{1}\left(\frac{\max_{i\le N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \ge M\right)\right]$$

$$\leq \mathbb{E}\left[\frac{Z_N - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\mathbb{1}\left(\frac{\max_{i\leq N}\sup_{s>0}(B_i(s) - (1-1/\sqrt{\log N})s) - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \geq M/2\right)\right]$$
$$+ \mathbb{E}\left[\frac{Z_N - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\mathbb{1}\left(\frac{\sup_{s>0}(B_A(s) - s/\sqrt{\log N})}{\sqrt{\log N}} \geq M/2\right)\right].$$

We have

$$\mathbb{E}\left[\frac{\sup_{s>0}(B_A(s) - s/\sqrt{\log N})}{\sqrt{\log N}}\right] = \frac{\sigma_A^2}{2},$$

and

$$\mathbb{E}\left[\frac{\max_{i\leq N}\sup_{s>0}(B_i(s) - (1-1/\sqrt{\log N})s) - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right] \overset{N\to\infty}{\longrightarrow} \frac{\sigma^2}{2}.$$

Furthermore, due to the memoryless property of exponential random variables, we have that

$$\mathbb{E}\left[\frac{\sup_{s>0}(B_A(s) - s/\sqrt{\log N})}{\sqrt{\log N}}\mathbb{1}\left(\frac{\sup_{s>0}(B_A(s) - s/\sqrt{\log N})}{\sqrt{\log N}} \geq M/2\right)\right]$$
$$= \exp(-M/\sigma_A^2)\left(\frac{M}{2} + \frac{\sigma_A^2}{2}\right) \overset{M\to\infty}{\longrightarrow} 0,$$

and

$$\mathbb{E}\left[\frac{\max_{i\leq N}\sup_{s>0}(B_i(s) - (1-1/\sqrt{\log N})s) - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right.$$
$$\left.\cdot\,\mathbb{1}\left(\frac{\max_{i\leq N}\sup_{s>0}(B_i(s) - (1-1/\sqrt{\log N})s) - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \geq M/2\right)\right]$$
$$\leq N\mathbb{E}\left[\frac{\sup_{s>0}(B_i(s) - (1-1/\sqrt{\log N})s) - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right.$$
$$\left.\cdot\,\mathbb{1}\left(\frac{\sup_{s>0}(B_i(s) - (1-1/\sqrt{\log N})s) - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \geq M/2\right)\right] \overset{N\to\infty}{\longrightarrow} 0,$$

for M large enough. From these results, it follows that,

$$\lim_{M\to\infty}\limsup_{N\to\infty}\mathbb{E}\left[\frac{\max_{i\leq N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\mathbb{1}\left(\frac{\max_{i\leq N} Y_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \geq M\right)\right] = 0,$$

the lemma follows. □

Chapter 4

Large deviations principle

4.1. Introduction

In this chapter, we study the Brownian fork-join queue. We consider a continuous-time model as described in Corollary 3.2 and Lemma 3.3 in Chapter 3. We model arrival and service processes directly by Brownian motions. We study the resulting maximum queue length, as we also do in Chapter 6.

Specifically, we model the delay in queue i by $Q_i^\beta = \sup_{s>0}(B_i(s) + B_A(s) - \beta s)$, where $(B_A(t), t \geq 0)$ is a Brownian motion term with standard deviation σ_A that represents the fluctuations in the arrival process, $(B_i(t), t \geq 0)$ is a Brownian motion term with standard deviation σ that represents the fluctuations in the service process, and $\beta > 0$ represents the drift of the queue. Furthermore, we assume that $(B_i, i \leq N)$ are i.i.d. Brownian motions, and for all i, the processes $(B_i(t), t \geq 0)$ and $(B_A(t), t \geq 0)$ are mutually independent. In this chapter, we write $\bar{Q}_N^\beta = \max_{i \leq N} Q_i^\beta$.

Under these assumptions, in Corollary 3.2 we have shown that $\bar{Q}_N^\beta$ is in the domain of attraction of the normal distribution:

$$\mathbb{P}\left(\bar{Q}_N^\beta > \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right) \overset{N\to\infty}{\longrightarrow} \mathbb{P}\left(\frac{\sigma\sigma_A}{\sqrt{2}\beta}X > x\right), \tag{4.1.1}$$

with $X \overset{d}{=} \mathcal{N}(0,1)$. This means that $\bar{Q}_N^\beta$ centers around $\frac{\sigma^2}{2\beta}\log N$ and deviates with order $\sqrt{\log N}$.

This convergence result provides a prediction of the typical maximum queue length. In assembly systems as discussed in Section 1.2, this maximum queue length determines the typical delay. So, one might also be interested in the question how likely it is that the delay will be much longer, as delays may cause large costs. Obviously, the probability $\mathbb{P}(\bar{Q}_N^\beta > y_N) \overset{N\to\infty}{\longrightarrow} 0$, when $y_N - \frac{\sigma^2}{2\beta}\log N$ grows to infinity at a rate faster than $\sqrt{\log N}$,

This chapter is based on [137].

but the question is how fast this probability converges to 0. In this chapter, we focus on the probability

$$\mathbb{P}\left(\bar{Q}_N^\beta > \left(\frac{\sigma^2}{2\beta} + a\right)\log N\right),$$

with $a > 0$. As we show later on, the exact behavior of this tail probability depends on the choice of a, where we can distinguish three regimes: $0 < a < a^\star$, $a = a^\star$, and $a > a^\star$, with $a^\star$ an explicitly identified constant in $(0, \infty)$. The logarithmic asymptotics for these three regimes are given in Theorem 4.1, while the sharper asymptotics for the cases $a > a^\star$, $a = a^\star$, and $0 < a < a^\star$ are given in Theorems 4.2, 4.3, and 4.4, respectively. It easily follows from the proofs that the convergence behavior of $\mathbb{P}(\bar{Q}_N^\beta > y_N)$ when y_N is of larger order than $\log N$, is the same as for the case $a > a^\star$; see Corollary 4.5.1.

The work in this chapter is related to the literature on extreme values of Gaussian processes. In this chapter, we examine exceedance probabilities of the order $(\frac{\sigma^2}{2\beta} + a)\log N$ with $a > 0$. More work has been done on joint suprema of Brownian motions. For instance, [90] gives the solution of the Laplace transform of joint first passage times in terms of the solution of a partial differential equation, where the Brownian motions are dependent. Further, [47] analyze the tail asymptotics of the all-time suprema of two dependent Brownian motions. The joint suprema of a finite number of Brownian motions is also studied [46], where the authors give tail asymptotics of the joint suprema of independent Gaussian processes over a finite time interval. These are just three examples – more results may be found in [102] and [126].

This chapter is organized as follows. In Section 4.2, we present our main results, which contain an interesting phase transition in the way a large supremum occurs depending on the value of a. We explain the reason behind this phase transition in detail. The rest of the chapter is devoted to proofs. In Section 4.3, we give the proof of Theorem 4.1, which focuses on logarithmic asymptotics. In Section 4.4, we present some auxiliary lemmas that allow us to provide the proofs of Theorems 4.2, 4.3, and 4.4 in Sections 4.5.1, 4.5.2, and 4.5.3, respectively, which deal with asymptotic estimates that are sharper than Theorem 4.1.

4.2. Main results

In this section, we present our main results and also provide some intuition. We first introduce some additional notation.

Definition 4.1. *The sequence $(B_i, i \leq N)$ is a sequence of i.i.d. Brownian motions with standard deviation σ, $(B_A(t), t \geq 0)$ is a Brownian motion with standard deviation σ_A, $(B_i(t), t \geq 0)$ and $(B_A(t), t \geq 0)$ are mutually independent for all i, the steady-state queue length in front of server i is given by*

$$Q_i^\beta := \sup_{s>0}(B_i(s) + B_A(s) - \beta s), \tag{4.2.1}$$

and the maximum queue length equals

$$\bar{Q}_N^\beta := \max_{i\le N} Q_i^\beta. \tag{4.2.2}$$

Next, we write the supremum of a Brownian motion $(B_i(t) + B_A(t) - \beta t, t \ge 0)$ *over an interval* (u, v) *as*

$$Q_i^\beta(u,v) := \sup_{u<s<v} (B_i(s) + B_A(s) - \beta s), \tag{4.2.3}$$

and the maximum of N *of these identically distributed random variables as*

$$\bar{Q}_N^\beta(u,v) := \max_{i\le N} Q_i^\beta(u,v). \tag{4.2.4}$$

Furthermore, we write $Q_i^\beta(u) = Q_i^\beta(u,\infty)$ *and* $\bar{Q}_N^\beta(u) = \bar{Q}_N^\beta(u,\infty)$.

We give additional shorthand notation that we use later on.

Definition 4.2.

$$f_N(a) := \left(\frac{\sigma^2}{2\beta} + a\right)\log N, \tag{4.2.5}$$

$$\lambda(a) := 1 - \sigma/\sqrt{2a\beta + \sigma^2}, \tag{4.2.6}$$

$$T_N(a,k) := f_N(a)/\beta + k\sqrt{\log N}, \tag{4.2.7}$$

$$T_N(a) := T_N(a,0). \tag{4.2.8}$$

Finally, we define

$$\gamma(a) := \begin{cases} \dfrac{2a\beta + 2\sigma^2 - 2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2} & \text{if } 0 < a < a^\star, \\ \dfrac{2a\beta - \sigma_A^2}{\sigma^2 + \sigma_A^2} & \text{if } a \ge a^\star, \end{cases} \tag{4.2.9}$$

with

$$a^\star := \frac{\sigma_A^4}{\sigma^2 2\beta} + \frac{\sigma_A^2}{\beta}.$$

The function $\gamma(a)$ appears in the limit of the logarithmic asymptotics of $\mathbb{P}(\bar{Q}_N^\beta > f_N(a))$. As can be seen from (4.2.9), from $a = a^\star$ onwards, the function $\gamma(a)$ is linear. Moreover, we see that $\gamma(a)$ is continuous everywhere, also for $a = a^\star$. In Figure 4.1, we plot $-\gamma(a)$ for certain choices of the parameters σ, σ_A, β, and $a^\star$.

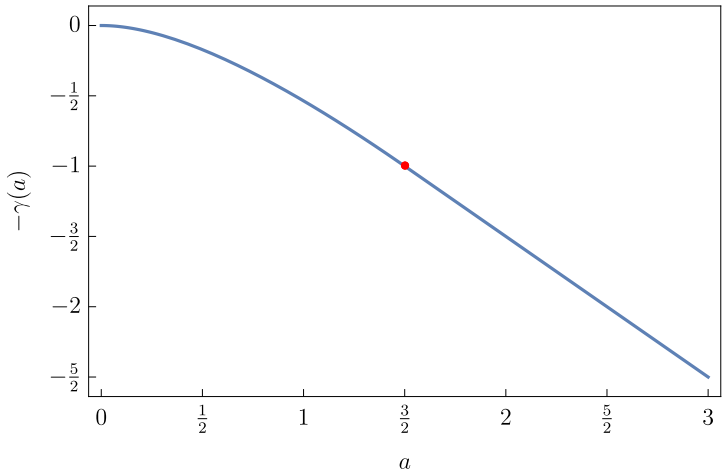


Figure 4.1 $\sigma = 1, \sigma_A = 1, \beta = 1, a^\star = 3/2$

Throughout this chapter, we analyze the fork-join queueing system as defined in Definitions 4.1 and 4.2. Our first result, Theorem 4.1, provides the logarithmic asymptotics of the tail probability of the maximum steady-state queue length $\mathbb{P}(\bar{Q}_N^\beta > f_N(a))$.

Theorem 4.1. *For the model given in Definition 4.1 with the additional notation given in Definition 4.2, and $a > 0$, we have that*

$$\frac{\log(\mathbb{P}(\bar{Q}_N^\beta > f_N(a)))}{\log N} \overset{N\to\infty}{\longrightarrow} -\gamma(a). \tag{4.2.10}$$

We give the proof of Theorem 4.1 in Section 4.3. To provide some intuition, the form of the function $\gamma(a)$ suggests there are at least two regimes: the case where $0 < a < a^\star$, and the case where $a \geq a^\star$. These two cases reveal interesting information on the tail behavior of the maximum queue length $\bar{Q}_N^\beta$.

Case $a > a^\star$. First, we give some intuitive explanation for the case $a > a^\star$. The maximum steady-state queue length is the maximum of N dependent exponentially distributed random variables. We can use the memoryless property of the exponential distribution to get some heuristic insights into the behavior of the maximum steady-state queue length. Define $\tau := \inf\{t > 0 : \max_{i\leq N} B_i(t) + B_A(t) - \beta t \geq f_N(a^\star)\}$ and $i^\star \in \{j \leq N : B_j(\tau) + B_A(\tau) - \beta\tau =$

$\max_{i\leq N} B_i(\tau) + B_A(\tau) - \beta\tau\}$. Then we get

$$\begin{aligned}
&\mathbb{P}(\bar{Q}_N^{\beta} > f_N(a)) \\
&\quad = \mathbb{P}\left(\max_{i\leq N}\sup_{s>0}\left(B_i(s) + B_A(s) - \beta s\right) > f_N(a)\right) \\
&\quad = \mathbb{P}\left(\max_{i\leq N}\sup_{s>0}\left(B_i(s) + B_A(s) - \beta s\right) > f_N(a) \;\middle|\; \tau < \infty\right)\mathbb{P}(\tau < \infty) \\
&\quad \geq \mathbb{P}\left(\sup_{s>\tau}\left(B_{i^\star}(s) + B_A(s) - \beta s\right) > f_N(a) \;\middle|\; \tau < \infty\right)\mathbb{P}(\tau < \infty).
\end{aligned} \tag{4.2.11}$$

Now, due to the fact that Brownian motions have independent increments, we can write $\sup_{s>\tau}(B_{i^\star}(s) + B_A(s) - \beta s) = B_{i^\star}(\tau) + B_A(\tau) - \beta\tau + \sup_{s>0}(\hat{B}_{i^\star}(s) + \hat{B}_A(s) - \beta s)$, with $(\hat{B}_{i^\star}(t), t \geq 0)$ and $(\hat{B}_A(t), t \geq 0)$ independent copies of $(B_{i^\star}(t), t \geq 0)$ and $(B_A(t), t \geq 0)$, respectively. Thus, the lower bound in (4.2.11) simplifies to

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\leq N}\sup_{s>0}\left(B_i(s) + B_A(s) - \beta s\right) > f_N(a^\star)\right) \\
&\qquad\qquad \cdot \mathbb{P}\left(\sup_{s>0}\left(B_{i^\star}(s) + B_A(s) - \beta s\right) > (a - a^\star)\log N\right).
\end{aligned}$$

Therefore, when we compare this lower bound with the convergence result given in (4.2.10), we get that

$$\begin{aligned}
&\mathbb{P}\Big(\bar{Q}_N^{\beta} > f_N(a)\Big) \\
&\quad = \mathbb{P}\left(\max_{i\leq N}\sup_{s>0}\left(B_i(s) + B_A(s) - \beta s\right) > f_N(a)\right) \\
&\quad \geq \mathbb{P}\left(\max_{i\leq N}\sup_{s>0}\left(B_i(s) + B_A(s) - \beta s\right) > f_N(a^\star)\right) \\
&\qquad \cdot \mathbb{P}\left(\sup_{s>0}\left(B_i(s) + B_A(s) - \beta s\right) > (a - a^\star)\log N\right) \\
&\quad = \mathbb{P}\left(\max_{i\leq N}\sup_{s>0}\left(B_i(s) + B_A(s) - \beta s\right) > f_N(a^\star)\right)\exp\left(-\frac{2\beta(a-a^\star)}{\sigma^2 + \sigma_A^2}\log N\right) \\
&\quad \approx N^{-\gamma(a^\star)}\exp\left(-\frac{2\beta(a-a^\star)}{\sigma^2 + \sigma_A^2}\log N\right) \\
&\quad = N^{-\gamma(a)},
\end{aligned} \tag{4.2.12}$$

with the $\approx$-sign indicating that we use the logarithmic asymptotics from (4.2.10), but we ignore lower order terms. Thus, we see that when we use the result from (4.2.10) for $a = a^\star$, then this lower bound is sharp for $a > a^\star$. From this lower bound (4.2.12) we can conclude that for $a > a^\star$, there is at most one Brownian motion $(B_i(t), t \geq 0)$ for which it holds that $\sup_{s>0}(B_i(s) + B_A(s) - \beta s) > f_N(a)$. The second intuitive observation is that for $a \geq a^\star$,

$N^{-\gamma(a)} = N\,\mathbb{P}(Q_i^\beta > f_N(a))$. Obviously, since $a \geq 0$, the union bound gives that

$$\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \leq N\,\mathbb{P}(Q_i^\beta > f_N(a)) = N^{-\frac{2a\beta-\sigma_A^2}{\sigma^2+\sigma_A^2}}. \tag{4.2.13}$$

The fact that the union bound is sharp when $a \geq a^\star$ indicates that for $a \geq a^\star$, the N queues are asymptotically independent; i.e.,

$$\mathbb{P}\left(\max_{i\leq N}\sup_{s>0}\left(B_i(s)+B_A(s)-\beta s\right) > f_N(a)\right) \approx \mathbb{P}\left(\max_{i\leq N}\sup_{s>0}\left(B_i(s)+B_{A,i}(s)-\beta s\right) > f_N(a)\right),$$

where the arrival processes $(B_{A,i}, i \leq N)$ are independent Brownian motions, and $(B_{A,i}(t), t \geq 0)$ and $(B_i(t), t \geq 0)$ are mutually independent. In Section 4.5.2, we see that the boundary case $a = a^\star$ does show some dependent behavior, but this dependence structure cannot be deduced from the logarithmic asymptotics.

Case $0 < a < a^\star$. Finally, the case $0 < a < a^\star$ is more involved. The function $\gamma(a)$ involves a in a nonlinear fashion. As we observe in Equation (4.2.13), due to the fact that the exponent of the tail probability of an exponentially distributed random variable is linear in a, we expect that the logarithmic asymptotics are also linear in a. Thus the structure of $\gamma(a)$ shows that the dependent part B_A influences the tail asymptotics, and we have that

$$\liminf_{N\to\infty}\mathbb{P}\left(\#\{j \leq N : \sup_{s>0}(B_j(s)+B_A(s)-\beta s) > f_N(a)\} > 1 \,\middle|\, \bar{Q}_N^\beta > f_N(a)\right) > 0.$$

The reason that we see this is that in order to get that the maximum steady-state queue length $\bar{Q}_N^\beta$ reaches the level $f_N(a)$, the arrival process $(B_A(t) - \lambda(a)\beta t, t \geq 0)$ must reach a high level around $\lambda(a)f_N(a)$, which is a rare event; see Equation (4.4.1). Furthermore, one of the N service processes needs to reach a level around $(1-\lambda(a))f_N(a)$; however, this is not a rare event. Even more, the event that a finite number of service processes reaches a level around $(1-\lambda(a))f_N(a)$ has a non-zero probability; see Equation (4.4.2).

The function $\gamma(a)$ has more characteristics that can be explained from Corollary 3.2. For example, $\gamma(0) = 0$, which is to be expected as we know from (4.1.1) and (4.2.5) that for $x = 0$

$$\mathbb{P}(\bar{Q}_N^\beta > f_N(0)) \overset{N\to\infty}{\longrightarrow} \frac{1}{2}.$$

We further have that $(\log N)\gamma(x/\sqrt{\log N}) \overset{N\to\infty}{\longrightarrow} \frac{x^2\beta^2}{\sigma^2\sigma_A^2}$. It thus follows that for N large,

$$N^{-\gamma(x/\sqrt{\log N})} \approx N^{-\frac{x^2\beta^2}{\sigma^2\sigma_A^2\log N}} = \exp\left(-\frac{x^2\beta^2}{\sigma^2\sigma_A^2}\right),$$

which is the exponent of the limiting distribution given in (4.1.1).

To prove the logarithmic asymptotics in Theorem 4.1, it suffices to look at random variables of the type $\max_{i\le N}(B_i(T_N)+B_A(T_N)-\beta T_N)$ instead of the random variable $\bar{Q}_N^{\beta}=\max_{i\le N}\sup_{s>0}(B_i(s)+B_A(s)-\beta s)$, where the appropriate choice of T_N is $T_N(a)$; see Equation (4.2.8). We show this in more detail in the proof of Lemma 4.1. For $a>a^\star$, the logarithmic asymptotics are relatively straightforward to derive because we see a notion of asymptotic independence, as explained above. In the proof of Lemma 4.1, we show that when $0<a\le a^\star$,

$$\begin{aligned}&\log(\mathbb{P}(\bar{Q}_N^{\beta}>f_N(a)))\\&\quad\approx\log(\mathbb{P}(\max_{i\le N}B_i\big(T_N(a)\big)-(1-\lambda(a))\beta T_N(a)>(1-\lambda(a))f_N(a)))\\&\qquad+\log(\mathbb{P}(B_A\big(T_N(a)\big)-\lambda(a)\beta T_N(a)>\lambda(a)f_N(a))),\end{aligned}\tag{4.2.14}$$

when N is large, and we show that the term $\log(\mathbb{P}(\max_{i\le N}B_i\big(T_N(a)\big)-(1-\lambda(a))\beta T_N(a)>(1-\lambda(a))f_N(a)))$ becomes negligible as $N\to\infty$.

We now turn to precise asymptotics, which are stated in Theorems 4.2, 4.3, and 4.4 below for the cases $a>a^\star$, $a=a^\star$, and $0<a<a^\star$, respectively. The proofs of these theorems can be found in Sections 4.5.1, 4.5.2, and 4.5.3.

Theorem 4.2. *For the model given in Definition 4.1 with the additional notation given in Definition 4.2, and $a>a^\star$, we have that*

$$N^{\gamma(a)}\,\mathbb{P}(\bar{Q}_N^{\beta}>f_N(a))\overset{N\to\infty}{\longrightarrow}1.\tag{4.2.15}$$

The theorem shows that for $a>a^\star$, the tail probability of the steady-state maximum queue length has the same asymptotic behavior as the one for independently and identically distributed arrival processes for each queue.

Theorem 4.3. *For the model given in Definition 4.1 with the additional notation given in Definition 4.2, and $a=a^\star$, we have that*

$$N^{\gamma(a^\star)}\,\mathbb{P}(\bar{Q}_N^{\beta}>f_N(a^\star))\overset{N\to\infty}{\longrightarrow}\frac{1}{2}.\tag{4.2.16}$$

To give a heuristic explanation of why we have a transition point at $a=a^\star$, we argue as follows. Because the all-time supremum of a Brownian motion is exponentially distributed it is easy to see that for $a=a^\star$,

$$\sup_{s>0}(B_A(s)-\lambda(a^\star)\beta s)\overset{d}{=}\sup_{s>0}(B_i(s)-(1-\lambda(a^\star))\beta s)\overset{d}{=}\sup_{s>0}(B_i(s)+B_A(s)-\beta s),$$

where $\lambda(a)$ is given in Equation (4.2.6). Similarly, after a straightforward calculation, we observe that for $0<a<a^\star$,

$$\sup_{s>0}(B_A(s)-\lambda(a)\beta s)\ge_{st.}\sup_{s>0}(B_i(s)-(1-\lambda(a))\beta s),$$

and for $a > a^\star$,

$$\sup_{s>0}(B_A(s) - \lambda(a)\beta s) \leq_{st.} \sup_{s>0}(B_i(s) - (1-\lambda(a))\beta s),$$

with $X \geq_{st.} Y$ meaning that $\mathbb{P}(X \geq x) \geq \mathbb{P}(Y \geq x)$ for all x. For $0 < a < a^\star$, large values of $\bar{Q}_N^\beta$ are predominantly caused by fluctuations of $(B_A(t) - \lambda(a)\beta t, t \geq 0)$; we show this rigorously in Section 4.5.3. In contrast, for $a > a^\star$, fluctuations are caused by a combination of the arrival process and one of the service processes, and therefore we see a notion of asymptotic independence.

To explain in more detail why we have a constant $1/2$ at the boundary case $a = a^\star$, we first let $\hat{Q}_i^\beta$ be an independent copy of Q_i^β. Furthermore, observe that since the all-time supremum of a Brownian motion with negative drift is exponentially distributed, $\mathbb{P}(\sup_{s>0}(B_A(s) - \lambda(a^\star)\beta s) > \lambda(a^\star)f_N(a^\star)) = N^{-\gamma(a^\star)}$. Moreover, if the event $\sup_{s>0}(B_A(s) - \lambda(a^\star)\beta s) > \lambda(a^\star)f_N(a^\star)$ happens, it most likely occurs at time $T_N(a^\star)$. By using the union bound and that all suprema follow the same distribution, we may therefore write

$$\begin{aligned}
&\mathbb{P}(\bar{Q}_N^\beta(T_N(a^\star)) > f_N(a^\star) \mid B_A\big(T_N(a^\star)\big) - \lambda(a^\star)\beta T_N(a^\star) = \lambda(a^\star)f_N(a^\star)) \\
&\quad = \mathbb{P}\left(\max_{i\leq N}\left(B_i\big(T_N(a^\star)\big) - (1-\lambda(a^\star))\beta T_N(a^\star) + \hat{Q}_i^\beta\right) > (1-\lambda(a^\star))f_N(a^\star)\right) \\
&\quad \approx N\,\mathbb{P}\Big(B_i\big(T_N(a^\star)\big) - (1-\lambda(a^\star))\beta T_N(a^\star) + \hat{Q}_i^\beta > (1-\lambda(a^\star))f_N(a^\star)\Big) \\
&\quad = N\,\mathbb{P}\left(\sup_{s>T_N(a^\star)}(B_i(s) - (1-\lambda(a^\star))\beta s) > (1-\lambda(a^\star))f_N(a^\star)\right) \overset{N\to\infty}{\longrightarrow} \frac{1}{2}.
\end{aligned}$$

The reason that we see a factor $1/2$ emerging in the limit, follows from the fact that we take the supremum over the set $(T_N(a^\star), \infty)$. As the all-time suprema of Brownian motions are exponentially distributed, it is easy to see that

$$N\,\mathbb{P}\left(\sup_{s>0}(B_i(s) - (1-\lambda(a^\star))\beta s) > (1-\lambda(a^\star))f_N(a^\star)\right) \overset{N\to\infty}{\longrightarrow} 1.$$

Typical hitting times of this supremum are of the form $T_N(a^\star) + k\sqrt{\log N}$, with $k \in \mathbb{R}$. We will see in the proofs that the density of these hitting times will deviate symmetrically around $T_N(a^\star)$; see Lemma 4.4. This heuristically explains that when we take the supremum over the set $(T_N(a^\star), \infty)$, we obtain the limit of $1/2$. If we condition on $\max_{i\leq N}\sup_{s>0}(B_i(s) - (1-\lambda(a^\star))\beta s) = (1-\lambda(a^\star))f_N(a^\star)$, we obtain the same expression after using the same heuristic argument.

Our final result is an improvement of the logarithmic asymptotics for the case $0 < a < a^\star$.

Theorem 4.4. *For the model given in Definition 4.1 with the additional notation given in Definition 4.2, and $0 < a < a^\star$, we have that*

$$\liminf_{N\to\infty} N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}\,\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) > 0, \tag{4.2.17}$$

and

$$\limsup_{N\to\infty} N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}\,\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) < \infty. \tag{4.2.18}$$

We give a proof of this result in Section 4.5.3. As already suggested in Theorem 4.1, for the case $0 < a < a^\star$ we observe more irregular behavior, which manifests itself already in the values of $\gamma(a)$. In Theorem 4.4, we observe that the second term is not a constant, as was the case for the values $a > a^\star$ and $a = a^\star$, but is $(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}$. To obtain heuristic insights, we argue that

$$\begin{aligned}\mathbb{P}\bigg(\sup_{s>0}(B_A(s)-\lambda(a)\beta s) > \lambda(a)f_N(a)+r_N\bigg) &= \exp\bigg(-\frac{2\lambda(a)\beta}{\sigma_A^2}(\lambda(a)f_N(a)+r_N)\bigg)\\ &= N^{-\gamma(a)}(\log N)^{-\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}},\end{aligned} \tag{4.2.19}$$

with $r_N = \frac{\sigma\sqrt{2a\beta+\sigma^2}}{4\beta}\log\log N$. Furthermore, we have for all k that

$$\mathbb{P}\bigg(\max_{i\le N} B_i\big(T_N(a,k)\big) - (1-\lambda(a))\beta T_N(a,k) > (1-\lambda(a))f_N(a) - r_N\bigg) = \Theta(1), \tag{4.2.20}$$

where $z_N = \Theta(1)$ means that $\liminf_{N\to\infty} z_N > 0$ and $\limsup_{N\to\infty} z_N < \infty$. Combining these two results together with the definition of $\bar{Q}_N^\beta$ in (4.2.2), we see that

$$\begin{aligned}&\mathbb{P}\Big(\bar{Q}_N^\beta > f_N(a)\Big)\\ &\quad\ge \mathbb{P}\bigg(\sup_{s>0}(B_A(s)-\lambda(a)\beta s) > \lambda(a)f_N(a)+r_N\\ &\qquad, \max_{i\le N} B_i(\tau^{(N)}) - (1-\lambda(a))\beta\tau^{(N)} > (1-\lambda(a))f_N(a) - r_N\bigg),\end{aligned} \tag{4.2.21}$$

where $\tau^{(N)} = \inf\{t \ge 0 : B_A(t) - \lambda(a)\beta t > \lambda(a)f_N(a) + r_N\}$. We show later on that $\tau^{(N)}$, conditioned on being finite, is of the form $T_N(a,K)$ with K being a random variable. Because

$$\begin{aligned}&\mathbb{P}\bigg(\sup_{s>0}(B_A(s)-\lambda(a)\beta s) > \lambda(a)f_N(a)+r_N\\ &\qquad, \max_{i\le N} B_i(\tau^{(N)}) - (1-\lambda(a))\beta\tau^{(N)} > (1-\lambda(a))f_N(a) - r_N\bigg)\\ &= \mathbb{P}\bigg(\sup_{s>0}(B_A(s)-\lambda(a)\beta s) > \lambda(a)f_N(a)+r_N\bigg)\\ &\qquad\cdot \mathbb{P}\bigg(\max_{i\le N} B_i(\tau^{(N)}) - (1-\lambda(a))\beta\tau^{(N)} > (1-\lambda(a))f_N(a) - r_N \,\bigg|\, \tau^{(N)} < \infty\bigg),\end{aligned} \tag{4.2.22}$$

we retrieve (4.2.17) after combining the results from (4.2.19)–(4.2.22). Thus, it turns out

that for $0 < a < a^\star$, r_N plays a key role. As explained in Section 4.5.2, in the case $0 < a < a^\star$, $(B_A(t) - \lambda(a)\beta t, t \geq 0)$ dominates, which explains why the tail asymptotics of the maximum queue length $\bar{Q}_N^\beta$ are the same as the tail asymptotics of $\sup_{s>0}(B_A(s) - \lambda(a)\beta s)$, and the behavior of $\max_{i\leq N} B_i\big(T_N(a,k)\big) - (1-\lambda(a))\beta T_N(a,k)$ is typical.

The main approach of proving the lower and upper bounds in (4.2.17) and (4.2.18), as well as the limits in (4.2.15) and (4.2.16), is by analyzing lower and upper bounds on the tail probability of the steady-state maximum queue length $\mathbb{P}(\bar{Q}_N^\beta > f_N(a))$. These bounds are derived by utilizing the union bound, Bonferroni's inequality, and a careful construction of hitting times. These hitting times are needed to estimate the time when the supremum most likely hits the desired level and to adequately separate the independent part B_i and the dependent part B_A from each other. We also rely on some existing asymptotic estimates in the literature from extreme-value theory, and on [47], which investigates the case $N = 2$. Finally, we develop a number of auxiliary technical estimates related to the asymptotic behavior of convolutions of normally and exponentially distributed random variables.

These techniques, when put together, are effective in the case $a = a^\star$ and $a > a^\star$ in order to obtain exact asymptotics. In the case $0 < a < a^\star$, we are able to improve upon Theorem 4.1 and characterize the asymptotic behavior of $\mathbb{P}(\bar{Q}_N^\beta > f_N(a))$ up to a constant. To derive precise asymptotics in this case seems beyond the scope of techniques developed in this chapter.

4.3. Proof of the logarithmic asymptotics

In this section, we give a proof of Theorem 4.1, establishing logarithmic asymptotics for the maximum queue length. Our approach is to derive logarithmic lower and upper bounds of the maximum queue length by using the heuristic idea given in (4.2.14), and show that they coincide. These bounds are presented in Lemmas 4.1 and 4.2 below.

Lemma 4.1. *For the model given in Definition 4.1 with the additional notation given in Definition 4.2, and $a > 0$, we have that*

$$\liminf_{N\to\infty} \frac{\log(\mathbb{P}(\bar{Q}_N^\beta > f_N(a)))}{\log N} \geq -\gamma(a). \tag{4.3.1}$$

Proof. Recall that $\lambda(a) = 1-\sigma/\sqrt{2a\beta+\sigma^2}$ and $T_N(a) = f_N(a)/\beta$. By choosing $s = f_N(a)/\beta$ and splitting $-\beta s$ into two terms, observe that

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\leq N}\sup_{s>0}\big(B_i(s) + B_A(s) - \beta s\big) > f_N(a)\right) \qquad (4.3.2)\\
&\geq \mathbb{P}\Bigg(\max_{i\leq N} B_i\big(T_N(a)\big) - (1-\lambda(a))\beta T_N(a) > (1-\lambda(a))f_N(a)\\
&\quad , B_A\big(T_N(a)\big) - \lambda(a)\beta T_N(a) > \lambda(a) f_N(a)\Bigg)
\end{aligned}$$

$$= \mathbb{P}\Big(\max_{i\le N} B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\Big)\,\mathbb{P}\Big(B_A\big(T_N(a)\big) > 2\lambda(a)f_N(a)\Big). \tag{4.3.3}$$

The expression in (4.3.3) is due to the fact that for all i, $(B_i(t), t \ge 0)$ and $(B_A(t), t \ge 0)$ are independent. We now analyze the two probabilities in (4.3.3) separately. Since $(B_i(t), t \ge 0)$ and $(B_j(t), t \ge 0)$ are i.i.d. for all i and j, for the first probability in (4.3.3) we get from Bonferroni's inequality that

$$\begin{aligned}\mathbb{P}\Big(\max_{i\le N} B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\Big) \ge & N\,\mathbb{P}\Big(B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\Big) \\ & -\binom{N}{2}\mathbb{P}\Big(B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\Big)^2.\end{aligned} \tag{4.3.4}$$

Furthermore, it is easy to see that

$$\mathbb{P}\Big(\sup_{s>0}(B_i(s) - (1-\lambda(a))\beta s) > (1-\lambda(a))f_N(a)\Big) = \frac{1}{N} \tag{4.3.5}$$

and that

$$\begin{aligned}&\mathbb{P}(B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)) \\ &\qquad \le \mathbb{P}\Big(\sup_{s>0}(B_i(s) - (1-\lambda(a))\beta s) > (1-\lambda(a))f_N(a)\Big),\end{aligned}$$

and therefore we bound the second term in (4.3.4) as

$$\begin{aligned}&\binom{N}{2}\mathbb{P}\Big(B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\Big)^2 \\ &\quad \le \frac{N^2}{2}\,\mathbb{P}\Big(\sup_{s>0}(B_i(s) - (1-\lambda(a))\beta s) > (1-\lambda(a))f_N(a)\Big) \\ &\qquad \cdot \mathbb{P}\Big(B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\Big) \\ &\quad = \frac{N}{2}\,\mathbb{P}\Big(B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\Big).\end{aligned}$$

Thus, the lower bound given in (4.3.4) can be further bounded to

$$\mathbb{P}\Big(\max_{i\le N} B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\Big) \ge \frac{N}{2}\,\mathbb{P}(B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)).$$

As we aim to derive logarithmic asymptotics, we see that

$$\log\left(\frac{N}{2}\,\mathbb{P}\Big(B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\Big)\right)$$

$$\sim \log N + \log\left(\mathbb{P}\Big(B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\Big)\right),$$

as $N \to \infty$, with $f(x) \sim g(x)$ as $x \to \infty$ meaning that $\lim_{x\to\infty} f(x)/g(x) = 1$. In addition, recall that for a normally distributed random variable X with standard deviation σ, $\log(\mathbb{P}(X > x)) \sim -x^2/(2\sigma^2)$, as $x \to \infty$. Thus, we get that

$$\log\left(\mathbb{P}\Big(B_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\Big)\right) \sim -\frac{(2(1-\lambda(a))f_N(a))^2}{2\sigma^2 T_N(a)} = -\log N,$$

as $N \to \infty$, following the definitions of $\lambda(a)$, $f_N(a)$, and $T_N(a)$. Concluding,

$$\liminf_{N\to\infty} \frac{\log\left(\mathbb{P}\Big(\max_{i\le N} B_i\big(T_N(a)\big) - (1-\lambda(a))\beta T_N(a) > (1-\lambda(a))f_N(a)\Big)\right)}{\log N} \ge 0. \quad (4.3.6)$$

For the second probability in (4.3.3), the logarithmic asymptotics can be easily computed since $B_A\big(f_N(a)\big)$ is normally distributed. We obtain that

$$\frac{\log\left(\mathbb{P}\Big(B_A\big(T_N(a)\big) > 2\lambda(a)f_N(a)\Big)\right)}{\log N} \xrightarrow{N\to\infty} -\frac{2a\beta + 2\sigma^2 - 2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2}. \quad (4.3.7)$$

Thus, after combining these two results in (4.3.6) and (4.3.7) with (4.3.3), we have that,

$$\liminf_{N\to\infty} \frac{\log\left(\mathbb{P}\Big(\max_{i\le N}\sup_{s>0}\big(B_i(s) + B_A(s) - \beta s\big) > f_N(a)\Big)\right)}{\log N} \ge -\frac{2a\beta + 2\sigma^2 - 2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2}, \quad (4.3.8)$$

irrespective of the choice of a. Now, observe that for $a > 0$,

$$\frac{2a\beta + 2\sigma^2 - 2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2} \ge \frac{2a\beta - \sigma_A^2}{\sigma^2 + \sigma_A^2},$$

with equality for $a = a^\star$. This means that only for $0 < a \le a^\star$, the lower bound in (4.3.8) is sharp enough. For $a > a^\star$, we apply the inequality in (4.2.12) to obtain for all $c > 0$ that

$$\mathbb{P}\Big(\max_{i\le N}\sup_{s>0}\big(B_i(s) + B_A(s) - \beta s\big) > f_N(a^\star + c)\Big) \ge \mathbb{P}\Big(\max_{i\le N}\sup_{s>0}\big(B_i(s) + B_A(s) - \beta s\big) > f_N(a^\star)\Big)\exp\left(-\frac{2\beta c\log N}{\sigma^2+\sigma_A^2}\right). \quad (4.3.9)$$

Combining this result with the inequality in (4.3.8), we get that for all $c > 0$,

$$\liminf_{N\to\infty} \frac{\log\left(\mathbb{P}\Big(\max_{i\le N}\sup_{s>0}\big(B_i(s)+B_A(s)-\beta s\big) > f_N(a^\star+c)\Big)\right)}{\log N} \ge -\gamma(a^\star) - \frac{2\beta c}{\sigma^2+\sigma_A^2} = -\gamma(a^\star+c).$$

Combining the lower bounds in (4.3.8) and (4.3.9) gives the lower bound in (4.3.1). □

Lemma 4.2. *For the model given in Definition 4.1 with the additional notation given in Definition 4.2, and $a > 0$, we have that*

$$\limsup_{N\to\infty} \frac{\log(\mathbb{P}(\bar{Q}_N^\beta > f_N(a)))}{\log N} \le -\gamma(a). \tag{4.3.10}$$

Proof. We have by the union bound in (4.2.13) that

$$\limsup_{N\to\infty} \frac{\log(\mathbb{P}(\bar{Q}_N^\beta > f_N(a)))}{\log N} \le -\frac{2a\beta-\sigma_A^2}{\sigma^2+\sigma_A^2}. \tag{4.3.11}$$

This upper bound implies the upper bound given in (4.3.10) for $a \ge a^\star$. Turning to the case $0 < a < a^\star$, we can bound the tail probability of the maximum queue length by using subadditivity, the union bound, and by integrating over possible values of $\sup_{s>0}(B_A(s) - \lambda(a)\beta s)$, and we obtain that

$$\mathbb{P}\Big(\bar{Q}_N^\beta > f_N(a)\Big) \tag{4.3.12}$$

$$\le \mathbb{P}\left(\max_{i\le N}\sup_{s>0}\big(B_i(s)-(1-\lambda(a))\beta s\big) + \sup_{s>0}\big(B_A(s)-\lambda(a)\beta s\big) > f_N(a)\right)$$

$$\le \int_0^{\lambda(a)(\frac{\sigma^2}{2\beta}+a)} \frac{2\lambda(a)\beta}{\sigma_A^2} N\log N\, \mathbb{P}\left(\sup_{s>0}\big(B_i(s)-(1-\lambda(a))\beta s\big) > f_N(a) - y\log N\right) \cdot \exp\left(-\frac{2\lambda(a)\beta y\log N}{\sigma_A^2}\right) dy$$

$$+ \mathbb{P}\left(\sup_{s>0}\big(B_A(s)-\lambda(a)\beta s\big) > \lambda(a) f_N(a)\right) \tag{4.3.13}$$

$$= \int_0^{\lambda(a)(\frac{\sigma^2}{2\beta}+a)} \frac{2\lambda(a)\beta}{\sigma_A^2} N\log N \cdot \exp\left(-\frac{2(1-\lambda(a))\beta}{\sigma^2}\big(f_N(a)-y\log N\big) - \frac{2\lambda(a)\beta y\log N}{\sigma_A^2}\right) dy \tag{4.3.14}$$

$$+ \mathbb{P}\left(\sup_{s>0}\big(B_A(s)-\lambda(a)\beta s\big) > \lambda(a) f_N(a)\right).$$

Because the function $\exp(-\frac{2(1-\lambda(a))\beta}{\sigma^2}(f_N(a)-y\log N) - \frac{2\lambda(a)\beta y\log N}{\sigma_A^2})$ with $y \in [0, \lambda(a)(\frac{\sigma^2}{2\beta}+$

$a)]$ is maximized when $y = \lambda(a)(\frac{\sigma^2}{2\beta} + a)$ and equals $N^{-\frac{2a\beta+2\sigma^2-2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2}-1}$, we get that

$$\begin{aligned}
&\limsup_{N\to\infty} \\
&\frac{\log\left(\int_0^{\lambda(a)(\frac{\sigma^2}{2\beta}+a)} \frac{2\lambda(a)\beta}{\sigma_A^2}\log N \cdot N \exp\left(-\frac{2(1-\lambda(a))\beta}{\sigma^2}\left(f_N(a) - y\log N\right) - \frac{2\lambda(a)\beta y\log N}{\sigma_A^2}\right) dy\right)}{\log N} \\
&= 1 + \limsup_{N\to\infty} \frac{\log\left(\int_0^{\lambda(a)(\frac{\sigma^2}{2\beta}+a)} \exp\left(-\frac{2(1-\lambda(a))\beta}{\sigma^2}\left(f_N(a) - y\log N\right) - \frac{2\lambda(a)\beta y\log N}{\sigma_A^2}\right) dy\right)}{\log N} \\
&\leq -\frac{2a\beta + 2\sigma^2 - 2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2}. \qquad (4.3.15)
\end{aligned}$$

Now that we have found an upper bound for the integral in (4.3.14), we are left with the expression $\mathbb{P}\Big(\sup_{s>0}\left(B_A(s) - \lambda(a)\beta s\right) > \lambda(a) f_N(a)\Big)$ in (4.3.14). For this expression, it holds that

$$\mathbb{P}\left(\sup_{s>0}\left(B_A(s) - \lambda(a)\beta s\right) > \lambda(a) f_N(a)\right) = N^{-\frac{2a\beta+2\sigma^2-2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2}}.$$

Combining the upper bounds in (4.3.11) and (4.3.14) gives the logarithmic upper bound on the maximum queue length in (4.3.10). □

4.4. Useful lemmas

In the previous section, we gave a proof of the logarithmic asymptotics for the maximum queue length $\bar{Q}_N^\beta$. In order to be able to prove sharper results on the tail asymptotics, we need some auxiliary results; the goal of this section is to derive these. We begin by giving an overview of the results in this section.

First, observe that for a Brownian motion $(B(t), t \geq 0)$, we have that

$$\sup_{s>T}(B(s) - \beta s) \stackrel{d}{=} B(T) - \beta T + \sup_{s>0}(\hat{B}(s) - \beta s),$$

where $(\hat{B}(t), t \geq 0)$ is an independent copy of $(B(t), t \geq 0)$. From this, it follows that if we take the supremum of a Brownian motion starting at a positive time, this is in distribution the same as adding a normally distributed random variable to an exponentially distributed random variable. The tail asymptotics of this convolution equal the tail asymptotics of the normally distributed part, the exponentially distributed part, or a more complicated mixture of the two, depending on the starting time T, the standard deviation of $B(s)$ and the drift β. In Lemma 4.3, these three cases are studied in more detail.

Second, our main strategy to investigate the tail asymptotics involves the use of hitting times. Observe that we have a maximum of N mutually dependent random variables. Based

on the results in Section 4.3, we are able to make an educated guess where the supremum is attained. Following the proof of Lemma 4.1, we see that for $T_N(a)$ given in (4.2.8),

$$\mathbb{P}\left(\max_{i\leq N}\sup_{s>0}\left(B_i(s)+B_A(s)-\beta s\right)>f_N(a)\right) \approx \mathbb{P}\left(\max_{i\leq N}\left(B_i(T_N(a))+B_A(T_N(a))-\beta T_N(a)\right)>f_N(a)\right).$$

So the hitting time, conditioned on being finite, is approximately equal to $T_N(a)$. Next, observe that for $0<a\leq a^\star$,

$$\mathbb{P}\left(\sup_{s>0}\left(B_A(s)-\lambda(a)\beta s\right)>\lambda(a)f_N(a)\right)=\exp\left(-\frac{2\lambda(a)\beta}{\sigma_A^2}\lambda(a)f_N(a)\right)=N^{-\gamma(a)}, \quad (4.4.1)$$

and

$$\begin{aligned}&\mathbb{P}\left(\max_{i\leq N}\sup_{s>0}\left(B_i(s)-(1-\lambda(a))\beta s\right)>(1-\lambda(a))f_N(a)\right)\\&=1-\left(1-\exp\left(-\frac{2(1-\lambda(a))\beta}{\sigma^2}(1-\lambda(a))f_N(a)\right)\right)^N\\&=\Theta(1). \end{aligned} \quad (4.4.2)$$

Since the expectation of the hitting time, conditioned on being finite, of a level x, equals this value x divided by the drift, it is easy to see that in both (4.4.1) and (4.4.2) the conditional expectation of the hitting time equals $T_N(a)$. Thus, this heuristically explains why the processes $(B_A(t)-\lambda(a)\beta t, t\geq 0)$ and $(B_i(t)-(1-\lambda(a))\beta t, t\geq 0)$ are important. In Definition 4.3 below, we define the hitting-time densities of these processes and in Lemma 4.4 we show that after proper scaling these densities converge to the densities of normally distributed random variables, corrected with a constant.

Finally, we need to analyze limits of the type

$$\lim_{N\to\infty}\int_{-\infty}^{\infty}\mathbb{P}\left(\sup_{s>\tau^{(N)}}X_i(s)>y_N\middle|\tau^{(N)}=t\right)f_{\tau^{(N)}}(t)dt, \quad (4.4.3)$$

where $\tau^{(N)}$ is a hitting time and $f_{\tau^{(N)}}$ its density, with $f_{\tau^{(N)}}(t)=0$ for $t<0$. In Lemma 4.5, we show that under certain assumptions, we can interchange the integral and the limit, when the integrand is a product of two functions, as is the case in (4.4.3). The proof of this interchange is similar to the proof of the dominated convergence theorem.

Lemma 4.3 (Convolution of normal and exponential distributions)**.** *Let $X\overset{d}{=}\mathcal{N}(0,1)$ and $E\overset{d}{=}Exp(1)$ be independent random variables. Let $(\eta_N, N\geq 1)$, $(x_N, N\geq 1)$ be sequences with $\eta_N>0$, $x_N\overset{N\to\infty}{\longrightarrow}\infty$, and $x_N/\eta_N\overset{N\to\infty}{\longrightarrow}\infty$. Furthermore, let $\mu>0$ and $c\in\mathbb{R}$. Then*

1. *if* $\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N} \overset{N\to\infty}{\longrightarrow} c$,

$$\mathbb{P}\left(\eta_N X+\frac{1}{\mu}E>x_N\right)\sim\frac{\eta_N e^{-\frac{x_N^2}{2\eta_N^2}}}{\sqrt{2\pi}x_N}+\frac{1}{2}e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}(1+erf(c)), \tag{4.4.4}$$

as $N\to\infty$, *with*

$$erf(c)=\frac{2}{\sqrt{\pi}}\int_0^c\exp(-t^2)dt.$$

2. *if* $\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N} \overset{N\to\infty}{\longrightarrow} \infty$,

$$\mathbb{P}\left(\eta_N X+\frac{1}{\mu}E>x_N\right)\sim\frac{\eta_N e^{-\frac{x_N^2}{2\eta_N^2}}}{\sqrt{2\pi}x_N}+e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}, \tag{4.4.5}$$

as $N\to\infty$,

3. *and if* $\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N} \overset{N\to\infty}{\longrightarrow} -\infty$,

$$\mathbb{P}\left(\eta_N X+\frac{1}{\mu}E>x_N\right)\sim\frac{\eta_N e^{-\frac{x_N^2}{2\eta_N^2}}}{\sqrt{2\pi}x_N}-\frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}\frac{\eta_N e^{-\frac{\left(x_N-\mu\eta_N^2\right)^2}{2\eta_N^2}}}{x_N-\mu\eta_N^2}, \tag{4.4.6}$$

as $N\to\infty$.

Proof. We have

$$\mathbb{P}\left(\eta_N X+\frac{1}{\mu}E>x_N\right)=\mathbb{P}(\eta_N X>x_N)+\int_{-\infty}^{x_N/\eta_N}\mathbb{P}\left(\frac{1}{\mu}E>x_N-\eta_N z\right)\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}dz. \tag{4.4.7}$$

The first term satisfies

$$\mathbb{P}(\eta_N X>x_N)\sim\frac{\eta_N e^{-\frac{x_N^2}{2\eta_N^2}}}{\sqrt{2\pi}x_N},$$

as $N\to\infty$. Furthermore,

$$\int_{-\infty}^{x_N/\eta_N}\mathbb{P}\left(\frac{1}{\mu}E>x_N-\eta_N z\right)\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}dz=\frac{1}{2}e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}\left(\operatorname{erf}\left(\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N}\right)+1\right).$$

Two standard results on the error function are that $\operatorname{erf}(z)\to 1$, as $z\to\infty$, and $1+\operatorname{erf}(-z)\sim\frac{e^{-z^2}}{\sqrt{\pi}z}$, as $z\to\infty$; see [2, 7.1.13, 7.1.16 & 7.1.23]. The lemma follows. □

For the remainder of this chapter, we use τ to indicate stochastic hitting times.

Definition 4.3. *For $a > 0$, $r \in \mathbb{R}$, and $i \in \{1, 2, \dots, N\}$, we define the random variable $\tau^{(N)}_{i,a,-r}$ by*

$$\tau^{(N)}_{i,a,-r} := \inf\{t \geq 0 : B_i(t) - (1-\lambda(a))\beta t > (1-\lambda(a))f_N(a) - r\},$$

and the function $f_{\tau^{(N)}_{i,a,-r}}$ as its density, with $f_{\tau^{(N)}_{i,a,-r}}(t) = 0$ for $t < 0$. Similarly, we define the random variable $\tilde{\tau}^{(N)}_{A,a,r}$ by

$$\tilde{\tau}^{(N)}_{A,a,r} := \inf\{t \geq 0 : B_A(t) - \lambda(a)\beta t > \lambda(a)f_N(a) + r\},$$

and the function $f_{\tilde{\tau}^{(N)}_{A,a,r}}$ as its density, with $f_{\tilde{\tau}^{(N)}_{A,a,r}}(t) = 0$ for $t < 0$.

Lemma 4.4 (Convergence of hitting-time density)**.** *For the density function $f_{\tau^{(N)}_{i,a,-r}}$ given in Definition 4.3 and $T_N(a,k)$ given in Equation* (4.2.7)*, we have that*

$$\begin{aligned} &N\sqrt{\log N} f_{\tau^{(N)}_{i,a,-r}}\left(T_N(a,k)\right) \\ &\xrightarrow{N\to\infty} \frac{\beta^2}{\sqrt{\pi}\left(2a\beta+\sigma^2\right)} \exp\left(\frac{\beta\left(8a^2\beta^2 r - \beta^3 k^2 \sigma\sqrt{2a\beta+\sigma^2} + 8a\beta r\sigma^2 + 2r\sigma^4\right)}{\sigma\left(2a\beta+\sigma^2\right)^{5/2}}\right). \end{aligned} \tag{4.4.8}$$

Proof. The density $f_{\tau^{(N)}_{i,a,-r}}(t)$ satisfies

$$f_{\tau^{(N)}_{i,a,-r}}(t) = \frac{(1-\lambda(a))f_N(a) - r}{\sqrt{2\pi}\sigma t^{3/2}} \exp\left(-\frac{((1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta t)^2}{2\sigma^2 t}\right),$$

for $t > 0$, and 0 otherwise; see [32, Eq. (2.0.2), p. 301]. Due to the fact that $T_N(a,k) = f_N(a)/\beta + k\sqrt{\log N}$, for all $k \in \mathbb{R}$, there exists N_k, such that for $N > N_k$, $T_N(a,k) > 0$. Following the notation given in Definition 4.2, we have that the prefactor of the density of the hitting time equals

$$\begin{aligned} \frac{(1-\lambda(a))f_N(a) - r}{\sqrt{2\pi}\sigma T_N(a,k)^{3/2}} &= \frac{\frac{\sigma}{\sqrt{2a\beta+\sigma^2}}(\frac{\sigma^2}{2\beta}+a)\log N - r}{\sqrt{2\pi}\sigma((\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta})\log N + k\sqrt{\log N})^{3/2}} \\ &\sim \frac{\frac{\sigma}{\sqrt{2a\beta+\sigma^2}}(\frac{\sigma^2}{2\beta}+a)\log N}{\sqrt{2\pi}\sigma((\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta})\log N)^{3/2}}, \end{aligned}$$

as $N \to \infty$. When we simplify this last term further, we get

$$\frac{\frac{\sigma}{\sqrt{2a\beta+\sigma^2}}(\frac{\sigma^2}{2\beta}+a)}{\sqrt{2\pi}\sigma((\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}))^{3/2}\sqrt{\log N}} = \frac{\frac{1}{\sqrt{2a\beta+\sigma^2}}}{\sqrt{2\pi}\frac{1}{\beta}\sqrt{\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}}\sqrt{\log N}}$$

$$= \frac{1}{\sqrt{2\pi}\frac{1}{\beta}\sqrt{2a\beta+\sigma^2}\sqrt{\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}}\sqrt{\log N}}.$$

Because we can write $\sqrt{\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}} = \frac{1}{\sqrt{2}\beta}\sqrt{2a\beta+\sigma^2}$, we get

$$\frac{1}{\sqrt{2\pi}\frac{1}{\beta}\sqrt{2a\beta+\sigma^2}\sqrt{\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}}\sqrt{\log N}} = \frac{\beta^2}{\sqrt{\pi}(2a\beta+\sigma^2)\sqrt{\log N}}.$$

So, we can conclude that $\sqrt{\log N}$ times the first term of the density converges to $\frac{\beta^2}{\sqrt{\pi}(2a\beta+\sigma^2)}$, as $N \to \infty$, which is the prefactor of the limit. So, in order to prove the limit in (4.4.8), we are left with proving that

$$N \exp\left(- \frac{((1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k))^2}{2\sigma^2 T_N(a,k)} \right) \overset{N\to\infty}{\longrightarrow} \exp\left(\frac{\beta\left(8a^2\beta^2 r - \beta^3 k^2 \sigma\sqrt{2a\beta+\sigma^2} + 8a\beta r\sigma^2 + 2r\sigma^4\right)}{\sigma\left(2a\beta+\sigma^2\right)^{5/2}} \right). \quad (4.4.9)$$

The numerator of the exponent on the left-hand side of (4.4.9) equals

$$((1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k))^2.$$

Because of the form of $f_N(a)$ and $T_N(a,k)$ as given in Definition 4.2, we can write

$$((1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k))^2 = c_1(\log N)^2 + c_2(\log N)^{3/2} + c_3 \log N + c_4\sqrt{\log N} + r^2, \quad (4.4.10)$$

with c_1, c_2, c_3, c_4 to be determined. In order to determine c_1 we should gather all the terms in

$$(1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k)$$

that scale as $\log N$. We have

$$\begin{aligned}
&(1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k) \\
&\quad = \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\log N - r + \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta\left(\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}\right)\log N \\
&\qquad\qquad + \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta k\sqrt{\log N} \\
&\quad = \frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\log N + o(\log N).
\end{aligned}$$

Therefore,

$$c_1 = \left(\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\right)^2 = \frac{4\sigma^2}{2a\beta+\sigma^2}\left(\frac{\sigma^2}{2\beta}+a\right)^2 = \frac{2\sigma^2}{\beta}\left(\frac{\sigma^2}{2\beta}+a\right).$$

Now, to determine c_2 in (4.4.10), we have

$$\begin{aligned}
&(1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k)\\
&\quad= \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\log N - r + \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta\left(\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}\right)\log N\\
&\qquad\qquad + \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta k\sqrt{\log N}\\
&\quad= \frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\log N + \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta k\sqrt{\log N} - r.
\end{aligned}$$

Therefore, c_2 equals

$$c_2 = 2\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta k = 4\frac{\sigma^2}{2a\beta+\sigma^2}\left(\frac{\sigma^2}{2\beta}+a\right)\beta k = 2\sigma^2 k.$$

Observe that

$$c_1(\log N)^2 + c_2(\log N)^{3/2} = 2\sigma^2 T_N(a,k)\log N.$$

Thus, the numerator of the exponent on the left-hand side of (4.4.9) can be rewritten as

$$-\frac{((1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k))^2}{2\sigma^2 T_N(a,k)} = -\log N + O(1),$$

and we can conclude that

$$N\exp\left(-\frac{((1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k))^2}{2\sigma^2 T_N(a,k)}\right) = N\exp(-\log N + O(1)) = O(1).$$

The only term in (4.4.10) that is still of importance, is the term c_3. We have

$$\begin{aligned}
&-((1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k))^2\\
&\qquad= -\left(\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\log N + \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta k\sqrt{\log N} - r\right)^2.
\end{aligned}$$

The terms that scale as $\log N$ are as follows:

$$c_3\log N = -\left(-2r\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right) + \frac{\sigma^2}{2a\beta+\sigma^2}\beta^2k^2\right)\log N.$$

Thus,

$$\frac{-(-2r\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}(\frac{\sigma^2}{2\beta}+a)+\frac{\sigma^2}{2a\beta+\sigma^2}\beta^2k^2)\log N}{2\sigma^2T_N(a,k)}$$
$$=\frac{-(-2r\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}(\frac{\sigma^2}{2\beta}+a)+\frac{\sigma^2}{2a\beta+\sigma^2}\beta^2k^2)\log N}{2\sigma^2((\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta})\log N+k\sqrt{\log N})}.$$

This expression converges to

$$\frac{-(-2r\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}(\frac{\sigma^2}{2\beta}+a)+\frac{\sigma^2}{2a\beta+\sigma^2}\beta^2k^2)}{2\sigma^2(\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta})}$$
$$=\frac{\beta\left(8a^2\beta^2r-\beta^3k^2\sigma\sqrt{2a\beta+\sigma^2}+8a\beta r\sigma^2+2r\sigma^4\right)}{\sigma\left(2a\beta+\sigma^2\right)^{5/2}},$$

as $N\to\infty$, which is exactly the exponent in the limit of (4.4.9). Putting everything together, the limit in (4.4.8) follows. □

Corollary 4.1. *For the density function $f_{\tau^{(N)}_{i,a,-r}}$ given in Definition 4.3 and $T_N(a,k)$ given in Equation* (4.2.7) *we have that*

$$\lim_{N\to\infty}\int_{-\infty}^{\infty}N\sqrt{\log N}f_{\tau^{(N)}_{i,a,-r}}\big(T_N(a,k)\big)dk=\int_{-\infty}^{\infty}\lim_{N\to\infty}N\sqrt{\log N}f_{\tau^{(N)}_{i,a,-r}}\big(T_N(a,k)\big)dk. \tag{4.4.11}$$

Proof. Observe that for N large enough such that $(1-\lambda(a))f_N(a)-r>0$,

$$\begin{aligned}&\int_{-\infty}^{\infty}N\sqrt{\log N}f_{\tau^{(N)}_{i,a,-r}}\big(T_N(a,k)\big)dk\\&=N\,\mathbb{P}\left(\sup_{s>0}(B_i(s)-(1-\lambda(a))\beta s)>(1-\lambda(a))f_N(a)-r\right)\\&=\exp\left(\frac{2(1-\lambda(a))\beta r}{\sigma^2}\right),\end{aligned}$$

due to the fact that $\sup_{s>0}(B_i(s)-(1-\lambda(a))\beta s)$ is exponentially distributed with parameter $2(1-\lambda(a))\beta/\sigma^2$. Additionally,

$$\int_{-\infty}^{\infty}\frac{\beta^2\exp\left(\frac{\beta\left(8a^2\beta^2r-\beta^3k^2\sigma\sqrt{2a\beta+\sigma^2}+8a\beta r\sigma^2+2r\sigma^4\right)}{\sigma\left(2a\beta+\sigma^2\right)^{5/2}}\right)}{\sqrt{\pi}\left(2a\beta+\sigma^2\right)}dk$$

$$= \int_{-\infty}^{\infty} \frac{\beta^2 \exp\left(-\frac{\beta^4 k^2}{(2a\beta+\sigma^2)^2}\right)}{\sqrt{\pi}\,(2a\beta+\sigma^2)} \exp\left(\frac{2(1-\lambda(a))\beta r}{\sigma^2}\right) dk.$$

The first term in this integral is the density of a normally distributed random variable. Therefore, we get that

$$\int_{-\infty}^{\infty} \frac{\beta^2 \exp\left(-\frac{\beta^4 k^2}{(2a\beta+\sigma^2)^2}\right)}{\sqrt{\pi}\,(2a\beta+\sigma^2)} \exp\left(\frac{2(1-\lambda(a))\beta r}{\sigma^2}\right) dk = \exp\left(\frac{2(1-\lambda(a))\beta r}{\sigma^2}\right).$$

□

Lemma 4.5 (Convergence of integrals of sequences of functions)**.** *Assume we have sequences of positive integrable functions $v_N(x)$ and $w_N(x)$ that satisfy the following:*

- $v_N(x) \overset{N\to\infty}{\longrightarrow} v(x)$,
- $\int_{-\infty}^{\infty} v_N(x)dx \overset{N\to\infty}{\longrightarrow} \int_{-\infty}^{\infty} v(x)dx$,
- $w_N(x) \overset{N\to\infty}{\longrightarrow} w(x)$,
- *There exists a constant $c > 0$ such that $w_N(x) < c$ for all x and N.*

Then

$$\int_{-\infty}^{\infty} v_N(x)w_N(x)dx \overset{N\to\infty}{\longrightarrow} \int_{-\infty}^{\infty} v(x)w(x)dx. \tag{4.4.12}$$

Proof. First, by using Fatou's lemma, we obtain that

$$\liminf_{N\to\infty} \int_{-\infty}^{\infty} v_N(x)w_N(x)dx \geq \int_{-\infty}^{\infty} v(x)w(x)dx.$$

Furthermore, observe that $v_N(x)c - v_N(x)w_N(x) > 0$ for all x and N. Now, from Fatou's lemma, it follows that

$$\liminf_{N\to\infty} \int_{-\infty}^{\infty} v_N(x)c - v_N(x)w_N(x)dx \geq \int_{-\infty}^{\infty} v(x)c - v(x)w(x)dx.$$

Because $\int_{-\infty}^{\infty} v_N(x)cdx \overset{N\to\infty}{\longrightarrow} \int_{-\infty}^{\infty} v(x)cdx$, we get that

$$\limsup_{N\to\infty} \int_{-\infty}^{\infty} v_N(x)w_N(x)dx \leq \int_{-\infty}^{\infty} v(x)w(x)dx.$$

The lemma follows. □

In Definition 4.4, we give shorthand notation of some probability measures that we use later on.

Definition 4.4.

$$P_{i,j}^{(N)} := \mathbb{P}\Big(\min(Q_i^\beta(\tau_{i,a^\star,0}^{(N)})\mathbb{1}(\tau_{i,a^\star,0}^{(N)}<\infty), Q_j^\beta(\tau_{j,a^\star,0}^{(N)})\mathbb{1}(\tau_{j,a^\star,0}^{(N)}<\infty)) > f_N(a)\Big), \quad (4.4.13)$$

$$\begin{aligned} &Q_{i,j}^{(N)}(k,l) \\ &\quad := \mathbb{P}\Big(\min(Q_i^\beta(\tau_{i,a^\star,0}^{(N)}), Q_j^\beta(\tau_{j,a^\star,0}^{(N)})) > f_N(a) \;\Big|\; \tau_{i,a^\star,0}^{(N)} = T_N(a^\star,k), \tau_{j,a^\star,0}^{(N)} = T_N(a^\star,l)\Big), \end{aligned} \quad (4.4.14)$$

$$\mathbb{P}^{(k<l)}(A) := \mathbb{P}\Big(A \;\Big|\; \tau_{i,a^\star,0}^{(N)} = T_N(a^\star,k) < \tau_{j,a^\star,0}^{(N)} = T_N(a^\star,l)\Big), \quad (4.4.15)$$

and

$$\mathbb{P}_{i,a,-r,k}^{(N)}(A) := \mathbb{P}(A \mid \tau_{i,a,-r}^{(N)} = T_N(a,k)). \quad (4.4.16)$$

4.5. Proofs of the sharper asymptotics

In this section, we prove sharper asymptotics of the tail behavior of $\mathbb{P}(\bar{Q}_N^\beta > f_N(a))$. Recall the definition of $\tau_{i,a,-r}^{(N)}$ and $\tilde{\tau}_{A,a,r}^{(N)}$ given in Definition 4.3, and observe that

$$\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) = \mathbb{P}(\max_{i\le N} Q_i^\beta(\tau_{i,a,-r}^{(N)} \wedge \tilde{\tau}_{A,a,r}^{(N)})\mathbb{1}(\tau_{i,a,-r}^{(N)} \wedge \tilde{\tau}_{A,a,r}^{(N)} < \infty) > f_N(a)). \quad (4.5.1)$$

This equation is valid, because for $0 < t < \tau_{i,a,-r}^{(N)} \wedge \tilde{\tau}_{A,a,r}^{(N)}$, we see that $B_i(t) - (1-\lambda(a))\beta t < (1-\lambda(a))f_N(a) - r$ and $B_A(t) - \lambda(a)\beta t < \lambda(a)f_N(a) + r$. Thus, $B_i(t) + B_A(t) - \beta t < f_N(a)$. Now, using (4.5.1), we obtain lower and upper bounds of the form

$$\begin{aligned} &\max\left(\mathbb{P}\big(\max_{i\le N} Q_i^\beta(\tau_{i,a,-r}^{(N)})\mathbb{1}(\tau_{i,a,-r}^{(N)}<\infty) > f_N(a)\big), \mathbb{P}\big(\bar{Q}_N^\beta(\tilde{\tau}_{A,a,r}^{(N)})\mathbb{1}(\tilde{\tau}_{A,a,r}^{(N)}<\infty) > f_N(a)\big)\right) \\ &\quad \le \mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \\ &\quad \le \mathbb{P}\big(\max_{i\le N} Q_i^\beta(\tau_{i,a,-r}^{(N)})\mathbb{1}(\tau_{i,a,-r}^{(N)}<\infty) > f_N(a)\big) + \mathbb{P}\big(\bar{Q}_N^\beta(\tilde{\tau}_{A,a,r}^{(N)})\mathbb{1}(\tilde{\tau}_{A,a,r}^{(N)}<\infty) > f_N(a)\big), \end{aligned} \quad (4.5.2)$$

which we can exploit. Other important inequalities that we use are the union bound and Bonferroni's inequality. In the case of identically distributed random variables X_i, these bounds simplify to

$$N\,\mathbb{P}(X_i > x) - \binom{N}{2}\mathbb{P}(\min(X_i,X_j) > x) \le \mathbb{P}(\max_{i\le N} X_i > x) \le N\,\mathbb{P}(X_i > x),$$

which is the case for our problem. Dębicki et al. [47] have derived the tail asymptotics of $\min(Q_i^\beta, Q_j^\beta)$. In Lemma 4.7, we show how we use [47, Thm. 2.3] on the tails of $\min(Q_i^\beta, Q_j^\beta)$ together with Bonferroni's inequality such that these are applicable in our proof of the case

$a > a^\star$.

Now that we can write upper and lower bounds in which hitting times play a role, we condition on the hitting times and get sequences of the form as given in (4.4.3). By using Lemma 4.5, we obtain that

$$\lim_{N\to\infty} \int_{-\infty}^{\infty} \mathbb{P}\left(\sup_{s>\tau^{(N)}} X_i(s) > y_N \Big| \tau^{(N)} = t \right) f_{\tau^{(N)}}(t)dt$$
$$= \int_{-\infty}^{\infty} \lim_{N\to\infty} \mathbb{P}\left(\sup_{s>\tau^{(N)}} X_i(s) > y_N \Big| \tau^{(N)} = t \right) f_{\tau^{(N)}}(t)dt.$$

To obtain limits of the form as given in (4.4.3), we use Lemmas 4.3 and 4.4.

4.5.1 The case $a > a^\star$

In this section, we prove Theorem 4.2 on exact asymptotics of the maximum queue length when $a > a^\star$. As is stated in (4.2.15), $\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \sim N^{-\gamma(a)}$, as $N \to \infty$, when $a > a^\star$. Since the union bound in (4.2.13) gives us that $N^{\gamma(a)}\,\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \le 1$, we only need to show that

$$\liminf_{N\to\infty} N^{\gamma(a)}\,\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \ge 1.$$

In order to prove the lim inf, we first observe that $\bar{Q}_N^\beta > \max_{i\le N} Q_i^\beta(\tau_{i,a^\star,0}^{(N)})\mathbb{1}(\tau_{i,a^\star,0}^{(N)} < \infty)$, and we know by using Bonferroni's inequality that

$$\begin{aligned}
&\mathbb{P}(\max_{i\le N} Q_i^\beta(\tau_{i,a^\star,0}^{(N)})\mathbb{1}(\tau_{i,a^\star,0}^{(N)} < \infty) > f_N(a)) \\
&\quad \ge N\,\mathbb{P}\Big(Q_1^\beta(\tau_{1,a^\star,0}^{(N)})\mathbb{1}(\tau_{1,a^\star,0}^{(N)} < \infty) > f_N(a)\Big) \\
&\qquad - \binom{N}{2}\,\mathbb{P}\Big(\min(Q_1^\beta(\tau_{1,a^\star,0}^{(N)})\mathbb{1}(\tau_{1,a^\star,0}^{(N)} < \infty), Q_2^\beta(\tau_{2,a^\star,0}^{(N)})\mathbb{1}(\tau_{2,a^\star,0}^{(N)} < \infty)) > f_N(a)\Big),
\end{aligned} \tag{4.5.3}$$

where $\tau_{i,a^\star,0}^{(N)}$ and $\tau_{j,a^\star,0}^{(N)}$ are hitting times defined in Lemma 4.4. In Lemma 4.7, we show that the first term is leading, and the second order term is of smaller order. In order to prove this, we first give a convenient upper bound for

$$\mathbb{P}^{(k<l)}\Big(\min(Q_i^\beta(\tau_{i,a^\star,0}^{(N)}), Q_j^\beta(\tau_{j,a^\star,0}^{(N)})) > f_N(a)\Big)$$

in Lemma 4.6, with $\mathbb{P}^{(k<l)}(A)$ given in Equation (4.4.15) in Definition 4.4.

For the remainder of this chapter, let $(\hat{B}(t), t \ge 0)$ be an independent copy of the Brownian motion $(B(t), t \ge 0)$, and $\hat{Q}_i^\beta(s,t)$ an independent copy of $Q_i^\beta(s,t)$.

Lemma 4.6. *Let $a > a^\star$ and $\mathbb{P}^{(k<l)}(A)$ be given in Equation (4.4.15). Furthermore, $\tau_{i,a^\star,0}^{(N)}$ is given in Definition 4.3 and $\hat{Q}_i^\beta$ is an independent copy of Q_i^β. Then for all $\delta > 0$ there*

exists an $N_\delta > 0$ *such that for all* $N \geq N_\delta$

$$\begin{aligned}&\mathbb{P}^{(k<l)}\Big(\min(Q_i^\beta(\tau_{i,a^\star,0}^{(N)}), Q_j^\beta(\tau_{j,a^\star,0}^{(N)})) > f_N(a)\Big)\\ &\leq 4\mathbb{P}^{(k<l)}\bigg((1+\delta)B_A(\tau_{i,a^\star,0}^{(N)}) + \min(\hat{Q}_i^\beta, \hat{Q}_j^\beta) > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau_{i,a^\star,0}^{(N)}\bigg).\end{aligned}$$

Proof. First, we have that

$$\begin{aligned}&\mathbb{P}^{(k<l)}\Big(\min(Q_i^\beta(\tau_{i,a^\star,0}^{(N)}), Q_j^\beta(\tau_{j,a^\star,0}^{(N)})) > f_N(a)\Big)\\ &\leq \mathbb{P}^{(k<l)}\Big(Q_i^\beta(\tau_{i,a^\star,0}^{(N)}, \tau_{j,a^\star,0}^{(N)}) > f_N(a)\Big) + \mathbb{P}^{(k<l)}\Big(\min(Q_i^\beta(\tau_{j,a^\star,0}^{(N)}), Q_j^\beta(\tau_{j,a^\star,0}^{(N)})) > f_N(a)\Big),\end{aligned} \tag{4.5.4}$$

because

$$\min(Q_i^\beta(\tau_{i,a^\star,0}^{(N)}), Q_j^\beta(\tau_{j,a^\star,0}^{(N)})) \;<\; \max(Q_i^\beta(\tau_{i,a^\star,0}^{(N)}, \tau_{j,a^\star,0}^{(N)}), \min(Q_i^\beta(\tau_{j,a^\star,0}^{(N)}), Q_j^\beta(\tau_{j,a^\star,0}^{(N)})))$$

when $\tau_{i,a^\star,0}^{(N)} < \tau_{j,a^\star,0}^{(N)} < \infty$. Now, recall from Definition 4.3 that

$$\begin{aligned}Q_i^\beta(\tau_{i,a^\star,0}^{(N)}, \tau_{j,a^\star,0}^{(N)}) = \sup_{\tau_{i,a^\star,0}^{(N)} < s < \tau_{j,a^\star,0}^{(N)}} &(B_i(s) + B_A(s) - \beta s) \overset{d}{=} (1-\lambda(a^\star))f_N(a^\star)\\ &+ B_A(\tau_{i,a^\star,0}^{(N)}) - \lambda(a^\star)\beta\tau_{i,a^\star,0}^{(N)} + \hat{Q}_i^\beta(0, \tau_{j,a^\star,0}^{(N)} - \tau_{i,a^\star,0}^{(N)}).\end{aligned}$$

Thus, for the first term in (4.5.4) we have

$$\begin{aligned}&\mathbb{P}^{(k<l)}\Big(Q_i^\beta(\tau_{i,a^\star,0}^{(N)}, \tau_{j,a^\star,0}^{(N)}) > f_N(a)\Big)\\ &= \mathbb{P}^{(k<l)}\bigg(B_A(\tau_{i,a^\star,0}^{(N)}) + \hat{Q}_i^\beta(0, \tau_{j,a^\star,0}^{(N)} - \tau_{i,a^\star,0}^{(N)}) > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau_{i,a^\star,0}^{(N)}\bigg)\\ &\leq \mathbb{P}^{(k<l)}\bigg(B_A(\tau_{i,a^\star,0}^{(N)}) + \Big|\hat{B}_i(\tau_{j,a^\star,0}^{(N)} - \tau_{i,a^\star,0}^{(N)}) + \hat{B}_A(\tau_{j,a^\star,0}^{(N)} - \tau_{i,a^\star,0}^{(N)})\Big|\\ &\qquad\qquad > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau_{i,a^\star,0}^{(N)}\bigg).\end{aligned} \tag{4.5.5}$$

For any x and y, it holds that $x + |y| = \max(x+y, x-y)$. Therefore, by the union bound, we can bound the probability in (4.5.5) as

$$\begin{aligned}&\mathbb{P}^{(k<l)}\bigg(B_A(\tau_{i,a^\star,0}^{(N)}) + \Big|\hat{B}_i(\tau_{j,a^\star,0}^{(N)} - \tau_{i,a^\star,0}^{(N)}) + \hat{B}_A(\tau_{j,a^\star,0}^{(N)} - \tau_{i,a^\star,0}^{(N)})\Big|\\ &\qquad\qquad > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau_{i,a^\star,0}^{(N)}\bigg)\end{aligned} \tag{4.5.6}$$

$$\leq 2\mathbb{P}^{(k<l)}\Bigg(B_A(\tau^{(N)}_{i,a^\star,0}) + \hat{B}_i(\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0}) + \hat{B}_A(\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0}) \\ > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{i,a^\star,0}\Bigg) \tag{4.5.7}$$

$$\leq 2\,\mathbb{P}^{(k<l)}\Big((1+\delta)B_A(\tau^{(N)}_{i,a^\star,0}) > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{i,a^\star,0}\Big) \tag{4.5.8}$$

$$\leq 2\mathbb{P}^{(k<l)}\Bigg((1+\delta)B_A(\tau^{(N)}_{i,a^\star,0}) + \min(\hat{Q}^\beta_i, \hat{Q}^\beta_j) \\ > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{i,a^\star,0}\Bigg), \tag{4.5.9}$$

for $\delta > 0$ and $N > N_\delta$. The upper bound in (4.5.8) holds since under the measure $\mathbb{P}^{(k<l)}$ given in (4.4.15), $\tau^{(N)}_{i,a^\star,0} = f_N(a^\star)/\beta + k\sqrt{\log N} \sim \left(\frac{\sigma^2}{2\beta^2} + \frac{a}{\beta}\right)\log N$ as $N \to \infty$, and $\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0} = (l-k)\sqrt{\log N} = O(\sqrt{\log N})$. The upper bound in (4.5.9) holds because we add a positive random variable. For the second term in (4.5.4), first observe that $\mathbb{P}(\min(X,Y) > z) = \mathbb{P}(X > z, Y > z)$. Second, under the assumption that $\tau^{(N)}_{i,a^\star,0} < \tau^{(N)}_{j,a^\star,0} < \infty$, we can write

$$Q^\beta_i(\tau^{(N)}_{j,a^\star,0}) \stackrel{d}{=} (1-\lambda(a^\star))f_N(a^\star) + B_i(\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0}) \\ - (1-\lambda(a^\star))\beta(\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0}) + B_A(\tau^{(N)}_{j,a^\star,0}) - \lambda(a^\star)\beta\tau^{(N)}_{j,a^\star,0} + \hat{Q}^\beta_i.$$

Thus, by applying similar techniques as for the analysis of the first term in (4.5.4), we obtain that

$$\mathbb{P}^{(k<l)}\Big(\min(Q^\beta_i(\tau^{(N)}_{j,a^\star,0}), Q^\beta_j(\tau^{(N)}_{j,a^\star,0})) > f_N(a)\Big) \\ = \mathbb{P}^{(k<l)}\Bigg(B_A(\tau^{(N)}_{j,a^\star,0}) + B_i(\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0}) - (1-\lambda(a^\star))\beta(\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0}) + \hat{Q}^\beta_i \\ > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{j,a^\star,0}, \\ B_A(\tau^{(N)}_{j,a^\star,0}) + \hat{Q}^\beta_j > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{j,a^\star,0}\Bigg).$$

This joint probability satisfies the following bound:

$$\mathbb{P}^{(k<l)}\Bigg(B_A(\tau^{(N)}_{j,a^\star,0}) + B_i(\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0}) - (1-\lambda(a^\star))\beta(\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0}) + \hat{Q}^\beta_i \\ > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{j,a^\star,0}, \\ B_A(\tau^{(N)}_{j,a^\star,0}) + \hat{Q}^\beta_j > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{j,a^\star,0}\Bigg)$$

$$\leq \mathbb{P}^{(k<l)}\Bigg(B_A(\tau^{(N)}_{j,a^\star,0}) + B_i(\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0}) + \hat{Q}^\beta_i \\ > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{j,a^\star,0}, \\ B_A(\tau^{(N)}_{j,a^\star,0}) + \hat{Q}^\beta_j > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{j,a^\star,0}\Bigg).$$

We can bound this further and get

$$\mathbb{P}^{(k<l)}\Bigg(B_A(\tau^{(N)}_{j,a^\star,0}) + B_i(\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0}) + \hat{Q}^\beta_i > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{j,a^\star,0}, \\ B_A(\tau^{(N)}_{j,a^\star,0}) + \hat{Q}^\beta_j > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{j,a^\star,0}\Bigg)$$

$$\leq \mathbb{P}^{(k<l)}\Bigg(B_A(\tau^{(N)}_{j,a^\star,0}) + \max(B_i(\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0}), 0) + \min(\hat{Q}^\beta_i, \hat{Q}^\beta_j) \\ > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{j,a^\star,0}\Bigg)$$

$$\leq 2\mathbb{P}^{(k<l)}\Bigg(B_A(\tau^{(N)}_{j,a^\star,0}) + B_i(\tau^{(N)}_{j,a^\star,0} - \tau^{(N)}_{i,a^\star,0}) + \min(\hat{Q}^\beta_i, \hat{Q}^\beta_j) \\ > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{j,a^\star,0}\Bigg)$$

$$\leq 2\mathbb{P}^{(k<l)}\Bigg((1+\delta)B_A(\tau^{(N)}_{j,a^\star,0}) + \min(\hat{Q}^\beta_i, \hat{Q}^\beta_j) > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{(N)}_{i,a^\star,0}\Bigg).$$

Combining this bound with the bound in (4.5.9) completes the proof of the lemma. □

Lemma 4.7. *For the model given in Definition 4.1 with the additional notation given in Definition 4.2, and $a > a^\star$, we have that*

$$\liminf_{N\to\infty} N^{\gamma(a)}\,\mathbb{P}(\bar{Q}^\beta_N > f_N(a)) \geq 1.$$

The general idea of the proof of Lemma 4.7 is to make rigorous that the lower bound on the maximum queue length $\bar{Q}^\beta_N$ given in (4.5.3) is approximately the same as $N\,\mathbb{P}(Q^\beta_i(\tau^{(N)}_{i,a^\star,0})\mathbb{1}(\tau^{(N)}_{i,a^\star,0} < \infty) > f_N(a))$ when N is large. Thus the last term in (4.5.3) is asymptotically negligible. We use the result from Lemma 4.6 to establish this. Observe now that, following Definition 4.3,

$$Q^\beta_i(\tau^{(N)}_{i,a^\star,0}) \stackrel{d}{=} B_i(\tau^{(N)}_{i,a^\star,0}) + B_A(\tau^{(N)}_{i,a^\star,0}) - \beta\tau^{(N)}_{i,a^\star,0} + \hat{Q}^\beta_i \\ = (1-\lambda(a^\star))f_N(a^\star) + B_A(\tau^{(N)}_{i,a^\star,0}) - \lambda(a^\star)\beta\tau^{(N)}_{i,a^\star,0} + \hat{Q}^\beta_i.$$

Furthermore, observe that due to Equation (4.3.5), $\mathbb{P}(\tau^{(N)}_{i,a^\star,0} < \infty) = 1/N$. From this, it follows that

$$N\,\mathbb{P}(Q^\beta_i(\tau^{(N)}_{i,a^\star,0})\mathbb{1}(\tau^{(N)}_{i,a^\star,0} < \infty) > f_N(a)) = \mathbb{P}(Q^\beta_i(\tau^{(N)}_{i,a^\star,0}) > f_N(a) \mid \tau^{(N)}_{i,a^\star,0} < \infty).$$

Therefore, in order to prove a sharp lower bound on the tail asymptotics of the maximum queue length, we prove by using Fatou's lemma that

$$\liminf_{N\to\infty} N^{\gamma(a)}\,\mathbb{P}(B_A(\tau^{(N)}_{i,a^\star,0})-\lambda(a^\star)\beta\tau^{(N)}_{i,a^\star,0}+\hat{Q}_i^\beta > f_N(a)-(1-\lambda(a^\star))f_N(a^\star) \mid \tau^{(N)}_{i,a^\star,0}<\infty) \geq 1.$$

In order to prove this, we show that $\hat{Q}_i^\beta$ is most likely to hit a level $g_N(a,x,k)$, and $B_A(\tau^{(N)}_{i,a^\star,0})-\lambda(a^\star)\beta\tau^{(N)}_{i,a^\star,0}$ is most likely to hit the level $f_N(a)-(1-\lambda(a^\star))f_N(a^\star)-g_N(a,x,k)$. We define the function $g_N(a,x,k)$ later on.

We now turn to a formal proof of Lemma 4.7.

Proof. Following Equation (4.4.13) in Definition 4.4, we can simplify the inequality in (4.5.3) to

$$\mathbb{P}(\max_{i\leq N} Q_i^\beta(\tau^{(N)}_{i,a^\star,0})\mathbb{1}(\tau^{(N)}_{i,a^\star,0}<\infty) > f_N(a)) \geq NP^{(N)}_{i,i} - \binom{N}{2}P^{(N)}_{i,j}. \tag{4.5.10}$$

Now, before we analyze (4.5.10) in more detail, observe that we can express $\mathbb{P}(\tau^{(N)}_{i,a^\star,0}<\infty,\tau^{(N)}_{j,a^\star,0}<\infty)$ as

$$\mathbb{P}(\tau^{(N)}_{i,a^\star,0}<\infty,\tau^{(N)}_{j,a^\star,0}<\infty) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\tau^{(N)}_{i,a^\star,0}}\big(T_N(a^\star,k)\big)f_{\tau^{(N)}_{j,a^\star,0}}\big(T_N(a^\star,l)\big)\log N dkdl = \frac{1}{N^2}.$$

Then, by using Equation (4.4.14) in Definition 4.4, we get that

$$\begin{aligned} NP^{(N)}_{i,i} &= N\int_{-\infty}^{\infty} f_{\tau^{(N)}_{i,a^\star,0}}\big(T_N(a^\star,k)\big)\sqrt{\log N}Q^{(N)}_{i,i}(k,k)dk \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\tau^{(N)}_{i,a^\star,0}}\big(T_N(a^\star,k)\big)f_{\tau^{(N)}_{j,a^\star,0}}\big(T_N(a^\star,l)\big)N^2\log N Q^{(N)}_{i,i}(k,k)dkdl. \end{aligned}$$

Also, observe that $\binom{N}{2} < N^2/2$, and that

$$\frac{N^2}{2}P^{(N)}_{i,j} = \frac{N^2}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\tau^{(N)}_{i,a^\star,0}}\big(T_N(a^\star,k)\big)f_{\tau^{(N)}_{j,a^\star,0}}\big(T_N(a^\star,l)\big)\log N Q^{(N)}_{i,j}(k,l)dkdl.$$

In conclusion, we can write the inequality in (4.5.10) as

$$\mathbb{P}(\max_{i\leq N} Q_i^\beta(\tau^{(N)}_{i,a^\star,0})\mathbb{1}(\tau^{(N)}_{i,a^\star,0}<\infty) > f_N(a)) \tag{4.5.11}$$

$$\geq \int_{\mathbb{R}^2} f_{\tau^{(N)}_{i,a^\star,0}}\big(T_N(a^\star,k)\big)f_{\tau^{(N)}_{j,a^\star,0}}\big(T_N(a^\star,l)\big)N^2\log N\left(Q^{(N)}_{i,i}(k,k)-\frac{Q^{(N)}_{i,j}(k,l)}{2}\right)dkdl \tag{4.5.12}$$

$$= \int_{\mathbb{R}\times(-\infty,l)} f_{\tau^{(N)}_{i,a^\star,0}}\big(T_N(a^\star,k)\big) f_{\tau^{(N)}_{j,a^\star,0}}\big(T_N(a^\star,l)\big) N^2 \log N \left(Q^{(N)}_{i,i}(k,k) - \frac{Q^{(N)}_{i,j}(k,l)}{2} \right) dkdl$$

$$+ \int_{\mathbb{R}\times[l,\infty)} f_{\tau^{(N)}_{i,a^\star,0}}\big(T_N(a^\star,k)\big) f_{\tau^{(N)}_{j,a^\star,0}}\big(T_N(a^\star,l)\big) N^2 \log N \left(Q^{(N)}_{i,i}(k,k) - \frac{Q^{(N)}_{i,j}(k,l)}{2} \right) dkdl.$$

Since we want to prove a sharp lower bound on the tail asymptotics of the maximum queue length $\bar{Q}^{\beta}_N$ we can use the expression in (4.5.12). We want to prove the convergence of a lower bound of this integral by using Fatou's lemma. Therefore, we focus on the integrand first and prove convergence for the integrand as $N \to \infty$. Assume that $k \leq l$, and observe that $Q^{(N)}_{i,i}(k,k) - Q^{(N)}_{i,j}(k,l)/2 > 0$. Thus,

$$Q^{(N)}_{i,i}(k,k) - \frac{1}{2} Q^{(N)}_{i,j}(k,l) = \left(Q^{(N)}_{i,i}(k,k) - \frac{Q^{(N)}_{i,j}(k,l)}{2} \right)^+.$$

The density of $B_A\big(T_N(a^\star,k)\big)$ equals

$$\frac{\exp\left(-x^2/(2\sigma_A^2 T_N(a^\star,k))\right)}{\sqrt{2\pi}\sigma_A \sqrt{T_N(a^\star,k)}}.$$

We write $a = a^\star + \epsilon$, with $\epsilon > 0$. Let

$$g_N(a,x,k) = f_N(a) - (1-\lambda(a^\star)) f_N(a^\star) + \lambda(a^\star)\beta T_N(a^\star,k) - \frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N - x\sqrt{\log N}.$$

Observe that

$$g_N(a,x,k) + \frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N} = f_N(a) - (1-\lambda(a^\star)) f_N(a^\star) + \lambda(a^\star)\beta T_N(a^\star,k).$$

Furthermore,

$$N^{\gamma(a)} Q^{(N)}_{i,i}(k,k) = N^{\gamma(a)} \mathbb{P}\left(B_A\big(T_N(a^\star,k)\big) + \hat{Q}^{\beta}_i > g_N(a,x,k) + \frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N} \right)$$

$$= \int_{-\infty}^{\infty} N^{\gamma(a)}\, \mathbb{P}(\hat{Q}^{\beta}_i > g_N(a,x,k)) \frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N} \right)^2}{2\sigma_A^2 T_N(a^\star,k)} \right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}} dx.$$

We can simplify this expression further and get with a similar analysis as given in the proof of Lemma 4.4, that

$$\begin{aligned}
&N^{\gamma(a)}\,\mathbb{P}(\hat{Q}_i^\beta > g_N(a,x,k))\frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}\\
&\quad= N^{\gamma(a)}\exp\left(-\frac{2\beta}{\sigma^2+\sigma_A^2}g_N(a,x,k)\right)\frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}\\
&\quad\overset{N\to\infty}{\longrightarrow}\frac{\beta\sigma}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)}\exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\beta k\sigma_A^2\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right).
\end{aligned}$$

Furthermore, following Lemma 4.4, we have that

$$\begin{aligned}
&f_{\tau_{i,a^\star,0}^{(N)}}\left(T_N(a^\star,k)\right)f_{\tau_{j,a^\star,0}^{(N)}}\left(T_N(a^\star,l)\right)N^2\log N\\
&\qquad\overset{N\to\infty}{\longrightarrow}\frac{\beta^2\exp\left(-\frac{\beta^4k^2}{\left(2a^\star\beta+\sigma^2\right)^2}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)}\frac{\beta^2\exp\left(-\frac{\beta^4l^2}{\left(2a^\star\beta+\sigma^2\right)^2}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)}.
\end{aligned}$$

Also, following Lemma 4.6, we have that

$$\begin{aligned}
&Q_{i,j}^{(N)}(k,l)\\
&\leq 4\mathbb{P}\Big((1+\delta)B_A\big(T_N(a^\star,k)\big)+\min(\hat{Q}_i^\beta,\hat{Q}_j^\beta) > f_N(a)-(1-\lambda(a^\star))f_N(a^\star)+\lambda(a^\star)\beta T_N(a^\star,k)\Big),
\end{aligned}$$

for all $\delta>0$ for $N>N_\delta$. Let $0<\delta<\frac{\beta\sigma^4\epsilon}{2\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^2}$ and let

$$\begin{aligned}
&h_N(a,x,k)\\
&=f_N(a)-(1-\lambda(a^\star))f_N(a^\star)+\lambda(a^\star)\beta T_N(a^\star,k)-(1+\delta)\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right).
\end{aligned}$$

From Dębicki et al. [47, Thm. 2.3], we know that

$$\mathbb{P}\Big(\min(\hat{Q}_i^\beta,\hat{Q}_j^\beta)>x\Big)\exp\left(\frac{2\beta}{\sigma^2/2+\sigma_A^2}x\right)\overset{x\to\infty}{\longrightarrow}0. \tag{4.5.13}$$

We have that

$$N^{\gamma(a)} \exp\left(-\frac{2\beta}{\sigma^2/2+\sigma_A^2} h_N(a,x,k)\right) \frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}} \overset{N\to\infty}{\longrightarrow} 0.$$

Thus, when $k \le l$, then

$$\begin{aligned}
&\liminf_{N\to\infty} N^{\gamma(a)} f_{\tau^{(N)}_{i,a^\star,0}}\left(T_N(a^\star,k)\right) f_{\tau^{(N)}_{j,a^\star,0}}\left(T_N(a^\star,l)\right) N^2 \log N \left(Q^{(N)}_{i,i}(k,k) - \frac{Q^{(N)}_{i,j}(k,l)}{2}\right)^+ \\
&\ge \frac{\beta^2 \exp\left(-\frac{\beta^4 k^2}{\left(2a^\star\beta+\sigma^2\right)^2}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)} \frac{\beta^2 \exp\left(-\frac{\beta^4 l^2}{\left(2a^\star\beta+\sigma^2\right)^2}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)} \frac{\beta\sigma \exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\beta k\sigma_A^2\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)}.
\end{aligned}$$

The case $k > l$ can be treated analogously. Finally, we have

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\beta^2 \exp\left(-\frac{\beta^4 k^2}{\left(2a^\star\beta+\sigma^2\right)^2}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)} \frac{\beta^2 \exp\left(-\frac{\beta^4 l^2}{\left(2a^\star\beta+\sigma^2\right)^2}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)} \\
&\qquad \cdot \frac{\beta\sigma \exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\beta k\sigma_A^2\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)} dx\,dk\,dl = 1,
\end{aligned}$$

because this is an integral over the whole domain of a product of three densities of normally distributed random variables. By applying Fatou's lemma, Lemma 4.7 follows. □

Corollary 4.2. *Let $(y_N, N \ge 1)$ be a sequence such that $\liminf_{N\to\infty} y_N/\log N = \infty$, then the tail probability of the steady-state maximum queue length satisfies*

$$\mathbb{P}(\bar{Q}^\beta_N > y_N) \sim N\,\mathbb{P}(Q^\beta_i > y_N),$$

as $N \to \infty$.

Proof. By using the union bound, we have that $\mathbb{P}(\bar{Q}^\beta_N > y_N) \le N\,\mathbb{P}(Q^\beta_i > y_N)$. Furthermore, by using Bonferroni's inequality, we obtain that $\mathbb{P}(\bar{Q}^\beta_N > y_N) \ge N\,\mathbb{P}(Q^\beta_i >$

$y_N) - N^2/2\,\mathbb{P}(Q_i^\beta > y_N, Q_j^\beta > y_N)$. Now, using the limit in (4.5.13), we see that

$$\limsup_{N\to\infty} \frac{N^2/2\,\mathbb{P}(Q_i^\beta > y_N, Q_j^\beta > y_N)}{N\,\mathbb{P}(Q_i^\beta > y_N)} \le \limsup_{N\to\infty} \frac{1}{2} \frac{N \exp\left(-\frac{2\beta}{\sigma^2/2+\sigma_A^2} y_N\right)}{\exp\left(-\frac{2\beta}{\sigma^2+\sigma_A^2} y_N\right)} = 0.$$

The corollary follows. □

4.5.2 The case $a = a^\star$

In Section 4.3, we showed that we have at least two regimes, namely $0 < a < a^\star$, and $a \ge a^\star$. It turns out, that when we investigate sharper asymptotics, the case $a = a^\star$ deserves special attention. In the present section, we establish that in the case $a = a^\star$, $\mathbb{P}(\bar{Q}_N^\beta > f_N(a^\star)) \sim \frac{1}{2} N^{-\gamma(a^\star)}$, thus the prefactor is $1/2$ instead of 1 as in the case $a > a^\star$. To make the heuristics given in Section 4.2 rigorous, we proceed by deriving asymptotic lower and upper bounds, in two separate lemmas. As in Section 4.5.1, we prove that the lim inf converges to the desired limit. We do this in Lemma 4.8. The proof of this Lemma is similar to the proof of Lemma 4.7. However, the simple union bound $N\,\mathbb{P}(Q_i^\beta > f_N(a^\star)) \sim N^{-\gamma(a^\star)}$ is not tight for $a = a^\star$. Thus, we also need to prove that the lim sup is tight. We provide this proof in Lemma 4.9.

Lemma 4.8. *For the model given in Definition 4.1 with the additional notation given in Definition 4.2, and $a = a^\star$, we have that*

$$\liminf_{N\to\infty} N^{\gamma(a^\star)}\,\mathbb{P}(\bar{Q}_N^\beta > f_N(a^\star)) \ge \frac{1}{2}.$$

Proof. First, we have the lower bound

$$\mathbb{P}(\bar{Q}_N^\beta > f_N(a^\star)) \ge \mathbb{P}(\max_{i\le N} Q_i^\beta(\tau_{i,a^\star,r}^{(N)}) \mathbb{1}(\tau_{i,a^\star,r}^{(N)} < \infty) > f_N(a^\star)).$$

As in (4.5.10) we can bound this further by Bonferroni's inequality to

$$\begin{aligned}
&N\,\mathbb{P}\Big(Q_i^\beta(\tau_{i,a^\star,r}^{(N)})\mathbb{1}(\tau_{i,a^\star,r}^{(N)} < \infty) > f_N(a^\star)\Big) \\
&- \binom{N}{2} \mathbb{P}\Big(\min(Q_i^\beta(\tau_{i,a^\star,r}^{(N)})\mathbb{1}(\tau_{i,a^\star,r}^{(N)} < \infty), Q_j^\beta(\tau_{j,a^\star,r}^{(N)})\mathbb{1}(\tau_{j,a^\star,r}^{(N)} < \infty)) > f_N(a^\star)\Big) \\
&\quad \ge \left(N - \frac{N^2}{2}\,\mathbb{P}\Big(\tau_{j,a^\star,r}^{(N)} < \infty\Big)\right) \mathbb{P}\Big(Q_i^\beta(\tau_{i,a^\star,r}^{(N)})\mathbb{1}(\tau_{i,a^\star,r}^{(N)} < \infty) > f_N(a^\star)\Big). \qquad (4.5.14)
\end{aligned}$$

The last step is true because

$$\begin{aligned}
&\mathbb{P}\Big(\min(Q_i^\beta(\tau_{i,a^\star,r}^{(N)})\mathbb{1}(\tau_{i,a^\star,r}^{(N)} < \infty), Q_j^\beta(\tau_{j,a^\star,r}^{(N)})\mathbb{1}(\tau_{j,a^\star,r}^{(N)} < \infty)) > f_N(a^\star)\Big) \\
&\quad = \mathbb{P}\Big(Q_i^\beta(\tau_{i,a^\star,r}^{(N)})\mathbb{1}(\tau_{i,a^\star,r}^{(N)} < \infty) > f_N(a^\star), Q_j^\beta(\tau_{j,a^\star,r}^{(N)})\mathbb{1}(\tau_{j,a^\star,r}^{(N)} < \infty) > f_N(a^\star)\Big) \\
&\quad \le \mathbb{P}\Big(Q_i^\beta(\tau_{i,a^\star,r}^{(N)})\mathbb{1}(\tau_{i,a^\star,r}^{(N)} < \infty) > f_N(a^\star), \tau_{j,a^\star,r}^{(N)} < \infty\Big)
\end{aligned}$$

$$= \mathbb{P}\Big(Q_i^{\beta}(\tau_{i,a^\star,r}^{(N)})\mathbb{1}(\tau_{i,a^\star,r}^{(N)} < \infty) > f_N(a^\star)\Big)\,\mathbb{P}\Big(\tau_{j,a^\star,r}^{(N)} < \infty\Big).$$

Since $\mathbb{P}(\tau_{j,a^\star,r}^{(N)} < \infty) = \exp(-2(1-\lambda(a^\star))\beta r/\sigma^2)/N$, we can simplify the expression in (4.5.14) to

$$\left(1 - \frac{\exp\left(-\frac{2(1-\lambda(a^\star))\beta r}{\sigma^2}\right)}{2}\right) N\,\mathbb{P}\Big(Q_i^{\beta}(\tau_{i,a^\star,r}^{(N)})\mathbb{1}(\tau_{i,a^\star,r}^{(N)} < \infty) > f_N(a^\star)\Big). \tag{4.5.15}$$

Following the same strategy as in the proof of Lemma 4.7, we have that

$$\begin{aligned} &g_N(a^\star,x,k)\\ &\quad= f_N(a^\star) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta T_N(a^\star,k) - \frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N - x\sqrt{\log N}\\ &\quad= \left(-x + \frac{\sigma_A^2\beta k}{\sigma^2+\sigma_A^2}\right)\sqrt{\log N}. \end{aligned}$$

Now, for $x < \sigma_A^2\beta k/(\sigma^2+\sigma_A^2)$, it follows that

$$N^{\gamma(a^\star)}\,\mathbb{P}(\hat{Q}_i^{\beta} > g_N(a^\star,x,k) - r)\frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}$$

$$= N^{\gamma(a^\star)} \cdot \frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)} - \frac{2\beta}{\sigma^2+\sigma_A^2}\left(\left(-x + \frac{\sigma_A^2\beta k}{\sigma^2+\sigma_A^2}\right)\sqrt{\log N} - r\right)\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}. \tag{4.5.16}$$

By using the definition of $T_N(a^\star,k)$ in (4.2.7), we see that $\sqrt{\log N}/(\sqrt{2\pi}\sigma_A T_N(a^\star,k)) \overset{N\to\infty}{\longrightarrow} \beta\sigma/(\sqrt{2\pi}\sigma_A((\sigma^2+\sigma_A^2))$. Furthermore, $\gamma(a^\star)\log N$ plus the exponent on the right-hand side of (4.5.16) equals

$$\gamma(a^\star)\log N - \frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)} - \frac{2\beta}{\sigma^2+\sigma_A^2}\left(\left(-x + \frac{\sigma_A^2\beta k}{\sigma^2+\sigma_A^2}\right)\sqrt{\log N} - r\right)$$

$$\overset{N\to\infty}{\longrightarrow} -\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\beta k\sigma_A^2\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}+\frac{2\beta r}{\sigma^2+\sigma_A^2},$$

with a similar proof as in the proof of Lemma 4.4. Thus,

$$N^{\gamma(a^\star)}\,\mathbb{P}(\hat{Q}_i^\beta > g_N(a^\star,x,k)-r)\frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}$$

$$\overset{N\to\infty}{\longrightarrow} \frac{\beta\sigma}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)}\exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\beta k\sigma_A^2\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right)\exp\left(\frac{2\beta r}{\sigma^2+\sigma_A^2}\right),$$

when $x<\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)$. When $x>\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)$, we see that $g_N(a^\star,x,k)=(-x+\sigma_A^2\beta k/(\sigma^2+\sigma_A^2))\sqrt{\log N}\overset{N\to\infty}{\longrightarrow}-\infty$, thus $\mathbb{P}(\hat{Q}_i^\beta>g_N(a^\star,x,k)-r)\overset{N\to\infty}{\longrightarrow}1$. In this case, we get that

$$\begin{aligned}
&N^{\gamma(a^\star)}\,\mathbb{P}(\hat{Q}_i^\beta > g_N(a^\star,x,k)-r)\frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}\\
&\quad= N^{\gamma(a^\star)}\frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}\\
&\quad= \frac{\sqrt{\log N}}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}\exp\left(-\frac{2\beta\sigma_A^2\log N\left(\sigma_A^2(x-\beta k)+\sigma^2 x\right)+\beta^2\sigma^2x^2\sqrt{\log N}}{\sigma_A^2\left(2\beta^2k\sigma^2+(\sigma^2+\sigma_A^2)^2\sqrt{\log N}\right)}\right)\\
&\quad\overset{N\to\infty}{\longrightarrow} 0,
\end{aligned}$$

for $x>\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)$.

Thus, by combining this result with the result from Lemma 4.4, for $x<\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)$,

$$f_{\tau_{i,a^\star,r}^{(N)}}\left(T_N(a^\star,k)\right)N\sqrt{\log N}N^{\gamma(a^\star)}\,\mathbb{P}\Big(\hat{Q}_i^\beta > g_N(a^\star,x,k)-r\Big)$$

$$
\begin{aligned}
&\cdot \frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}} \\
&\xrightarrow{N\to\infty} \frac{\beta^2\sigma^2\exp\left(-\frac{\beta\left(\beta^3k^2\sigma^4+2r\left(\sigma^2+\sigma_A^2\right)^3\right)}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}\frac{\beta\sigma\exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\beta k\sigma_A^2\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)} \\
&\quad\cdot\exp\left(\frac{2\beta r}{\sigma^2+\sigma_A^2}\right) \\
&=: L_1(x,k).
\end{aligned}
$$

The function $L_1(x,k)$ satisfies

$$
\begin{aligned}
&L_1(x,k) \\
&= \frac{\beta^2\sigma^2\exp\left(-\frac{\beta\left(\beta^3k^2\sigma^4+2r\left(\sigma^2+\sigma_A^2\right)^3\right)}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}\frac{\beta\sigma\exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\sigma_A^2\beta k\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)} \\
&\quad\cdot\exp\left(\frac{2\beta r}{\sigma^2+\sigma_A^2}\right) \\
&= \frac{\beta^2\sigma^2\exp\left(-\frac{\beta^4k^2\sigma^4}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}\frac{\beta\sigma\exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\sigma_A^2\beta k\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)} \\
&= \frac{\beta^2\sigma^2\exp\left(-\frac{\beta^4k^2\sigma^4}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}\frac{\beta\sigma\exp\left(-\frac{\beta^2\sigma^2\left(x-2\sigma_A^2\beta k/\left(\sigma^2+\sigma_A^2\right)\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^2}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)}. \qquad (4.5.17)
\end{aligned}
$$

Thus, $L_1(x,k)$ can be written as a product of two densities of normally distributed random variables. When we consider the last term in (4.5.17) as a function of x, we get that the function

$$
\frac{\beta\sigma\exp\left(-\frac{\beta^2\sigma^2\left(x-2\sigma_A^2\beta k/\left(\sigma^2+\sigma_A^2\right)\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^2}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)}
$$

is the density of a normally distributed random variable with mean $2\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)$ and standard deviation $\sigma_A(\sigma^2+\sigma_A^2)/(\sqrt{2}\beta\sigma)$. From this, it follows that

$$\int_{-\infty}^{\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)} \frac{\beta\sigma \exp\left(-\frac{\beta^2\sigma^2\left(x-2\beta k\sigma_A^2/(\sigma^2+\sigma_A^2)\right)^2}{\sigma_A^2(\sigma^2+\sigma_A^2)^2}\right)}{\sqrt{\pi}\sigma_A(\sigma^2+\sigma_A^2)}dx$$
$$= \mathbb{P}\left(\frac{\sigma_A(\sigma^2+\sigma_A^2)}{\sqrt{2}\beta\sigma}X_1 + \frac{2\sigma_A^2\beta k}{\sigma^2+\sigma_A^2} \leq \frac{\sigma_A^2\beta k}{\sigma^2+\sigma_A^2}\right) = \mathbb{P}\left(\frac{\sigma_A(\sigma^2+\sigma_A^2)}{\sqrt{2}\beta\sigma}X_1 \leq -\frac{\sigma_A^2\beta k}{\sigma^2+\sigma_A^2}\right),$$

with X_1 standard normally distributed. Furthermore, when we consider the first term in (4.5.17) as a function of k, we get that the function

$$\frac{\beta^2\sigma^2 \exp\left(-\frac{\beta^4k^2\sigma^4}{(\sigma^2+\sigma_A^2)^4}\right)}{\sqrt{\pi}(\sigma^2+\sigma_A^2)^2}$$

is the density of a normally distributed random variable with mean 0 and standard deviation $(\sigma^2+\sigma_A^2)^2/(\sqrt{2}\beta^2\sigma^2)$. Therefore, we can conclude that the integral

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)} L_1(x,k)dxdk \\
&= \int_{-\infty}^{\infty} \mathbb{P}\left(\frac{\sigma_A(\sigma^2+\sigma_A^2)}{\sqrt{2}\beta\sigma}X_1 \leq -\frac{\sigma_A^2\beta k}{\sigma^2+\sigma_A^2}\right)\frac{\beta^2\sigma^2\exp\left(-\frac{\beta^4k^2\sigma^4}{(\sigma^2+\sigma_A^2)^4}\right)}{\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}dk \\
&= \mathbb{P}\left(\frac{\sigma_A(\sigma^2+\sigma_A^2)}{\sqrt{2}\beta\sigma}X_1 \leq -\frac{\sigma_A^2\beta}{\sigma^2+\sigma_A^2}\frac{(\sigma^2+\sigma_A^2)^2}{\sqrt{2}\beta^2\sigma^2}X_2\right) \\
&= \frac{1}{2},
\end{aligned}$$

with X_2 standard normally distributed, and X_1 and X_2 mutually independent. Now, by applying Fatou's lemma, we have that

$$\begin{aligned}
\liminf_{N\to\infty} N^{\gamma(a^\star)}N\,\mathbb{P}\Big(Q_i^\beta(\tau_{i,a^\star,r}^{(N)})\mathbb{1}(\tau_{i,a^\star,r}^{(N)}<\infty) > f_N(a^\star)\Big) \\
\geq \int_{-\infty}^{\infty}\int_{-\infty}^{\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)} L_1(x,k)dxdk = \frac{1}{2}.
\end{aligned}$$

Thus, by applying this result to the expression in (4.5.15), we get that

$$\liminf_{N\to\infty} N^{\gamma(a^\star)}\,\mathbb{P}(\bar{Q}_N^\beta > f_N(a^\star)) \geq \frac{1}{2}\left(1 - \frac{\exp\left(-\frac{2(1-\lambda(a^\star))\beta r}{\sigma^2}\right)}{2}\right) \overset{r\to\infty}{\longrightarrow} \frac{1}{2}.$$

□

Lemma 4.9. *For the model given in Definition 4.1 with the additional notation given in*

Definition 4.2, and $a = a^\star$, we have that

$$\limsup_{N\to\infty} N^{\gamma(a^\star)}\,\mathbb{P}(\bar{Q}_N^\beta > f_N(a^\star)) \le \frac{1}{2}.$$

Proof. Let $\tilde{\tau}^{(N)}_{A,a^\star,r} = \inf\{t : B_A(t) - \lambda(a^\star)\beta t > \lambda(a^\star) f_N(a^\star) + r\}$. Following Equation (4.5.1) and the upper bound in (4.5.2), we have that

$$\begin{aligned}\mathbb{P}(\bar{Q}_N^\beta > f_N(a^\star)) \le{}& \mathbb{P}(\bar{Q}_N^\beta(\tilde{\tau}^{(N)}_{A,a^\star,r})\mathbb{1}(\tilde{\tau}^{(N)}_{A,a^\star,r} < \infty) > f_N(a^\star)) \\ &+ \mathbb{P}(\max_{i\le N} Q_i^\beta(\tau^{(N)}_{i,a^\star,-r})\mathbb{1}(\tau^{(N)}_{i,a^\star,-r} < \infty) > f_N(a^\star)). \quad (4.5.18)\end{aligned}$$

Now, observe that we can bound the first term in (4.5.18) as

$$\mathbb{P}(\bar{Q}_N^\beta(\tilde{\tau}^{(N)}_{A,a^\star,r})\mathbb{1}(\tilde{\tau}^{(N)}_{A,a^\star,r} < \infty) > f_N(a^\star)) \le \mathbb{P}(\tilde{\tau}^{(N)}_{A,a^\star,r} < \infty) = N^{-\gamma(a^\star)} \exp\left(-\frac{2\lambda(a^\star)\beta r}{\sigma_A^2}\right). \quad (4.5.19)$$

Furthermore, by using Equation (4.4.16) in Definition 4.4, we can bound the second term in (4.5.18) as

$$\begin{aligned}&N^{\gamma(a^\star)}\,\mathbb{P}(\max_{i\le N} Q_i^\beta(\tau^{(N)}_{i,a^\star,-r})\mathbb{1}(\tau^{(N)}_{i,a^\star,-r} < \infty) > f_N(a^\star)) \\ &\le N^{\gamma(a^\star)} N\,\mathbb{P}(Q_i^\beta(\tau^{(N)}_{i,a^\star,-r})\mathbb{1}(\tau^{(N)}_{i,a^\star,-r} < \infty) > f_N(a^\star)) \\ &= \int_{-\infty}^{\infty} N^{\gamma(a^\star)} N\,\mathbb{P}^{(N)}_{i,a^\star,-r,k}(Q_i^\beta(\tau^{(N)}_{i,a^\star,-r}) > f_N(a^\star)) f_{\tau^{(N)}_{i,a^\star,-r}}(T_N(a^\star,k))\sqrt{\log N}dk. \quad (4.5.20)\end{aligned}$$

Now, we examine the parts of the integrand of this integral separately. First, note that, following Definition 4.3,

$$\begin{aligned}&\mathbb{P}^{(N)}_{i,a^\star,-r,k}(Q_i^\beta(\tau^{(N)}_{i,a^\star,-r}) > f_N(a^\star)) \\ &\qquad = \mathbb{P}^{(N)}_{i,a^\star,-r,k}(B_A(\tau^{(N)}_{i,a^\star,-r}) + \hat{Q}_i^\beta > \lambda(a^\star) f_N(a^\star) + r + \lambda(a^\star)\beta\tau^{(N)}_{i,a^\star,-r}).\end{aligned}$$

We can analyze this probability using Lemma 4.3 by taking $x_N = 2\lambda(a^\star) f_N(a^\star) + \lambda(a^\star)\beta k\sqrt{\log N} + r$, $\eta_N = \sigma_A\sqrt{T_N(a^\star,k)}$, and $\mu = 2\beta/(\sigma^2+\sigma_A^2)$. Write

$$\begin{aligned}\frac{x_N - \mu\eta_N^2}{\sqrt{2}\eta_N} &= \frac{2\lambda(a^\star) f_N(a^\star) + \lambda(a^\star)\beta k\sqrt{\log N} + r - \frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2 T_N(a^\star,k)}{\sqrt{2}\sqrt{\sigma_A^2 T_N(a^\star,k)}} \\ &= \frac{r - \lambda(a^\star)\beta k\sqrt{\log N}}{\sqrt{2}\sqrt{\sigma_A^2 T_N(a^\star,k)}} \overset{N\to\infty}{\longrightarrow} -\frac{\beta^2\sigma\sigma_A k}{(\sigma^2+\sigma_A^2)^2}.\end{aligned}$$

The first term in (4.4.4) of Lemma 4.3 satisfies

$$\frac{\eta_N e^{-\frac{x_N^2}{2\eta_N^2}}}{\sqrt{2\pi}x_N} \sim \frac{\sigma \exp\left(-\frac{\beta\left(\beta^3 k^2\sigma_A^2\sigma^2+2r\left(\sigma^2+\sigma_A^2\right)^3\right)}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{2\sqrt{\pi}\sigma_A}\frac{N^{-\gamma(a^\star)}}{\sqrt{\log N}},$$

and the second term satisfies

$$\frac{1}{2}e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}\left(1+\operatorname{erf}\left(-\frac{\beta^2\sigma\sigma_A k}{(\sigma^2+\sigma_A^2)^2}\right)\right) \sim \frac{1}{2}\exp\left(-\frac{2\beta r}{\sigma^2+\sigma_A^2}\right)\left(1+\operatorname{erf}\left(-\frac{\beta^2\sigma\sigma_A k}{(\sigma^2+\sigma_A^2)^2}\right)\right)N^{-\gamma(a^\star)},$$

as $N\to\infty$. So, we can conclude that

$$\mathbb{P}^{(N)}_{i,a^\star,-r,k}(Q_i^\beta(\tau^{(N)}_{i,a^\star,-r}) > f_N(a^\star)) \sim \frac{1}{2}\exp\left(-\frac{2\beta r}{\sigma^2+\sigma_A^2}\right)\left(1+\operatorname{erf}\left(-\frac{\beta^2\sigma\sigma_A k}{(\sigma^2+\sigma_A^2)^2}\right)\right)N^{-\gamma(a^\star)}$$

as $N\to\infty$. Second, following Lemma 4.4, the density of the hitting time $\tau^{(N)}_{i,a^\star,-r}$ appears in the integrand in (4.5.20), and satisfies

$$\begin{aligned} N f_{\tau^{(N)}_{i,a^\star,-r}}\left(T_N(a^\star,k)\right)\sqrt{\log N} &\overset{N\to\infty}{\longrightarrow} \frac{\beta^2\exp\left(\frac{\beta\left(8a^{\star 2}\beta^2 r-\beta^3k^2\sigma\sqrt{2a^\star\beta+\sigma^2}+8a^\star\beta r\sigma^2+2r\sigma^4\right)}{\sigma\left(2a^\star\beta+\sigma^2\right)^{5/2}}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)} \\ &= \frac{\beta^2\sigma^2\exp\left(\frac{\beta\left(2r\left(\sigma^2+\sigma_A^2\right)^3-\beta^3k^2\sigma^4\right)}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}. \end{aligned}$$

Thus, for the integrand in (4.5.20) we have that

$$\begin{aligned} &N^{\gamma(a^\star)}N\,\mathbb{P}^{(N)}_{i,a^\star,-r,k}(Q_i^\beta(\tau^{(N)}_{i,a^\star,-r}) > f_N(a^\star)) f_{\tau^{(N)}_{i,a^\star,-r}}\left(T_N(a^\star,k)\right)\sqrt{\log N} \\ &\qquad\overset{N\to\infty}{\longrightarrow} \frac{\beta^2\sigma^2\left(1+\operatorname{erf}\left(-\frac{\beta^2\sigma\sigma_A k}{(\sigma^2+\sigma_A^2)^2}\right)\right)\exp\left(\frac{\beta\left(2r\left(\sigma^2+\sigma_A^2\right)^3-\beta^3k^2\sigma^4\right)}{\left(\sigma^2+\sigma_A^2\right)^4}-\frac{2\beta r}{\sigma^2+\sigma_A^2}\right)}{2\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}. \end{aligned}$$

When we integrate this result we get

$$\int_{-\infty}^{\infty} \frac{\beta^2\sigma^2\left(1+\operatorname{erf}\left(-\frac{\beta^2\sigma\sigma_A k}{(\sigma^2+\sigma_A^2)^2}\right)\right)\exp\left(\frac{\beta\left(2r\left(\sigma^2+\sigma_A^2\right)^3-\beta^3k^2\sigma^4\right)}{\left(\sigma^2+\sigma_A^2\right)^4}-\frac{2\beta r}{\sigma^2+\sigma_A^2}\right)}{2\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}dk=\frac{1}{2}.$$

For $a=a^\star$, we have that

$$\sup_{s>0}(B_A(s)-\lambda(a^\star)\beta s)\overset{d}{=}\sup_{s>0}(B_i(s)-(1-\lambda(a^\star))\beta s)\overset{d}{=}\sup_{s>0}(B_i(s)+B_A(s)-\beta s).$$

Thus,

$$\begin{aligned}
&N^{\gamma(a^\star)}\,\mathbb{P}^{(N)}_{i,a^\star,-r,k}(Q_i^\beta(\tau^{(N)}_{i,a^\star,-r})>f_N(a^\star))\\
&\quad=N^{\gamma(a^\star)}\,\mathbb{P}^{(N)}_{i,a^\star,-r,k}\left(\sup_{s>\tau^{(N)}_{i,a^\star,-r}}(B_i(s)+B_A(s)-\beta s)>f_N(a^\star)\right)\\
&\quad=N^{\gamma(a^\star)}\\
&\qquad\cdot\mathbb{P}^{(N)}_{i,a^\star,-r,k}\left(B_A(\tau^{(N)}_{i,a^\star,-r})-\lambda(a^\star)\beta\tau^{(N)}_{i,a^\star,-r}+\sup_{s>0}(\hat{B}_i(s)+\hat{B}_A(s)-\beta s)>\lambda(a^\star)f_N(a^\star)+r\right)\\
&\quad=N^{\gamma(a^\star)}\,\mathbb{P}^{(N)}_{i,a^\star,-r,k}\left(\sup_{s>\tau^{(N)}_{i,a^\star,-r}}(\hat{B}_A(s)-\lambda(a^\star)\beta s)>\lambda(a^\star)f_N(a^\star)+r\right)\\
&\quad\le N^{\gamma(a^\star)}\,\mathbb{P}\left(\sup_{s>0}\left(B_A(s)-\lambda(a^\star)\beta s\right)>\lambda(a^\star)f_N(a^\star)+r\right)\\
&\quad=N^{\gamma(a^\star)}\exp\left(-\frac{2\lambda(a^\star)\beta}{\sigma_A^2}(\lambda(a^\star)f_N(a^\star)+r)\right)=\exp\left(-\frac{2\lambda(a^\star)\beta r}{\sigma_A^2}\right).
\end{aligned}$$

Furthermore, we have that

$$\lim_{N\to\infty}\int_{-\infty}^{\infty}Nf_{\tau^{(N)}_{i,a^\star,-r}}\left(T_N(a^\star,k)\right)\sqrt{\log N}dk=\int_{-\infty}^{\infty}\lim_{N\to\infty}Nf_{\tau^{(N)}_{i,a^\star,-r}}\left(T_N(a^\star,k)\right)\sqrt{\log N}dk.$$

We can use Lemma 4.5 to conclude that

$$\limsup_{N\to\infty}N^{\gamma(a^\star)}N\,\mathbb{P}\Big(Q_i^\beta(\tau^{(N)}_{i,a^\star,-r})\mathbb{1}(\tau^{(N)}_{i,a^\star,-r}<\infty)>f_N(a^\star)\Big)\le\frac{1}{2}. \tag{4.5.21}$$

Now, after combining the bounds in (4.5.19) and (4.5.21),

$$\limsup_{N\to\infty}N^{\gamma(a^\star)}\,\mathbb{P}(\bar{Q}_N^\beta>f_N(a^\star))\le\frac{1}{2}+\exp\left(-\frac{2\lambda(a^\star)\beta r}{\sigma_A^2}\right)\overset{r\to\infty}{\longrightarrow}\frac{1}{2}.$$

□

4.5.3 The case $0 < a < a^\star$

As we have proven the exact asymptotics for the cases $a > a^\star$ and $a = a^\star$ in Theorems 4.2 and 4.3, respectively, we now turn to the proof of Theorem 4.4. In Theorem 4.1 we have shown that $\gamma(a) = \frac{2a\beta+2\sigma^2-2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2}$, thus we expect highly dependent behavior because this indicates that the union upper bound $\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \leq N\,\mathbb{P}(Q_i^\beta > f_N(a))$ is not sharp when $0 < a < a^\star$, as is explained in the proof of Lemma 4.2.

Proof of Theorem 4.4. First, we prove Equation (4.2.17). We write

$$r_N := \frac{\sigma\sqrt{2a\beta+\sigma^2}}{4\beta}\log\log N.$$

Let $\tilde{\tau}_{A,a,r_N}^{(N)} = \inf\{t \geq 0 : B_A(t) - \lambda(a)\beta t > \lambda(a) f_N(a) + r_N\}$. Let $f_{\tilde{\tau}_{A,a,r_N}^{(N)}}$ be its density. Observe that

$$\begin{aligned}
&\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \\
&\quad \geq \mathbb{P}(\bar{Q}_N^\beta(\tilde{\tau}_{A,a,r_N}^{(N)}, \tilde{\tau}_{A,a,r_N}^{(N)})\mathbb{1}(\tilde{\tau}_{A,a,r_N}^{(N)} < \infty) > f_N(a)) \\
&\quad = \int_{-\infty}^{\infty} \mathbb{P}\left(\max_{i\leq N} B_i\big(T_N(a,k)\big) - (1-\lambda(a))\beta T_N(a,k) > (1-\lambda(a)) f_N(a) - r_N\right) \\
&\qquad\qquad \cdot f_{\tilde{\tau}_{A,a,r_N}^{(N)}}\big(T_N(a,k)\big)\sqrt{\log N}dk.
\end{aligned} \tag{4.5.22}$$

As in the proof of Lemma 4.9, we analyze the components of the integrand of (4.5.22) separately. By following a similar derivation as in Lemma 4.4, we see that the hitting-time density $f_{\tilde{\tau}_{A,a,r_N}^{(N)}}\big(T_N(a,k)\big)$ in (4.5.22), with $\tilde{\tau}_{A,a,r_N}^{(N)}$ defined in Definition 4.3 and the hitting-time density given in [32, Eq. (2.0.2), p. 301], satisfies

$$\begin{aligned}
&N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} f_{\tilde{\tau}_{A,a,r_N}^{(N)}}\big(T_N(a,k)\big)\sqrt{\log N} \\
&\quad = N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} \frac{\lambda(a) f_N(a) + r_N}{\sqrt{2\pi}\sigma_A T_N(a,k)^{3/2}} \\
&\qquad \cdot \exp\left(-\frac{(\lambda(a) f_N(a) + r_N + \lambda(a)\beta T_N(a,k))^2}{2\sigma_A^2 T_N(a,k)}\right)\sqrt{\log N} \\
&\quad \xrightarrow{N\to\infty} \frac{\beta^2\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)\exp\left(-\frac{\beta^4 k^2\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)^2}{\sigma_A^2\left(2a\beta+\sigma^2\right)^2}\right)}{\sqrt{\pi}\sigma_A\left(2a\beta+\sigma^2\right)}.
\end{aligned} \tag{4.5.23}$$

Moreover, a result in extreme-value theory states that for

$$b_N = \sqrt{2\log N} - \frac{\log(4\pi\log N)}{2\sqrt{2\log N}},$$

we have that

$$b_N\left(\frac{\max_{i\le N} B_i(d\log N)}{\sigma\sqrt{d\log N}} - b_N\right) \xrightarrow{d} G,$$

as $N\to\infty$, with $G\sim$ Gumbel; see [67, Ex. 1.1.7, p. 11] for a proof. From this, it follows that the term $\mathbb{P}\Big(\max_{i\le N} B_i\big(T_N(a,k)\big) - (1-\lambda(a))\beta T_N(a,k) > (1-\lambda(a))f_N(a) - r_N\Big)$ in (4.5.22) satisfies

$$\mathbb{P}\left(\max_{i\le N} B_i\big(T_N(a,k)\big) - (1-\lambda(a))\beta T_N(a,k) > (1-\lambda(a))f_N(a) - r_N\right) \overset{N\to\infty}{\to} 1 - \exp\left(-\frac{\exp\left(-\frac{\beta^4k^2}{\left(2a\beta+\sigma^2\right)^2}\right)}{2\sqrt{\pi}}\right). \quad (4.5.24)$$

Thus, the product of the limits in (4.5.23) and (4.5.24) gives the tail asymptotics of the integrand in (4.5.22). Now, by applying Fatou's lemma, we obtain a sharper than logarithmic lower bound on the asymptotics for the maximum queue length, and is given in (4.2.17).

In order to prove (4.2.18), we use the upper bound given in (4.5.2) and observe that

$$\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \le \mathbb{P}(\bar{Q}_N^\beta(\tilde{\tau}_{A,a,r_N}^{(N)})\mathbb{1}(\tilde{\tau}_{A,a,r_N}^{(N)} < \infty) > f_N(a)) \quad (4.5.25)$$

$$+ \mathbb{P}(\max_{i\le N} Q_i^\beta(\tau_{i,a,-r_N}^{(N)})\mathbb{1}(\tau_{i,a,-r_N}^{(N)} < \infty) > f_N(a)). \quad (4.5.26)$$

We can bound the expression in (4.5.25) as follows:

$$\mathbb{P}(\bar{Q}_N^\beta(\tilde{\tau}_{A,a,r_N}^{(N)})\mathbb{1}(\tilde{\tau}_{A,a,r_N}^{(N)} < \infty) > f_N(a)) \le \mathbb{P}(\tilde{\tau}_{A,a,r_N}^{(N)} < \infty) = N^{-\gamma(a)}(\log N)^{-\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}. \quad (4.5.27)$$

Therefore,

$$\limsup_{N\to\infty} N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}\ \mathbb{P}(\bar{Q}_N^\beta(\tilde{\tau}_{A,a,r_N}^{(N)})\mathbb{1}(\tilde{\tau}_{A,a,r_N}^{(N)} < \infty) > f_N(a)) \le 1.$$

Thus, because of the bounds given in (4.5.25) and (4.5.26), to prove that (4.2.18) holds, it is left to show that

$$\limsup_{N\to\infty} N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}\ \mathbb{P}(\max_{i\le N} Q_i^\beta(\tau_{i,a,-r_N}^{(N)})\mathbb{1}(\tau_{i,a,-r_N}^{(N)} < \infty) > f_N(a)) < \infty.$$

To prove this, observe that, by using the union bound and by conditioning on the hitting

time $\tau^{(N)}_{i,a,-r_N}$ the expression in (4.5.26) satisfies

$$\begin{aligned}
&\mathbb{P}(\max_{i\leq N} Q_i^{\beta}(\tau^{(N)}_{i,a,-r_N})\mathbb{1}(\tau^{(N)}_{i,a,-r_N}<\infty) > f_N(a)) \\
&\quad \leq N\,\mathbb{P}(Q_i^{\beta}(\tau^{(N)}_{i,a,-r_N})\mathbb{1}(\tau^{(N)}_{i,a,-r_N}<\infty) > f_N(a)) \\
&\quad = \int_{-\infty}^{\infty} N\,\mathbb{P}^{(N)}_{i,a,-r_N,k}(Q_i^{\beta}(\tau^{(N)}_{i,a,-r_N}) > f_N(a)) f_{\tau^{(N)}_{i,a,-r_N}}\big(T_N(a,k)\big)\sqrt{\log N}dk. \qquad (4.5.28)
\end{aligned}$$

Now, we can use Lemma 4.5 to show convergence of the integral in (4.5.28). By following a similar analysis as in Lemma 4.4 and by using the expression of the hitting-time density given in [32, Eq. (2.0.2), p. 301], we have that

$$\begin{aligned}
&N\frac{1}{\sqrt{\log N}}\sqrt{\log N} f_{\tau^{(N)}_{i,a,-r_N}}\big(T_N(a,k)\big) \\
&\quad = N\frac{(1-\lambda(a))f_N(a)-r_N}{\sqrt{2\pi}\sigma T_N(a,k)^{3/2}}\exp\left(-\frac{((1-\lambda(a))f_N(a)-r_N+(1-\lambda(a))\beta T_N(a,k))^2}{2\sigma^2 T_N(a,k)}\right) \\
&\quad \xrightarrow{N\to\infty} \frac{\beta^2\exp\left(-\frac{\beta^4k^2}{(2a\beta+\sigma^2)^2}\right)}{\sqrt{\pi}\,(2a\beta+\sigma^2)}.
\end{aligned}$$

Furthermore,

$$\int_{-\infty}^{\infty}\frac{\beta^2 e^{-\frac{\beta^4k^2}{(2a\beta+\sigma^2)^2}}}{\sqrt{\pi}\,(2a\beta+\sigma^2)}dk = \int_{-\infty}^{\infty}\frac{N}{\sqrt{\log N}}\sqrt{\log N} f_{\tau^{(N)}_{i,a,-r_N}}\big(T_N(a,k)\big)dk = 1. \qquad (4.5.29)$$

Thus, the first and second condition in Lemma 4.5 hold. Thus, we now only need to analyze

$$\begin{aligned}
&\mathbb{P}^{(N)}_{i,a,-r_N,k}(Q_i^{\beta}(\tau^{(N)}_{i,a,-r_N}) > f_N(a)) \\
&\qquad = \mathbb{P}^{(N)}_{i,a,-r_N,k}(B_A(\tau^{(N)}_{i,a,-r_N}) + \hat{Q}_i^{\beta} > \lambda(a)f_N(a) + r_N + \lambda(a)\beta\tau^{(N)}_{i,a,-r_N}), \qquad (4.5.30)
\end{aligned}$$

which is a component in the integrand in (4.5.28). We show that this expression satisfies the third and fourth condition of Lemma 4.5 by proving pointwise convergence and by proving that this probability is uniformly bounded by a constant. To do this, first observe that the random variable in (4.5.30) has the form of the sum of a normally distributed random variable and an exponentially distributed random variable. Hence we can follow the framework of Lemma 4.3 in order to analyze this probability. We take $x_N = 2\lambda(a)f_N(a) + \lambda(a)\beta k\sqrt{\log N} + r_N$, $\eta_N = \sigma_A\sqrt{T_N(a,k)}$, and $\mu = 2\beta/(\sigma^2+\sigma_A^2)$. Now, the expression in (4.5.30) can be written in the form of Equation (4.4.7). Furthermore, observe that

$$\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N} = \frac{2\lambda(a)f_N(a)+\lambda(a)\beta k\sqrt{\log N}+r_N-\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2 T_N(a,k)}{\sqrt{2}\sqrt{\sigma_A^2 T_N(a,k)}} \xrightarrow{N\to\infty} -\infty.$$

Thus, for $0 < a < a^{\star}$, we are in the third situation of Lemma 4.3. Following the same

analysis as in the proof of Lemma 4.4, we see that the first term in (4.4.6) satisfies

$$\frac{\eta_N e^{-\frac{x_N^2}{2\eta_N^2}}}{\sqrt{2\pi}x_N} \sim \frac{\sigma_A \exp\left(-\frac{\beta^4 k^2\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)^2}{\sigma_A^2\left(2a\beta+\sigma^2\right)^2}\right)}{2\sqrt{\pi}\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)}(\log N)^{-\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} N^{-\gamma(a)}\frac{1}{\sqrt{\log N}},$$

as $N \to \infty$. Furthermore, we have for all $t > 0$ that

$$\mathbb{P}\big(B_A(t) - \lambda(a)\beta t > x\big) \le \mathbb{P}\big(B_A(x/(\lambda(a)\beta)) > 2x\big).$$

From this, it follows that the first part in (4.4.7) satisfies

$$\begin{aligned}
&\mathbb{P}(\eta_N X > x_N)\\
&\quad= \mathbb{P}\left(B_A(\tau^{(N)}_{i,a,-r_N}) > \lambda(a)f_N(a) + r_N + \lambda(a)\beta\tau^{(N)}_{i,a,-r_N}\,\middle|\,\tau^{(N)}_{i,a,-r_N} = T_N(a,k)\right)\\
&\quad\le \mathbb{P}\left(B_A(\tau^{(N)}_{i,a,-r_N}) > \lambda(a)f_N(a) + r_N + \lambda(a)\beta\tau^{(N)}_{i,a,-r_N}\,\middle|\,\tau^{(N)}_{i,a,-r_N} = \frac{f_N(a)}{\beta} + \frac{r_N}{\lambda(a)\beta}\right)\\
&\quad\sim \frac{\sigma_A}{2\sqrt{\pi}\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)}(\log N)^{-\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} N^{-\gamma(a)}\frac{1}{\sqrt{\log N}},
\end{aligned}$$

as $N \to \infty$. So there exists an $\epsilon > 0$ and an N_ϵ such that for $N > N_\epsilon$ and all $k > -f_N(a)/(\beta\sqrt{\log N})$,

$$\begin{aligned}
&(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} N^{\gamma(a)}\sqrt{\log N}\\
&\qquad\cdot \mathbb{P}\left(B_A(\tau^{(N)}_{i,a,-r_N}) > \lambda(a)f_N(a) + r_N + \lambda(a)\beta\tau^{(N)}_{i,a,-r_N}\,\middle|\,\tau^{(N)}_{i,a,-r_N} = T_N(a,k)\right) \qquad (4.5.31)\\
&\quad\le \frac{\sigma_A}{2\sqrt{\pi}\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)} + \epsilon.
\end{aligned}$$

The second term in (4.4.6) satisfies

$$
\begin{aligned}
&-\frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)} \frac{\eta_N e^{-\frac{\left(x_N-\mu\eta_N^2\right)^2}{2\eta_N^2}}}{x_N-\mu\eta_N^2} \\
&\quad \sim \frac{\sigma_A\left(\sigma^2+\sigma_A^2\right)\exp\left(-\frac{2\beta^4k^2\left(\sigma^2\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)+a\beta\left(\sqrt{2a\beta+\sigma^2}-2\sigma\right)\right)}{\sigma_A^2\left(2a\beta+\sigma^2\right)^{5/2}}\right)}{2\sqrt{\pi}\sigma\left(\sigma\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)+\sigma_A^2\right)} \\
&\qquad\qquad \cdot (\log N)^{-\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} N^{-\gamma(a)} \frac{1}{\sqrt{\log N}},
\end{aligned}
\tag{4.5.32}
$$

as $N\to\infty$. In this case, first observe that in Equation (4.4.7) the exact expression of the convolution term equals

$$
\int_{-\infty}^{x_N/\eta_N} \mathbb{P}\left(\frac{1}{\mu}E > x_N-\eta_N z\right)\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}dz = \frac{1}{2}\left(\operatorname{erf}\left(\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N}\right)+1\right) e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}.
$$

Second, observe that this can be further rewritten into

$$
\begin{aligned}
&\frac{1}{2}\left(\operatorname{erf}\left(\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N}\right)+1\right) e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)} \\
&= \mathbb{P}^{(N)}_{i,a,-r_N,k}\left(B_A(\tau^{(N)}_{i,a,-r_N}) > \frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2\tau^{(N)}_{i,a,-r_N} - \lambda(a)f_N(a) - r_N - \lambda(a)\beta\tau^{(N)}_{i,a,-r_N}\right) \\
&\quad \cdot \exp\left(\frac{1}{2}\frac{2\beta}{\sigma^2+\sigma_A^2}\left(\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2 T_N(a,k) - 2\lambda(a)f_N(a) - 2\lambda(a)\beta T_N(a,k) - 2r_N\right)\right).
\end{aligned}
$$

Thus, the expression that we are investigating is a product of a tail probability of a Gaussian random variable and an exponential function. As for the first term in (4.4.6), we need to prove that the last condition of Lemma 4.5 holds for the sequence of functions $(J_N(t), N\geq 1, t\geq 0)$ with

$$
\begin{aligned}
&J_N(t) \\
&:= (\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} N^{\gamma(a)}\sqrt{\log N}\,\mathbb{P}\left(B_A(t) > \frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2 t - \lambda(a)f_N(a) - r_N - \lambda(a)\beta t\right) \\
&\quad \cdot \exp\left(\frac{1}{2}\frac{2\beta}{\sigma^2+\sigma_A^2}\left(\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2 t - 2\lambda(a)f_N(a) - 2\lambda(a)\beta t - 2r_N\right)\right).
\end{aligned}
\tag{4.5.33}
$$

In order to order to prove that the sequence $(J_N(t), N\geq 1, t\geq 0)$ is uniformly bounded, we

first observe that for X standard normally distributed and $x > 0$, we have that

$$\mathbb{P}(X > x) \leq \frac{\exp(-x^2/2)}{\sqrt{2\pi}x}; \tag{4.5.34}$$

see [4, Eq. (2.1.1), p. 49]. We call

$$\delta_1 = \frac{\sigma\left(2a\beta+\sigma^2\right)\left(\sigma\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)+\sigma_A^2\right)}{\beta^2\left(\sigma_A^2\left(\sqrt{2a\beta+\sigma^2}+\sigma\right)+\sigma^2\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)\right)}$$

Now, let $0 < \delta < \delta_1$. Then, for $t > T_N(a) - \delta \log N$, the prefactor $1/(\sqrt{2\pi}x)$ in (4.5.34) becomes

$$\frac{\sigma_A\sqrt{t}}{\sqrt{2\pi}(\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2 t - \lambda(a)f_N(a) - r_N - \lambda(a)\beta t)}.$$

This function is decreasing in t for $t \geq T_N(a) - \delta \log N$. Thus,

$$\begin{aligned}
&\limsup_{N\to\infty} \sup_{t\geq T_N(a)-\delta\log N} \sqrt{\log N}\frac{\sigma_A\sqrt{t}}{\sqrt{2\pi}(\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2 t - \lambda(a)f_N(a) - r_N - \lambda(a)\beta t)} \\
&\quad\leq \lim_{N\to\infty} \sqrt{\log N}\frac{\sigma_A\sqrt{T_N(a)-\delta\log N}}{\sqrt{2\pi}(\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2 (T_N(a)-\delta\log N) - \lambda(a)f_N(a) - r_N - \lambda(a)\beta(T_N(a)-\delta\log N))} \\
&\quad\leq C_\delta,
\end{aligned}$$

with $C_\delta > 0$ a constant depending on δ. By using this bound and replacing the term

$$\sqrt{\log N}\,\mathbb{P}\left(B_A(t) > \frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2 t - \lambda(a)f_N(a) - r_N - \lambda(a)\beta t\right)$$

in (4.5.33) with

$$C_\delta \exp\left(-\frac{\left(\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2 t - \lambda(a)f_N(a) - r_N - \lambda(a)\beta t\right)^2}{2\sigma_A^2 t}\right),$$

we get that $t = T_N(a)$ gives an asymptotic upper bound for $J_N(t)$, for $t \geq T_N(a) - \delta \log N$.

For the case that $t < T_N(a) - \delta \log N$, we argue as follows: we can bound $J_N(t)$ as defined in (4.5.33) by

$$\begin{aligned}
J_N(t) \leq (\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} N^{\gamma(a)}\sqrt{\log N} \\
\cdot \exp\left(\frac{1}{2}\frac{2\beta}{\sigma^2+\sigma_A^2}\left(\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2 t - 2\lambda(a)f_N(a) - 2\lambda(a)\beta t - 2r_N\right)\right).
\end{aligned}$$

The upper bound is increasing in t. Thus,

$$\sup_{t\in[0,T_N(a)-\delta\log N]} J_N(t) \le (\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} N^{\gamma(a)}\sqrt{\log N}$$
$$\cdot\exp\left(\frac{1}{2}\frac{2\beta}{\sigma^2+\sigma_A^2}\left(\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2(T_N(a)-\delta\log N)-2\lambda(a)f_N(a)-2\lambda(a)\beta(T_N(a)-\delta\log N)-2r_N\right)\right). \tag{4.5.35}$$

We call

$$\delta_2 = \frac{\sigma\left(\sigma_A^2\sqrt{2a\beta+\sigma^2}+\sigma^2\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)-2a\beta\sigma\right)}{2\beta^2\sigma_A^2}.$$

We obtain that for $t = T_N(a) - \delta_2\log N$, the exponential term in (4.5.33) equals

$$\exp\left(\frac{1}{2}\frac{2\beta}{\sigma^2+\sigma_A^2}\left(\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2(T_N(a)-\delta_2\log N)-2\lambda(a)f_N(a)-2\lambda(a)\beta(T_N(a)-\delta_2\log N)-2r_N\right)\right)$$
$$= N^{-\gamma(a)}\exp\left(-\frac{2\beta}{\sigma^2+\sigma_A^2}r_N\right).$$

Furthermore, we have that $\delta_2 < \delta_1$. Thus, when we take $0 < \delta < \delta_2$, we obtain by the upper bound in (4.5.35) that $\sup_{t\in[0,T_N(a)-\delta\log N]} J_N(t)$ is uniformly bounded by a sequence that converges to 0. Thus, the sequence $(J_N(t), N \ge 1, t \ge 0)$ is uniformly bounded.

Hence, due to the upper bounds for (4.5.31) and (4.5.33), we have that the third and fourth condition of Lemma 4.5 are satisfied. Thus, in the end, we know that

$$(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} N^{\gamma(a)} N\,\mathbb{P}^{(N)}_{i,a,-r_N,k}(Q_i^\beta(\tau^{(N)}_{i,a,-r_N}) > f_N(a)) f_{\tau^{(N)}_{i,a,-r_N}}\left(T_N(a,k)\right)\sqrt{\log N}$$
$$\overset{N\to\infty}{\longrightarrow} \left(\frac{\sigma_A\exp\left(-\frac{\beta^4k^2\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)^2}{\sigma_A^2\left(2a\beta+\sigma^2\right)^2}\right)}{2\sqrt{\pi}\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)}\right.$$
$$\left. + \frac{\sigma_A\left(\sigma^2+\sigma_A^2\right)\exp\left(-\frac{2\beta^4k^2\left(\sigma^2\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)+a\beta\left(\sqrt{2a\beta+\sigma^2}-2\sigma\right)\right)}{\sigma_A^2\left(2a\beta+\sigma^2\right)^{5/2}}\right)}{2\sqrt{\pi}\sigma\left(\sigma\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)+\sigma_A^2\right)}\right) \cdot \frac{\beta^2 e^{-\frac{\beta^4k^2}{\left(2a\beta+\sigma^2\right)^2}}}{\sqrt{\pi}\left(2a\beta+\sigma^2\right)}$$
$$=: L_2(k),$$

and we apply Lemma 4.5 to conclude that (4.2.18) holds. □

Remark 4.1. *We have stated in Theorem 4.4 that we can prove lower and upper bounds that are sharper than logarithmic. However, we do not specify these bounds, but from the*

proof of Theorem 4.4 it becomes clear that

$$\begin{aligned}
&\liminf_{N\to\infty} N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}\,\mathbb{P}(\bar{Q}_N^{\beta} > f_N(a)) \\
&\quad\geq \int_{-\infty}^{\infty} \frac{\beta^2\left(\sigma\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)+2a\beta\right)\exp\left(-\frac{\beta^4k^2\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)^2}{\sigma_A^2\left(2a\beta+\sigma^2\right)^2}\right)}{\sqrt{\pi}\sigma_A\left(2a\beta+\sigma^2\right)^{3/2}} \\
&\qquad\qquad \cdot\left(1-\exp\left(-\frac{\exp\left(-\frac{\beta^4k^2}{\left(2a\beta+\sigma^2\right)^2}\right)}{2\sqrt{\pi}}\right)\right)dk,
\end{aligned}$$

and

$$\limsup_{N\to\infty} N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}\,\mathbb{P}(\bar{Q}_N^{\beta} > f_N(a)) \leq \int_{-\infty}^{\infty} L_2(k)dk + 1.$$

Chapter 5

Heavy-tailed services

5.1. Introduction

In this chapter, we investigate the longest waiting times in the N-server fork-join queue, as we did in Chapter 3 as well. In Chapter 3, our analysis heavily relied on the fact that service times are light-tailed, and we used Cramér-Lundberg theory. We were able to prove convergence results for a general class of light-tailed service times, that is characterized by the properties given in Assumption 3.1. We saw that the resulting longest waiting time scaled with order of magnitude of $\log N$ as $N \to \infty$. In this chapter, we abandon the assumption of light-tailed services, and we will look at a specific class of fork-join queueing systems with heavy-tailed service times. We will see that, in contrast with the results obtained in Chapter 3, the scaling of the longest waiting time will differ when we consider different heavy-tailed service times.

Applications of these heavy-tailed fork-join queues are usually found in parallel computing. Companies such as Google, Microsoft, and Alibaba have data centers with thousands of servers that are available for cloud computing, where there is often a form of parallel scheduling. Jobs in these systems have typically large sizes and are often heavy tailed. However, most literature on parallel queueing theory assumes service times to be light tailed; see the survey [70]. This motivates the analysis of parallel queueing networks with heavy-tailed job sizes.

We assume that service times are mutually dependent between servers, and can be written as a product of two random variables, where one term is independent and identically distributed for all servers, and has a Weibull-like tail, while the other term is the same for all servers and has a regularly varying tail. This describes the situation that if a job has a large size, all the subtasks also have a large size, where the fluctuation is described by the Weibull-like distributed random variable. The reason that we focus on the Weibull distribution is that we can exploit specific properties of this distribution, which we will explain in more

This chapter is based on [136].

detail in Section 5.2.3.

We obtain a convergence result for the rescaled transient longest waiting time $\max_{i\leq N} W_i(tc_N)/c_N$ as $N \to \infty$, after choosing the proper temporal and spatial scaling $(c_N, N \geq 1)$. This longest waiting time converges in distribution to a process which is the supremum of Fréchet-distributed random variables minus a drift term. The temporal and spatial scaling c_N depends on the extreme-value scaling of N independent Weibull-distributed random variables, a slowly varying function, and the index of regular variation. Hence, to obtain this result, a mixture of classic extreme-value theory and analysis of heavy tails is needed. To state our result in more detail, we show that this rescaled longest waiting time process $(\max_{i\leq N} W_i(tc_N)/c_N, t \in [0,T])$ converges as a process in $D[0,T]$ to an extremal process $(\sup_{s\in[0,t]}(X_{(s,t)} - \mu(t-s)), t \in [0,T])$ with Fréchet marginals, with $D[0,T]$ the space of càdlàg functions on $[0,T]$, which we equip with the d° metric [28, Eq. (12.16)], under which $D[0,T]$ is separable and complete. Finally, we prove steady-state convergence of $\max_{i\leq N} W_i(\infty)/c_N$ to $\lim_{t\to\infty} \sup_{s\in[0,t]}(X_{(s,t)} - \mu(t-s))$.

The work in this chapter is connected to the literature on heavy-tailed phenomena; cf. [116] for a summary. Specific results on the interplay between fork-join queues and heavy-tailed services can be found in [129, 156, 157]. In [129, Thm. 2], asymptotic lower and upper bounds for the tail probability of the longest waiting time in steady state are given; however these bounds are not sharp when N is large. In [156] and [157], the authors investigate the fork-join queue with heavy-tailed services under a blocking mechanism. This chapter contributes to the existing literature, as we give sharp convergence results for the longest waiting time with heavy-tailed service times, where the number of servers N grows large.

As mentioned before, the limiting process in this chapter is an extremal process with negative drift. Several papers have been written on extremal processes; see [21, 22, 51, 52, 133]. These extremal processes are, among others, used and applied on the analysis of records in sport, cf. [21, 22]. For example, in [21], a model is used to analyze the times in the mile run.

This chapter is organized as follows. We present our model in Section 5.2 and our main results in Theorems 5.1, 5.2, 5.3, and Proposition 5.1. We give a heuristic analysis of our results in Section 5.2.1. In Section 5.2.2, we present some simulations. In Section 5.2.3, we discuss other modeling choices. In Section 5.3, we present some auxiliary results. We prove process convergence in Section 5.4. We prove our main results in Section 5.6.

The model that we study is very specific. In Section 5.7, we deviate from this model and look at two other N-server parallel-server systems with i.i.d. regularly varying service times, and give extreme-value results of the longest waiting time in steady state.

5.2. Model and main results

In this chapter, we analyze a fork-join queue with a common arrival process, and a service process that consists of a Weibull-like i.i.d. part and a regularly varying part that is the same among all servers. This models the situation that if a job has a large size, then all the subtasks have a large size, with some variability. We show in Section 5.2.1 that the Weibull

distribution has convenient properties that we exploit in this chapter; in Section 5.2.3 we briefly discuss what happens when the i.i.d. part of the service process has a lighter tail. We write the random variable $S_i^{(1)}(j)S^{(2)}(j)$ as the representation of a service time at server i of the subtask of the j-th job, while the random variable $A(j)$ is the interarrival time between the j-th and $(j+1)$-st job. Now, by Lindley's recursion, the waiting time at server i upon arrival of the $(n+1)$-st job equals

$$W_i(n) = \sup_{0\le k\le n} \sum_{j=k+1}^{n} (S_i^{(1)}(j)S^{(2)}(j) - A(j)), \tag{5.2.1}$$

with $W_i(0) = 0$. Moreover, we write $W_i(t) = W_i(\lfloor t \rfloor)$. Furthermore, the maximum of the N waiting times equals

$$\max_{i\le N} W_i(n) = \max_{i\le N} \sup_{0\le k\le n} \sum_{j=k+1}^{n} (S_i^{(1)}(j)S^{(2)}(j) - A(j)). \tag{5.2.2}$$

We assume that the sequence of random variables $(S^{(2)}(j), j \ge 1)$ are independent random variables that satisfy

$$\mathbb{P}(S^{(2)}(j) > x) = \frac{\ell(x)}{x^\beta}, \tag{5.2.3}$$

with $\ell(x)$ a slowly varying function and $\beta > 1$. A positive function ℓ is slowly varying if and only if $\lim_{x\to\infty} \ell(ax)/\ell(x) = 1$ for all $a > 0$ [116, Def. 2.6]. We let the i.i.d. random variables $(S_i^{(1)}(j), i \ge 1, j \ge 1)$ satisfy

$$\log \mathbb{P}(S_i^{(1)}(j) > x) \sim -qx^\alpha, \tag{5.2.4}$$

as $x \to \infty$, with $0 < \alpha < 1$ and $q > 0$. Thus, the random variable $S_i^{(1)}(j)$ has the same logarithmic tail asymptotics as the Weibull distribution. Let $b_N = \left(\log N/q\right)^{1/\alpha}$. Then, we know from standard extreme-value theory [67, Thm. 5.4.1, p. 188] that

$$\frac{\max_{i\le N} S_i^{(1)}}{b_N} \xrightarrow{\mathbb{P}} 1, \tag{5.2.5}$$

as $N \to \infty$. Thus, the number b_N indicates the approximate size of the largest of N independent Weibull-distributed random variables. Furthermore, we have independent and identically distributed random variables $(A(j), j \ge 1)$, such that

$$\mathbb{E}[S_i^{(1)}(j)S^{(2)}(j) - A(j)] = -\mu, \tag{5.2.6}$$

with $\mu > 0$.

In this chapter, we prove process convergence of the scaled longest waiting time over N servers in Theorem 5.2; cf. Theorem 2.1 in Chapter 2 for a similar result for fork-join queues with light-tailed services. In order to achieve this result, we need to scale the number of arriving jobs and the longest waiting time with a sequence $(c_N, N \ge 1)$, where the sequence

$(c_N, N \geq 1)$ satisfies

$$c_N \sim \frac{(c_N/b_N)^\beta}{\ell(c_N/b_N)}, \tag{5.2.7}$$

as $N \to \infty$, with $c_N/b_N \overset{N\to\infty}{\longrightarrow} \infty$ and $f(x) \sim g(x)$ as $x \to \infty$ meaning that $\lim_{x\to\infty} f(x)/g(x) = 1$. We explain in Section 5.2.1 in more detail why this sequence scales as given in (5.2.7). Following standard arguments on generalized inverses of regularly varying functions; see [132, Prop. 2.6 (v,vi,vii)] and [29, Thm. 1.5.12], we can solve the right-hand side of (5.2.7) and get that $c_N/b_N \sim c_N^{1/\beta}/\hat{\ell}(c_N)$, with $\hat{\ell}$ a slowly varying function. From this, it follows that $b_N \sim \hat{\ell}(c_N)c_N^{(\beta-1)/\beta}$. Now, we define the sequence $(c_N, N \geq 1)$ as

$$c_N := \tilde{\ell}(b_N)b_N^{\beta/(\beta-1)} \tag{5.2.8}$$

where $\tilde{\ell}$ is a slowly varying function that equals

$$\tilde{\ell}(x)x^{\beta/(\beta-1)} = (((x/((x^\beta/\ell(x))^\leftarrow))^*)^\leftarrow)^* \tag{5.2.9}$$

with $H(y)^\leftarrow = \inf\{s : H(s) \geq y\}$ and $f(x)^*$ a monotone function with the property that $f(x)^* \sim f(x)$ as $x \to \infty$. Thus, the sequence $(c_N, N \geq 1)$ satisfies the relation described in (5.2.7). More precise properties of the function $\tilde{\ell}$ are given in Lemma 5.1.

As we have a proper scaling of the number of arriving jobs and the longest waiting time by a sequence $(c_N, N \geq 1)$, the scaled longest waiting time has the form

$$\frac{\max_{i\leq N} W_i(tc_N)}{c_N} = \sup_{s\in[0,t]} \frac{\max_{i\leq N} \sum_{j=\lfloor sc_N\rfloor+1}^{\lfloor tc_N\rfloor} (S_i^{(1)}(j)S^{(2)}(j) - A(j))}{c_N}. \tag{5.2.10}$$

Notice that

$$\sup_{s\in[0,t]} \frac{\max_{i\leq N} \sum_{j=\lfloor sc_N\rfloor+1}^{\lfloor tc_N\rfloor} (S_i^{(1)}(j)S^{(2)}(j) - A(j))}{c_N} \overset{d}{=} \sup_{s\in[0,t]} \frac{\max_{i\leq N} \sum_{j=1}^{\lfloor sc_N\rfloor} (S_i^{(1)}(j)S^{(2)}(j) - A(j))}{c_N}. \tag{5.2.11}$$

Thus, to prove convergence of a single random variable $\max_{i\leq N} W_i(tc_N)/c_N$, it suffices to prove convergence of the right-hand side in Equation (5.2.11). However, the processes

$$\left(\sup_{s\in[0,t]} \frac{\max_{i\leq N} \sum_{j=1}^{\lfloor sc_N\rfloor} (S_i^{(1)}(j)S^{(2)}(j) - A(j))}{c_N}, t \in [0,T] \right) \tag{5.2.12}$$

and

$$\left(\frac{\max_{i\leq N} W_i(tc_N)}{c_N}, t \in [0,T] \right) \tag{5.2.13}$$

are not equal in distribution. For instance, the process in (5.2.12), which we will refer to as the *auxiliary process*, is non-decreasing in t and the longest waiting time process in (5.2.13) is not non-decreasing in t. In Theorem 5.1, we show that this auxiliary process converges in distribution to a limiting process;

$$\left(\frac{\max_{i \leq N} \sup_{s \in [0,t]} \sum_{j=1}^{\lfloor sc_N \rfloor} (S_i^{(1)}(j) S^{(2)}(j) - A(j))}{c_N}, t \in [0,T] \right) \xrightarrow{d} \left(\sup_{s \in [0,t]} (X_s - \mu s), t \in [0,T] \right),$$

as $N \to \infty$. The process $(X_t, t \in [0,T])$ is a stochastic process with Fréchet-distributed marginals. This process has cumulative distribution function $\mathbb{P}(X_t \leq x) = \exp(-t/x^\beta)$ for $x > 0$. Furthermore, $X_{t+s} = \max(X_t, \hat{X}_s)$, where $\hat{X}_s$ is an independent copy of X_s, because $\mathbb{P}(X_{t+s} < x) = \mathbb{P}(X_t < x)\,\mathbb{P}(\hat{X}_s < x) = \exp(-t/x^\beta)\exp(-s/x^\beta) = \exp(-(t+s)/x^\beta)$. Thus, the process $(X_t, t \in [0,T])$ is a function in $D[0,T]$ and is called an extremal process [133]. It is easy to see that $(\sup_{s \in [0,t]}(X_s - \mu s), t \in [0,T])$ is also non-decreasing in t. The limiting process of $(\max_{i \leq N} W_i(tc_N)/c_N, t \in [0,T])$ has the same marginals as the process $(\sup_{s \in [0,t]}(X_s - \mu s), t \in [0,T])$, but is not non-decreasing. We write the limiting process of the longest waiting time as $(\sup_{s \in [0,t]}(X_{(s,t)} - \mu(t-s)), t \in [0,T])$, with $X_{(s,t)} \stackrel{d}{=} X_{t-s}$. For $r < s < t$, we have that $X_{(r,t)} = \max(X_{(r,s)}, X_{(s,t)})$, and we have that $X_{(s,t)}$ and $X_{(u,v)}$ are independent if and only if the intervals (s,t) and (u,v) are disjoint. We write $X_t := X_{(0,t)}$. In conclusion, the random variable X_t involves a single time parameter, while the random variable $X_{(s,t)}$ is defined by two time parameters, which complicates the proof. There is a clear connection between the stochastic processes however, and in this chapter, we first prove convergence of the non-decreasing process $(\max_{i \leq N} \sup_{s \in [0,t]} \sum_{j=1}^{\lfloor sc_N \rfloor} (S_i^{(1)}(j) S^{(2)}(j) - A(j))/c_N, t \in [0,T])$ and we use this result with some additional steps to prove process convergence of the scaled longest waiting time $(\max_{i \leq N} W_i(tc_N)/c_N, t \in [0,T])$.

Definition 5.1. *We write*

$$R_i(k) := \sum_{j=1}^{k} (S_i^{(1)}(j) S^{(2)}(j) - A(j)), \tag{5.2.14}$$

$$R_i(l,k) := \sum_{j=l}^{k} (S_i^{(1)}(j) S^{(2)}(j) - A(j)), \tag{5.2.15}$$

with $R_i(s,t) = R_i(\lfloor s \rfloor, \lfloor t \rfloor)$*, and*

$$\tilde{W}_i(n) := \sup_{0 \leq k \leq n} R_i(k), \tag{5.2.16}$$

with $\tilde{W}_i(t) = \tilde{W}_i(\lfloor t \rfloor)$.

To summarize the model, we have that the waiting time of subtasks in front of the i-th server is given in Equation (5.2.1), the i.i.d. random variables $(S_i^{(1)}(j), i \geq 1, j \geq 1)$ satisfy (5.2.4), and the i.i.d. random variables $(S^{(2)}(j), j \geq 1)$ satisfy (5.2.3), with $\ell(x)$ a slowly varying function. The slowly varying function satisfies (5.2.9). Furthermore, the sequence of i.i.d. random variables $(A(j), j \geq 1)$ satisfies (5.2.6). Additionally, all random variables $S_i^{(1)}(j_1)$, $S^{(2)}(j_2)$, and $A(j_3)$ are mutually independent.

Moreover, we have a scaling sequence $(b_N, N \geq 1)$ with $b_N = (\log N/q)^{1/\alpha}$, and we have a scaling sequence $(c_N, N \geq 1)$ that satisfies (5.2.7) and (5.2.8).

Finally, we have a limiting process $(X_{(s,t)}, t \in [0,T])$ with Fréchet-distributed marginals. For $r < s < t$, we have that $X_{(r,t)} = \max(X_{(r,s)}, X_{(s,t)})$, and we have that $X_{(s,t)}$ and $X_{(u,v)}$ are independent if and only if the intervals (s,t) and (u,v) are disjoint. We write $X_t := X_{(0,t)}$ and we have that $X_{(s,t)} \stackrel{d}{=} X_{t-s}$. Furthermore, $\mathbb{P}(X_t \leq x) := \exp(-t/x^\beta)$ for $x > 0$.

Now, we give the auxiliary and main results in Theorems 5.1–5.3 and in Proposition 5.1.

Theorem 5.1. *For the sequence of random variables* $(\tilde{W}_i(n), i \geq 1, n \geq 1)$ *given in* (5.2.16)*, we have that*

$$\left(\frac{\max_{i\leq N} \tilde{W}_i(tc_N)}{c_N}, t \in [0,T]\right) \xrightarrow{d} \left(\sup_{s\in[0,t]} (X_s - \mu s), t \in [0,T]\right) \tag{5.2.17}$$

as $N \to \infty$.

The main result proven in this chapter is given in Theorem 5.2.

Theorem 5.2. *We have for all* $T > 0$ *that*

$$\left(\frac{\max_{i\leq N} W_i(tc_N)}{c_N}, t \in [0,T]\right) \xrightarrow{d} \left(\sup_{s\in[0,t]} (X_{(s,t)} - \mu(t-s)), t \in [0,T]\right) \tag{5.2.18}$$

as $N \to \infty$.

When $t \to \infty$ in (5.2.18), we expect that the longest steady-state waiting time satisfies $\mathbb{P}\big(\max_{i\leq N} W_i(\infty) > xc_N\big) \stackrel{N\to\infty}{\longrightarrow} \mathbb{P}\big(\sup_{t>0}(X_t - \mu t) > x\big)$. Though this does not trivially follow from Theorem 5.2, it is indeed true, and we prove this in Theorem 5.3.

Theorem 5.3. *The longest steady-state waiting time satisfies*

$$\mathbb{P}\left(\max_{i\leq N} W_i(\infty) > xc_N\right) \stackrel{N\to\infty}{\longrightarrow} \mathbb{P}\left(\sup_{t>0}(X_t - \mu t) > x\right). \tag{5.2.19}$$

We can write the limiting probabilities explicitly.

Proposition 5.1. *We have that*

$$\mathbb{P}\left(\sup_{t>0}(X_t - \mu t) > x\right) = 1 - \exp\left(-\frac{1}{\mu(\beta-1)x^{\beta-1}}\right), \tag{5.2.20}$$

and

$$\mathbb{P}\left(\sup_{s\in[0,t]}(X_{(s,t)}-\mu(t-s))>x\right)$$
$$=1-\exp\left(-\frac{1}{\mu^{\beta}(\beta-1)}\left(\frac{1}{(x/\mu)^{\beta-1}}-\frac{1}{(x/\mu+t)^{\beta-1}}\right)\right). \quad (5.2.21)$$

5.2.1 Main ideas for the proofs

To prove Theorem 5.2 directly is challenging, since the limiting random variable $X_{(s,t)}$ depends on two parameters and cannot be written as a difference of the form $Y_t - Y_s$, as is the case in standard queueing theory. However, the marginal distributions of $X_{(s,t)}$ and X_{t-s} are the same. Thus, we first prove Theorem 5.1, after which we prove Theorem 5.2 using some auxiliary results on bounds on tail probabilities, convergence rates of sums of Weibull-distributed random variables, and auxiliary results on process convergence in $D[0,T]$; see Section 5.3. To get a better understanding of the convergence result in Theorem 5.1, it benefits to first examine the process

$$\left(\frac{\max_{i\le N} R_i(tc_N)}{c_N}, t\in[0,T]\right), \quad (5.2.22)$$

so we remove the supremum term from the expression on the left-hand side of (5.2.17) and we are left with a maximum of N random walks. We can however apply the continuous mapping theorem on this stochastic process and obtain the result in Theorem 5.1 because the supremum is a continuous functional; see [154, Sec. 6]. Obviously, the law of large numbers implies that

$$\frac{R_i(tc_N)}{c_N} \xrightarrow{\mathbb{P}} -\mu t, \quad (5.2.23)$$

as $N\to\infty$. However, when we investigate the largest of N of these random variables, we obtain that

$$\frac{\max_{i\le N} R_i(tc_N)}{c_N} \xrightarrow{d} X_t - \mu t, \quad (5.2.24)$$

as $N\to\infty$. The fact that we see this limiting behavior has two main reasons; first, a standard result is that for i.i.d. regularly varying $(S^{(2)}(j), j\ge 1)$, the tail behavior of a finite sum is the same as the tail behavior of the largest regularly varying random variable. Second, for Weibull-distributed random variables and a deterministic sequence $(b_j, j\ge 1)$, we have that $\max_{i\le N}\sum_{j=1}^{n} A_{i,j}b_j/b_N \xrightarrow{\mathbb{P}} \max_{j\le n} b_j$, as $N\to\infty$, which follows from Lemma 2.13. Therefore,

$$\max_{i\le N}\sum_{j=1}^{n} S_i^{(1)}(j)S^{(2)}(j) \approx \max_{i\le N} S_i^{(1)}\cdot\max_{j\le n} S^{(2)}(j) + \mathbb{E}[S_i^{(1)}(j)S^{(2)}(j)](n-1)$$

for N large. Thus, we can conclude that for N large,

$$\frac{\max_{i\le N} R_i(tc_N)}{c_N} \approx \frac{\max_{i\le N} S_i^{(1)}}{b_N}\frac{\max_{j\le\lfloor tc_N\rfloor} S^{(2)}(j)}{c_N/b_N} + \frac{\sum_{j=1}^{\lfloor tc_N\rfloor}(S_i^{(1)}(j)S^{(2)}(j) - A(j))}{c_N} \tag{5.2.25}$$

$$\approx \frac{\max_{i\le N} S_i^{(1)}}{b_N}\frac{\max_{j\le\lfloor tc_N\rfloor} S^{(2)}(j)}{c_N/b_N} - \mu t \tag{5.2.26}$$

$$\approx \frac{\max_{j\le\lfloor tc_N\rfloor} S^{(2)}(j)}{c_N/b_N} - \mu t. \tag{5.2.27}$$

We see that the largest regularly varying random variable $\max_{j\le\lfloor tc_N\rfloor} S^{(2)}(j)$ determines the stochastic part in the limit, and is of order c_N/b_N. Now, it is easy to see that

$$\mathbb{P}\left(\max_{j\le\lfloor tc_N\rfloor} S^{(2)}(j) \le (x+\mu t)\frac{c_N}{b_N}\right) = \mathbb{P}\left(S^{(2)}(j) \le (x+\mu t)\frac{c_N}{b_N}\right)^{\lfloor tc_N\rfloor}$$
$$\sim \left(1 - \frac{\ell((x+\mu t)c_N/b_N)}{\big((x+\mu t)c_N/b_N\big)^\beta}\right)^{\lfloor tc_N\rfloor}.$$

Because we have defined c_N as having the relation $c_N \sim (c_N/b_N)^\beta/\ell(c_N/b_N)$ as $N\to\infty$, we get that

$$\left(1 - \frac{\ell((x+\mu t)c_N/b_N)}{\big((x+\mu t)c_N/b_N\big)^\beta}\right)^{\lfloor tc_N\rfloor} \sim \left(1 - \frac{1}{(x+\mu t)^\beta c_N}\right)^{\lfloor tc_N\rfloor}$$
$$\overset{N\to\infty}{\longrightarrow} \exp\left(-\frac{t}{(x+\mu t)^\beta}\right).$$

In conclusion, the limiting distribution of $\max_{i\le N} R_i(tc_N)/c_N$ is a Fréchet-distributed random variable with a negative drift term. We also see that we can approximate $\max_{i\le N} R_i(tc_N)/c_N$ with $\max_{j\le\lfloor tc_N\rfloor} S^{(2)}(j)/(c_N/b_N) - \mu t$ as N is large. This approximating process has convenient properties since the stochastic term is non-decreasing in t. Therefore, to prove process convergence of $\big(\max_{i\le N} R_i(tc_N)/c_N, t\in[0,T]\big)$ to $(X_t-\mu t, t\in[0,T])$, we first prove that $(\max_{j\le\lfloor tc_N\rfloor} S^{(2)}(j)/(c_N/b_N) - \mu t, t\in[0,T])$ converges to $(X_t-\mu t, t\in[0,T])$. Furthermore, we prove in Lemma 5.9 that for all $\epsilon>0$,

$$\mathbb{P}\left(\sup_{t\in[0,T]}\left|\frac{\max_{i\le N} R_i(tc_N)}{c_N} - \left(\frac{\max_{j\le\lfloor tc_N\rfloor} S^{(2)}(j)}{c_N/b_N} - \mu t\right)\right| > \epsilon\right) \overset{N\to\infty}{\longrightarrow} 0.$$

After applying the triangle inequality, we obtain that

$$\left(\frac{\max_{i\leq N} R_i(tc_N)}{c_N}, t\in[0,T]\right) \xrightarrow{d} \left(X_t-\mu t, t\in[0,T]\right),$$

as $N\to\infty$. Now, by applying the continuous mapping theorem, we obtain the result of Theorem 5.1. This is still an auxiliary result because the process on the left-hand side of (5.2.17) is not the longest waiting time process. We can however prove the process convergence of the longest waiting time in Theorem 5.2 by using some additional results, as the marginals of the processes on the left side of the limit in Theorems 5.1 and 5.2 are the same, and the marginals of the limiting processes in Theorems 5.1 and 5.2 are the same. Thus, we already know that pointwise convergence holds.

In order to prove the convergence of the finite-dimensional distributions for the longest waiting time process, we show that we can decompose the joint probabilities of both the longest waiting time process and the limiting process into an operation of marginal probabilities, and thus, the convergence of finite-dimensional distributions follows from pointwise convergence. For example, for $x_2+\mu t_2 > x_1+\mu t_1$,

$$\begin{aligned}&\mathbb{P}\left(\sup_{s\in[0,t_1]}(X_{(s,t_1)}-\mu(t_1-s))<x_1,\ \sup_{s\in[0,t_2]}(X_{(s,t_2)}-\mu(t_2-s))<x_2\right)\\&=\frac{\mathbb{P}\left(\sup_{s\in[0,t_1]}(X_{(s,t_1)}-\mu(t_1-s))<x_1\right)}{\mathbb{P}\left(\sup_{s\in[0,t_1]}(X_{(s,t_1)}-\mu(t_1-s))<x_2+\mu(t_2-t_1)\right)}\,\mathbb{P}\left(\sup_{s\in[0,t_2]}(X_{(s,t_2)}-\mu(t_2-s))<x_2\right).\end{aligned}$$

An analogous equation holds for the process

$$\left(\sup_{s\in[0,t]}\left(\max_{\lfloor sc_N\rfloor\leq j\leq\lfloor tc_N\rfloor}\frac{S^{(2)}(j)}{(c_N/b_N)}-\mu(t-s)\right), t\in[0,T]\right).$$

Now, as an abbreviation, we write

$$\tilde{S}(s,t):=\max_{\lfloor s\rfloor\leq j\leq\lfloor t\rfloor} S^{(2)}(j) \tag{5.2.28}$$

and

$$\tilde{S}(t):=\max_{j\leq\lfloor t\rfloor} S^{(2)}(j). \tag{5.2.29}$$

In Lemma 5.10, we prove that the longest waiting time in (5.2.10) satisfies

$$\mathbb{P}\left(\sup_{t\in[0,T]}\left|\sup_{s\in[0,t]}\frac{\max_{i\leq N}R_i(sc_N,tc_N)}{c_N}-\sup_{s\in[0,t]}\left(\frac{\tilde{S}(sc_N,tc_N)}{c_N/b_N}-\mu(t-s)\right)\right|>\epsilon\right)\xrightarrow{N\to\infty}0,$$

by using similar techniques as in Lemma 5.9. Finally, we show in the proof of Theorem 5.2

that

$$\left(\sup_{s\in[0,t]} \left(\frac{\tilde{S}(sc_N, tc_N)}{(c_N/b_N)} - \mu(t-s) \right), t \in [0,T] \right) \xrightarrow{d} \left(\sup_{s\in[0,t]} (X_{(s,t)} - \mu(t-s)), t \in [0,T] \right),$$

as $N \to \infty$, by using the earlier results together with [28, Thm. 13.3].

In summary, we prove process convergence of the longest waiting time through three steps; first, pointwise convergence follows from Theorem 5.1; second, we show in Lemma 5.10 that the longest waiting process is asymptotically equivalent to an extremal process that only depends on the regularly varying random variables, and finally, we prove process convergence for this latter process in Theorem 5.2.

In Section 5.6, we show that the cumulative distribution function of the limiting longest steady-state waiting time converges to $\mathbb{P}(\sup_{t>0}(X_t - \mu t) < x)$. This means that the limiting cumulative distribution function of the longest steady-state waiting time is the same as $\lim_{T\to\infty} \mathbb{P}(\sup_{s\in[0,T]}(X_{(s,T)} - \mu(T-s)) < x)$, thus the steady-state behavior of the limiting process of $(\sup_{s\in[0,t]}(X_{(s,t)} - \mu(t-s)), t \in [0,T])$ is the same as the extreme-value limit of the longest steady-state waiting time, which is not a trivial result.

5.2.2 Numerical examples

In Figure 5.1, we give four examples of the evolution of the longest waiting time. The unbroken line indicates the rescaled longest waiting time per arriving job $(\max_{i\leq N} W_i(k)/c_N, 0 \leq k \leq n)$, while the dashed line indicates the sample path of the rescaled auxiliary process $(\sup_{0\leq j\leq k}(\tilde{S}(j,k)/(c_N/b_N) - \mu(k-j)/c_N, 0 \leq k \leq n)$ on the same probability space. We set $\mathbb{P}(S_i^{(1)}(j) > x) = \exp(-\sqrt{x})$ for $x > 0$, $\mathbb{P}(S^{(2)}(j) > x) = 1/x^2$ for $x > 1$, and $A(j) = 5$. In this case, we have that $\alpha = 1/2$, $\beta = 2$, $\mu = 1$, $b_N = (\log N)^2$, and $c_N = (\log N)^4$. Furthermore, we let $n = c_N$.

As we can see in Figure 5.1, the rescaled longest waiting time converges to a process that has a negative drift and has jumps of random sizes at random moments in time. We see that for $N = 1000$ and $N = 10,000$, and k small, the rescaled longest waiting time and the rescaled auxiliary process follow approximately the same trajectory. However, when k grows, the error becomes larger. We can explain this fact when we look at the heuristic calculations in (5.2.25)–(5.2.27): in order to derive the auxiliary process, we replace the quantity $\max_{i\leq N} S_i/b_N$ with 1, because $\max_{i\leq N} S_i/b_N \xrightarrow{\mathbb{P}} 1$, as $N \to \infty$. To give a more precise approximation, we look at the lower bound

$$\frac{\max_{i\leq N} R_i(k)}{c_N} \geq \frac{\max_{i\leq Nk} S_i^{(1)}(j^*)}{b_N} \frac{\max_{j\leq k} S^{(2)}(j)}{c_N/b_N} + \frac{\sum_{j=1, j\neq j^*}^{k} (S_{i^*}^{(1)}(j) S^{(2)}(j) - A(j))}{c_N},$$

with $S^{(2)}(j^*) = \max_{j\leq k} S^{(2)}(j)$ and $S_{i^*}^{(1)}(j^*) = \max_{i\leq Nk} S_i^{(1)}(j^*)$. Because $b_N/b_{Nc_N} \overset{N\to\infty}{\longrightarrow} 1$, we also have that $\max_{i\leq Nk} S_i/b_N \xrightarrow{\mathbb{P}} 1$, as $N \to \infty$ for all $1 < k < c_N$. However, for

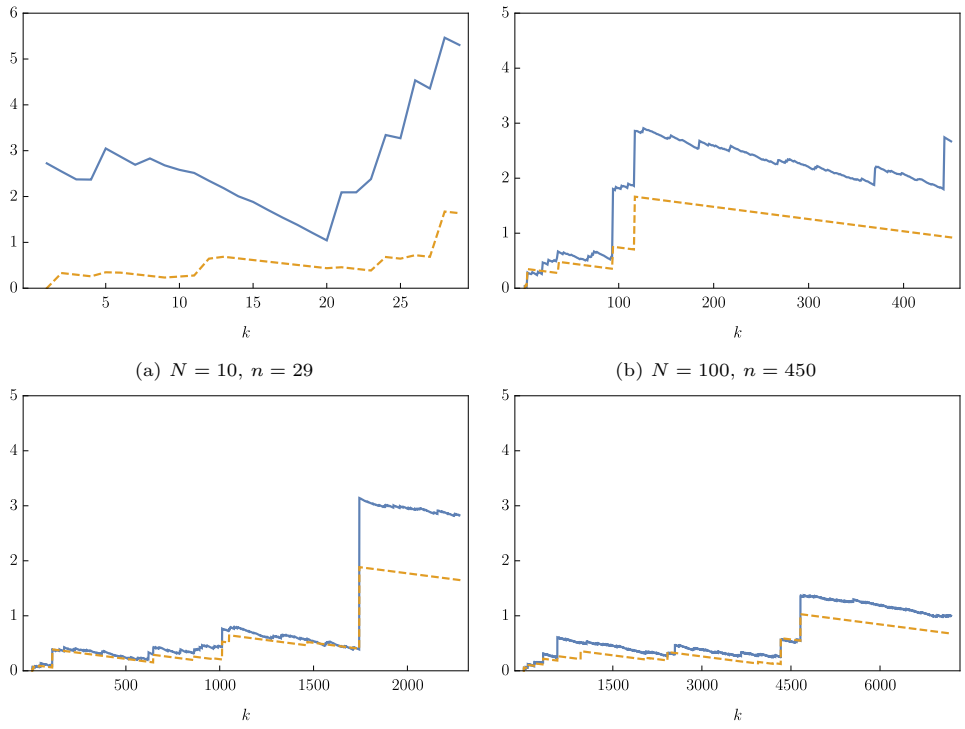
$N = 10,000$ and $k = 6000$, we have that

$$\mathbb{E}\left[\frac{\max_{i\leq Nk} S_i}{b_N}\right] = 4.05.$$

Moreover, we have that

$$\mathbb{E}\left[\frac{\max_{i\leq N} S_i}{b_N}\right] = 1.15.$$

Thus, this explains that the jump sizes of the rescaled longest waiting time and the rescaled auxiliary process differ. More specifically, we see that $\mathbb{E}[\max_{i\leq Nk} S_i/b_N] \geq 1$ for all k; this explains that the plots of the rescaled longest waiting time lie above the plots of the rescaled auxiliary process.



(a) $N = 10$, $n = 29$

(b) $N = 100$, $n = 450$

(c) $N = 1000$, $n = 2277$

(d) $N = 10,000$, $n = 7197$

Figure 5.1 Longest waiting time and auxiliary process

5.2.3 Other choices for $S_i^{(1)}(j)$

Our main result in Theorem 5.2 heavily relies on the fact that $S_i^{(1)}(j)$ is Weibull-like, and we are able to derive general results. Furthermore, when $S_i^{(1)}(j)$ has finite support, we are also able to derive general results. Under the assumption that $S_i^{(1)}(j)$ has a finite right endpoint

b, then from this, it follows that $b_N = b$. Furthermore, in Lemma 2.13, we have shown that $\max_{i\leq N} \sum_{j=1}^{n} x_j S_i^{(1)}(j)/b \xrightarrow{\mathbb{P}} \sum_{j=1}^{n} x_j$, as $N \to \infty$. Then,

$$\mathbb{P}\left(\max_{i\leq N} R_i(k) > x\right) \overset{N\to\infty}{\longrightarrow} \mathbb{P}\left(\sum_{j=1}^{k}(bS^{(2)}(j) - A(j)) > x\right). \tag{5.2.30}$$

Furthermore, if $\mathbb{E}[bS^{(2)}(j) - A(j)] < 0$, we get that

$$\mathbb{P}\left(\max_{i\leq N}\sup_{k\geq 0} R_i(k) > x\right) \overset{N\to\infty}{\longrightarrow} \mathbb{P}\left(\sup_{k\geq 0}\sum_{j=1}^{k}(bS^{(2)}(j) - A(j)) > x\right). \tag{5.2.31}$$

In general, when $S_i^{(1)}(j)$ has unbounded support, it follows from Lemma 2.13 that

$$\mathbb{P}\left(\max_{i\leq N} \tilde{W}_i(l) > xb_N\right) \overset{N\to\infty}{\longrightarrow} \mathbb{P}\left(\max_{z}\left\{\sum_{j=1}^{l} z_j S^{(2)}(j) : \sum_{j=1}^{l} z_j^{\alpha} \leq 1, 0 \leq z_j \leq 1\right\} > x\right). \tag{5.2.32}$$

For the heavy-tailed case, i.e., when $0 < \alpha \leq 1$, we have that

$$\max_{z}\left\{\sum_{j=1}^{l} z_j S^{(2)}(j) : \sum_{j=1}^{l} z_j^{\alpha} \leq 1, 0 \leq z_j \leq 1\right\} = \max_{j\leq l} S^{(2)}(j).$$

In contrast, for the light-tailed case, i.e., when $\alpha > 1$, we get non-trivial results depending on α. For instance for $\alpha = 2$, we obtain

$$\max_{z}\left\{\sum_{j=1}^{l} z_j S^{(2)}(j) : \sum_{j=1}^{l} z_j^{\alpha} \leq 1, 0 \leq z_j \leq 1\right\} = \sqrt{S^{(2)}(1)^2 + \cdots + S^{(2)}(l)^2}.$$

In conclusion, when $\alpha > 1$, we cannot rely on the property described in (5.2.25)–(5.2.27) that follows from Weibull-distributed random variables. Therefore, in this chapter, we will limit ourselves to the case $0 < \alpha < 1$. The case $\alpha = 1$ lies on the boundary between these two regimes; this case needs a separate analysis, beyond the scope of our methods.

5.3. Preliminary results

In Section 5.2.1, we gave the ideas behind our proofs. In order to be able to make these rigorous, we need some auxiliary lemmas.

In (5.2.7), (5.2.8), and (5.2.9), we heuristically describe the behavior of the sequence $(c_N, N \geq 1)$ and the slowly varying function $\tilde{\ell}$ given a sequence $(b_N, N \geq 1)$ and a slowly varying function ℓ. An unanswered question is whether this sequence $(c_N, N \geq 1)$ and this

function $\tilde{\ell}$ exist. In Lemma 5.1, we show how ℓ and $\tilde{\ell}$ are asymptotically related. Their asymptotic relation resembles the asymptotic relation between a slowly varying function ℓ and its de Bruijn conjugate $\ell^{\#}$; cf. [29, Thm. 1.5.13].

Lemma 5.1 (Asymptotic behavior of $\tilde{\ell}(x)$)**.** *The sequence* $(c_N, N \geq 1)$ *is given as* $c_N = \tilde{\ell}(b_N) b_N^{\beta/(\beta-1)}$ *where the function* $\tilde{\ell}$ *satisfies the relation*

$$\tilde{\ell}(x) \sim \ell(\tilde{\ell}(x) x^{1/(\beta-1)})^{1/(\beta-1)}, \tag{5.3.1}$$

as $x \to \infty$.

Proof. We write $x = b_N$, then the relation in (5.2.7) can be rewritten to

$$\tilde{\ell}(x) x^{\beta/(\beta-1)} \sim \frac{(\tilde{\ell}(x) x^{\beta/(\beta-1)}/x)^{\beta}}{\ell(\tilde{\ell}(x) x^{\beta/(\beta-1)}/x)},$$

as $x \to \infty$. This simplifies to

$$\tilde{\ell}(x) \sim \frac{\tilde{\ell}(x)^{\beta}}{\ell(\tilde{\ell}(x) x^{1/(\beta-1)})},$$

as $x \to \infty$. The lemma follows. □

Remark 5.1 (Asymptotic solutions of $\tilde{\ell}(x)$)**.** *It is not trivial to find functions* $\tilde{\ell}(x)$ *that have the asymptotic relation described in* (5.3.1)*, since* $\tilde{\ell}$ *appears both on the left-hand and the right-hand side of the equation. However, we know that* $\tilde{\ell}$ *is slowly varying; thus the term* $x^{1/(\beta-1)}$ *is dominant in* $\ell(\tilde{\ell}(x) x^{1/(\beta-1)})^{1/(\beta-1)}$*, so we can remove* $\tilde{\ell}$ *from the right-hand side in* (5.3.1) *and look at a function* $\tilde{\ell}^{(1)}$ *that equals*

$$\tilde{\ell}^{(1)}(x) = \ell(x^{1/(\beta-1)})^{1/(\beta-1)}.$$

For example, when $\ell(x) = \log x$*,* $\tilde{\ell}^{(1)}$ *satisfies* (5.3.1)*. However, there are also examples where* $\tilde{\ell}^{(1)}$ *does not satisfy the relation in* (5.3.1)*, for example, when* $\ell(x) = \exp(\sqrt{\log x})$*. Still, we are able to find candidates that satisfy the relation in* (5.3.1)*. First, we see that the relation in* (5.3.1) *is actually an iterative relation. Thus, we can rewrite* (5.3.1) *to*

$$\tilde{\ell}(x) \sim \ell(\ell(\tilde{\ell}(x) x^{1/(\beta-1)})^{1/(\beta-1)} x^{1/(\beta-1)})^{1/(\beta-1)},$$

as $x \to \infty$*. Now, with the same reasoning as before, we define*

$$\tilde{\ell}^{(2)}(x) = \ell(\ell(x^{1/(\beta-1)})^{1/(\beta-1)} x^{1/(\beta-1)})^{1/(\beta-1)}.$$

The function $\tilde{\ell}^{(2)}$ *satisfies the relation in* (5.3.1) *when* $\ell(x) = \exp(\sqrt{\log x})$ *and is a slowly varying function itself.*

In order to prove that the heuristic approximations in Equations (5.2.25)–(5.2.27) are correct, we need to prove two things; first, that the largest regularly varying random variable determines the stochastic part of the limit, and second, that the other random variables satisfy the law of large numbers. To prove this second property, we use Bennett's inequality

as stated below. In Corollary 5.1, we state a simplified version of this inequality which we use in our proofs.

Lemma 5.2 (Bennett's inequality [25])**.** *Let $Y_1, \ldots, Y_n$ be independent random variables, $\mathbb{E}[Y_i] = 0$, $\mathbb{E}[Y_i^2] = \sigma_i^2$, and $|Y_i| < M$ almost surely. Then for $y > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i > y\right) \leq \exp\left(-\frac{\sum_{i=1}^{n}\sigma_i^2}{M^2} h\left(\frac{yM}{\sum_{i=1}^{n}\sigma_i^2}\right)\right),$$

with $h(x) = (1+x)\log(1+x) - x$.

For a proof, see [158].

Corollary 5.1. *Let $Y_1, \ldots, Y_n$ be independent random variables, $\mathbb{E}[Y_i] = 0$, $\mathbb{E}[Y_i^2] = \sigma_i^2$, and $|Y_i| < M$ almost surely. Then for $y > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i > y\right) \leq \exp\left(-\frac{y}{M}\left(\log\left(1 + \frac{yM}{\sum_{i=1}^{n}\sigma_i^2}\right) - 1\right)\right).$$

Proof. Observe that for $x > 0$ we get that $h(x) > x(\log(1+x) - 1)$. Now, the corollary follows from Lemma 5.2. □

Though in Lemma 2.13 it is proven that for Weibull-distributed random variables $\max_{i\leq N}\sum_{j=1}^{n} x_j S_i^{(1)}(j)/b_N \xrightarrow{\mathbb{P}} \max_{j\leq n} x_j$, as $N \to \infty$, which heuristically explains the nature of our main result, our approximations in Equations (5.2.25)–(5.2.27) suggest that we should take the sum of $\lfloor tc_N \rfloor$ random variables. In Lemma 2.13 however, n does not depend on N. Thus, we cannot resort to this lemma in our proofs. However, in Lemma 5.3, a result is presented that we can use in this chapter and proves the approximations in Equations (5.2.25)–(5.2.27).

Lemma 5.3 ([35, Thm. 2])**.** *Let $Y_1, \ldots, Y_n$ be independent random variables with $\log \mathbb{P}(Y_i > x) \sim -qx^{\alpha}$, as $x \to \infty$, with $0 < \alpha < 1$ and $q > 0$. Let $(x_n, n \geq 1)$ be a sequence such that $\lim_{n\to\infty} x_n/n^{1/(2-\alpha)} = \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{x_n^{\alpha}} \log \mathbb{P}\left(\sum_{i=1}^{n} Y_i > x_n\right) = -q.$$

We want to prove process convergence of the longest waiting time to a limiting process, this limiting process is a function in $D[0,T]$. In [28, Thm. 13.3], a result is given that guarantees the convergence of a process in $D[0,T]$ when three conditions are satisfied, which we will apply.

Lemma 5.4 ([28, Thm. 13.3])**.** *Assume a sequence of processes $(Y^{(N)}(t), t \in [0,T])$ and a process $(Y(t), t \in [0,T])$ in $D[0,T]$, equipped with the d° metric, satisfy the following conditions:*

1. *For all* $\{t_1,\dots,t_k\} \subseteq [0,T]$*:* $(Y^{(N)}(t_1),\dots,Y^{(N)}(t_k)) \xrightarrow{d} (Y(t_1),\dots,Y(t_k))$ *as* $N \to \infty$*.*
2. $Y(T) - Y(T-\delta) \xrightarrow{\mathbb{P}} 0$ *as* $\delta \downarrow 0$*, and*
3. *For* $0 < r < s < t < T$*,* $\epsilon, \eta > 0$ *there exists* $N_0 \geq 1$ *and* $\delta > 0$ *such that*

$$\mathbb{P}\left(\sup_{s\in[r,t],t-r<\delta} \min\left(\left|Y^{(N)}(s) - Y^{(N)}(r)\right|, \left|Y^{(N)}(t) - Y^{(N)}(s)\right| \right) > \epsilon \right) \leq \eta, \quad N \geq N_0.$$

Then $(Y^{(N)}(t), t \in [0,T]) \xrightarrow{d} (Y(t), t \in [0,T])$ *as* $N \to \infty$*.*

Finally, to prove pointwise convergence of the longest waiting time process in (5.2.18) to the limiting random variable, we need to pay special attention to the case that $S^{(2)}$ is a regularly varying random variable with $1 < \beta \leq 2$, since in this case the second moment of $S^{(2)}$ is not finite. In Lemma 5.5, we give a useful convergence result of the second moment of $S^{(2)}$ conditioned on $S^{(2)}$ being bounded.

Lemma 5.5. *Let* S *be a positive random variable that satisfies* $\mathbb{P}(S > x) = \ell(x)/x^\beta$*, with* $\ell(x)$ *a slowly varying function and* $1 < \beta \leq 2$*. Then,*

$$\frac{\mathbb{E}[S^2 | S < r]}{r} \xrightarrow{r\to\infty} 0.$$

Proof. Choose $0 < \epsilon < \beta - 1$. Because $\mathbb{P}(S > x) = \ell(x)/x^\beta$ we have that $\mathbb{E}[S^{\beta-\epsilon}] < \infty$. Therefore,

$$\frac{\mathbb{E}[S^2 | S < r]}{r} \leq \frac{r^{2-(\beta-\epsilon)}}{r} \mathbb{E}[S^{\beta-\epsilon}] \xrightarrow{r\to\infty} 0.$$

□

5.4. Convergence of the auxiliary process in $D[0,T]$

In this section, we prove Theorem 5.1. As explained in Section 5.2.1, we first remove the supremum functional from the random variable on the left-hand side in (5.2.17) and prove convergence of the process $(\max_{i\leq N} R_i(tc_N)/c_N, t \in [0,T])$ to $(X_t - \mu t, t \in [0,T])$. To do so, we first show pointwise convergence in Lemma 5.6; afterwards, we prove process convergence in Lemma 5.7. In order to prove Lemma 5.7, we need two auxiliary results, which are given in Lemmas 5.8 and 5.9. By using the continuous mapping theorem, Theorem 5.1 follows.

Lemma 5.6. *For the sequence of random variables* $(R_i(k), i \geq 1, k \geq 1)$ *given in* (5.2.14)*, we have for all* $t > 0$ *and* $x > 0$*, that*

$$\mathbb{P}\left(\max_{i\leq N} R_i(tc_N) > xc_N \right) \xrightarrow{N\to\infty} 1 - \exp\left(-\frac{t}{(x+\mu t)^\beta} \right). \tag{5.4.1}$$

Proof. The approach to prove this lemma is by analyzing upper and lower bounds of the probability given in (5.4.1) and by proving that these bounds are sharp as $N \to \infty$. Thus,

first we see that

$$\mathbb{P}\left(\max_{i\le N} R_i(tc_N) > xc_N\right) \tag{5.4.2}$$

$$\begin{aligned}
&= \mathbb{P}\left(\max_{i\le N} R_i(tc_N) > xc_N \;\middle|\; \tilde{S}(tc_N) > (x+\mu t-\delta)\frac{c_N}{b_N}\right)\mathbb{P}\left(\tilde{S}(tc_N) > (x+\mu t-\delta)\frac{c_N}{b_N}\right)\\
&\quad + \mathbb{P}\left(\max_{i\le N} R_i(tc_N) > xc_N \;\middle|\; \tilde{S}(tc_N) \le (x+\mu t-\delta)\frac{c_N}{b_N}\right)\mathbb{P}\left(\tilde{S}(tc_N) \le (x+\mu t-\delta)\frac{c_N}{b_N}\right)\\
&\le \mathbb{P}\left(\tilde{S}(tc_N) > (x+\mu t-\delta)\frac{c_N}{b_N}\right) + \mathbb{P}\left(\max_{i\le N} R_i(tc_N) > xc_N \;\middle|\; \tilde{S}(tc_N) \le (x+\mu t-\delta)\frac{c_N}{b_N}\right).
\end{aligned} \tag{5.4.3}$$

The first term in (5.4.3) yields

$$\begin{aligned}
\mathbb{P}\left(\tilde{S}(tc_N) > (x+\mu t-\delta)\frac{c_N}{b_N}\right) &\sim 1 - \left(1 - \frac{\ell((x+\mu t-\delta)c_N/b_N)}{\left((x+\mu t-\delta)c_N/b_N\right)^{\beta}}\right)^{\lfloor tc_N\rfloor}\\
&\overset{N\to\infty}{\longrightarrow} 1-\exp\left(-\frac{t}{(x+\mu t-\delta)^{\beta}}\right)\\
&\overset{\delta\downarrow 0}{\longrightarrow} 1-\exp\left(-\frac{t}{(x+\mu t)^{\beta}}\right).
\end{aligned}$$

Hence, in order to prove that the upper bound of (5.4.1) is asymptotically sharp, we are left with proving that the second term in (5.4.3) vanishes as $N\to\infty$. We analyze this term as follows: first, we have that $(x+\mu t-\delta/2)/(x+\mu t-\delta) > 1$ for δ small enough, thus we write $(x+\mu t-\delta/2)/(x+\mu t-\delta) = 1+\epsilon$ with $\epsilon > 0$. Second, in order to bound the second term in (5.4.3), we first write $\mathbb{P}^t(A) = \mathbb{P}(A \mid \tilde{S}(tc_N) \le (x+\mu t-\delta)c_N/b_N)$. Then we can bound the second term in (5.4.3) as

$$\mathbb{P}^t\left(\max_{i\le N} R_i(tc_N) > xc_N\right) \tag{5.4.4}$$

$$\begin{aligned}
&\le \mathbb{P}^t\Bigg(\max_{i\le N}\sum_{j=1}^{\lfloor tc_N\rfloor}(S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j) < (1+\epsilon)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j) - A(j))\\
&\quad + \max_{i\le N}\sum_{j=1}^{\lfloor tc_N\rfloor} S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j) \ge (1+\epsilon)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j) > xc_N\Bigg).
\end{aligned} \tag{5.4.5}$$

The upper bound in (5.4.5) holds because for $(S_i^{(1)}(j), i\ge 1, j\ge 1)$, we have that

$$\max_{i\le N}\left(\sum_{j=1}^{k} S_i^{(1)}(j)\right) \le \max_{i\le N}\sum_{j=1}^{k} S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j) < c) + \max_{i\le N} S_i^{(1)}(j)\sum_{j=1}^{k}\mathbb{1}(S_i^{(1)}(j) > c).$$

Now, we can further bound the expression in (5.4.5) as follows:

$$\begin{aligned}
&\mathbb{P}^t\Bigg(\max_{i\le N}\sum_{j=1}^{\lfloor tc_N\rfloor}(S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)<(1+\epsilon)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j)-A(j))\\
&\quad+\max_{i\le N}\sum_{j=1}^{\lfloor tc_N\rfloor}S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)\ge(1+\epsilon)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j)>xc_N\Bigg)\\
&\le\mathbb{P}^t\left(\max_{i\le N}\sum_{j=1}^{\lfloor tc_N\rfloor}S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)<(1+\epsilon)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j)>\left(\mathbb{E}[S_i^{(1)}(j)S^{(2)}(j)]t+\frac{\delta}{4}\right)c_N\right) \qquad (5.4.6)\\
&\quad+\mathbb{P}\left(\sum_{j=1}^{\lfloor tc_N\rfloor}-A(j)>\left(-\mathbb{E}[A(j)]t+\frac{\delta}{4}\right)c_N\right) \qquad (5.4.7)\\
&\quad+\mathbb{P}^t\left(\max_{i\le N}\sum_{j=1}^{\lfloor tc_N\rfloor}S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)\ge(1+\epsilon)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j)>\left(x+\mu t-\frac{\delta}{2}\right)c_N\right). \qquad (5.4.8)
\end{aligned}$$

The upper bound from (5.4.5) to (5.4.6), (5.4.7), and (5.4.8) holds because of the union bound. The term in (5.4.7) converges to 0 due to the law of large numbers. For the term in (5.4.6), we know by the union bound that

$$\begin{aligned}
&\mathbb{P}^t\left(\max_{i\le N}\sum_{j=1}^{\lfloor tc_N\rfloor}S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)<(1+\epsilon)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j)>\left(\mathbb{E}[S_i^{(1)}(j)S^{(2)}(j)]t+\frac{\delta}{4}\right)c_N\right)\\
&\le N\mathbb{P}^t\left(\sum_{j=1}^{\lfloor tc_N\rfloor}S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)<(1+\epsilon)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j)>\left(\mathbb{E}[S_i^{(1)}(j)S^{(2)}(j)]t+\frac{\delta}{4}\right)c_N\right). \qquad (5.4.9)
\end{aligned}$$

Now, since we have a probability of sums of almost surely bounded random variables, we can apply Bennett's inequality with the setting given in Lemma 5.2 and Corollary 5.1. We see that $\mathbb{E}[S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)<(1+\epsilon)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j)\mid S^{(2)}(j)\le(x+\mu t-\delta)c_N/b_N]<\mathbb{E}[S_i^{(1)}(j)S^{(2)}(j)]$. Furthermore, we can choose M as $M=(x+\mu t-\delta)(1+\epsilon)^{1-\alpha}b_N^{1-\alpha}c_N/b_N$, and $y=\delta/4c_N$. Thus,

$$\frac{y}{M}=\frac{\delta}{4(x+\mu t-\delta)(1+\epsilon)^{1-\alpha}}b_N^{\alpha}=\frac{\delta}{4(x+\mu t-\delta)(1+\epsilon)^{1-\alpha}q}\log N.$$

It is important to note here, that y/M equals a constant times $\log N$. We now add a subscript N to the variables y, M, and σ_i to indicate sequences that change with N. Now, for $\beta>2$, $\limsup_{N\to\infty}\sigma_{i,N}^2<\infty$. Thus

$$\frac{y_N M_N}{\sum_{j=1}^{\lfloor tc_N\rfloor}\sigma_{i,N}^2}\overset{N\to\infty}{\longrightarrow}\infty.$$

Therefore, using the information that y/M equals a constant times $\log N$ and by using Corollary 5.1, we see that the exponent in Corollary 5.1 grows faster to infinity than $\log N$. Thus, by applying Bennett's inequality, we get that the expression in (5.4.9) converges to 0 as $N \to \infty$. When $1 < \beta \leq 2$, $\sigma_{i,N}^2 \overset{N\to\infty}{\longrightarrow} \infty$, however, from Lemma 5.5 it follows that $\sigma_{i,N}^2/(c_N/b_N) \overset{N\to\infty}{\longrightarrow} 0$. Therefore, $y_N M_N / \sum_{j=1}^{\lfloor tc_N \rfloor} \sigma_{i,N}^2 \overset{N\to\infty}{\longrightarrow} \infty$. Concluding, from Corollary 5.1 we again get that the expression in (5.4.9) and therefore the expression in (5.4.6) converges to 0.

Furthermore, for the term in (5.4.8) we have that

$$\mathbb{P}^t\left(\max_{i\leq N}\sum_{j=1}^{\lfloor tc_N\rfloor} S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)\geq(1+\epsilon)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j) > \left(x+\mu t-\frac{\delta}{2}\right)c_N\right)$$
$$\leq \mathbb{P}\left(\max_{i\leq N}\sum_{j=1}^{\lfloor tc_N\rfloor} S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)\geq(1+\epsilon)^{1-\alpha}b_N^{1-\alpha}) > \frac{x+\mu t-\delta/2}{x+\mu t-\delta}b_N\right).$$

We have $(x+\mu t-\delta/2)/(x+\mu t-\delta) = 1+\epsilon$ with $\epsilon > 0$; thus we can further simplify this probability. We write

$$A_{\epsilon,t}^{(N)} := \left\{\max_{i\leq N}\max_{j\leq\lfloor tc_N\rfloor} S_i^{(1)}(j) > (1+\epsilon)b_N\right\}.$$

Then

$$\mathbb{P}\left(\max_{i\leq N}\sum_{j=1}^{\lfloor tc_N\rfloor} S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)\geq(1+\epsilon)^{1-\alpha}b_N^{1-\alpha}) > (1+\epsilon)b_N\right)$$
$$= \mathbb{P}\left(\max_{i\leq N}\sum_{j=1}^{\lfloor tc_N\rfloor} S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)\geq(1+\epsilon)^{1-\alpha}b_N^{1-\alpha}) > (1+\epsilon)b_N \cap A_{\epsilon,t}^{(N)}\right) \tag{5.4.10}$$
$$+ \mathbb{P}\left(\max_{i\leq N}\sum_{j=1}^{\lfloor tc_N\rfloor} S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)\geq(1+\epsilon)^{1-\alpha}b_N^{1-\alpha}) > (1+\epsilon)b_N \cap \neg A_{\epsilon,t}^{(N)}\right). \tag{5.4.11}$$

Since $c_N = \tilde{\ell}(b_N)b_N^{\beta/(\beta-1)}$ with $b_N = (\log N/q)^{1/\alpha}$, it follows that $b_N/b_{Ntc_N} \overset{N\to\infty}{\longrightarrow} 1$, therefore, we have that

$$\frac{\max_{i\leq N}\max_{j\leq\lfloor tc_N\rfloor} S_i^{(1)}(j)}{b_N} \overset{\mathbb{P}}{\longrightarrow} 1,$$

as $N \to \infty$. From this, it follows that the term in (5.4.10) converges to 0 as $N \to \infty$, and we only need to focus on the term in (5.4.11). Observe that by the union bound,

$$\mathbb{P}\left(\max_{i\leq N}\sum_{j=1}^{\lfloor tc_N\rfloor} S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)\geq(1+\epsilon)^{1-\alpha}b_N^{1-\alpha}) > (1+\epsilon)b_N \cap \neg A_{\epsilon,t}^{(N)}\right)$$

$$\leq N\,\mathbb{P}\left(\sum_{j=1}^{\lfloor tc_N\rfloor} S_i^{(1)}(j)\mathbb{1}((1+\epsilon)^{1-\alpha}b_N^{1-\alpha}\leq S_i^{(1)}(j)\leq(1+\epsilon)b_N)>(1+\epsilon)b_N\right). \tag{5.4.12}$$

Following the proof given in [35, Lem. 8], we assume without loss of generality that $q=1$ and choose $1/(1+\epsilon)^{\alpha}<q'<1$ and $q'<q''<1$. Now, we have by using Chernoff's bound, that for $\theta>0$,

$$N\,\mathbb{P}\left(\sum_{j=1}^{\lfloor tc_N\rfloor} S_i^{(1)}(j)\mathbb{1}((1+\epsilon)^{1-\alpha}b_N^{1-\alpha}\leq S_i^{(1)}(j)\leq(1+\epsilon)b_N)>(1+\epsilon)b_N\right)$$
$$\leq N\left(1+\mathbb{E}\left[\exp\left(\theta S_i^{(1)}(j)\right)\mathbb{1}((1+\epsilon)^{1-\alpha}b_N^{1-\alpha}\leq S_i^{(1)}(j)\leq(1+\epsilon)b_N)\right]\right)^{\lfloor tc_N\rfloor}$$
$$\cdot\exp(-\theta(1+\epsilon)b_N).$$

Then, for $\theta=q'(1+\epsilon)^{\alpha-1}b_N^{\alpha-1}$, in [35, Lem. 8] it is proven that for N large enough

$$\mathbb{E}\left[\exp\left(q'(1+\epsilon)^{\alpha-1}b_N^{\alpha-1}S_i^{(1)}(j)\right)\mathbb{1}((1+\epsilon)^{1-\alpha}b_N^{1-\alpha}\leq S_i^{(1)}(j)\leq(1+\epsilon)b_N)\right]$$
$$\leq(1+q'(1+\epsilon)^{\alpha}b_N^{\alpha})\exp(q'-q''(1+\epsilon)^{\alpha(1-\alpha)}b_N^{\alpha(1-\alpha)}).$$

Now, by using the fact that $x>0$ we have the simple bound $1+x\leq\exp(x)$, and that $c_N=\tilde{\ell}(b_N)b_N^{\beta/(\beta-1)}$, it is easy to see that

$$\left(1+(1+q'(1+\epsilon)^{\alpha}b_N^{\alpha})\exp(q'-q''(1+\epsilon)^{\alpha(1-\alpha)}b_N^{\alpha(1-\alpha)})\right)^{\lfloor tc_N\rfloor}\overset{N\to\infty}{\longrightarrow}1.$$

Therefore, we know that Chernoff's bound with $\theta=q'(1+\epsilon)^{\alpha-1}b_N^{\alpha-1}$ applied to the expression in (5.4.12) satisfies

$$\limsup_{N\to\infty}N\Big(1+\mathbb{E}\Big[\exp\left(q'(1+\epsilon)^{\alpha-1}b_N^{\alpha-1}S_i^{(1)}(j)\right)\mathbb{1}((1+\epsilon)^{1-\alpha}b_N^{1-\alpha}\leq S_i^{(1)}(j)\leq(1+\epsilon)b_N)\Big]\Big)^{\lfloor tc_N\rfloor}$$
$$\cdot\exp(-q'(1+\epsilon)^{\alpha-1}b_N^{\alpha-1}(1+\epsilon)b_N)\leq\limsup_{N\to\infty}N\exp(-q'(1+\epsilon)^{\alpha}b_N^{\alpha}).$$

Since $q'>1/(1+\epsilon)^{\alpha}$, we have that $q'(1+\epsilon)^{\alpha}b_N^{\alpha}>\log N$ and therefore that $N\exp(-q'(1+\epsilon)^{\alpha}b_N^{\alpha})\overset{N\to\infty}{\longrightarrow}0$. Thus, we can conclude that the expression in (5.4.12) converges to 0 as $N\to\infty$. From this, it follows that the term in (5.4.8) converges to 0 as $N\to\infty$ as well, and we can conclude that the upper bound proposed in (5.4.3) is asymptotically sharp.

To prove a sharp lower bound for the probability in (5.4.1), observe that, because for a sequence $(a_i(j), i\geq1, j\geq1)$ we have that $\max_{i\leq N}\sum_{j=1}^{k}a_i(j)\geq\max_{i\leq N}\max_{j\leq k}a_i(j)+\sum_{j=1,j\neq j^*}^{k}a_{i^*}(j)$, with $j^*\in\arg\max\{j:\max_{i\leq N}a_i(j)=\max_{i\leq N}\max_{l\leq k}a_i(l)\}$ and $i^*\in\arg\max\{i:a_i(j^*)=\max_{m\leq N}a_m(j^*)\}$. Then,

$$\liminf_{N\to\infty}\mathbb{P}\left(\max_{i\leq N}R_i(tc_N)>xc_N\right)$$

$$\geq \liminf_{N\to\infty} \mathbb{P}\Bigg(\max_{i\leq N} S_i^{(1)}(j^*(tc_N))\tilde{S}(tc_N) - A(j^*(tc_N)) + \sum_{j=1, j\neq j^*(tc_N)}^{\lfloor tc_N \rfloor} (S^{(1)}_{i^*(tc_N)}(j)S^{(2)}(j) - A(j)) > xc_N \Bigg), \tag{5.4.13}$$

with $j^*(tc_N) \in \arg\max\{j : S^{(2)}(j) = \tilde{S}(tc_N)\}$ and $i^*(tc_N) \in \arg\max\{i : S_i^{(1)}(j^*(tc_N)) = \max_{m\leq N} S_m^{(1)}(j^*(tc_N))\}$. Because $\tilde{S}(tc_N)$ scales as c_N/b_N, we get that $\mathbb{E}[S^{(1)}]\tilde{S}(tc_N)/c_N \xrightarrow{\mathbb{P}} 0$, as $N \to \infty$. Since $S^{(1)}_{i^*(tc_N)}(j)$ with $j \neq j^*(tc_N)$ and $S^{(1)}_{i^*(tc_N)}(j^*(tc_N))$ are independent, we have that $\sum_{j=1, j\neq j^*(tc_N)}^{\lfloor tc_N \rfloor} (S^{(1)}_{i^*(tc_N)}(j)S^{(2)}(j) - A(j))/c_N \xrightarrow{\mathbb{P}} -\mu t$; cf. [79, Thm. 1]. Furthermore,

$$\mathbb{P}\left(\max_{i\leq N} \max_{j\leq \lfloor tc_N \rfloor} (S_i^{(1)}(j)S^{(2)}(j))/c_N > x + \mu t \right) \overset{N\to\infty}{\longrightarrow} 1 - \exp\left(-\frac{t}{(x+\mu t)^\beta} \right)$$

and $A(j^*(tc_N))/c_N \xrightarrow{\mathbb{P}} 0$ as $N \to \infty$. In conclusion, the lower bound in (5.4.13) is sharp, as the limit is the same as the limit in (5.4.1). □

We have established pointwise convergence of the process $(\max_{i\leq N} R_i(tc_N)/c_N, t \in [0,T])$ to a Fréchet-distributed random variable. In Lemma 5.7, we prove convergence in $D[0,T]$.

Lemma 5.7. *For the sequence of random variables $(R_i(k), i \geq 1, k \geq 1)$ given in (5.2.14), we have for all $T > 0$, that*

$$\left(\frac{\max_{i\leq N} R_i(tc_N)}{c_N}, t \in [0,T] \right) \xrightarrow{d} \left(X_t - \mu t, t \in [0,T] \right), \tag{5.4.14}$$

as $N \to \infty$.

This lemma follows from the two results stated in Lemma 5.8 and 5.9.

Lemma 5.8. *For the sequence of random variables $(\tilde{S}(k), k \geq 1)$ given in (5.2.29), we have for all $T > 0$, that*

$$\left(\frac{\tilde{S}(tc_N)}{c_N/b_N}, t \in [0,T] \right) \xrightarrow{d} \left(X_t, t \in [0,T] \right), \tag{5.4.15}$$

as $N \to \infty$.

Lemma 5.9. *For the sequences of random variables $(R_i(k), i \geq 1, k \geq 1)$ and $(\tilde{S}(k), k \geq 1)$ given in (5.2.14) and (5.2.29), we have for all $T > 0$ and $\epsilon > 0$, that*

$$\mathbb{P}\left(\sup_{t\in[0,T]} \left| \frac{\max_{i\leq N} R_i(tc_N)}{c_N} - \left(\frac{\tilde{S}(tc_N)}{c_N/b_N} - \mu t \right) \right| > \epsilon \right) \overset{N\to\infty}{\longrightarrow} 0. \tag{5.4.16}$$

Using the triangle inequality, we get that (5.4.14) follows from (5.4.15) and (5.4.16).

Proof of Lemma 5.8. In this proof, we use Lemma 5.4, thus we need to prove the three conditions stated in Lemma 5.4. First, we need to prove that

$$\left(\frac{\tilde{S}(t_1 c_N)}{c_N/b_N}, \ldots, \frac{\tilde{S}(t_m c_N)}{c_N/b_N}\right) \xrightarrow{d} (X_{t_1}, \ldots, X_{t_m}),$$

as $N \to \infty$. Let us assume that $m = 2$ and $t_2 > t_1$. If $x_2 \leq x_1$, because $\tilde{S}(k)$ is increasing in k, we have that

$$\begin{aligned}\mathbb{P}\left(\frac{\tilde{S}(t_1 c_N)}{c_N/b_N} \leq x_1, \frac{\tilde{S}(t_2 c_N)}{c_N/b_N} \leq x_2\right) \\ = \mathbb{P}\left(\frac{\tilde{S}(t_2 c_N)}{c_N/b_N} \leq x_2\right) \overset{N\to\infty}{\longrightarrow} \mathbb{P}(X_{t_2} \leq x_2) = \mathbb{P}(X_{t_1} \leq x_1, X_{t_2} \leq x_2). \quad (5.4.17)\end{aligned}$$

When $x_2 > x_1$, we have that

$$\begin{aligned}&\mathbb{P}\left(\frac{\tilde{S}(t_1 c_N)}{c_N/b_N} \leq x_1, \frac{\tilde{S}(t_2 c_N)}{c_N/b_N} \leq x_2\right) \\ &= \mathbb{P}\left(\frac{\tilde{S}(t_2 c_N)}{c_N/b_N} \leq x_2 \,\middle|\, \frac{\tilde{S}(t_1 c_N)}{c_N/b_N} \leq x_1\right) \mathbb{P}\left(\frac{\tilde{S}(t_1 c_N)}{c_N/b_N} \leq x_1\right) \\ &= \mathbb{P}\left(\frac{\tilde{S}(\lfloor t_2 c_N \rfloor - \lfloor t_1 c_N \rfloor)}{c_N/b_N} \leq x_2\right) \mathbb{P}\left(\frac{\tilde{S}(t_1 c_N)}{c_N/b_N} \leq x_1\right) \\ &\overset{N\to\infty}{\longrightarrow} \mathbb{P}(X_{t_2 - t_1} \leq x_2)\, \mathbb{P}(X_{t_1} \leq x_1) = \mathbb{P}(X_{t_1} \leq x_1, X_{t_2} \leq x_2).\end{aligned} \quad (5.4.18)$$

For $m > 2$ but finite, we can prove by induction that all the finite-dimensional distributions converge. Assume that the m-dimensional distributions converge. Consider $t_1 < t_2 < \ldots < t_m < t_{m+1}$, and $x_1, \ldots, x_{m+1}$. Now, by the induction hypothesis, we know that

$$\mathbb{P}\left(\frac{\tilde{S}(t_1 c_N)}{c_N/b_N} \leq x_1, \ldots, \frac{\tilde{S}(t_m c_N)}{c_N/b_N} \leq x_m\right) \overset{N\to\infty}{\longrightarrow} \mathbb{P}(X_{t_1} \leq x_1, \ldots, X_{t_m} \leq x_m).$$

To prove that the $(m+1)$-dimensional distributions also converge, we need to distinguish two cases: first, the case that $x_{m+1} < \max(x_1, \ldots, x_m)$. Because $\tilde{S}(t)$ is non-decreasing in t, the joint probability

$$\mathbb{P}\left(\frac{\tilde{S}(t_1 c_N)}{c_N/b_N} \leq x_1, \ldots, \frac{\tilde{S}(t_{m+1} c_N)}{c_N/b_N} \leq x_{m+1}\right)$$

reduces to a joint probability of at most m events, similar to (5.4.17). We know that the m-dimensional distributions converge, and thus, by the same argument as in (5.4.17), the $(m+1)$-dimensional distribution also converges. The second case is that $x_{m+1} \geq \max(x_1, \ldots, x_m)$.

Similarly to (5.4.18), we have that

$$\begin{aligned}
&\mathbb{P}\left(\frac{\tilde{S}(t_1c_N)}{c_N/b_N} \le x_1, \ldots, \frac{\tilde{S}(t_{m+1}c_N)}{c_N/b_N} \le x_{m+1}\right)\\
&= \mathbb{P}\left(\frac{\tilde{S}(t_{m+1}c_N)}{c_N/b_N} \le x_{m+1} \middle| \frac{\tilde{S}(t_1c_N)}{c_N/b_N} \le x_1, \ldots, \frac{\tilde{S}(t_mc_N)}{c_N/b_N} \le x_m\right)\\
&\quad \cdot \mathbb{P}\left(\frac{\tilde{S}(t_1c_N)}{c_N/b_N} \le x_1, \ldots, \frac{\tilde{S}(t_mc_N)}{c_N/b_N} \le x_m\right)\\
&= \mathbb{P}\left(\frac{\tilde{S}(\lfloor t_{m+1}c_N\rfloor - \lfloor t_mc_N\rfloor)}{c_N/b_N} \le x_{m+1}\right)\mathbb{P}\left(\frac{\tilde{S}(t_1c_N)}{c_N/b_N} \le x_1, \ldots, \frac{\tilde{S}(t_mc_N)}{c_N/b_N} \le x_m\right)\\
&\overset{N\to\infty}{\longrightarrow} \mathbb{P}(X_{t_{m+1}-t_m} \le x_{m+1})\,\mathbb{P}(X_{t_1} \le x_1, \ldots, X_{t_m} \le x_m)\\
&= \mathbb{P}(X_{t_1} \le x_1, \ldots, X_{t_{m+1}} \le x_{m+1}).
\end{aligned}$$

Second, we need to prove that

$$X_T - X_{T-\delta} \xrightarrow{\mathbb{P}} 0,$$

as $\delta \downarrow 0$. We can write $X_T = \max(X_{T-\delta}, \hat{X}_\delta)$ with $\hat{X}_\delta$ an independent copy of X_δ. Therefore, $X_T - X_{T-\delta} \le \hat{X}_\delta$. Let $\epsilon > 0$, then

$$\mathbb{P}(X_T - X_{T-\delta} > \epsilon) \le \mathbb{P}(\hat{X}_\delta > \epsilon) = 1 - \exp\left(-\frac{\delta}{\epsilon^\beta}\right) \xrightarrow{\delta\downarrow 0} 0.$$

Finally, we show that the process $\left(\tilde{S}(tc_N)/(c_N/b_N), t \in [0,T]\right)$ satisfies the third condition in Lemma 5.4. The random variable $\tilde{S}(k)$ is increasing with k. Furthermore, the minimum of two numbers is bounded from above by the average. Also, because for $k > l$, $\tilde{S}(k) - \tilde{S}(l) = \max(\tilde{S}(l), \tilde{S}(l+1,k)) - \tilde{S}(l)$, we can bound

$$\frac{\tilde{S}(sc_N) - \tilde{S}(rc_N)}{c_N/b_N} \le_{st.} \frac{\max_{j \le \lfloor sc_N\rfloor - \lfloor rc_N\rfloor} \hat{S}^{(2)}(j)}{c_N/b_N},$$

where $\hat{S}^{(2)}$ is an independent copy of $S^{(2)}$. Therefore, we have that

$$\begin{aligned}
&\sup_{s\in[r,t]} \min\left|\frac{\tilde{S}(sc_N) - \tilde{S}(rc_N)}{c_N/b_N}, \frac{\tilde{S}(tc_N) - \tilde{S}(sc_N)}{c_N/b_N}\right|\\
&= \sup_{s\in[r,t]} \min\left(\frac{\tilde{S}(sc_N) - \tilde{S}(rc_N)}{c_N/b_N}, \frac{\tilde{S}(tc_N) - \tilde{S}(sc_N)}{c_N/b_N}\right)\\
&\le \frac{\tilde{S}(tc_N) - \tilde{S}(rc_N)}{2c_N/b_N}\\
&\le_{st.} \frac{\max_{j \le \lfloor tc_N\rfloor - \lfloor rc_N\rfloor} \hat{S}^{(2)}(j)}{2c_N/b_N}.
\end{aligned}$$

Thus, using the expression in the third condition of Lemma 5.4, we obtain that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[r,t],t-r<\delta}\min\left|\frac{\tilde{S}(sc_N)-\tilde{S}(rc_N)}{c_N/b_N},\frac{\tilde{S}(tc_N)-\tilde{S}(sc_N)}{c_N/b_N}\right|>\epsilon\right)\\
&\quad\le\mathbb{P}\left(\frac{\max_{j\le\lfloor\delta c_N\rfloor}\hat{S}^{(2)}(j)}{c_N/b_N}>2\epsilon\right)\\
&\quad\le\lfloor\delta c_N\rfloor\frac{\ell(2\epsilon c_N/b_N)}{(2\epsilon c_N/b_N)^\beta}.
\end{aligned}$$

We have that $c_N\ell(2\epsilon c_N/b_N)/(c_N/b_N)^\beta \overset{N\to\infty}{\longrightarrow} 1$ because $c_N\sim(c_N/b_N)^\beta/\ell(c_N/b_N)$ as $N\to\infty$ and ℓ is a slowly varying function, so we choose $N_0>1$ such that

$$\lfloor\delta c_N\rfloor\frac{\ell(2\epsilon c_N/b_N)}{(2\epsilon c_N/b_N)^\beta}<(1+\epsilon)\frac{\delta}{(2\epsilon)^\beta}$$

for all $N>N_0$. Now, choose $0<\delta<\eta(2\epsilon)^\beta/(1+\epsilon)$ and we get

$$\mathbb{P}\left(\sup_{s\in[r,t],t-r<\delta}\min\left|\frac{\tilde{S}(sc_N)-\tilde{S}(rc_N)}{c_N/b_N},\frac{\tilde{S}(tc_N)-\tilde{S}(sc_N)}{c_N/b_N}\right|>\epsilon\right)<\eta,\quad N>N_0.$$

Hence, the process $\left(\tilde{S}(tc_N)/(c_N/b_N),t\in[0,T]\right)$ also satisfies the third condition in Lemma 5.4 and the result follows. □

We have proven process convergence of $\left(\tilde{S}(tc_N)/(c_N/b_N),t\in[0,T]\right)$ to $(X_t,t\in[0,T])$. Now, in order to prove Lemma 5.7, we are left with proving that the convergence result in (5.4.16) holds. We do this in Lemma 5.9.

Proof of Lemma 5.9. The random variable in (5.4.16) has the form of a supremum of the absolute value of a stochastic process. We know that $|X|=\max(X,-X)$. Then, by applying the union bound, we get that $\mathbb{P}(|X|>x)\le\mathbb{P}(X>x)+\mathbb{P}(-X>x)$. Thus, to prove the convergence result in (5.4.16), we can remove the absolute value and prove that the probability of a supremum of a stochastic process converges to 0 as $N\to\infty$; see (5.4.19). Then, we need to prove that the probability of the supremum of the mirrored process converges to 0 as $N\to\infty$; cf. (5.4.28).

We first prove that

$$\mathbb{P}\left(\sup_{t\in[0,T]}\left(-\frac{\max_{i\le N}R_i(tc_N)}{c_N}+\left(\frac{\tilde{S}(tc_N)}{c_N/b_N}-\mu t\right)\right)>\epsilon\right)\tag{5.4.19}$$

converges to 0 as $N\to\infty$. We have, by using $i^*(tc_N)$ and $j^*(tc_N)$, as defined in Lemma 5.6,

that

$$\mathbb{P}\left(\sup_{t\in[0,T]}\left(-\frac{\max_{i\le N} R_i(tc_N)}{c_N}+\left(\frac{\tilde{S}(tc_N)}{c_N/b_N}-\mu t\right)\right)>\epsilon\right)$$
$$\le \mathbb{P}\Bigg(\sup_{t\in[0,T]}\Bigg(-\mu t-\frac{\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}(S^{(1)}_{i^*(tc_N)}(j)S^{(2)}(j)-A(j))-A(j^*(tc_N))}{c_N}$$
$$+\frac{\tilde{S}(tc_N)}{c_N/b_N}-\frac{\max_{i\le N} S_i^{(1)}(j^*(tc_N))\tilde{S}(tc_N)}{c_N}\Bigg)>\epsilon\Bigg)$$
$$\le \mathbb{P}\left(\sup_{t\in[0,T]}\left(-\mu t-\frac{\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}(S^{(1)}_{i^*(tc_N)}(j)S^{(2)}(j)-A(j))-A(j^*(tc_N))}{c_N}\right)>\frac{\epsilon}{2}\right) \tag{5.4.20}$$
$$+\mathbb{P}\left(\sup_{t\in[0,T]}\left(\frac{\tilde{S}(tc_N)}{c_N/b_N}-\frac{\max_{i\le N} S_i^{(1)}(j^*(tc_N))\tilde{S}(tc_N)}{c_N}\right)>\frac{\epsilon}{2}\right). \tag{5.4.21}$$

For the term in (5.4.20), we use the union bound to obtain that

$$\mathbb{P}\left(\sup_{t\in[0,T]}\left(-\mu t-\frac{\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}(S^{(1)}_{i^*(tc_N)}(j)S^{(2)}(j)-A(j))-A(j^*(tc_N))}{c_N}\right)>\frac{\epsilon}{2}\right)$$
$$\le \mathbb{P}\left(\sup_{t\in[0,\frac{\epsilon}{4\mathbb{E}[A(j)]}]}\left(-\mu t-\frac{\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}(S^{(1)}_{i^*(tc_N)}(j)S^{(2)}(j)-A(j))-A(j^*(tc_N))}{c_N}\right)>\frac{\epsilon}{2}\right) \tag{5.4.22}$$
$$+\mathbb{P}\left(\sup_{t\in[\frac{\epsilon}{4\mathbb{E}[A(j)]},T]}\left(-\mu t-\frac{\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}(S^{(1)}_{i^*(tc_N)}(j)S^{(2)}(j)-A(j))-A(j^*(tc_N))}{c_N}\right)>\frac{\epsilon}{2}\right). \tag{5.4.23}$$

Because all random variables $S_i^{(1)}(j)$, $S^{(2)}(j)$, and $A(j)$ are positive, it is easy to see that the term in (5.4.22) can be upper bounded by

$$\mathbb{P}\left(\sup_{t\in[0,\frac{\epsilon}{4\mathbb{E}[A(j)]}]}\left(\frac{\sum_{j=1}^{\lfloor tc_N\rfloor}A(j)}{c_N}\right)>\frac{\epsilon}{2}\right)=\mathbb{P}\left(\frac{\sum_{j=1}^{\lfloor \frac{\epsilon}{4\mathbb{E}[A(j)]}c_N\rfloor}A(j)}{c_N}>\frac{\epsilon}{2}\right)\overset{N\to\infty}{\longrightarrow}0,$$

as we can conclude from the law of large numbers that $\sum_{j=1}^{\lfloor \frac{\epsilon}{4\mathbb{E}[A(j)]}c_N\rfloor}A(j)/c_N \overset{\mathbb{P}}{\longrightarrow} \epsilon/4$ as $N\to\infty$. For the term in (5.4.23) we have for $0<\delta<1$, since all random variables are

positive, that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[\frac{\epsilon}{4\mathbb{E}[A(j)]},T]} \left(-\mu t - \frac{\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}(S^{(1)}_{i^*(tc_N)}(j)S^{(2)}(j) - A(j)) - A(j^*(tc_N))}{c_N} \right) > \frac{\epsilon}{2} \right) \\
&\quad \leq \sup_{t\in[\frac{\epsilon}{4\mathbb{E}[A(j)]},T]} \frac{1}{\delta} \\
&\qquad \cdot \mathbb{P}\left(\sup_{s\in[t,t+\delta]} \left(-\mu s - \frac{\sum_{j=1,j\neq j^*(sc_N)}^{\lfloor sc_N\rfloor}(S^{(1)}_{i^*(sc_N)}(j)S^{(2)}(j) - A(j)) - A(j^*(sc_N))}{c_N} \right) > \frac{\epsilon}{2} \right) \\
&\quad \leq \sup_{t\in[\frac{\epsilon}{4\mathbb{E}[A(j)]},T]} \frac{1}{\delta} \\
&\qquad \cdot \mathbb{P}\left(\left(-\mu t - \frac{\inf_{s\in[t,t+\delta]}\sum_{j=1,j\neq j^*(sc_N)}^{\lfloor sc_N\rfloor} S^{(1)}_{i^*(sc_N)}(j)S^{(2)}(j)}{c_N} + \frac{\sum_{j=1}^{\lfloor (t+\delta)c_N\rfloor} A(j)}{c_N} \right) > \frac{\epsilon}{2} \right).
\end{aligned} \tag{5.4.24}$$

To bound the term in (5.4.24), we argue as follows: we have that

$$\inf_{s\in[t,t+\delta]} \sum_{j=1,j\neq j^*(sc_N)}^{\lfloor sc_N\rfloor} S^{(1)}_{i^*(sc_N)}(j)S^{(2)}(j) \geq \inf_{s\in[t,t+\delta]} \sum_{j=1,j\neq j^*(sc_N)}^{\lfloor tc_N\rfloor} S^{(1)}_{i^*(sc_N)}(j)S^{(2)}(j).$$

Due to the definition of $j^*(sc_N)$ in Lemma 5.6 we know that if the process $(\tilde{S}(k), k \geq 0)$ achieves a new extreme at time $\lfloor \tau c_N\rfloor$ in the interval $[tc_N, (t+\delta)c_N]$, then that means that $j^*(\tau c_N) > tc_N$. When the process $(\tilde{S}(k), k \geq 0)$ does not achieve a new extreme in the interval $[tc_N, (t+\delta)c_N]$, that means that $j^*(sc_N) = j^*(tc_N)$, for all $s \in [t, t+\delta]$. In either case, we have that

$$\inf_{s\in[t,t+\delta]} \sum_{j=1,j\neq j^*(sc_N)}^{\lfloor tc_N\rfloor} S^{(1)}_{i^*(sc_N)}(j)S^{(2)}(j) \geq \inf_{s\in[t,t+\delta]} \sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor} S^{(1)}_{i^*(sc_N)}(j)S^{(2)}(j).$$

The latter lower bound is an infimum over a stochastic number of sums, and this number is determined by the number of new extremes of the process $(\tilde{S}(k), k \geq 0)$ in the interval $[tc_N, (t+\delta)c_N]$. We use the result from [65, Eq. (6)] that the expected number of new extremes of the process $(\tilde{S}(k), k \geq 0)$ in the interval $[tc_N, (t+\delta)c_N]$ equals $\sum_{j=\lfloor tc_N\rfloor}^{\lfloor (t+\delta)c_N\rfloor} 1/j \overset{N\to\infty}{\longrightarrow} \log((t+\delta)/t)$. Therefore, we can conclude that the number of different instances of $i^*(sc_N)$ when $s \in [t, t+\delta]$ is asymptotically finite, with probability converging to 1. Therefore, we have that

$$\begin{aligned}
&\mathbb{P}\left(\left(-\mu t - \frac{\inf_{s\in[t,t+\delta]}\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor} S^{(1)}_{i^*(sc_N)}(j)S^{(2)}(j)}{c_N} + \frac{\sum_{j=1}^{\lfloor (t+\delta)c_N\rfloor} A(j)}{c_N} \right) > \frac{\epsilon}{2} \right) \\
&\quad \leq \mathbb{P}\left(\#\text{ new extremes of } \left(\tilde{S}(sc_N), s \geq 0\right) \text{ in } [t, t+\delta] \geq \left\lceil \frac{1}{\delta^2} \right\rceil \right)
\end{aligned} \tag{5.4.25}$$

$$+\mathbb{P}\Bigg(\Bigg(-\mu t-\frac{\inf_{s\in[t,t+\delta]}\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}S^{(1)}_{i^*(sc_N)}(j)S^{(2)}(j)}{c_N}+\frac{\sum_{j=1}^{\lfloor(t+\delta)c_N\rfloor}A(j)}{c_N}\Bigg)>\frac{\epsilon}{2},$$
$$\#\text{ new extremes of }\left(\tilde{S}(sc_N),s\geq 0\right)\text{ in }[t,t+\delta]<\left\lceil\frac{1}{\delta^2}\right\rceil\Bigg). \tag{5.4.26}$$

We can bound the term in (5.4.25) by using Markov's inequality:

$$\limsup_{N\to\infty}\mathbb{P}\left(\#\text{ new extremes of }\left(\tilde{S}(sc_N),s\geq 0\right)\text{ in }[t,t+\delta]\geq\left\lceil\frac{1}{\delta^2}\right\rceil\right)\leq\delta^2\log\left(\frac{t+\delta}{t}\right).$$

For the term in (5.4.26) we can apply the union bound and get

$$\begin{aligned}
&\mathbb{P}\Bigg(\Bigg(-\mu t-\frac{\inf_{s\in[t,t+\delta]}\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}S^{(1)}_{i^*(sc_N)}(j)S^{(2)}(j)}{c_N}+\frac{\sum_{j=1}^{\lfloor(t+\delta)c_N\rfloor}A(j)}{c_N}\Bigg)>\frac{\epsilon}{2},\\
&\qquad\#\text{ new extremes of }\left(\tilde{S}(sc_N),s\geq 0\right)\text{ in }[t,t+\delta]<\left\lceil\frac{1}{\delta^2}\right\rceil\Bigg)\\
&\leq\mathbb{P}\Bigg(\Bigg(-\mu t-\frac{\min_{i\leq\lceil 1/\delta^2\rceil}\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}S^{(1)}_{i}(j)S^{(2)}(j)}{c_N}+\frac{\sum_{j=1}^{\lfloor(t+\delta)c_N\rfloor}A(j)}{c_N}\Bigg)>\frac{\epsilon}{2}\Bigg)\\
&\leq\left\lceil\frac{1}{\delta^2}\right\rceil\mathbb{P}\Bigg(\Bigg(-\mu t-\frac{\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}S^{(1)}_{i}(j)S^{(2)}(j)}{c_N}+\frac{\sum_{j=1}^{\lfloor(t+\delta)c_N\rfloor}A(j)}{c_N}\Bigg)>\frac{\epsilon}{2}\Bigg).
\end{aligned}$$

Obviously, the distribution of the random variables $(S^{(2)}(j), j\leq\lfloor tc_N\rfloor, j\neq j^*(tc_N))$ depends on the value of $S^{(2)}(j^*(tc_N))$, because we know that $S^{(2)}(j)\leq S^{(2)}(j^*(tc_N))$ for $j\leq\lfloor tc_N\rfloor$ and $j\neq j^*(tc_N)$. Observe that for $y<z$,

$$\mathbb{P}(S^{(2)}(j)\leq x\mid S^{(2)}(j)\leq y)\geq\mathbb{P}(S^{(2)}(j)\leq x\mid S^{(2)}(j)\leq z).$$

Now, we can choose $x_\delta=(t/\log(1/\delta^3))^{1/\beta}$ such that $\limsup_{N\to\infty}\mathbb{P}\big(\tilde{S}(tc_N)\leq x_\delta c_N/b_N\big)=\delta^3$. Then, we can bound

$$\begin{aligned}
&\mathbb{P}\Bigg(\Bigg(-\mu t-\frac{\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}S^{(1)}_{i}(j)S^{(2)}(j)}{c_N}+\frac{\sum_{j=1}^{\lfloor(t+\delta)c_N\rfloor}A(j)}{c_N}\Bigg)>\frac{\epsilon}{2}\Bigg)\\
&\leq\mathbb{P}\left(\tilde{S}(tc_N)\leq x_\delta\frac{c_N}{b_N}\right)\\
&\quad+\mathbb{P}\left(\tilde{S}(tc_N)\geq x_\delta\frac{c_N}{b_N}\right)\\
&\qquad\cdot\mathbb{P}\Bigg(\Bigg(-\mu t-\frac{\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}S^{(1)}_{i}(j)S^{(2)}(j)}{c_N}+\frac{\sum_{j=1}^{\lfloor(t+\delta)c_N\rfloor}A(j)}{c_N}\Bigg)>\frac{\epsilon}{2}\Bigg|\tilde{S}(tc_N)\geq x_\delta\frac{c_N}{b_N}\Bigg)\\
&\leq\delta^3+\mathbb{P}\Bigg(\Bigg(-\mu t-\frac{\sum_{j=1,j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor}S^{(1)}_{i}(j)S^{(2)}(j)}{c_N}+\frac{\sum_{j=1}^{\lfloor(t+\delta)c_N\rfloor}A(j)}{c_N}\Bigg)>\frac{\epsilon}{2}\Bigg|\tilde{S}(tc_N)=x_\delta\frac{c_N}{b_N}\Bigg).
\end{aligned}$$

Because $\mathbb{E}[S^{(2)}(j) \mid S^{(2)}(j) \leq x_\delta c_N/b_N] \overset{N\to\infty}{\longrightarrow} \mathbb{E}[S^{(2)}(j)]$, we get by the law of large numbers that

$$\begin{aligned}-\mu t - \frac{\sum_{j=1, j\neq j^*(tc_N)}^{\lfloor tc_N\rfloor} S_i^{(1)}(j)S^{(2)}(j)}{c_N} + \frac{\sum_{j=1}^{\lfloor (t+\delta)c_N\rfloor} A(j)}{c_N} \\ \overset{\mathbb{P}}{\longrightarrow} -\mu t - \mathbb{E}[S_i^{(1)}(j)S^{(2)}(j)]t + (t+\delta)\mathbb{E}[A(j)] = \mathbb{E}[A(j)]\delta,\end{aligned}$$

as $N \to \infty$. Thus, we can conclude that when we take δ small enough compared to ϵ, the expression in (5.4.24), and therefore also in (5.4.20), converge to 0 as $N \to \infty$.

For the term in (5.4.21), we have that

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\in[0,T]}\left(\frac{\tilde{S}(tc_N)}{c_N/b_N} - \frac{\max_{i\leq N} S_i^{(1)}(j^*(tc_N))\tilde{S}(tc_N)}{c_N}\right) > \frac{\epsilon}{2}\right) \\ &\qquad \leq \mathbb{P}\left(\sup_{t\in[0,T]}\left(1 - \frac{\max_{i\leq N} S_i^{(1)}(\lfloor tc_N\rfloor)}{b_N}\right)\frac{\tilde{S}(Tc_N)}{c_N/b_N} > \frac{\epsilon}{2}\right). \quad (5.4.27)\end{aligned}$$

This tail probability converges to 0 as $N \to \infty$, since, we know that $\tilde{S}(Tc_N)/(c_N/b_N)$ converges in distribution to a Fréchet random variable as $N \to \infty$, and $\sup_{t\in[0,T]}(1 - \max_{i\leq N} S_i^{(1)}(\lfloor tc_N\rfloor)/b_N) \overset{\mathbb{P}}{\longrightarrow} 0$, as $N \to \infty$. To see this, we first bound

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[0,T]}\left(1 - \frac{\max_{i\leq N} S_i^{(1)}(\lfloor tc_N\rfloor)}{b_N}\right) > \frac{\epsilon}{2}\right) &= \mathbb{P}\left(\inf_{t\in[0,T]}\frac{\max_{i\leq N} S_i^{(1)}(\lfloor tc_N\rfloor)}{b_N} < 1 - \frac{\epsilon}{2}\right) \\ &\leq \lfloor Tc_N\rfloor\,\mathbb{P}\left(\frac{\max_{i\leq N} S_i^{(1)}(1)}{b_N} < 1 - \frac{\epsilon}{2}\right).\end{aligned}$$

Now, we have that

$$\mathbb{P}\left(\frac{\max_{i\leq N} S_i^{(1)}(1)}{b_N} < 1 - \frac{\epsilon}{2}\right) \leq \exp\left(-N\,\mathbb{P}\left(\frac{S_i^{(1)}(1)}{b_N} > 1 - \frac{\epsilon}{2}\right)\right),$$

see the proof of [67, Thm. 5.4.4, p. 192]. Thus

$$\begin{aligned}\lfloor Tc_N\rfloor\,\mathbb{P}\left(\frac{\max_{i\leq N} S_i^{(1)}(1)}{b_N} < 1 - \frac{\epsilon}{2}\right) &\leq \lfloor Tc_N\rfloor \exp\left(-N\,\mathbb{P}\left(\frac{S_i^{(1)}(1)}{b_N} > 1 - \frac{\epsilon}{2}\right)\right) \\ &= \lfloor Tc_N\rfloor \exp\left(-N\exp(-(1+o(1))(1-\epsilon/2)^\alpha \log N)\right) \\ &= \lfloor Tc_N\rfloor \exp\left(-N^{1-(1+o(1))(1-\epsilon/2)^\alpha}\right) \\ &\overset{N\to\infty}{\longrightarrow} 0.\end{aligned}$$

Hence, the upper bound in (5.4.27) converges to 0 as $N \to \infty$. These results together give that the tail probability in (5.4.19) converges to 0 as $N \to \infty$.

To prove the convergence result in (5.4.16), we are left with proving that the probability

$$\mathbb{P}\left(\sup_{t\in[0,T]} \left(\frac{\max_{i\le N} R_i(tc_N)}{c_N} - \left(\frac{\tilde{S}(tc_N)}{c_N/b_N} - \mu t \right) \right) > \epsilon \right) \tag{5.4.28}$$

converges to 0 as $N \to \infty$. In order to do so, we have $\eta > 0$ and use the upper bound

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[0,T]} \left(\frac{\max_{i\le N} R_i(tc_N)}{c_N} - \left(\frac{\tilde{S}(tc_N)}{c_N/b_N} - \mu t \right) \right) > \epsilon \right) \\
&\le \sup_{t\in[0,T]} \frac{1}{\delta} \mathbb{P}\Bigg(\sup_{s\in[t,t+\delta]} \Bigg(\frac{\max_{i\le N} \sum_{j=1}^{\lfloor sc_N \rfloor} (S_i^{(1)}(j) \mathbb{1}(S_i^{(1)}(j) < (1+\eta)^{1-\alpha} b_N^{1-\alpha}) S^{(2)}(j) - A(j))}{c_N} \\
&\qquad\qquad + \mu s \Bigg) > \frac{\epsilon}{2} \Bigg) \qquad (5.4.29) \\
&\quad + \mathbb{P}\left(\sup_{t\in[0,T]} \left(\frac{\max_{i\le N} \sum_{j=1}^{\lfloor tc_N \rfloor} S_i^{(1)}(j) \mathbb{1}(S_i^{(1)}(j) \ge (1+\eta)^{1-\alpha} b_N^{1-\alpha}) S^{(2)}(j)}{c_N} - \frac{\tilde{S}(tc_N)}{c_N/b_N} \right) > \frac{\epsilon}{2} \right), \qquad (5.4.30)
\end{aligned}$$

with $0 < \delta < 1$. For the term in (5.4.29), we argue as follows:

$$\begin{aligned}
&\sup_{s\in[t,t+\delta]} \left(\frac{\max_{i\le N} \sum_{j=1}^{\lfloor sc_N \rfloor} (S_i^{(1)}(j) \mathbb{1}(S_i^{(1)}(j) < (1+\eta)^{1-\alpha} b_N^{1-\alpha}) S^{(2)}(j) - A(j))}{c_N} + \mu s \right) \\
&\le \frac{\max_{i\le N} \sum_{j=1}^{\lfloor (t+\delta)c_N \rfloor} S_i^{(1)}(j) \mathbb{1}(S_i^{(1)}(j) < (1+\eta)^{1-\alpha} b_N^{1-\alpha}) S^{(2)}(j) - \sum_{j=1}^{\lfloor tc_N \rfloor} A(j)}{c_N} + \mu(t+\delta).
\end{aligned}$$

This last expression converges in probability to $\delta \mathbb{E}[A]$ as $N \to \infty$, which follows from the proof of Lemma 5.6. Therefore, the expression in (5.4.29) asymptotically vanishes by taking δ small enough compared to ϵ. For the term in (5.4.30), we know from Lemma 5.6 that

$$\left(\frac{\max_{i\le N} \sum_{j=1}^{\lfloor Tc_N \rfloor} S_i^{(1)}(j) \mathbb{1}(S_i^{(1)}(j) \ge (1+\eta)^{1-\alpha} b_N^{1-\alpha})}{b_N} - 1 \right) \xrightarrow{\mathbb{P}} 0,$$

as $N \to \infty$. We also know from Lemma 5.6 that $\tilde{S}(Tc_N)/(c_N/b_N)$ converges in distribution to X_T as $N \to \infty$. Now, we can conclude from a similar proof as in Lemma 5.6 that

$$\mathbb{P}\left(\sup_{t\in[0,T]} \left(\frac{\max_{i\le N} \sum_{j=1}^{\lfloor tc_N \rfloor} S_i^{(1)}(j) \mathbb{1}(S_i^{(1)}(j) \ge (1+\eta)^{1-\alpha} b_N^{1-\alpha}) S^{(2)}(j)}{c_N} - \frac{\tilde{S}(tc_N)}{c_N/b_N} \right) > \frac{\epsilon}{2} \right)$$

$$\leq \mathbb{P}\left(\sup_{t\in[0,T]}\left(\frac{\max_{i\leq N}\sum_{j=1}^{\lfloor tc_N\rfloor} S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)\geq(1+\eta)^{1-\alpha}b_N^{1-\alpha})}{c_N}\frac{\tilde{S}(tc_N)}{c_N/b_N}-\frac{\tilde{S}(tc_N)}{c_N/b_N}\right)>\frac{\epsilon}{2}\right)$$

$$\leq \mathbb{P}\left(\left(\frac{\max_{i\leq N}\sum_{j=1}^{\lfloor Tc_N\rfloor} S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)\geq(1+\eta)^{1-\alpha}b_N^{1-\alpha})}{b_N}-1\right)\frac{\tilde{S}(Tc_N)}{c_N/b_N}>\frac{\epsilon}{2}\right)$$

$$\leq \mathbb{P}\left(\frac{\max_{i\leq N}\sum_{j=1}^{\lfloor Tc_N\rfloor} S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)\geq(1+\eta)^{1-\alpha}b_N^{1-\alpha})}{b_N}-1>\eta\right)+\mathbb{P}\left(\frac{\tilde{S}(Tc_N)}{c_N/b_N}>\frac{\epsilon}{2\eta}\right)$$

$$\overset{N\to\infty}{\longrightarrow} \mathbb{P}(X_T>\epsilon/(2\eta)) \overset{\eta\downarrow 0}{\longrightarrow} 0.$$

The last bound holds due to the union bound: for random variables X and Y with $Y\geq 0$ and $x,y>0$, we have that $\mathbb{P}(XY\geq xy)\leq \mathbb{P}(X\geq x)+\mathbb{P}(Y\geq y)$. Now that we have established that the probabilities in (5.4.19) and (5.4.28) converge to 0 as $N\to\infty$, the result follows. □

From the results in Lemmas 5.7 and 5.8, we can conclude that the convergence result in (5.4.14) holds. Furthermore, by applying the continuous mapping theorem, Theorem 5.1 follows.

5.5. Process convergence of the longest waiting time in $D[0,T]$

At this point, we have proven the convergence result of an auxiliary process whose marginals are the same as the marginals of the longest waiting time. Now, we can extend these results to prove convergence of the longest waiting time $(\max_{i\leq N} W_i(tc_N)/c_N, t\in[0,T])$ to the process $(\sup_{s\in[0,t]}(X_{(s,t)}-\mu(t-s)), t\in[0,T])$ as $N\to\infty$. We first show in Lemma 5.10 that the longest waiting time can be approximated by an auxiliary process, as we did in Lemma 5.9, and then we prove the main result described in Theorem 5.2.

Lemma 5.10. *For the sequences of random variables $(R_i(l,k), i\geq 1, k\geq l\geq 1)$ and $(\tilde{S}(l,k), k\geq l\geq 1)$ given in (5.2.15) and (5.2.28), we have for all $T>0$ and $\epsilon>0$, that*

$$\mathbb{P}\left(\sup_{t\in[0,T]}\left|\sup_{s\in[0,t]}\frac{\max_{i\leq N}R_i(sc_N,tc_N)}{c_N}-\sup_{s\in[0,t]}\left(\frac{\tilde{S}(sc_N,tc_N)}{c_N/b_N}-\mu(t-s)\right)\right|>\epsilon\right)$$
$$\overset{N\to\infty}{\longrightarrow} 0. \quad (5.5.1)$$

Proof. As in Lemma 5.9, we first use that $|X|=\max(X,-X)$. Then, by applying the union bound we get that $\mathbb{P}(|X|>x)\leq\mathbb{P}(X>x)+\mathbb{P}(-X>x)$. Now, we have that

$$\sup_{t\in[0,T]}\left(\sup_{s\in[0,t]}\frac{\max_{i\leq N}R_i(sc_N,tc_N)}{c_N}-\sup_{s\in[0,t]}\left(\frac{\tilde{S}(sc_N,tc_N)}{c_N/b_N}-\mu(t-s)\right)\right)$$

$$\leq \sup_{t\in[0,T]} \sup_{s\in[0,t]} \left(\frac{\max_{i\leq N} R_i(sc_N, tc_N)}{c_N} - \left(\frac{\tilde{S}(sc_N, tc_N)}{c_N/b_N} - \mu(t-s) \right) \right).$$

Similarly,

$$\sup_{t\in[0,T]} \left(\sup_{s\in[0,t]} \left(\frac{\tilde{S}(sc_N, tc_N)}{c_N/b_N} - \mu(t-s) \right) - \sup_{s\in[0,t]} \frac{\max_{i\leq N} R_i(sc_N, tc_N)}{c_N} \right)$$
$$\leq \sup_{t\in[0,T]} \sup_{s\in[0,t]} \left(\left(\frac{\tilde{S}(sc_N, tc_N)}{c_N/b_N} - \mu(t-s) \right) - \frac{\max_{i\leq N} R_i(sc_N, tc_N)}{c_N} \right).$$

Therefore,

$$\mathbb{P}\left(\sup_{t\in[0,T]} \left| \sup_{s\in[0,t]} \frac{\max_{i\leq N} R_i(sc_N, tc_N)}{c_N} - \sup_{s\in[0,t]} \left(\frac{\tilde{S}(sc_N, tc_N)}{c_N/b_N} - \mu(t-s) \right) \right| > \epsilon \right)$$
$$\leq 2\,\mathbb{P}\left(\sup_{t\in[0,T]} \sup_{s\in[0,t]} \left| \frac{\max_{i\leq N} R_i(sc_N, tc_N)}{c_N} - \left(\frac{\tilde{S}(sc_N, tc_N)}{c_N/b_N} - \mu(t-s) \right) \right| > \epsilon \right).$$

Now, we use the same approach as in Lemma 5.9, with the somewhat different upper bound:

$$\mathbb{P}\left(\sup_{t\in[0,T]} \sup_{s\in[0,t]} \left(\frac{\max_{i\leq N} R_i(sc_N, tc_N)}{c_N} - \left(\frac{\tilde{S}(sc_N, tc_N)}{c_N/b_N} - \mu(t-s) \right) \right) > \epsilon \right) \leq$$
$$\sup_{t\in[0,T]} \sup_{s\in[0,t]} \frac{1}{\delta^2}\,\mathbb{P}\left(\sup_{r\in[t,t+\delta]} \sup_{q\in[s-\delta,s]} \left(\frac{\max_{i\leq N} R_i(qc_N, rc_N)}{c_N} - \frac{\tilde{S}(qc_N, rc_N)}{c_N/b_N} + \mu(r-q) \right) > \epsilon \right). \tag{5.5.2}$$

Also,

$$\mathbb{P}\left(\sup_{t\in[0,T]} \sup_{s\in[0,t]} \left(-\frac{\max_{i\leq N} R_i(sc_N, tc_N)}{c_N} + \left(\frac{\tilde{S}(sc_N, tc_N)}{c_N/b_N} - \mu(t-s) \right) \right) > \epsilon \right) \leq$$
$$\sup_{t\in[0,T]} \sup_{s\in[0,t]} \frac{1}{\delta^2}\,\mathbb{P}\left(\sup_{r\in[t,t+\delta]} \sup_{q\in[s-\delta,s]} \left(-\frac{\max_{i\leq N} R_i(qc_N, rc_N)}{c_N} + \frac{\tilde{S}(qc_N, rc_N)}{c_N/b_N} - \mu(r-q) \right) > \epsilon \right). \tag{5.5.3}$$

We can use the same arguments as in the proof of Lemma 5.9. For the expression in (5.5.2), we have by the subadditivity property of the sup operators that

$$\sup_{r\in[t,t+\delta]} \sup_{q\in[s-\delta,s]} \left(\frac{\max_{i\leq N} \sum_{j=\lfloor qc_N \rfloor}^{\lfloor rc_N \rfloor} (S_i^{(1)}(j) S^{(2)}(j) - A(j))}{c_N} - \left(\frac{\tilde{S}(qc_N, rc_N)}{c_N/b_N} - \mu(r-q) \right) \right) \tag{5.5.4}$$

$$\leq \sup_{r\in[t,t+\delta]}\sup_{q\in[s-\delta,s]}\left(\frac{\max_{i\leq N}\sum_{j=\lfloor qc_N\rfloor}^{\lfloor rc_N\rfloor}(S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)<(1+\eta)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j)-A(j))}{c_N}+\mu(r-q)\right)$$
$$+\sup_{r\in[t,t+\delta]}\sup_{q\in[s-\delta,s]}\left(\frac{\max_{i\leq N}\sum_{j=\lfloor qc_N\rfloor}^{\lfloor rc_N\rfloor}S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)\geq(1+\eta)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j)}{c_N}-\frac{\tilde{S}(qc_N,rc_N)}{c_N/b_N}\right).$$

Because the random variables $S_i^{(1)}(j)$, $S^{(2)}(j)$, and $A(j)$ are positive, we can bound these two expressions on the right-hand side in the same way as in Lemma 5.9, and thus remove the two sup operators. For the first term on the right-hand side of (5.5.4), we have that

$$\sup_{r\in[t,t+\delta]}\sup_{q\in[s-\delta,s]}\left(\frac{\max_{i\leq N}\sum_{j=\lfloor qc_N\rfloor}^{\lfloor rc_N\rfloor}(S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)<(1+\eta)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j)-A(j))}{c_N}+\mu(r-q)\right)$$
$$\leq \frac{\max_{i\leq N}\sum_{j=\lfloor(s-\delta)c_N\rfloor}^{\lfloor(t+\delta)c_N\rfloor}S_i^{(1)}(j)\mathbb{1}(S_i^{(1)}(j)<(1+\eta)^{1-\alpha}b_N^{1-\alpha})S^{(2)}(j)-\sum_{j=\lfloor sc_N\rfloor}^{\lfloor tc_N\rfloor}A(j)}{c_N}+\mu(t-s+2\delta).$$

A similar analysis holds for the second term on the right-hand side of (5.5.4). Thus, by following the same reasoning as in the proof of Lemma 5.9, we have that the term in (5.5.2) converges to 0 when δ is small enough.

For the probability in (5.5.3), we also follow the same steps as in Lemma 5.9. □

Proof of Theorem 5.2. We have proven in Lemma 5.10 that the longest waiting time can be approximated with the process $(\sup_{s\in[0,t]}(\tilde{S}(sc_N,tc_N)/(c_N/b_N)-\mu(t-s)),t\in[0,T])$. Therefore, in order to prove convergence of the longest waiting time to the process $(\sup_{s\in[0,t]}(X_{(s,t)}-\mu(t-s)),t\in[0,T])$ in $D[0,T]$, it suffices to prove convergence of the process $(\sup_{s\in[0,t]}(\tilde{S}(sc_N,tc_N)/(c_N/b_N)-\mu(t-s)),t\in[0,T])$ to the process $(\sup_{s\in[0,t]}(X_{(s,t)}-\mu(t-s)),t\in[0,T])$ in $D[0,T]$. As in Lemma 5.8, we again check the conditions given in Lemma 5.4.

We start with proving the convergence of finite-dimensional distributions. To do this, we show that the joint probabilities of these processes can be written as operations of marginal probabilities, and therefore, convergence of finite-dimensional distributions follows

from convergence of the one-dimensional distributions. Thus, we can write

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[0,t_1]}(X_{(s,t_1)}-\mu(t_1-s))<x_1,\ \sup_{s\in[0,t_2]}(X_{(s,t_2)}-\mu(t_2-s))<x_2\right)\\
&\quad=\mathbb{P}\left(\sup_{s\in[0,t_1]}(X_{(s,t_1)}+\mu s)<x_1+\mu t_1\,\middle|\,\sup_{s\in[0,t_2]}(X_{(s,t_2)}+\mu s)<x_2+\mu t_2\right)\\
&\qquad\cdot\mathbb{P}\left(\sup_{s\in[0,t_2]}(X_{(s,t_2)}+\mu s)<x_2+\mu t_2\right).
\end{aligned}\tag{5.5.5}$$

Now, we can further rewrite the event

$$\begin{aligned}
&\Big\{\sup_{s\in[0,t_2]}(X_{(s,t_2)}+\mu s)<x_2+\mu t_2\Big\}\\
&\quad=\Big\{\sup_{s\in[0,t_1]}(X_{(s,t_1)}+\mu s)<x_2+\mu t_2\Big\}\cap\Big\{X_{(t_1,t_2)}+\mu t_1<x_2+\mu t_2\Big\}\\
&\qquad\qquad\cap\Big\{\sup_{s\in(t_1,t_2]}(X_{(s,t_2)}+\mu s)<x_2+\mu t_2\Big\}.
\end{aligned}$$

Thus, when $x_2+\mu t_2\le x_1+\mu t_1$, then

$$\mathbb{P}\left(\sup_{s\in[0,t_1]}(X_{(s,t_1)}+\mu s)<x_1+\mu t_1\,\middle|\,\sup_{s\in[0,t_2]}(X_{(s,t_2)}+\mu s)<x_2+\mu t_2\right)=1,$$

and when $x_2+\mu t_2>x_1+\mu t_1$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[0,t_1]}(X_{(s,t_1)}+\mu s)<x_1+\mu t_1\,\middle|\,\sup_{s\in[0,t_2]}(X_{(s,t_2)}+\mu s)<x_2+\mu t_2\right)\\
&\qquad\qquad=\frac{\mathbb{P}\big(\sup_{s\in[0,t_1]}(X_{(s,t_1)}+\mu s)<x_1+\mu t_1\big)}{\mathbb{P}\big(\sup_{s\in[0,t_1]}(X_{(s,t_1)}+\mu s)<x_2+\mu t_2\big)}.
\end{aligned}$$

From now on, we focus on the case $x_2+\mu t_2>x_1+\mu t_1$. The proof of the case $x_2+\mu t_2\le x_1+\mu t_1$ is analogous. For the case $x_2+\mu t_2>x_1+\mu t_1$, we have that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[0,t_1]}(X_{(s,t_1)}-\mu(t_1-s))<x_1,\ \sup_{s\in[0,t_2]}(X_{(s,t_2)}-\mu(t_2-s))<x_2\right)\\
&\quad=\frac{\mathbb{P}\big(\sup_{s\in[0,t_1]}(X_{(s,t_1)}-\mu(t_1-s))<x_1\big)}{\mathbb{P}\big(\sup_{s\in[0,t_1]}(X_{(s,t_1)}-\mu(t_1-s))<x_2+\mu(t_2-t_1)\big)}\\
&\qquad\cdot\mathbb{P}\left(\sup_{s\in[0,t_2]}(X_{(s,t_2)}-\mu(t_2-s))<x_2\right).
\end{aligned}\tag{5.5.6}$$

Thus, we can write the joint probability in (5.5.5) as an operation of marginal probabilities.

We can do the same for the process $(\sup_{s\in[0,t]}\left(\tilde{S}(sc_N,tc_N)/(c_N/b_N)-\mu(t-s)\right),t\in[0,T])$;

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[0,t_1]}\left(\frac{\tilde{S}(sc_N,t_1c_N)}{c_N/b_N}-\mu(t_1-s)\right)<x_1,\ \sup_{s\in[0,t_2]}\left(\frac{\tilde{S}(sc_N,t_2c_N)}{c_N/b_N}-\mu(t_2-s)\right)<x_2\right)\\
&\quad=\frac{\mathbb{P}\big(\sup_{s\in[0,t_1]}(\tilde{S}(sc_N,t_1c_N)/(c_N/b_N)-\mu(t_1-s))<x_1\big)}{\mathbb{P}\big(\sup_{s\in[0,t_1]}(\tilde{S}(sc_N,t_1c_N)/(c_N/b_N)-\mu(t_1-s))<x_2+\mu(t_2-t_1)\big)}\\
&\qquad\cdot\mathbb{P}\left(\sup_{s\in[0,t_2]}\left(\frac{\tilde{S}(sc_N,t_2c_N)}{c_N/b_N}-\mu(t_2-s)\right)<x_2\right).
\end{aligned}\tag{5.5.7}$$

By the same induction arguments as in Lemma 5.8, the same result holds for all finite-dimensional distributions.

Using Lemma 5.8 and the decomposition of a joint probability into marginal probabilities, we establish that the joint probability in (5.5.7) converges to the joint probability in (5.5.6) as $N\to\infty$. Analogous extensions hold for higher dimensional distributions. Hence, the convergence of finite-dimensional distributions follows. To prove process convergence of $(\sup_{s\in[0,t]}\left(\tilde{S}(sc_N,tc_N)/(c_N/b_N)-\mu(t-s)\right),t\in[0,T])$, we show that the second and third condition of Lemma 5.4 also hold. To establish that the second condition holds, we bound

$$\begin{aligned}
&\mathbb{P}\left(\left|\sup_{s\in[0,T]}(X_{(s,T)}-\mu(T-s))-\sup_{s\in[0,T-\delta]}(X_{(s,T-\delta)}-\mu(T-\delta-s))\right|>\epsilon\right)\\
&\qquad\qquad\leq\mathbb{P}\left(\sup_{s\in[0,T]}(X_{(s,T)}+\mu s)-\sup_{s\in[0,T-\delta]}(X_{(s,T-\delta)}+\mu s)+\mu\delta>\epsilon\right).
\end{aligned}$$

Now, we can further bound this as follows:

$$\begin{aligned}
&\sup_{s\in[0,T]}(X_{(s,T)}+\mu s)-\sup_{s\in[0,T-\delta]}(X_{(s,T-\delta)}+\mu s)+\mu\delta\\
&\quad=\max\left(\sup_{s\in[0,T-\delta]}(X_{(s,T)}+\mu s),\sup_{s\in[T-\delta,T]}(X_{(s,T)}+\mu s)\right)\\
&\qquad-\sup_{s\in[0,T-\delta]}(X_{(s,T-\delta)}+\mu s)+\mu\delta\\
&\quad\leq\max\left(\sup_{s\in[0,T-\delta]}(X_{(s,T)}-X_{(s,T-\delta)}),X_{(T-\delta,T)}+\mu T-\sup_{s\in[0,T-\delta]}(X_{(s,T-\delta)}+\mu s)\right)\\
&\qquad+\mu\delta\\
&\quad\leq\max\left(\sup_{s\in[0,T-\delta]}(X_{(s,T)}-X_{(s,T-\delta)}),X_{(T-\delta,T)}+\mu T-\mu(T-\delta)\right)+\mu\delta\\
&\quad=X_{(T-\delta,T)}+2\mu\delta.
\end{aligned}\tag{5.5.8}$$

We have that

$$\mathbb{P}(X_{(T-\delta,T)} + 2\mu\delta > \epsilon) = 1 - \exp\left(-\frac{\delta}{(\epsilon - 2\mu\delta)^\beta}\right) \xrightarrow{\delta\downarrow 0} 0.$$

To establish that the third condition of Lemma 5.4 holds, we first observe that for $r \leq t$

$$\begin{aligned}\left|\sup_{u\in[0,t]}\left(\frac{\tilde{S}(uc_N, tc_N)}{c_N/b_N} - \mu(t-u)\right) - \sup_{u\in[0,r]}\left(\frac{\tilde{S}(uc_N, rc_N)}{c_N/b_N} - \mu(r-u)\right)\right| \\ \leq \sup_{u\in[0,t]}\left(\frac{\tilde{S}(uc_N, tc_N)}{c_N/b_N} + \mu u\right) - \sup_{u\in[0,r]}\left(\frac{\tilde{S}(uc_N, rc_N)}{c_N/b_N} + \mu u\right) + \mu(t-r).\end{aligned}$$

Thus, due to the fact that for $x, y > 0$, $\min(x, y) \leq (x + y)/2$, we have for $r < s < t$, that

$$\begin{aligned}&\min\left(\left|\sup_{u\in[0,s]}\left(\frac{\tilde{S}(uc_N, sc_N)}{c_N/b_N} - \mu(s-u)\right) - \sup_{u\in[0,r]}\left(\frac{\tilde{S}(uc_N, rc_N)}{c_N/b_N} - \mu(r-u)\right)\right|,\right. \\ &\qquad\left.\left|\sup_{u\in[0,t]}\left(\frac{\tilde{S}(uc_N, tc_N)}{c_N/b_N} - \mu(t-u)\right) - \sup_{u\in[0,s]}\left(\frac{\tilde{S}(uc_N, sc_N)}{c_N/b_N} - \mu(s-u)\right)\right|\right) \\ &\leq \frac{1}{2}\left|\sup_{u\in[0,s]}\left(\frac{\tilde{S}(uc_N, sc_N)}{c_N/b_N} - \mu(s-u)\right) - \sup_{u\in[0,r]}\left(\frac{\tilde{S}(uc_N, rc_N)}{c_N/b_N} - \mu(r-u)\right)\right| \\ &\quad + \frac{1}{2}\left|\sup_{u\in[0,t]}\left(\frac{\tilde{S}(uc_N, tc_N)}{c_N/b_N} - \mu(t-u)\right) - \sup_{u\in[0,s]}\left(\frac{\tilde{S}(uc_N, sc_N)}{c_N/b_N} - \mu(s-u)\right)\right| \\ &\leq \frac{1}{2}\left(\sup_{u\in[0,t]}\left(\frac{\tilde{S}(uc_N, tc_N)}{c_N/b_N} + \mu u\right) - \sup_{u\in[0,r]}\left(\frac{\tilde{S}(uc_N, rc_N)}{c_N/b_N} + \mu u\right) + \mu(t-r)\right).\end{aligned}$$

For $t - r < \delta$, we have by using the same bounds as in (5.5.8), that

$$\begin{aligned}\frac{1}{2}\left(\sup_{u\in[0,t]}\left(\frac{\tilde{S}(uc_N, tc_N)}{c_N/b_N} + \mu u\right) - \sup_{u\in[0,r]}\left(\frac{\tilde{S}(uc_N, rc_N)}{c_N/b_N} + \mu u\right) + \mu(t-r)\right) \\ \leq_{st.} \frac{1}{2}\frac{\tilde{S}(\delta c_N)}{c_N/b_N} + \mu\delta.\end{aligned}$$

By taking $\delta > 0$ small enough, we see that the third condition of Lemma 5.4 holds. Thus, we have process convergence of the longest waiting time to $(\sup_{s\in[0,t]}(X_{(s,t)} - \mu(t-s)), t \in [0, T])$. □

5.6. Steady-state convergence of the longest waiting time

Finally, we prove steady-state convergence of the longest of the N waiting times. We give lower and upper bounds of $\mathbb{P}(\max_{i\leq N} W_i(\infty) > xc_N)$ and show that these are asymptotically sharp.

Proof of Theorem 5.3. To prove a sharp lower bound, we first use that $\max_{i\leq N} W_i(\infty) \stackrel{d}{=}$

$\max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}(S_i^{(1)}(j)S^{(2)}(j)-A(j))$; see Lemma 3.1. Thus, the longest steady-state waiting time satisfies the lower bound

$$\max_{i\le N} W_i(\infty) \ge_{st.} \max_{i\le N}\sup_{0\le k\le l}\sum_{j=1}^{k}(S_i^{(1)}(j)S^{(2)}(j)-A(j))$$

with $l>0$. Thus, by using the convergence result in (5.2.17) in Theorem 5.1, we know that

$$\begin{aligned}\liminf_{N\to\infty}\mathbb{P}\left(\max_{i\le N} W_i(\infty)>xc_N\right) &\ge \lim_{N\to\infty}\mathbb{P}\left(\max_{i\le N}\tilde{W}_i(Mc_N)>xc_N\right)\\ &=\mathbb{P}\left(\sup_{t\in[0,M]}(X_t-\mu t)>x\right)\\ &\overset{M\to\infty}{\longrightarrow}\mathbb{P}\left(\sup_{t>0}(X_t-\mu t)>x\right).\end{aligned}$$

The last limit follows from the monotone convergence theorem. Thus, we have a tight lower bound.

We now want to find a tight upper bound for the tail probability of the longest steady-state waiting time. We have that

$$\begin{aligned}&\mathbb{P}\left(\max_{i\le N} W_i(\infty)>xc_N\right)\\ &=\mathbb{P}\left(\max_{i\le N}\sup_{k\ge 0} R_i(k)>xc_N\right)\\ &=\mathbb{P}\left(\max\left(\max_{i\le N}\tilde{W}_i(Mc_N),\max_{i\le N}\sup_{t>M}R_i(tc_N)\right)>xc_N\right)\\ &\le\mathbb{P}\left(\max_{i\le N}\tilde{W}_i(Mc_N)>xc_N\right)+\mathbb{P}\left(\max_{i\le N}\sup_{t>M}R_i(tc_N)>xc_N\right).\end{aligned}\tag{5.6.1}$$

For the first term in (5.6.1), we obtain that

$$\begin{aligned}\mathbb{P}\left(\max_{i\le N}\tilde{W}_i(Mc_N)>xc_N\right)&\overset{N\to\infty}{\longrightarrow}\mathbb{P}\left(\sup_{t\in[0,M]}(X_t-\mu t)>x\right)\\ &\overset{M\to\infty}{\longrightarrow}1-\exp\left(-\frac{1}{\mu(\beta-1)x^{\beta-1}}\right).\end{aligned}$$

Thus, we need to prove that the second term in (5.6.1) asymptotically vanishes when $N,M\to\infty$. Let $\hat{S}_i^{(1)}(j)$ and $\hat{S}^{(2)}(j)$ be independent copies of $S_i^{(1)}(j)$ and $S^{(2)}(j)$, respectively. Then we can bound the second term in (5.6.1) as follows:

$$\mathbb{P}\left(\max_{i\le N}\sup_{t>M}R_i(tc_N)>xc_N\right)$$

$$
\begin{aligned}
&= \mathbb{P}\left(\max_{i\le N} \left(R_i(Mc_N) + \sup_{k\ge 0} \sum_{j=1}^{k} (\hat{S}_i^{(1)}(j)\hat{S}^{(2)}(j) - \hat{A}(j)) \right) > xc_N \right) \\
&\le \mathbb{P}\left(\max_{i\le N} R_i(Mc_N) + \max_{i\le N} \sup_{k\ge 0} \sum_{j=1}^{k} (\hat{S}_i^{(1)}(j)\hat{S}^{(2)}(j) - \hat{A}(j)) > xc_N \right) \qquad (5.6.2) \\
&\le \mathbb{P}\left(\max_{i\le N} R_i(Mc_N) > -\frac{\mu}{2} Mc_N \right) \\
&\quad + \mathbb{P}\left(\max_{i\le N} \sup_{k\ge 0} \sum_{j=1}^{k} (\hat{S}_i^{(1)}(j)\hat{S}^{(2)}(j) - \hat{A}(j)) > \left(x + \frac{\mu}{2}M\right) c_N \right). \qquad (5.6.3)
\end{aligned}
$$

The bound in (5.6.2) holds because $\max_{i\le N}(a_i + b_i) \le \max_{i\le N} a_i + \max_{i\le N} b_i$, the bound in (5.6.3) follows from the union bound. For the first term in (5.6.3), we have that

$$
\mathbb{P}\left(\max_{i\le N} R_i(Mc_N) > -\frac{\mu}{2} Mc_N \right) \overset{N\to\infty}{\longrightarrow} 1 - \exp\left(\frac{-M}{(\mu M/2)^\beta} \right) \overset{M\to\infty}{\longrightarrow} 0.
$$

In order to analyze the second term in (5.6.3), we use the fact that $\mathbb{E}[S_i^{(1)}(j)S^{(2)}(j) - A(j)] = -\mu < 0$. From this, it follows that there exists a $\gamma > 1$, such that $\mathbb{E}[S_i^{(1)}(j)S^{(2)}(j)] < \mathbb{E}[A(j)]/\gamma$. We write $\gamma\mathbb{E}[S_i^{(1)}(j)S^{(2)}(j)] - \mathbb{E}[A(j)] = -\mu_\gamma < 0$. Then,

$$
\begin{aligned}
&\mathbb{P}\left(\max_{i\le N} \sup_{k\ge 0} \sum_{j=1}^{k} (\hat{S}_i^{(1)}(j)\hat{S}^{(2)}(j) - \hat{A}(j)) > \left(x + \frac{\mu}{2}M\right) c_N \right) \\
&\le \mathbb{P}\left(\max_{i\le N} \sup_{k\in[0,\lfloor c_N \rfloor]} \sum_{j=1}^{k} (\hat{S}_i^{(1)}(j)\hat{S}^{(2)}(j) - \hat{A}(j)) > \left(x + \frac{\mu}{2}M\right) c_N \right) \\
&\quad + \sum_{n=0}^{\infty} \mathbb{P}\left(\max_{i\le N} \sup_{k\in[\lfloor \gamma^n c_N \rfloor, \lfloor \gamma^{n+1} c_N \rfloor]} \sum_{j=1}^{k} (\hat{S}_i^{(1)}(j)\hat{S}^{(2)}(j) - \hat{A}(j)) > \left(x + \frac{\mu}{2}M\right) c_N \right) \\
&\le \mathbb{P}\left(\max_{i\le N} \sup_{k\in[0,\lfloor c_N \rfloor]} \sum_{j=1}^{k} (\hat{S}_i^{(1)}(j)\hat{S}^{(2)}(j) - \hat{A}(j)) > \left(x + \frac{\mu}{2}M\right) c_N \right) \\
&\quad + \sum_{n=0}^{\infty} \mathbb{P}\left(\max_{i\le N} \sum_{j=1}^{\lfloor \gamma^{n+1} c_N \rfloor} \hat{S}_i^{(1)}(j)\hat{S}^{(2)}(j) - \sum_{j=1}^{\lfloor \gamma^n c_N \rfloor} \hat{A}(j) > \left(x + \frac{\mu}{2}M\right) c_N \right) \\
&\overset{N\to\infty}{\longrightarrow} \mathbb{P}\left(\sup_{t\in[0,1]} (X_t - \mu t) > x + \frac{\mu}{2}M \right) + \sum_{n=0}^{\infty} \left(1 - \exp\left(-\frac{\gamma^{n+1}}{(x + \mu M/2 + \gamma^n \mu_\gamma)^\beta} \right) \right). \qquad (5.6.4)
\end{aligned}
$$

It is clear that $\mathbb{P}(\sup_{t\in[0,1]}(X_t-\mu t)>x+\mu M/2)\overset{M\to\infty}{\longrightarrow}0$. The sum in (5.6.4) is finite and also converges to 0 as $M\to\infty$, as the ratio test gives us that

$$\lim_{n\to\infty}\frac{\left(1-\exp\left(-\gamma^{n+2}/(x+\mu M/2+\gamma^{n+1}\mu_\gamma)^\beta\right)\right)}{\left(1-\exp\left(-\gamma^{n+1}/(x+\mu M/2+\gamma^{n}\mu_\gamma)^\beta\right)\right)}=\frac{1}{\gamma^{\beta-1}}<1.$$

Hence, we can choose for all $\epsilon>0$ a K large enough such that

$$\sum_{n=K}^{\infty}\left(1-\exp\left(-\frac{\gamma^{n+1}}{(x+\mu M/2+\gamma^n\mu_\gamma)^\beta}\right)\right)<\epsilon,$$

and it is obvious that

$$\sum_{n=0}^{K}\left(1-\exp\left(-\frac{\gamma^{n+1}}{(x+\mu M/2+\gamma^n\mu_\gamma)^\beta}\right)\right)\overset{M\to\infty}{\longrightarrow}0.$$

Thus, we can conclude that both terms in (5.6.4) converge to 0 as $M\to\infty$, and consequently, both terms in (5.6.3) asymptotically vanish. Returning to the upper bound for the steady-state tail probability of the longest waiting time given in (5.6.1), we can conclude that

$$\limsup_{N\to\infty}\mathbb{P}\left(\max_{i\leq N}W_i(\infty)>xc_N\right)\leq\mathbb{P}\left(\sup_{t>0}(X_t-\mu t)>x\right).$$

□

We have proven process convergence of the maximum transient waiting time and we have proven steady state convergence. The limiting processes have the form of a supremum of Fréchet-distributed random variables with a negative drift. We now give an explicit expression of the cumulative distribution function.

Proof of Proposition 5.1. To prove Equation (5.2.20), we provide sharp lower and upper bounds of $\mathbb{P}(\sup_{t>0}(X_t-\mu t)<x)$. First, let $\delta>0$. We have that

$$\mathbb{P}(X_\delta-\mu\delta<x)=\exp\left(-\frac{\delta}{(x+\mu\delta)^\beta}\right).$$

Obviously, we can bound $\mathbb{P}\big(\sup_{t>0}(X_t-\mu t)<x\big)$ from above as

$$\mathbb{P}\left(\sup_{t>0}(X_t-\mu t)<x\right)<\mathbb{P}\big(\cap_{i=1}^{\infty}X_{i\delta}-\mu i\delta<x\big).$$

We can write $X_{2\delta}=\max(\hat{X}_\delta,X_\delta)$, with $\hat{X}_\delta$ an independent copy of X_δ. From this relation we know that, if $X_\delta-\delta<x$, then $X_{2\delta}-2\delta<x$ if and only if $\hat{X}_\delta-2\delta<x$. Therefore,

$$\mathbb{P}(X_\delta-\delta<x,X_{2\delta}-2\delta<x)=\mathbb{P}(X_\delta-\delta<x,\hat{X}_\delta-2\delta<x)$$

$$= \mathbb{P}(X_\delta - \delta < x)\,\mathbb{P}(\hat{X}_\delta - 2\delta < x).$$

Thus, in general, the cumulative distribution function of $\sup_{t>0}(X_t - \mu t)$ is bounded from above as

$$\mathbb{P}\left(\sup_{t>0}(X_t - \mu t) < x\right) < \mathbb{P}\big(\cap_{i=1}^{\infty} X_{i\delta} - \mu i\delta < x\big) = \prod_{i=1}^{\infty} \exp\left(-\frac{\delta}{(x+\mu i\delta)^\beta}\right). \tag{5.6.5}$$

We can find a lower bound as well. Because both X_t and μt are non-decreasing in t, we know that $\sup_{s\in((i-1)\delta,i\delta]}(X_s - \mu s) \le X_{i\delta} - \mu(i-1)\delta$. Therefore,

$$\mathbb{P}\left(\sup_{t>0}(X_t - \mu t) < x\right) = \mathbb{P}\left(\cap_{i=1}^{\infty} \sup_{s\in((i-1)\delta,i\delta]} (X_s - \mu s) < x\right)$$
$$> \mathbb{P}\big(\cap_{i=1}^{\infty} X_{i\delta} - \mu(i-1)\delta < x\big).$$

With a similar derivation as before, we have that

$$\mathbb{P}\big(\cap_{i=1}^{\infty} X_{i\delta} - \mu(i-1)\delta < x\big) = \prod_{i=1}^{\infty} \exp\left(-\frac{\delta}{(x+\mu(i-1)\delta)^\beta}\right).$$

Now, we can rewrite this expression as

$$\prod_{i=0}^{\infty} \exp\left(-\frac{\delta}{(x+\mu i\delta)^\beta}\right) = \exp\left(-\frac{\delta}{(\mu\delta)^\beta}\sum_{i=0}^{\infty}\frac{1}{\big(x/(\mu\delta)+i\big)^\beta}\right)$$
$$= \exp\left(-\frac{\delta}{(\mu\delta)^\beta}\zeta\left(\beta, \frac{x}{\mu\delta}\right)\right),$$

where $\zeta(\beta, x)$ is the Hurwitz zeta function [3, Eq. (1.10)]. We have that

$$\lim_{\delta\downarrow 0} \frac{\delta}{(\mu\delta)^\beta}\zeta\left(\beta, \frac{x}{\mu\delta}\right) = \frac{1}{\mu(\beta-1)x^{\beta-1}},$$

which follows directly from [76, Thm. 2]. The same limit holds for the upper bound in (5.6.5); thus Equation (5.2.20) follows.

The proof of Equation (5.2.21) is analogous and follows from the fact that

$$\lim_{\delta\downarrow 0} \frac{\delta}{(\mu\delta)^\beta}\sum_{i=0}^{\lfloor t/\delta\rfloor}\frac{1}{\big(x/(\mu\delta)+i\big)^\beta} = \lim_{\delta\downarrow 0}\frac{\delta}{(\mu\delta)^\beta}\left(\zeta\left(\beta, \frac{x}{\mu\delta}\right) - \zeta\left(\beta, \frac{x}{\mu\delta} + \left\lfloor\frac{t}{\delta}\right\rfloor + 1\right)\right)$$
$$= \frac{1}{\mu^\beta(\beta-1)}\left(\frac{1}{(x/\mu)^{\beta-1}} - \frac{1}{(x/\mu+t)^{\beta-1}}\right).$$

□

5.7. Other results

In this section, we deviate from our original model and present some other results on heavy-tailed parallel-server systems. The first result we derive is a convergence result of the longest steady-state waiting time with i.i.d. regularly varying service times. In contrast to the fork-join queue described in Section 5.2, the service times are mutually independent between the different servers.

Proposition 5.2. *Let $(S_i(j), i \geq 1, j \geq 1)$ be a sequence of i.i.d. regularly varying random variables, i.e., $\mathbb{P}(S_i(j) > x) = \ell(x)/x^{\beta}$, with $\ell(x)$ a slowly varying function. Moreover, we assume that $\beta > 1$. We also define a slowly varying function $\tilde{\ell}(x)$ which satisfies that $((x^{\beta-1}/\ell(x))^{\leftarrow})^* = \tilde{\ell}(x)x^{1/(\beta-1)}$. Furthermore, let $(A(j), j \geq 1)$ be a sequence of i.i.d. random variables with the property that $A(j)$ and $S_i(k)$ are independent for all i, j, and k, and $\mathbb{E}[S_i(j) - A(j)] < 0$. Then*

$$\mathbb{P}\left(\max_{i\leq N}\sup_{k\geq 0}\sum_{j=1}^{k}(S_i(j) - A(j)) < x\tilde{\ell}(N)N^{1/(\beta-1)}\right) \overset{N\to\infty}{\longrightarrow} \exp\left(-\frac{1}{x^{\beta-1}\mathbb{E}[A(1) - S_i(1)](\beta-1)}\right). \quad (5.7.1)$$

Proof. Let $\epsilon > 0$ such that $\mathbb{E}[S_i(j) - (1-\epsilon)A(j)] < 0$. Then, we have by the subadditivity property of the sup operator that

$$\max_{i\leq N}\sup_{k\geq 0}\sum_{j=1}^{k}(S_i(j) - A(j)) \leq \max_{i\leq N}\sup_{k\geq 0}\sum_{j=1}^{k}(S_i(j) - (1-\epsilon)\mathbb{E}[A(j)]) + \sup_{k\geq 0}\sum_{j=1}^{k}((1-\epsilon)\mathbb{E}[A(j)] - A(j)).$$

Following [38, Thm. 1] and [116, Thm. 7.6], we have that

$$\mathbb{P}\left(\sup_{k\geq 0}\sum_{j=1}^{k}(S_i(j) - (1-\epsilon)\mathbb{E}[A(j)]) > x\right) \sim \frac{1}{\mathbb{E}[(1-\epsilon)A(1) - S_i(1)](\beta-1)}x\,\mathbb{P}\big(S_i(1) > x\big) = \frac{1}{\mathbb{E}[(1-\epsilon)A(1) - S_i(1)](\beta-1)}\frac{\ell(x)}{x^{\beta-1}},$$

as $x \to \infty$. From [132, Prop. 2.6 (v,vi,vii)] and [29, Thm. 1.5.12], we know that we can find a slowly varying function $\tilde{\ell}(x)$ such that $((x^{\beta-1}/\ell(x))^{\leftarrow})^* = \tilde{\ell}(x)x^{1/(\beta-1)}$. From this, it follows that

$$N\,\mathbb{P}\left(\sup_{k\geq 0}\sum_{j=1}^{k}(S_i(j) - (1-\epsilon)\mathbb{E}[A(j)]) > x\tilde{\ell}(N)N^{1/(\beta-1)}\right)$$

$$\xrightarrow{N\to\infty} \frac{1}{x^{\beta-1}\mathbb{E}[(1-\epsilon)A(1)-S_i(1)](\beta-1)}.$$

Therefore,

$$\mathbb{P}\left(\max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}(S_i(j)-(1-\epsilon)\mathbb{E}[A(j)])<x\tilde{\ell}(N)N^{1/(\beta-1)}\right)$$
$$\xrightarrow{N\to\infty} \exp\left(-\frac{1}{x^{\beta-1}\mathbb{E}[(1-\epsilon)A(1)-S_i(1)](\beta-1)}\right).$$

Furthermore, it is easy to see that

$$\frac{\sup_{k\ge 0}\sum_{j=1}^{k}((1-\epsilon)\mathbb{E}[A(j)]-A(j))}{\tilde{\ell}(N)N^{1/(\beta-1)}} \xrightarrow{\mathbb{P}} 0,$$

as $N\to\infty$. Now, we can conclude that

$$\liminf_{N\to\infty}\mathbb{P}\left(\max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}(S_i(j)-A(j))<x\tilde{\ell}(N)N^{1/(\beta-1)}\right)$$
$$\ge \exp\left(-\frac{1}{x^{\beta-1}\mathbb{E}[(1-\epsilon)A(1)-S_i(1)](\beta-1)}\right)$$
$$\xrightarrow{\epsilon\downarrow 0} \exp\left(-\frac{1}{x^{\beta-1}\mathbb{E}[A(1)-S_i(1)](\beta-1)}\right).$$

With a similar analysis, we see that the lower bound

$$\max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}(S_i(j)-A(j))$$
$$\ge \max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}(S_i(j)-(1+\epsilon)\mathbb{E}[A(j)])+\inf_{k\ge 0}\sum_{j=1}^{k}((1+\epsilon)\mathbb{E}[A(j)]-A(j))$$

is asymptotically sharp. □

After replacing the common arrival sequence $(A(j), j\ge 1)$ in Proposition 5.2 with N i.i.d. arrival sequences $(A_i(j), i\ge 1, j\ge 1)$, the result still holds. We prove this in Proposition 5.3.

Proposition 5.3. *Let $(S_i(j), i\ge 1, j\ge 1)$ be a sequence of i.i.d. regularly varying random variables, i.e., $\mathbb{P}(S_i(j)>x)=\ell(x)/x^{\beta}$, with $\ell(x)$ a slowly varying function. Moreover, we assume that $\beta>1$. We also define a slowly varying function $\tilde{\ell}(x)$ which satisfies that $((x^{\beta-1}/\ell(x))^{\leftarrow})^{*}=\tilde{\ell}(x)x^{1/(\beta-1)}$. Furthermore, let $(A_i(j), i\ge 1, j\ge 1)$ be a sequence of i.i.d. random variables with the property that $A_i(j)$ and $S_k(l)$ are independent for all i, j, k, and*

l, and $\mathbb{E}[S_i(j) - A_i(j)] < 0$. *Then*

$$\mathbb{P}\left(\max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}(S_i(j)-A_i(j)) < x\tilde{\ell}(N)N^{1/(\beta-1)}\right) \overset{N\to\infty}{\longrightarrow} \exp\left(-\frac{1}{x^{\beta-1}\mathbb{E}[A_i(1)-S_i(1)](\beta-1)}\right). \quad (5.7.2)$$

Proof. Let $\epsilon > 0$ such that $\mathbb{E}[S_i(j) - (1-\epsilon)A_i(j)] < 0$. Then, we have by the subadditivity property of the sup operator that

$$\max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}(S_i(j)-A_i(j)) \le \max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}(S_i(j)-(1-\epsilon)\mathbb{E}[A_i(j)]) + \max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}((1-\epsilon)\mathbb{E}[A_i(j)]-A_i(j)).$$

The analysis of the first term on the right-hand side is identical to the analysis in the proof of Proposition 5.2.

For the second term on the right-hand side, we observe that the random variable $(1-\epsilon)\mathbb{E}[A_i(j)] - A_i(j)$ is light-tailed. From Chapter 3, we know that $\max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}((1-\epsilon)\mathbb{E}[A_i(j)] - A_i(j))$ scales like $\log N$. Thus,

$$\frac{\max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}((1-\epsilon)\mathbb{E}[A_i(j)]-A_i(j))}{\tilde{\ell}(N)N^{1/(\beta-1)}} \xrightarrow{\mathbb{P}} 0,$$

as $N\to\infty$. With a similar analysis, we see that the lower bound

$$\max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}(S_i(j)-A_i(j)) \ge \max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}(S_i(j)-(1+\epsilon)\mathbb{E}[A_i(j)]) + \inf_{k\ge 0}\sum_{j=1}^{k}((1+\epsilon)\mathbb{E}[A_i(j)]-A_{i^*}(j))$$

is asymptotically sharp, with i^* satisfying $\sup_{k\ge 0}\sum_{j=1}^{k}(S_{i^*}(j) - (1+\epsilon)\mathbb{E}[A_i(j)]) = \max_{i\le N}\sup_{k\ge 0}\sum_{j=1}^{k}(S_i(j) - (1+\epsilon)\mathbb{E}[A_i(j)])$. □

Chapter 6

Centralized optimization

6.1. Introduction

In this chapter, we consider the Brownian fork-join queue that we gave in Definition 4.1. We examine different scenarios: we distinguish between stochastic and deterministic arrivals. We use this fork-join queueing system to model key features of high-tech manufacturing. The focus lies on a specific type of assembly system where many suppliers produce components that need to be assembled into a final product. Due to delays, the system faces backorder costs. The main question is how suppliers should balance between optimizing a base stock of components and optimizing their production capacity in order to minimize the expected total costs in the system. This is an important question, as in these high-tech assembly systems, the total costs due to backorders are high [144]. This is also a non-trivial question, as the number of suppliers is usually high [12, p. 53]. The results yield new insights into the joint optimization of capacity and inventory for large-scale assembly systems.

We now provide a model description. We study an assembly system with N servers, where N is large, and the amount of demand and the number of produced components are deterministic with some random perturbation, which is assumed to be normally distributed. Thus, the total delay for one supplier in steady state can be modeled by the all-time supremum of a Brownian motion. We consider the model in which the manufacturer sends orders to all suppliers at the same time. We can therefore model the system as a fork-join queue. In such an assembly system, the total delay is determined by the slowest supplier, since a final product can only be assembled when all components are finished. Thus, we model the total delay with the maximum queue length in the fork-join queue.

Next, we give an overview of our results. When interarrival times are deterministic, the cumulative distribution function of the maximum queue length has a complicated structure due to the fact that N is large. We derive a simple approximation of this cumulative distribution function using basic extreme-value results. In the case that interarrival times

This chapter is based on [107].

are stochastic, the cumulative distribution function of the maximum queue length is not known, due to the dependence among different queues. We therefore use the second-order convergence result from Corollary 3.2 to estimate the longest queue, and we give a close-to-optimal approximation of the expected total costs in the system using this approximation. Thus, we obtain a base-stock level and a capacity that minimizes an approximation of the cost function. In Theorems 6.1 and 6.2, we analyze the model with deterministic and stochastic arrivals, respectively. In these theorems, we show how much the total costs under this base-stock level and capacity differ from the minimal costs.

The work in this chapter generates new insights in fork-join queues that lead to new analytical results for an important class of assembly systems: this chapter is the first to consider simultaneous optimization of inventory and capacity in a multi-component assembly system with dependent delays. Due to the dependencies in delays, evaluating such a system with fixed capacity and inventory is already a difficult problem. One possible way to do this would be to resort to simulation, but we will see in Section 6.4.2 that, due to dependencies, simulating this system is hard. We provide several asymptotically optimal expressions for capacity and inventory that are either in closed form or can easily be computed numerically.

Our results may help high-tech equipment manufacturers (OEMs) to optimally allocate budget to capacity and inventory, to cost-efficiently ensure timely deliveries to their customers. OEMs namely spend billions of dollars on spare component production capacity and component inventories in the hope of guaranteeing a reliable production system [12]. However, despite decades of research in inventory management, the joint optimization of production capacity and inventory remains a challenge [34], and there is a lack of analytical results that may aid OEMs in analyzing the crucial trade-offs that underlie the outcome of their investments. Indeed, while the topic has increasingly been studied, see for example [131], the focus of analysis has been on problems with a single component. We consider the much more common situation of assembling a system from many components, and we aim to choose capacity and inventory levels that minimize the sum of holding, capacity, and backorder costs.

We explore repercussions of our results for OEMs, for example Airbus and ASML. The production system of these OEMs consists of roughly two stages: 1) Component production; and 2) assembly/integration of components. This setup is crucial to enable the modular design, production and testing of components, and substantial value is added in both stages. For these reasons system integration is only initiated after customers have committed to purchasing the system. We consider a manufacturing system in which a manufacturer assembles a final product from N common components, where N is a large number, meaning that all components are required whenever a product is assembled. Each component is produced on a single production line that involves highly skilled staff and specialized equipment. In anticipation of uncertain demand, an inventory buffer is built up: production continues until a target inventory position is reached, after which production is switched off until the inventory position drops below this target. Such base-stock policies are widely used for modeling component inventories, e.g. [6, 30, 77]. Also in a high-tech manufacturing environment, where capacity mainly refers to people working in cleanrooms

that can be at work or have a day off instead of expensive machines with high start-up costs, such policies are suitable. Despite these inventory buffers, random delays may occur in the production process for each of the components.

6.1.1 Literature review

Simultaneous optimization of capacity and inventory is an important problem in supply chain management, but the literature on this topic is limited due to the complexity of the problem [34]. Considering the interaction between a manufacturer and a single supplier, Chaturvedi and Martínez-de-Albéniz [36] discuss the trade-off between inventory and capacity and how properly diversifying supply sources can reduce inventory and capacity investments. Sleptchenko et al. [143] study simultaneous optimization of spare-part inventory and repair capacity. In the last decade, simultaneous optimization of capacity and inventory in a single supplier-manufacturer relationship has been studied increasingly, cf. [130, 131]. Reed and Zhang [131] show that the square-root staffing rule of [69] is a valuable tool in optimizing inventory and capacity in a multi-server make-to-stock queue. Altendorfer and Minner [7] study simultaneous optimization of inventory and planned lead-time and [106] study the joint optimization of inventory and temporarily available additional capacity. Our work differs fundamentally from these studies, as we consider the assembly of multiple components that face the same (stochastic) demand.

The literature concerning simultaneous optimization of capacity and inventory in single-sourced assembly (or assembly-to-order) systems with multiple components is also limited. Zou et al. [160] study how supply chain efficiency can be increased by synchronizing processing times and delivery quantities. Pan and So [121] consider the simultaneous optimization of component prices and production quantities in a two-supplier setting where one supplier has uncertainty in the yield. Our main contribution compared to the work of [160] and [121] is that we provide approximations of the optimal capacity and base-stock levels that only require two moments.

Our work is related to [63], who provides approximations for setting base-stock levels in single-stage and multi-stage systems that are asymptotically exact as the target service level or the backorder penalty becomes large. For single-product-lost-sales inventory systems under periodic review, Huh et al. [73] show that order-up-to policies are asymptotically optimal when the lost-sales penalty is large compared to the holding cost. Bijvank et al. [27] show the robustness of this result when using the optimal base-stock levels of the corresponding backorder system instead of those of the lost-sales system. The asymptotic analysis in this chapter has also been influenced by related problems for queues with many servers, inspired by agent staffing problems in call centers; we refer to [33, 61] and [95] for background.

We use the Brownian fork-join queue to model delays in large-scale assembly systems. Brownian motion models are common in the literature on inventory control. Optimal control of inventory that can be described by a Brownian motion is described in [72, Par. 7], in which optimality conditions for both discounted and average cost criteria are provided. Closely related to our work is the Brownian motion model presented in [34, Par. 3] to study the

trade-off between capacity and inventory. They provide closed-form approximations to the optimal capacity and base-stock levels in a system with a single item. We consider an assembly system in which multiple components are merged into one end-product. This is an essential difference, since in our model inventory does not only buffer against uncertain demand, but a component may also need to be stored when other components are not yet available.

6.1.2 Overview of results

Now, we give an overview of the results that we obtained in this chapter. First, we investigate a base model, in which arrivals are not stochastic. Thus, we analyze a fork-join queue with a deterministic arrival stream. Extremes for this network as $N \to \infty$ are obtained using extreme-value theory. Based on those results, in Section 6.3 we derive easy-to-calculate expressions for capacity and inventory that are asymptotically optimal as the number of components grows large. We provide order bounds between the costs under optimal and approximate inventory and capacity. In particular, inspired by the literature on call centers [33, 61, 95], we distinguish three regimes, which depend on the growth rates of cost parameters and are determined by the probability γ_N of not having enough inventory. Given that $\gamma_N \to \gamma$, we say that the regime is *balanced* if $\gamma \in (0, 1)$. We are in the quality-driven regime if $\gamma = 0$ and in the efficiency-driven regime if $\gamma = 1$. For the base model, we establish asymptotic cost optimality in all three regimes. For the balanced, quality-driven, and efficiency-driven regimes, we have convergence rates of $1/(N \log N), \gamma_N/(N \log(N/\gamma_N))$ and $1/\log N$ respectively.

Next, we analyze the model in which arrivals are stochastic. In Section 6.4, we assume that the cumulative stochastic demand for systems is modeled by a Brownian motion; see [34] for a single-component manufacturing system. This implies that the demand over any finite time period is a normal variable, which is a standard assumption in literature, cf. [15, 85]. In high-tech manufacturing, normally distributed demand is a suitable assumption, especially when considering longer time periods, but it is also a reasonable approximation for shorter periods. As a consequence of these demand variations, component delays become *dependent*, since they face the same stochastic demands from system assembly. The question is now how this affects the maximum delay as the number of queues/components $N \to \infty$. We use the second-order convergence results that we derived in Corollary 3.2 in Chapter 3 as an approximation, and we especially use the convergence result in (3.2.8). This implies that, with proper scaling of holding and backorder costs, the optimal inventory for stochastic demand converges to a scaled version of the quantile function of the normal distribution, while this quantile function also appears in the limit of the optimal capacity.

Further, we test the validity of the approximation from Corollary 3.2 in Chapter 3 through simulations. In Section 6.4.2, numerical experiments show that we typically are most of the times 10% off the optimum (e.g., when N is in the range from 10 to 100); see Tables 6.6 and 6.7. Naturally, the difference goes to 0 as $N \to \infty$; see Theorem 6.2. We give an improvement of this approximation by combining our results for deterministic demand and stochastic demand. Based on this approximation, we optimize the capacity

and inventory decisions and we test the quality of these approximations through numerical experiments. It turns out that these approximations perform well already when considering a limited number of components, and are typically less than 2% off the optimum.

The remainder of this chapter is organized as follows. We introduce the general mathematical model in Section 6.2.1. As we distinguish between two cases; deterministic and stochastic arrivals, we first present general results that hold for both cases in Section 6.2.2. We study the assembly system with deterministic demand in Section 6.3. We provide explicit expressions and approximations for optimal inventory and capacity. The stochastic demand case, with solutions to the minimization problem and convergence results, is studied in more detail in Section 6.4. A refinement of the approximations from Section 6.4 is provided in Section 6.5, where we combine the lessons learned in Sections 6.3 and 6.4 to obtain better approximations for optimal capacity and inventory. In Section 6.6, we briefly touch upon the case of asymmetric systems and demonstrate that even in these settings our result for symmetric systems remains useful. We give a summary and conclusions in Section 6.7.

6.2. Model

In Section 6.2.1, we first define in Definition 6.1 the queueing process that models the delays in a high-tech assembly system. We then define the variables denoting the base-stock levels and the capacities, and the total inventory per server in steady state. We do this in Definition 6.2. In Definition 6.3, we present the resulting cost function. This cost function depends on the holding and backorder costs per item. In this chapter, we assume that these costs per item are the same for each server and can depend on N. These costs per item are given by h_N and b_N, and we show that the asymptotic behavior of the ratio $Nh_N/(Nh_N+b_N)$ determines three possible regimes. We therefore define a sequence $(\gamma_N, N \geq 1)$ with $\gamma_N = Nh_N/(Nh_N + b_N)$.

In Section 6.2.2, we present results on this cost function which hold for the system with deterministic and stochastic arrivals. We show in Lemmas 6.1 and 6.2 that the complexity of the minimization problem can be significantly reduced. Furthermore, we give expressions of the optimal base-stock levels, capacities, and optimal costs in Lemmas 6.3 and 6.4. In Lemmas 6.5 and 6.6 and in Corollary 6.1, we derive first-order approximations of the maximum steady-state queue length and apply these to give approximations of the optimal costs in the system.

6.2.1 Cost function

We examine the Brownian fork-join queue as defined in Definition 4.1. In Chapter 4, we considered the Brownian fork-join queue with a fixed $\sigma_A > 0$. In this chapter, we distinguish between the case $\sigma_A = 0$ and $\sigma_A > 0$. Therefore, we give a similar notation for the queue lengths per server and the maximum steady-state queue length as in Definition 4.1. Contrary to Chapter 4, in this chapter, we only focus on steady-state behavior. Thus, we only give the definition of the steady-state queue length.

Definition 6.1. *The sequence $(B_i, i \leq N)$ is a sequence of i.i.d. Brownian motions with standard deviation σ, $(B_A(t), t \geq 0)$ is a Brownian motion with standard deviation σ_A,*

$(B_i(t), t \geq 0)$ and $(B_A(t), t \geq 0)$ are mutually independent for all i, and the drift parameter β_i is positive, then the steady-state queue length in front of server i is given by

$$Q_i^{\beta_i,\sigma_A} := \sup_{s>0}(B_i(s) + B_A(s) - \beta_i s). \tag{6.2.1}$$

In the case that $\sigma_A = 0$, Equation (6.2.1) reduces to

$$Q_i^{\beta_i,0} := \sup_{s>0}(B_i(s) - \beta_i s).$$

In case that $\beta_i = \beta$ for all $i \leq N$, we write the maximum queue steady-state queue length as

$$\bar{Q}_N^{\beta,\sigma_A} := \max_{i\leq N} Q_i^{\beta,\sigma_A}. \tag{6.2.2}$$

We use the fork-join queueing system given in Definition 6.1 to define a large-scale assembly system.

Definition 6.2. *We define the following quantities:*

1. *The parameter I_i is the base-stock level for server i.*
2. *The parameter β_i is the capacity for server i.*
3. *The total inventory of server i in steady state is given by*

$$I_i - Q_i^{\beta_i,\sigma_A} + \max_{j\leq N}(Q_j^{\beta_j,\sigma_A} - I_j)^+. \tag{6.2.3}$$

Furthermore, we define in Definition 6.3 the total costs due to having backorders, due to having an inventory of components, and due to investing in capacity.

Definition 6.3. *The sequences $(h_N, N \geq 1)$ and $(b_N, N \geq 1)$ denote the holding costs and backorder costs per item, respectively, which may depend on N. Furthermore, the sequence $(\gamma_N, N \geq 1)$ is defined as*

$$\gamma_N := \frac{Nh_N}{Nh_N + b_N}. \tag{6.2.4}$$

Given the steady-state queue lengths per server given in Definition 6.1, the base-stock levels $(I_i, i \leq N)$ and capacities $(\beta_i, i \leq N)$ given in Definition 6.2, we define the expected total costs in the system; i.e., the sum of the expected holding, backorder, and capacity costs, as

$$\sum_{i=1}^{N} h_N \,\mathbb{E}\left[I_i - Q_i^{\beta_i,\sigma_A} + \max_{j\leq N}(Q_j^{\beta_j,\sigma_A} - I_j)^+\right] + b_N \,\mathbb{E}\left[\max_{i\leq N}(Q_i^{\beta_i,\sigma_A} - I_i)^+\right] + \sum_{i=1}^{N}\beta_i. \tag{6.2.5}$$

The cost function $C_N(I, \beta)$ denotes the expected total holding and backorder costs in the system when all servers have the same base-stock level I, and the same capacity β, and

equals

$$\begin{aligned} C_N(I,\beta) :=& \mathbb{E}\left[\sum_{i=1}^{N}\left[h_N(I - Q_i^{\beta,\sigma_A} + (\bar{Q}_N^{\beta,\sigma_A} - I)^+)\right] + b_N(\bar{Q}_N^{\beta,\sigma_A} - I)^+\right] \\ =& \mathbb{E}\Big[Nh_N(I - Q_i^{\beta,\sigma_A}) + (Nh_N + b_N)(\bar{Q}_N^{\beta,\sigma_A} - I)^+\Big]. \end{aligned} \tag{6.2.6}$$

Additionally, we write

$$C_N(I) := C_N(I,1). \tag{6.2.7}$$

Furthermore, we define the expected total costs in the system $F_N(I,\beta)$ when all servers have the same base-stock level I, and the same capacity β, as

$$F_N(I,\beta) := C_N(I,\beta) + \beta N. \tag{6.2.8}$$

The steady-state queue length given in (6.2.1) models the steady-state delays per supplier in a high-tech assembly system. Demand is represented by the common arrival process of jobs going to each server; each server, with independent, identical service processes, represents production of a component. Furthermore, we assume that production capacity in every finite time interval is normally distributed, meaning that cumulative production is a Brownian motion with drift. The backorder of each component is represented by a queue of jobs that have not been served yet.

After completion of a job, the finished component is stored in a warehouse. When all servers have a finished component in their warehouse, the end-product can be assembled. This system is visualized in Figure 1.4.

We look at this system in an equilibrium state, where the total backorder is determined by the slowest supplier. We subsequently find a trade-off between investing in the base-stock buffer and investing in capacity. To efficiently satisfy demand of the end-product, we must decide how much capacity to establish for each component and how many finished components to keep in inventory. Even though it is costly to establish capacity and to hold inventory, not being able to satisfy demand gives rise to backorder costs. Therefore, we need to find capacity and inventory levels that minimize the expected total costs.

The inventory of server i consists of two parts: first, the excess supply that works as a buffer against uncertain demand; second, the committed inventory that consists of items that are committed to realized demand but put aside because other components are not yet available. I.e., the excess supply of server i is given by $(I_i - Q_i^{\beta_i,\sigma_A})^+$, with I_i the base-stock level of server i and $Q_i^{\beta_i,\sigma_A}$ the steady-state queue length. Moreover, the number of backorders for server i is equal to $(Q_i^{\beta_i,\sigma_A} - I_i)^+$, since for $Q_i^{\beta_i,\sigma_A} \leq I_i$ the shortage is compensated by inventory I_i and only the part of $Q_i^{\beta_i,\sigma_A}$ exceeding I_i represents actual backorders that cannot be satisfied. Since all components need to be available to assemble the final product, the number of backorders in the system is equal to the number of backorders of the component with the largest backlog and is thus given by $\max_{j\leq N}(Q_j^{\beta_j,\sigma_A} - I_j)^+$.

Therefore, the committed inventory of server i equals the number of backorders in the system minus its own backlog and can be expressed as $\max_{j\leq N}(Q_j^{\beta_j,\sigma_A}-I_j)^+-(Q_i^{\beta_i,\sigma_A}-I_i)^+$. The total inventory of server i in steady state is thus given by

$$(I_i-Q_i^{\beta_i,\sigma_A})^+ + \max_{j\leq N}(Q_j^{\beta_j,\sigma_A}-I_j)^+-(Q_i^{\beta_i,\sigma_A}-I_i)^+,$$

which equals the expression in (6.2.3) in Definition 6.2.

Example 6.1. *We give an example of the evolution over time of the inventory on hand, the number of completed components, the number of components in service, and the number of assembled products, given a common demand and independent service speeds.*

Time	Demand	Q_1	# finished 1	Inv. pos. 1	Q_2	# finished 2	Inv. pos. 2	# assembl. prod.
0	0	0	0	2	0	0	2	0
1	1	1	0	1	1	0	1	1
2	0	1	0	1	0	1	2	0
3	2	3	0	0	2	0	1	1
4	0	3	0	0	0	2	3	0
5	0	2	1	0	0	0	2	1
6	0	0	2	2	0	0	2	0

Table 6.1 Fork-join queue with two servers, $I=2$.

In Table 6.1, we see that at time 1 there is a demand of one product. As both servers have two components in stock, this product is assembled immediately. Afterwards, both servers aim to produce one component to get their target inventory. At time 3, there is a demand of two products. Server 1 had only one component in stock while server 2 had two components in stock. Therefore, one product is assembled. However, in order to assemble the other product, server 1 needs to produce a component first. At time 4, server 2 has completed two components; now it has three components in inventory, two components to reach the target buffer, and one component, as there is an outstanding demand of one product, and server 1 has not produced its component yet. It is easy to see that the inventory on hand at any time equals the expression in Equation (6.2.3).

The total inventory of server i in steady state is given by (6.2.3). As the holding cost per item equals h_N, the total expected holding cost in the system equals

$$\sum_{i=1}^{N} h_N\,\mathbb{E}\left[I_i-Q_i^{\beta_i,\sigma_A}+\max_{j\leq N}(Q_j^{\beta_j,\sigma_A}-I_j)^+\right].$$

Furthermore, the number of backorders for server i is equal to $(Q_i^{\beta_i,\sigma_A}-I_i)^+$. Because the delay for the manufacturer equals the delay of the slowest server, the number of backorders for the manufacturer equals $\max_{i\leq N}(Q_i^{\beta_i,\sigma_A}-I_i)^+$. The backorder cost per item equals b_N,

thus the expected backorder cost equals

$$b_N \, \mathbb{E}\left[\max_{i \leq N} (Q_i^{\beta_i, \sigma_A} - I_i)^+\right].$$

Finally, we assume that the capacity costs per server equals the drift term β_i. Therefore, the total capacity cost equals $\sum_{i=1}^{N} \beta_i$. Now, the expected total cost in the system equals the expression in (6.2.5).

When we simplify the problem and only allow the servers to choose the same capacity $\beta_i = \beta$ and the same $I_i = I$ for given β and I, then the expected total costs given in (6.2.5) simplify to $C_N(I, \beta) + \beta N$. In the centralized optimization problem, this expression is minimized with respect to I and β. In Lemma 6.1, we show that it suffices to consider symmetric solutions where both I_i and β_i are constant in i when we consider the independent random variables $(Q_i^{\beta_i,0}, i \leq N)$, or when we consider the dependent random variables $(Q_i^{\beta,\sigma_A}, i \leq N)$; thus, we minimize over only one drift parameter. For these two cases, we exploit the self-similarity property of Brownian motions, which makes it more convenient to simplify $C_N(I, \beta)$. Due to the self-similarity of Brownian motion, we can write

$$\beta \max_{i \leq N} \sup_{s>0} (B_i(s) - \beta s) = \beta \max_{i \leq N} \sup_{t>0} \left(B_i\left(\frac{t}{\beta^2}\right) - \beta \frac{t}{\beta^2} \right) \stackrel{d}{=} \max_{i \leq N} \sup_{t>0} (B_i(t) - t).$$

This means that $\bar{Q}_N^{\beta,0} \stackrel{d}{=} \frac{1}{\beta} \bar{Q}_N^{1,0}$. Therefore, after rescaling the variable I, we can write

$$\min_{(I,\beta)} \left(C_N(I, \beta) + \beta N \right) = \min_{(I,\beta)} \left(\frac{1}{\beta} C_N(I\beta, 1) + \beta N \right) = \min_{(I,\beta)} \left(\frac{1}{\beta} C_N(I, 1) + \beta N \right). \quad (6.2.9)$$

In the last part of Equation (6.2.9), I has the interpretation of the base-stock level where the net capacity $\beta = 1$. Therefore, from now on, the actual number of products in stock at time 0 equals I/β. Similarly, the actual unsatisfied demands of component i equals $Q_i^{1,0}/\beta$. This allows us to write the cost function $F_N(I, \beta)$ in Definition 6.3 as

$$F_N(I, \beta) = \frac{1}{\beta} C_N(I) + \beta N.$$

6.2.2 General results

6.2.2.1 Simplifying the minimization problem

Our goal is to solve $\min_{(I,\beta)} F_N(I, \beta)$, focusing on the case where N is large. Before we focus on this regime, we first derive some additional properties of this problem, which are valid for each N. First, we show in Lemma 6.1 that we do not lose generality with our choice of the same base-stock level I for all servers. Furthermore, we prove that when $\sigma_A = 0$, we also do not lose generality with our choice of the same net capacity β for all servers.

Lemma 6.1. *(i) In the case that $\sigma_A = 0$, when we minimize the expected total costs over N base-stock levels and N capacities, we have that the minimizing base-stock levels are all the*

same, and that the minimizing capacities are all the same. Thus;

$$\min_{(I_1\ldots,I_N),(\beta_1,\ldots,\beta_N)} \sum_{i=1}^{N} \mathbb{E}\left[h_N(I_i - Q_i^{\beta_i,0}) + \beta_i\right] + (Nh_N + b_N)\mathbb{E}\left[\max_{j\leq N}(Q_j^{\beta_j,0} - I_j)^+\right]$$
$$= \min_{(I,\beta)} \mathbb{E}\left[Nh_N(I - Q_i^{\beta,0})\right] + \beta N + (Nh_N + b_N)\mathbb{E}\left[\max_{j\leq N}(Q_j^{\beta,0} - I)^+\right].$$

(ii) In the case that $\sigma_A > 0$, when we minimize the expected total costs over N base-stock levels and the same capacity for each server, we have that the minimizing base-stock levels are all the same. Thus;

$$\min_{(I_1,I_2,\ldots,I_N),\beta} \sum_{i=1}^{N} \mathbb{E}\left[h_N(I_i - Q_i^{\beta,\sigma_A}) + \beta\right] + (Nh_N + b_N)\mathbb{E}\left[\max_{j\leq N}(Q_j^{\beta,\sigma_A} - I_j)^+\right]$$
$$= \min_{(I,\beta)} \mathbb{E}\left[Nh_N(I - Q_i^{\beta,\sigma_A})\right] + \beta N + (Nh_N + b_N)\mathbb{E}\left[\max_{j\leq N}(Q_j^{\beta,\sigma_A} - I)^+\right].$$

In the proof of this lemma, we exploit the self-similarity property of Brownian motion. The proofs of this section can be found in Section 6.8.1.

In the next lemma, we show that minimizing the total costs $F_N(I,\beta)$ over the base-stock level I and capacity β can be simplified to two separate minimization problems; one in which we only need to minimize over the base-stock level, and one in which we only need to minimize over the capacity.

Lemma 6.2. *Let $(b_N, N \geq 1)$, $(h_N, N \geq 1)$ be sequences such that $h_N > 0$ and $b_N > 0$ for all N. Let (I_N, β_N) minimize the expected total costs $F_N(I,\beta)$ given in Definition 6.3. Then the optimal base-stock level I_N minimizes $C_N(I)$ and the optimal β_N minimizes $\frac{1}{\beta}C_N(I_N) + \beta N$. Furthermore, the function $C_N(I)$ is convex with respect to I and the function $\frac{1}{\beta}C_N(I) + \beta N$ is convex with respect to β.*

6.2.2.2 Expressions for optimal quantities

Using Lemma 6.2, we can characterize the optimal net capacity and base-stock level. In Lemma 6.3, we provide expressions for the optimal net capacity and costs in terms of the optimal base-stock level, which is given in Lemma 6.4.

Lemma 6.3. *Given $I_N^* = \arg\min_I C_N(I)$, minimizing $F_N(I,\beta)$ given in Definition 6.3 with respect to β yields $\beta_N^* = \sqrt{\frac{C_N(I_N^*)}{N}}$. Furthermore, the corresponding expected costs are $F_N(I_N^*, \beta_N^*) = 2N\beta_N^* = 2\sqrt{C_N(I_N^*)N}$.*

The optimal value of I can be expressed as a quantile of the distribution of $\bar{Q}_N^{1,\sigma_A}$:

Lemma 6.4. *The optimal base-stock level I_N^* is the unique solution of*

$$\mathbb{P}(\bar{Q}_N^{1,\sigma_A} \leq I_N^*) = 1 - \gamma_N, \tag{6.2.10}$$

with γ_N given in (6.2.4).

6.2.2.3 First-order approximations

The main technical issue is that the distribution of this maximum is in general not very tractable, as the random variable $\bar{Q}_N^{1,\sigma_A}$ is a maximum of N dependent random variable. Thus, solving Equation (6.2.10) is not possible when $\sigma_A > 0$. Therefore, we consider approximations of this distribution using extreme-value theory, to analyze their quality if N is large.

To explain our ideas, we mention the following first-order approximation of $\bar{Q}_N^{1,\sigma_A}$:

Lemma 6.5. *The maximum queue length $\bar{Q}_N^{1,\sigma_A}$ satisfies the first-order approximation*

$$\frac{\bar{Q}_N^{1,\sigma_A}}{\log N} \xrightarrow{L_1} \frac{\sigma^2}{2},$$

as $N \to \infty$.

This first-order approximation is valid regardless of whether $\sigma_A = 0$ or $\sigma_A > 0$. For $\sigma_A = 0$, a proof is given in [123, Thm. 3.1, p. 888], and for $\sigma_A > 0$, the result trivially follows from Lemma 3.3. In the subsequent two sections, we consider more refined extreme-value-theory approximations covering both cases. It turns out that the second-order behavior of the maximum is qualitatively different when σ_A becomes strictly positive. This has, in turn, an impact on the structure of the optimal solution of our cost minimization problem when N grows large.

To better understand this structure, we heuristically analyze the first-order approximation of the cost minimization problem and apply it to approximate I_N^* and β_N^*. First, we use the approximation $\bar{Q}_N^{1,\sigma_A} \approx \frac{\sigma^2}{2}\log N$ to write

$$C_N(I) \approx \bar{C}_N(I) = Nh_N\left(I - \frac{\sigma^2+\sigma_A^2}{2}\right) + (Nh_N + b_N)\left(\frac{\sigma^2}{2}\log N - I\right)^+.$$

The optimal value $\bar{I}_N$ for the associated first-order minimization problem $\min_I \bar{C}_N(I)$ is given by $\bar{I}_N = \frac{\sigma^2}{2}\log N$, since $b_N > 0$. Using this approximation, we see that $C_N(\bar{I}_N) \approx \bar{C}_N(\bar{I}_N) = (1+o(1))\frac{\sigma^2}{2}Nh_N\log N$, $\bar{\beta}_N = \sqrt{\bar{C}_N(\bar{I}_N)/N} = (1+o(1))\sqrt{\frac{\sigma^2}{2}h_N\log N}$, and $F_N(\bar{I}_N,\bar{\beta}_N) \approx 2\sqrt{N}\sqrt{\frac{\sigma^2}{2}Nh_N\log N}$. These results can be made rigorous and the decision rule $\bar{I}_N$ can be shown to be asymptotically optimal, i.e., $F_N(\bar{I}_N,\bar{\beta}_N) = F_N(I_N^*,\beta_N^*)(1+o(1))$. To prove this, we need to specify how the cost parameters h_N and b_N scale with N. For this, we consider three regimes. These regimes relate to the quantile $1-\gamma_N$ of $\bar{Q}_N^{1,0}$ at which I_N^* attains its optimal solution, with γ_N given in (6.2.4). Assume that $1-\gamma_N$ converges to a constant $1-\gamma$. We classify the three regimes in a similar way as is done in the analysis of large call centers; cf. [33]:

- we are in the *balanced regime* if $\gamma \in (0,1)$,

- if $\gamma = 0$, for large systems, the inventory is always sufficiently high to ensure that the manufacturer can assemble the end-product. We call this the *quality-driven regime*,
- finally, if $\gamma = 1$, inventories are much lower, and we call this the *efficiency-driven regime*.

When we are in the balanced or efficiency-driven regime, we can prove how far the costs under the first-order approximation are from the real optimal costs. This is established in Lemma 6.6.

Lemma 6.6. *Given that γ_N in (6.2.4), either satisfies $\gamma_N = \gamma \in (0,1)$ or $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$, then*

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\bar{I}_N, \bar{\beta}_N)} = 1 - o(1).$$

In the next two sections, we carry out a more elaborate program using more refined extreme-value estimates of $\bar{Q}_N^{1,0}$. This analysis gives sharper bounds than those given in Lemma 6.6. In particular, in the following sections, we consider the minimization in two distinct cases. First, in Section 6.3, we look at the case where demand is assumed to be deterministic, such that $\sigma_A = 0$. Thereafter, in Section 6.4, we consider the stochastic demand case. In the former case, we utilize existing results in extreme-value theory, while the latter case requires the development of a novel limit theorem. Furthermore, we use the result given in Corollary 6.1; this corollary shows how the ratio between the optimal costs and approximate costs can be represented, when the approximate base-stock level and net capacity are solutions to a minimization problem as well. This corollary follows trivially from Lemma 6.3.

Corollary 6.1. *Assume we have a function $\tilde{F}_N(I,\beta) : (0,\infty) \times (0,\infty) \to \mathbb{R}$. Furthermore, assume that the function $\tilde{F}_N$ has the form*

$$\tilde{F}_N(I,\beta) = \frac{1}{\beta}\tilde{C}_N(I) + \beta N,$$

where $\tilde{C}_N$ is a positive function with domain $(0,\infty)$. Moreover, assume that the minimum value $\tilde{F}_N(\tilde{I}_N, \tilde{\beta}_N) = 2N\tilde{\beta}_N = 2\sqrt{\tilde{C}_N(\tilde{I}_N)N}$, where $\tilde{I}_N$ and $\tilde{\beta}_N$ are minimizers, then

$$\frac{F(I_N^*, \beta_N^*)}{F(\tilde{I}_N, \tilde{\beta}_N)} = \frac{2\sqrt{C_N(I_N^*)}\sqrt{\tilde{C}_N(\tilde{I}_N)}}{C_N(\tilde{I}_N) + \tilde{C}_N(\tilde{I}_N)}.$$

6.3. The basic model: deterministic arrival stream

6.3.1 Solution and convergence of the minimization problem

We now analyze the minimization of the cost function described in Definition 6.3 for the special case with $\sigma_A = 0$ representing deterministic demand. Although we can simplify the minimization problem significantly, by using the self-similarity of Brownian motions and by writing the minimization problem as two separate minimization problems, as shown in

Lemma 6.2, the function F_N still has a difficult form, since we have the expression $\bar{Q}_N^{1,0}$ in this function. In Lemma 6.7, we give the optimal base-stock level that minimizes the costs. We assume that the holding and backorder costs per item $(h_N, N \geq 1)$ and $(b_N, N \geq 1)$ are positive sequences, and we distinguish three cases. First, we consider the balanced regime $\gamma_N = Nh_N/(Nh_N + b_N) = \gamma \in (0,1)$ for all $N > 0$. Second, we consider the quality-driven regime, where $\gamma_N \overset{N\to\infty}{\longrightarrow} 0$. Finally, we investigate the efficiency-driven regime, where $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$. All proofs for this section can be found in Section 6.8.2. We present numerical results for the three regimes in Section 6.3.2.

Lemma 6.7. *Let $Q_i^{1,0}$ be given in Definition 6.1, with $(B_i, 1 \leq i \leq N)$ independent Brownian motions with mean 0 and variance σ^2. Let $(h_N, N \geq 1)$ and $(b_N, N \geq 1)$ be positive sequences. In order to minimize $F_N(I,\beta)$ given in Definition 6.3, the optimal base-stock level I_N^* satisfies,*

$$I_N^* = P_N^{-1}(1-\gamma_N) = \frac{\sigma^2}{2}\log\left(\frac{1}{1-(1-\gamma_N)^{\frac{1}{N}}}\right), \tag{6.3.1}$$

with P_N^{-1} the quantile function of $\mathbb{P}(\bar{Q}_N^{1,0} < x)$ and γ_N given in (6.2.4).

To get a better understanding of the limiting behavior of the solution to $\min_{(I,\beta)} F_N(I,\beta)$, we would like to approximate the function F_N. Since $(Q_i^{1,0}, i \leq N)$ are independent and exponentially distributed, we know by standard extreme-value theory [67, Thm. 1.2.1, p. 19] that $\frac{2}{\sigma^2}\bar{Q}_N^{1,0} - \log N \overset{d}{\longrightarrow} G$, as $N \to \infty$, with $G \sim$ Gumbel. Therefore, for N large, $\bar{Q}_N^{1,0} \overset{d}{\approx} \frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log N$. We get a new minimization problem when we replace $\bar{Q}_N^{1,0}$ with this approximation $\frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log N$. In Definition 6.4, we give the resulting function $\hat{F}_N(I,\beta)$ that is to be minimized.

Definition 6.4. *The cost function $\hat{C}_N(I)$ denotes the approximation of the cost function $C_N(I)$ given in Definition 6.3, as we replace the random variable $\bar{Q}_N^{1,0}$ in $C_N(I)$ with $\frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log N$, where G follows a Gumbel distribution. The resulting approximation for the backorder and holding costs satisfies*

$$\hat{C}_N(I) := \mathbb{E}\left[Nh_N\left(I - Q_i^{1,0}\right) + \left(Nh_N + b_N\right)\left(\frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log N - I\right)^+\right]. \tag{6.3.2}$$

The approximation of the backorder, holding, and capacity costs in the system equals

$$\hat{F}_N(I,\beta) := \frac{1}{\beta}\hat{C}_N(I) + \beta N. \tag{6.3.3}$$

In the remainder of this section, we investigate whether minimizing $\hat{F}_N(I,\beta)$ results in costs that are close to those when we minimize $F_N(I,\beta)$. Note that we write (I_N^*, β_N^*) for the minimizers of the cost function F_N defined in Definition 6.3, and we write $(\hat{I}_N, \hat{\beta}_N)$ for the minimizers of the cost function $\hat{F}_N$ defined in Definition 6.4. Throughout this chapter, we indicate second-order approximations by the $\wedge$ symbol.

In Proposition 6.1, we present the base-stock level that minimizes $\hat{F}_N$. This base-stock level turns out to be a quantile of $\frac{\sigma^2}{2}G$ added to $\frac{\sigma^2}{2}\log N$.

Proposition 6.1 (Approximation of the optimal base-stock level)**.** *Minimizing $\hat{F}_N(I,\beta)$ given in Definition 6.4, gives the solution $(\hat{I}_N,\hat{\beta}_N,\hat{F}_N(\hat{I}_N,\hat{\beta}_N))$, with*

$$\hat{I}_N = \frac{\sigma^2}{2}\log N - \frac{\sigma^2}{2}\log\left(-\log\left(1-\gamma_N\right)\right), \tag{6.3.4}$$

and

$$\begin{aligned}\hat{C}_N(\hat{I}_N) = Nh_N&\left(\hat{I}_N - \frac{\sigma^2}{2}\right)\\ &+ (Nh_N + b_N)\frac{\sigma^2}{2}\left(\int_{-\log(1-\gamma_N)}^{\infty}\frac{e^{-t}}{t}dt + \Gamma + \log\left(-\log\left(1-\gamma_N\right)\right)\right),\end{aligned} \tag{6.3.5}$$

where $\Gamma \approx 0.577$ is Euler's constant and γ_N is given in (6.2.4)*.*

Combining Equations (6.3.4) and (6.3.5) with the results in Lemma 6.3 gives the solution $(\hat{I}_N,\hat{\beta}_N,\hat{F}_N(\hat{I}_N,\hat{\beta}_N))$.

We compare the costs under the optimal base-stock level and net capacity with the costs under the approximate base-stock level and net capacity. We distinguish the balanced regime, quality-driven regime, and efficiency-driven regime.

By using the results from Lemmas 6.11 and 6.12 in Section 6.8.2, we prove the order bounds in the balanced, quality-driven, and efficiency-driven regime in Theorem 6.1. In the efficiency-driven regime, we impose the additional condition $\gamma_N < 1-\exp(-N)$ needed to make sure that $\hat{I}_N > 0$. Namely, if we choose $\gamma_N > 1-\exp(-N)$, we get that $\hat{I}_N < 0$, which is not feasible, because $\hat{I}_N$ has the physical meaning of the number of items that need to be stored.

Theorem 6.1 (Order bounds between the optimal and approximate costs)**.** *For $(\gamma_N, N \geq 1)$ given in Definition 6.3, with $\gamma_N = \gamma \in (0,1)$, we have that*

$$\frac{F_N(I_N^*,\beta_N^*)}{F_N(\hat{I}_N,\hat{\beta}_N)} = 1 - O(1/(N\log N)). \tag{6.3.6}$$

For $(\gamma_N, N \geq 1)$ given in Definition 6.3, with $\gamma_N \overset{N\to\infty}{\longrightarrow} 0$, we have that

$$\frac{F_N(I_N^*,\beta_N^*)}{F_N(\hat{I}_N,\hat{\beta}_N)} = 1 - O(\gamma_N/(N\log(N/\gamma_N))). \tag{6.3.7}$$

For $(\gamma_N, N \geq 1)$ given in Definition 6.3, with $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$ and $\gamma_N < 1-\exp(-N)$, we have that

$$\frac{F_N(I_N^*,\beta_N^*)}{F_N(\hat{I}_N,\hat{\beta}_N)} = 1 - O(1/\log N). \tag{6.3.8}$$

Using the order bounds given in Theorem 6.1, we can establish for the three different

regimes how $F_N(I_N^*, \beta_N^*)$ scales with N as N becomes large.

Lemma 6.8. *Given γ_N in (6.2.4), if $\gamma_N = \gamma \in (0,1)$ in the balanced regime, then*

$$F_N(I_N^*, \beta_N^*) = 2\sqrt{N}\Bigg(Nh_N \frac{\sigma^2}{2}\Big(\log N - \log(-\log(1-\gamma)) - 1\Big) + (Nh_N + b_N)\frac{\sigma^2}{2}\,\mathbb{E}\Big[\Big(G + \log(-\log(1-\gamma))\Big)^+\Big]\Bigg)^{\frac{1}{2}} + O(\sqrt{h_N}/\sqrt{\log N}). \tag{6.3.9}$$

If $\gamma_N \overset{N\to\infty}{\longrightarrow} 0$ in the quality-driven regime, then

$$F_N(I_N^*, \beta_N^*) = 2\sqrt{N}\sqrt{Nh_N \frac{\sigma^2}{2}\Big(\log(N/\gamma_N) - 1\Big) + (Nh_N + b_N)\frac{\sigma^2}{2}\gamma_N} + O(\gamma_N\sqrt{h_N}/\sqrt{\log(N/\gamma_N)}). \tag{6.3.10}$$

Last, if $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$ and $\gamma_N < 1 - \exp(-N)$ in the efficiency-driven regime, then

$$F_N(I_N^*, \beta_N^*) = 2\sqrt{N}\sqrt{Nh_N \frac{\sigma^2}{2}(\log N - 1) + b_N \frac{\sigma^2}{2}\log(-\log(1-\gamma_N))} + O(N\sqrt{h_N}/\sqrt{\log N}). \tag{6.3.11}$$

The results given in Theorem 6.1 and Lemma 6.8 are obtained by using the properties stated in Lemmas 6.11 and 6.12. In Lemma 6.11, we show that we can write a Gumbel-distributed random variable that is on the same probability space as $\bar{Q}_N^{1,0}$. This gives us a very powerful result; namely, that $\bar{Q}_N^{1,0}$ and G_N are ordered and that their difference decreases as $\bar{Q}_N^{1,0}$ becomes large. Consequently, we obtain very sharp bounds on $|C_N(I_N^*) - C_N(\hat{I}_N)|$ and $|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)|$ in Lemma 6.12 which leads to sharp results in Theorem 6.1 and Lemma 6.8.

6.3.2 Numerical experiments

We now provide some numerical results to illustrate the solutions to the minimization problem and their characteristics discussed in Section 6.3.1. In all experiments, we let $\sigma = 1$ and let N vary from 10 to 1000. The results for the balanced regime, quality-driven regime, and efficiency-driven regime are given in Tables 6.2, 6.3, and 6.4, respectively. We can observe that in all regimes the approximate solutions are close to the optimal solutions. Most importantly, already for small N, the fraction of the costs corresponding to the optimal solution over the costs corresponding to the approximate solution nearly equals 1.

N	I_N^*	β_N^*	$F_N(I_N^*, \beta_N^*)$	$\hat{I}_N$	$\hat{\beta}_N$	$F_N(\hat{I}_N, \hat{\beta}_N)$	$\left(1 - \frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)}\right) N \log N$
10	1.35178	1.19648	23.9296	1.33455	1.19328	23.9315	0.001807
50	2.14273	1.49338	149.338	2.13927	1.49286	149.338	0.000379
100	2.48757	1.60499	320.997	2.48584	1.60475	320.997	0.000192
200	2.83328	1.70944	683.775	2.83242	1.70932	683.775	$9.68 \cdot 10^{-5}$
500	3.29091	1.8385	1838.5	3.29056	1.83846	1838.5	$3.91 \cdot 10^{-5}$
1000	3.63731	1.93044	3860.87	3.63713	1.93042	3860.87	$1.97 \cdot 10^{-5}$

Table 6.2 Balanced regime, $h_N = 1, b_N = N$ such that $\gamma_N = \frac{1}{2}$.

N	I_N^*	β_N^*	$F_N(I_N^*, \beta_N^*)$	$\hat{I}_N$	$\hat{\beta}_N$	$F_N(\hat{I}_N, \hat{\beta}_N)$	$\left(1 - \frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)}\right) \frac{N}{\gamma_N} \log \frac{N}{\gamma_N}$
10	2.32898	1.52962	30.5925	2.3266	1.52924	30.5925	0.000617
50	3.91708	1.97978	197.978	3.91698	1.97976	197.978	$2.52 \cdot 10^{-5}$
100	4.60768	2.14684	429.368	4.60766	2.14684	429.368	$6.31162 \cdot 10^{-6}$
200	5.29957	2.30221	920.886	5.29956	2.30221	920.886	$1.21801 \cdot 10^{-6}$
500	6.21511	2.49306	2493.06	6.21511	2.49306	2493.06	$5.51467 \cdot 10^{-6}$
1000	6.90801	2.62833	5256.66	6.90801	2.62833	5256.66	0.000176

Table 6.3 Quality-driven regime, $h_N = 1, b_N = N^2$ such that $\gamma_N = \frac{1}{1+N}$.

N	I_N^*	β_N^*	$F_N(I_N^*, \beta_N^*)$	$\hat{I}_N$	$\hat{\beta}_N$	$F_N(\hat{I}_N, \hat{\beta}_N)$	$\left(1 - \frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)}\right) \log N$
10	0.497572	3.12224	62.4448	0.386624	3.08439	62.4616	0.000797
50	0.965997	9.35451	935.451	0.927385	9.34122	935.452	$8.65678 \cdot 10^{-6}$
100	1.21527	14.4701	2894.02	1.19242	14.4615	2894.02	$1.30518 \cdot 10^{-6}$
200	1.48208	22.0864	8834.57	1.46889	22.0808	8834.57	$2.20863 \cdot 10^{-7}$
500	1.85348	38.0553	38055.3	1.84728	38.0521	38055.3	$2.51171 \cdot 10^{-8}$
1000	2.14443	56.945	113890	2.14098	56.9428	113890	$5.30189 \cdot 10^{-9}$

Table 6.4 Efficiency-driven regime, $h_N = N, b_N = 1$ such that $\gamma_N = \frac{N^2}{N^2+1}$.

6.4. Stochastic demand

We now extend our framework to the case where demand is stochastic. This means that stochasticity not only arises from the production process of the individual components but also results from uncertain demands. Consequently, delays may no longer only be caused by low production of a specific component, but may also occur when there is a sudden peak in demand. Since all components need to be available to assemble the end-product and satisfy demand, delays of the different components are now correlated. We use the same strategy when demand is stochastic as in the basic model with deterministic demand. However, we can no longer approximate the maximum queue length distribution with the Gumbel distribution. From Corollary 3.2 in Chapter 3 it follows that for N large, $\bar{Q}_N^{1,\sigma_A} \approx \frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X$ with X a standard normal random variable. Using this approximation, we obtain a new minimization problem, in which we minimize $\hat{F}_N^A(I, \beta)$ as given in Definition 6.5 with respect

to I and β.

Definition 6.5. *The cost function $\hat{C}_N^A(I)$ denotes the approximation of the cost function $C_N(I)$ given in Definition 6.3, as we replace the random variable $\bar{Q}_N^{1,\sigma_A}$ in $C_N(I)$ with $\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X$, where X follows a normal distribution. The resulting approximation for the backorder and holding costs satisfies*

$$\hat{C}_N^A(I) = \mathbb{E}\left[Nh_N\left(I - Q_i^{1,\sigma_A}\right) + \left(Nh_N + b_N\right)\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X - I\right)^+\right].$$

The approximation of the backorder, holding, and capacity costs in the system equals

$$\hat{F}_N^A(I,\beta) = \frac{1}{\beta}\hat{C}_N^A(I) + \beta N.$$

In Section 6.4.1, we elaborate on the solution and the convergence of the optimal value of $F_N(I,\beta)$ with $\sigma_A > 0$.

6.4.1 Solution and convergence of the minimization problem

We can use the convergence result proven in Corollary 3.2 to prove asymptotics of the optimal value of the function $F_N(I,\beta)$. Since $\frac{\sqrt{2}\beta}{\sigma\sigma_A}\frac{\bar{Q}_N^{\beta,\sigma_A} - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}}$ is a continuous random variable, we know that its quantile function converges to the quantile function of a standard normal random variable; see [148, Lem. 21.2, p. 305]. We can use this to derive asymptotics of the minimization problem of F_N.

Using $P_N^A(z)$ as described in Definition 6.6, we can solve the minimization problem $\min_{(I,\beta)} F_N(I,\beta)$, which yields the optimal base-stock level and net capacity given in Lemma 6.9. The proofs concerning the solution and subsequent convergence results are provided in Section 6.8.3.

Definition 6.6. *We define the cumulative distribution function of the rescaled maximum queue length as*

$$P_N^A(z) := \mathbb{P}\left(\frac{\sqrt{2}}{\sigma\sigma_A}\frac{\bar{Q}_N^{1,\sigma_A} - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \le z\right),$$

with $\bar{Q}_N^{1,\sigma_A}$ given in Definition 6.1.

Lemma 6.9. *Let $(b_N, N \ge 1)$, $(h_N, N \ge 1)$ be sequences such that $h_N > 0$ and $b_N > 0$ for all N, and γ_N given in (6.2.4). Let $\left(\beta_N^A, I_N^A\right)$ minimize $F_N(I,\beta)$ given in Definition 6.3. Then the optimal base-stock level I_N^A equals*

$$I_N^A = \frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}P_N^{A^{-1}}(1-\gamma_N)\sqrt{\log N}. \tag{6.4.1}$$

When we are in the balanced regime, we can approximate the minimization problem given in Definition 6.5, using the convergence result in Corollary 3.2, and prove how far

the approximate solution is from the optimal solution. This is done in Proposition 6.2 and Theorem 6.2. In Lemma 6.10, we show how the optimal costs scale with N when we are in the balanced regime. The proofs are given in Section 6.8.3.

Proposition 6.2. *For $(b_N, N \geq 1)$, $(h_N, N \geq 1)$ and γ_N given in* (6.2.4)*, the base-stock level $\hat{I}_N^A$ minimizes the function $\hat{F}_N^A(I, \beta)$ given in Definition 6.5, and equals*

$$\hat{I}_N^A = \frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log N} \Phi^{-1}\left(1 - \gamma_N\right). \tag{6.4.2}$$

Furthermore, the approximation of the backorder and holding cost equals

$$\begin{aligned} &\hat{C}_N^A(\hat{I}_N^A) \\ &\quad = Nh_N \left(\frac{\sigma^2}{2} \log N - \frac{\sigma^2 + \sigma_A^2}{2} \right) + (Nh_N + b_N) \frac{\sigma\sigma_A \sqrt{\log N} e^{-\frac{1}{2}\Phi^{-1}(1-\gamma_N)^2}}{2\sqrt{\pi}}. \end{aligned} \tag{6.4.3}$$

Theorem 6.2 (Order bound between the optimal and approximate costs)**.** *We assume that γ_N given in* (6.2.4) *satisfies $\gamma_N = \gamma \in (0, 1)$. Then*

$$\left| \frac{F_N(I_N^A, \beta_N^A)}{F_N(\hat{I}_N^A, \hat{\beta}_N^A)} - 1 \right| = o\left(\frac{1}{\sqrt{\log N}} \right).$$

Lemma 6.10 (Balanced regime)**.** *We assume that γ_N given in* (6.2.4) *satisfies $\gamma_N = \gamma \in (0, 1)$. Then the optimal base-stock level is given by*

$$I_N^A = \frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log N} \Phi^{-1}\left(1 - \gamma\right) + o(\sqrt{\log N}), \tag{6.4.4}$$

and the optimal cost for the system is given by

$$F_N(I_N^A, \beta_N^A) = 2\sqrt{N} \sqrt{\hat{C}_N^A(\hat{I}_N^A)} + o(N\sqrt{h_N}). \tag{6.4.5}$$

The result in Lemma 6.10 only holds for the balanced regime; a natural question is thus the performance of the system in the efficiency and the quality-driven regimes. As is shown in Lemma 6.6, in the efficiency-driven regime, the first-order approximation $\bar{I}_N = \frac{\sigma^2}{2} \log N$ gives that the ratio of the approximate costs and the optimal costs converge to 1. Thus, we expect the approximation given in (6.4.2) will also satisfy this convergence result. In order to determine whether this approximation also satisfies the order bound given in Theorem 6.2, further analysis is needed. The analysis we provide for the balanced regime heavily relies on [148, Lem. 21.2, p. 305], which says that if $Y_N \xrightarrow{d} Y$, then for $\gamma \in (0, 1)$, $P_{Y_N}^{-1}(\gamma) \overset{N\to\infty}{\longrightarrow} P_Y^{-1}(\gamma)$. This gives us the convergence result (6.4.4) of the inventory in the balanced regime. In order to be able to prove a similar result for the efficiency-driven regime, we need an improvement of [148, Lem. 21.2, p. 305] which also holds when $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$.

However, for the quality-driven regime, this convergence result does not hold, because we see in Lemma 6.8 that $I_N^A \approx \frac{\sigma^2}{2} \log(N/\gamma_N)$. In order to find a sharp order bound such as

given in Theorem 6.2 we should resort to the analysis of tail asymptotics, which is beyond the scope of this study.

6.4.2 Numerical experiments

In Section 6.4.1, we provided expressions to calculate the asymptotically optimal net capacity and base-stock level. The question remains how large the number of components has to be for these approximations to be of use. Therefore, we now examine the expected costs under both the optimal net capacity and base-stock level and under these asymptotic approximations. Since it is not straightforward to calculate $\mathbb{E}[(\bar{Q}_N^{1,\sigma_A} - I)^+]$, as $\bar{Q}_N^{1,\sigma_A}$ is a maximum of dependent random variables, to evaluate the cost function given in Definition 6.3 we resort to simulation. First, we explain the details of our simulation experiment, after which we discuss the numerical results.

In our simulation, we aim to determine the maximum queue length $\bar{Q}_N^{1,\sigma_A}$. For this, we use the algorithm proposed in [14, Par. 4.5], who describe an exact algorithm for simulating a reflected Brownian motion at the grid points. At every grid point, we draw normal random variables with the required drift and variance for the supply and demand processes and update the maximum. We use a step size of 0.001 for the grid points. Since we cannot simulate over an infinite horizon, we have to determine when to terminate the simulation. The maximum value is expected to be attained at a time which is smaller than $\hat{t} = \frac{\sigma^2+\sigma_A^2}{2} \sum_{j=1}^{N} \frac{1}{j}$. To simulate well beyond this point, we run the simulation until $t = 2\hat{t}$.

Using the above method to simulate $\bar{Q}_N^{1,\sigma_A}$, we can estimate ${P_N^A}^{-1}(1-\gamma_N)$ with $P_N^A(z)$ as described in Definition 6.6. To obtain a median-unbiased estimate of the quantile, we use the approach suggested in [159, p. 982–983]. For this, we sample $\bar{Q}_N^{1,\sigma_A}$ 100 times and randomly choose between the observations $(1-\gamma_N)\cdot 100$ and $(1-\gamma_N)\cdot 100+1$, with weights depending on the value of the fractile. Our estimate is equal to the median over 100 iterations. Once we have our estimate of ${P_N^A}^{-1}(1-\gamma_N)$, we determine the value of the optimal base-stock level as given in Equation (6.4.1). Using the optimal base-stock level we determine the optimal net capacity given in Lemma 6.3. Since this also requires the expectation of $(\bar{Q}_N^{1,\sigma_A} - I)^+$, we determine this value by taking the average based on 10,000 simulations.

Next, we compare the costs under our asymptotic approximations of the net capacity and base-stock level (provided in Proposition 6.2) to the costs under the optimal net capacity and base-stock level obtained from the simulation. We again sample $(\bar{Q}_N^{1,\sigma_A} - I)^+$ based on 10,000 new simulations and determine the costs of the different policies using cost function $F_N(I,\beta)$.

The procedure described above is applicable for N in the order of hundreds. However, it is close to impossible to provide a fast simulation for N in the order of thousands. Hence, to give a useful approximation of the optimal capacity and base-stock level in these cases, we need to use the limit we derived in Corollary 3.2.

In order to assess the performance of the approximations and their sensitivity to various model parameters, we perform a full factorial experiment. In our experiment, we vary the number of components, demand variability, and backorder costs. The setup of the experiment

is given in Table 6.5. We set the holding costs per item $h_N = 1$ and $\sigma = 1$ in all experiments. In total, we have 24 instances. The results are given in Tables 6.6 and 6.7 with the backorder costs per item $b_N = N$ and $b_N = 3N$, respectively.

Parameter	Values
Number of components N	10, 50, 100
Standard deviation of arrivals σ_A	0.1, 0.5, 0.75, 1
Backorder costs per item b_N	N, $3N$

Table 6.5 Parameter settings for experiments

N	σ_A	I_N^A	β_N^A	$F_N(I_N^A, \beta_N^A)$	$\hat{I}_N^A$	$\hat{\beta}_N^A$	$F_N(\hat{I}_N^A, \hat{\beta}_N^A)$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(\hat{I}_N^A, \hat{\beta}_N^A)}\right)\sqrt{\log N}$
10	0.1	1.327	1.1583	23.1894	1.151	0.855514	24.5143	0.0820
50	0.1	2.122	1.47611	147.534	1.956	1.25004	150.337	0.0369
100	0.1	2.455	1.58865	318.588	2.303	1.38516	322.994	0.0293
10	0.5	1.486	1.25448	25.333	1.151	0.976909	26.9363	0.0903
50	0.5	2.338	1.59412	159.934	1.956	1.3744	164.689	0.0571
100	0.5	2.715	1.71664	343.937	2.303	1.51094	352.91	0.0546
10	0.75	1.714	1.36908	27.191	1.151	1.00605	29.7614	0.1311
50	0.75	2.638	1.70591	171.443	1.956	1.41834	180.556	0.0998
100	0.75	2.980	1.83438	367.348	2.303	1.55865	383.319	0.0894
10	1	1.990	1.47358	29.8393	1.151	1.0037	34.6552	0.2109
50	1	3.006	1.84276	185.25	1.956	1.43941	201.314	0.1578
100	1	3.394	1.97602	393.668	2.303	1.58534	421.505	0.1417

Table 6.6 Comparison of costs approximate solution for $h_N = 1$, $b_N = N$

N	σ_A	I_N^A	β_N^A	$F_N(I_N^A, \beta_N^A)$	$\hat{I}_N^A$	$\hat{\beta}_N^A$	$F_N(\hat{I}_N^A, \hat{\beta}_N^A)$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(\hat{I}_N^A, \hat{\beta}_N^A)}\right)\sqrt{\log N}$
10	0.1	1.726	1.31058	25.9539	1.224	0.884692	31.2239	0.2561
50	0.1	2.533	1.5931	159.026	2.050	1.27624	173.141	0.1612
100	0.1	2.883	1.69656	341.44	2.405	1.41084	367.575	0.1526
10	0.5	2.067	1.43331	28.3311	1.513	1.0992	31.2606	0.1422
50	0.5	2.987	1.74381	173.875	2.428	1.48993	183.166	0.1003
100	0.5	3.370	1.86469	371.779	2.814	1.62542	387.809	0.0887
10	0.75	2.449	1.57036	31.4004	1.694	1.18023	35.5139	0.1758
50	0.75	3.418	1.89842	190.571	2.664	1.58369	205.174	0.1408
100	0.75	3.899	2.01955	404.306	3.070	1.72277	429.58	0.1263
10	1	2.913	1.72878	34.6096	1.875	1.23092	40.7704	0.2293
50	1	4.158	2.06968	207.553	2.899	1.65341	230.281	0.1952
100	1	4.567	2.20696	439.681	3.326	1.79761	479.663	0.1789

Table 6.7 Comparison of costs approximate solution for $h_N = 1$, $b_N = 3N$

There are several important observations to be made from Table 6.6. First, we can observe that for $N = 10$ the difference in costs between the simulated optimal solution and

the asymptotic solution is around 10% for most cases, the case $N = 10$ and $\sigma_A = 1$ is an outlier, where the difference is around 15%. As N increases to 50, the difference decreases. Furthermore, the difference becomes larger when σ increases. In the last column, we verify the convergence result from Theorem 6.2. We observe that the difference decreases as N increases, and that increasing σ_A causes the difference to increase.

When we consider the results for $b_N = 3N$ given in Table 6.7, we observe that the difference between the asymptotic and optimal costs is considerably higher than for $b_N = N$. Especially for $N = 10$, the difference is around 15% of the optimum, except for $N = 10$ and $\sigma_A = 0.1$, where the difference is around 20%. However, for a larger number of components, the difference is around 10% of the optimum. Interestingly, for the case $\sigma_A = 1$, the difference between $b_N = N$ and $b_N = 3N$ is relatively small.

Overall, in most of our experiments, the difference between the costs under the optimal base-stock level and net capacity and the costs under the approximations are around 10%. Furthermore, we can conclude that for small variations in demand and low backorder costs, the asymptotic approach performs well in terms of costs already for a reasonable number of components. Also, the performance improves by increasing N. Finally, the performance of the approximations highly depends on the backorder costs relative to the holding costs.

6.5. Mixed-behavior approximations

The numerical results in Section 6.4.2 show that the approximations are in most of the cases around 10-15% off the optimal value. In this section, we show how we can further improve the approximations.

Under deterministic demand and stochastic demand, the approximate problems are given in Definition 6.4 and Definition 6.5. If σ_A is small, then we know that on the one hand,

$$\bar{Q}_N^{1,\sigma_A} \overset{d}{\approx} \frac{\sigma^2}{2} G + \frac{\sigma^2}{2} \log N,$$

because Q_i^{1,σ_A} and Q_j^{1,σ_A} are only slightly correlated. But on the other hand,

$$\bar{Q}_N^{1,\sigma_A} \overset{d}{\approx} \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log N} X + \frac{\sigma^2}{2} \log N \approx \frac{\sigma^2}{2} \log N.$$

Since the Gumbel term is missing here, this could be the reason that this approximation is not working well for small N. Thus, it could be beneficial to look at the combination of these two approximations. Then, we have

$$\bar{Q}_N^{1,\sigma_A} \overset{d}{\approx} \frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log N} X + \frac{\sigma^2}{2} G. \tag{6.5.1}$$

When we replace $\bar{Q}_N^{1,\sigma_A}$ with Equation (6.5.1) in the minimization problem, we get

$$\min_{(I,\beta)} \left(\frac{1}{\beta} \mathbb{E}\left[N h_N (I - Q_i^{1,\sigma_A}) + (N h_N + b_N) \left(\frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X + \frac{\sigma^2}{2} G - I \right)^+ \right] + \beta N \right).$$

The optimal I_N^M satisfies $\mathbb{P}\left(\frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X + \frac{\sigma^2}{2} G < I_N^M \right) = 1 - \gamma_N$. Thus,

$$\int_{-\infty}^{\infty} \exp\left(- \exp\left(- \frac{2}{\sigma^2} \left(I_N^M - \frac{\sigma^2}{2} \log N - \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} x \right) \right) \right) \phi(x) dx = 1 - \gamma_N. \quad (6.5.2)$$

Now, I_N^M can be computed through standard numerical methods such as the bisection method. Furthermore, the optimal net capacity β_N^M satisfies

$$\beta_N^M = \frac{\sqrt{\mathbb{E}\left[N h_N (I_N^M - Q_i^{1,\sigma_A}) + (N h_N + b_N) \left(\frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X + \frac{\sigma^2}{2} G - I_N^M \right)^+ \right]}}{\sqrt{N}}. \quad (6.5.3)$$

The relevant expectations in this symbolic expression can be computed numerically; see Section 6.9 for details.

6.5.1 Numerical results for mixed-behavior approximations

Using the same simulation procedure as described in Section 6.4.2, we evaluate the performance of these adjusted approximations. The results for the cases of $h_N = 1$, $b_N = N$ and $h_N = 1$, $b_N = 3N$ are given in Tables 6.8 and 6.9, respectively.

From the simulation results, we can conclude that these adjusted approximations result in costs that are much closer to the optimal costs, already for small N. When comparing the last two columns, where the last column repeats the results from Section 6.4.2, we observe that the mixed-behavior approximations show better convergence, also when σ_A is larger. Furthermore, where we saw in Section 6.4.2 that the cost difference increased considerably with the change in b_N, we now do see an increase, but the difference is still small for a larger value of b_N. Therefore, we can conclude that these mixed-behavior approximations perform well especially when demand variations are no more than 75% of the variations in component production, even with a small number of components.

N	σ_A	I_N^M	β_N^M	$F_N(I_N^M, \beta_N^M)$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(I_N^M, \beta_N^M)}\right)\sqrt{\log N}$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(I_N^A, \beta_N^A)}\right)\sqrt{\log N}$
10	0.1	1.33785	1.1945	23.2022	0.000837	0.082011
50	0.1	2.14487	1.49567	147.567	0.000442	0.036877
100	0.1	2.49244	1.60808	318.638	0.000337	0.029273
10	0.5	1.38072	1.21129	25.4342	0.006038	0.090320
50	0.5	2.19829	1.53814	160.497	0.006938	0.057107
100	0.5	2.54871	1.65808	345.247	0.008143	0.054563
10	0.75	1.40013	1.2128	27.6956	0.027647	0.131055
50	0.75	2.216	1.56166	174.269	0.032074	0.099827
100	0.75	2.5656	1.68745	372.643	0.030493	0.089412
10	1	1.41255	1.19665	31.5428	0.081950	0.210871
50	1	2.22627	1.57136	192.722	0.076684	0.157827
100	1	2.57434	1.70384	407.343	0.072043	0.141724

Table 6.8 Comparison of costs convergence of the mixed-behavior approximation with the second-order approximation for $h_N = 1$, $b_N = N$

N	σ_A	I_N^M	β_N^M	$F_N(I_N^M, \beta_N^M)$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(I_N^M, \beta_N^M)}\right)\sqrt{\log N}$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(I_N^A, \beta_N^A)}\right)\sqrt{\log N}$
10	0.1	1.78238	1.34746	25.9965	0.002487	0.256113
50	0.1	2.59271	1.62088	159.162	0.001690	0.161243
100	0.1	2.94168	1.72533	341.49	0.000314	0.152581
10	0.5	1.94345	1.38309	28.3671	0.001926	0.142201
50	0.5	2.83775	1.68955	174.284	0.004642	0.100327
100	0.5	3.21861	1.8044	372.617	0.004826	0.088703
10	0.75	2.09429	1.41142	32.0055	0.028689	0.175760
50	0.75	3.04648	1.74512	193.854	0.033496	0.140773
100	0.75	3.44819	1.86761	410.624	0.033019	0.126256
10	1	2.25658	1.43095	36.5165	0.079240	0.229298
50	1	3.26538	1.79271	216.91	0.085321	0.195211
100	1	3.68765	1.92281	456.859	0.080689	0.178876

Table 6.9 Comparison of costs convergence of the mixed-behavior approximation with the second-order approximation for $h_N = 1$, $b_N = 3N$

6.6. Analyzing asymmetric systems

So far in this chapter, we have derived several new, analytic results for joint capacity and inventory optimization for large-scale, symmetric assembly systems. In this section, we provide an informal discussion of the application of such results in asymmetric settings.

For ease of exposition, consider a case where different components have different holding costs — for other parameters, our assumptions remain in place. In practical settings, component prices might range from a few thousand euros to hundreds of thousands of euros. Companies seeking to apply advanced methods for optimizing capacity and inventory investments typically focus on the most expensive components: for inexpensive components some coarse heuristics suffice.

Suppose the company seeks to derive separate safety stock and capacity rules for two groups of components: expensive and very expensive components. This yields $k = 2$ groups of components. We seek to apply our results on extremes as the total number of components N in these two groups grows large; we keep k and the ratio of components in the two groups fixed. Also, since we seek to derive rules at the group-level, and following Lemma 6.1, it makes sense to assume symmetry within groups, i.e., by averaging cost parameters within the groups. For example, consider the following: $N/2$ servers have a holding cost $h_N^{(1)}$ per item and $N/2$ servers have a holding cost $h_N^{(2)}$ per item. Then, we need to minimize

$$\frac{N}{2}\left(h_N^{(1)}\frac{1}{\beta_1}\left(I_1-\frac{\sigma^2}{2}\right)+\beta_1\right)+\frac{N}{2}\left(h_N^{(2)}\frac{1}{\beta_2}\left(I_2-\frac{\sigma^2}{2}\right)+\beta_2\right)$$
$$+\left(\frac{N}{2}h_N^{(1)}+\frac{N}{2}h_N^{(2)}+b_N\right)\mathbb{E}\left[\max\left(\frac{1}{\beta_1}\max_{i\le N/2}(Q_i^{1,0}-I_1),\frac{1}{\beta_2}\max_{N/2+1\le i\le N}(Q_i^{1,0}-I_2)\right)^+\right]. \tag{6.6.1}$$

over $(I_1, I_2, \beta_1, \beta_2)$. The expectation in (6.6.1) is an expectation of a maximum of N positive random variables. Therefore, we can bound

$$\mathbb{E}\left[\max\left(\frac{1}{\beta_1}\max_{i\le N/2}(Q_i^{1,0}-I_1),\frac{1}{\beta_2}\max_{N/2+1\le i\le N}(Q_i^{1,0}-I_2)\right)^+\right]$$
$$\le \mathbb{E}\left[\frac{1}{\beta_1}\max_{i\le N/2}(Q_i^{1,0}-I_1)^+\right]+\mathbb{E}\left[\frac{1}{\beta_2}\max_{i\le N/2}(Q_i^{1,0}-I_2)^+\right].$$

Therefore, the cost function in (6.6.1) can be bounded from above by

$$\frac{N}{2}\left(h_N^{(1)}\frac{1}{\beta_1}\left(I_1-\frac{\sigma^2}{2}\right)+\beta_1\right)+\frac{N}{2}\left(h_N^{(2)}\frac{1}{\beta_2}\left(I_2-\frac{\sigma^2}{2}\right)+\beta_2\right) \tag{6.6.2}$$
$$+\left(\frac{N}{2}h_N^{(1)}+\frac{N}{2}h_N^{(2)}+b_N\right)\left(\mathbb{E}\left[\frac{1}{\beta_1}\max_{i\le N/2}(Q_i^{1,0}-I_1)^+\right]+\mathbb{E}\left[\frac{1}{\beta_2}\max_{i\le N/2}(Q_i^{1,0}-I_2)^+\right]\right). \tag{6.6.3}$$

Our analytic results enable us to minimize this upper bound; for instance, by choosing $\tilde{h}_N^{(1,2)} = h_N^{(1,2)}$ and $\tilde{b}_N^{(1,2)} = \frac{N}{2}h_N^{(2,1)} + b_N$, we can rewrite the upper bound in (6.6.2) as follows:

$$\frac{N}{2}\left(h_N^{(1)}\frac{1}{\beta_1}\left(I_1-\frac{\sigma^2}{2}\right)+\beta_1\right)+\frac{N}{2}\left(h_N^{(2)}\frac{1}{\beta_2}\left(I_2-\frac{\sigma^2}{2}\right)+\beta_2\right)$$
$$+\left(\frac{N}{2}h_N^{(1)}+\frac{N}{2}h_N^{(2)}+b_N\right)\left(\mathbb{E}\left[\frac{1}{\beta_1}\max_{i\le N/2}(Q_i^{1,0}-I_1)^+\right]+\mathbb{E}\left[\frac{1}{\beta_2}\max_{i\le N/2}(Q_i^{1,0}-I_2)^+\right]\right)$$

$$= \frac{N}{2}\left(\tilde{h}_N^{(1)}\frac{1}{\beta_1}\left(I_1 - \frac{\sigma^2}{2}\right) + \beta_1\right) + \left(\frac{N}{2}\tilde{h}_N^{(1)} + \tilde{b}_N^{(1)}\right)\mathbb{E}\left[\frac{1}{\beta_1}\max_{i\leq N/2}(Q_i^{1,0} - I_1)^+\right]$$
$$+ \frac{N}{2}\left(\tilde{h}_N^{(2)}\frac{1}{\beta_2}\left(I_2 - \frac{\sigma^2}{2}\right) + \beta_2\right) + \left(\frac{N}{2}\tilde{h}_N^{(2)} + \tilde{b}_N^{(2)}\right)\mathbb{E}\left[\frac{1}{\beta_2}\max_{i\leq N/2}(Q_i^{1,0} - I_2)^+\right].$$

This is the sum of two functions that can be minimized using the exact solutions that we derived. In Table 6.10 we compare numerically the actual costs under the capacity and base-stock level that are obtained by minimizing this upper bound with the costs under the actual optimal capacity and base-stock level. In this table, the ratio indicates how many servers have a holding cost $h_N^{(1)}$ per item, and how many servers have a holding cost $h_N^{(2)}$ per item; the 1:1 ratio corresponds to the above example while the 1:3 ratio can be treated similarly. The table demonstrates that our asymptotic results may be useful when optimizing asymmetric systems as well as symmetric systems.

N	$h_N^{(1)}$	$h_N^{(2)}$	Ratio	b_N	Optimal	Heuristic	Diff.
10	1	10	1:1	10	42.3 ± 0.1	42.9 ± 0.1	0.14 %
100	1	10	1:1	100	615.6 ± 1.2	617.4 ± 1.0	0.3 %
1000	1	10	1:1	1000	7597.9 ± 8.2	7643.0 ± 7.8	0.6 %
10	10	100	1:1	1	126.0 ± 0.4	127.0 ± 0.4	0.7 %
100	100	1000	1:1	1	5967±10.9	6002 ± 9.6	0.6 %
1000	1000	10000	1:1	1	236063 ± 256	236402± 233	0.1 %
10	1	10	1:3	10	53.1±0.2	53.2 ± 0.2	0.2 %
100	1	10	1:3	100	770.5±1.3	772.9± 1.2	0.3 %
1000	1	10	1:3	1000	9551.1±10.7	9581.6± 9.5	0.3 %

Table 6.10 Comparison of optimal costs and costs under upper bound heuristic, $\sigma = 1$, $\sigma_A = 0$.

6.7. Summary of results

In this chapter, we defined a large-scale assembly system in which N components are assembled into a final product. We studied an assembly system with linear demand and production, subjected to some random noise. Thus, we imposed the natural assumption that this noise is normally distributed. Hence, delays per component are written as an all-time supremum of a Brownian motion minus a drift term. We aimed to minimize the expected total costs in the system with respect to the inventory and net capacity per component. The costs in the system consist of inventory holding costs for each component and penalty costs for delay of assembling the final product, which is equal to the delay of the slowest produced component. Before attempting to solve the minimization problem, we simplified the minimization problem, using the self-similarity property of a Brownian motion, into two separate minimization problems. We distinguished two cases: first, we covered the case of deterministic demand, resulting in all delays being independent; second, we investigated the case that demand is stochastic and consequently delays of the components are dependent.

For the deterministic demand scenario, we proved order bounds for three different regimes: balanced, quality driven, and efficiency driven. Additionally, we verified numerically that already for a limited number of components, our approximations result in costs that are very close to the costs corresponding to the optimal solution. For the stochastic demand scenario, we developed a limit theorem that we use to obtain approximate solutions. We showed numerically that even though theoretically these approximations perform well, for practical situations there is still room for improvement. However, this limit theorem is still necessary for systems with N of the order of thousands, because it is close to impossible to simulate these systems fast. Therefore, we provided additional approximations for a mixed-behavior regime, where we use a combination of the approximations for the deterministic and stochastic demand scenarios. We demonstrated numerically that these approximations perform very well already for a practical number of components.

Future work could extend the model to a decentralized minimization problem, where suppliers have their own objectives, which results in an asymptotic analysis of a game theoretical equilibrium, cf. [66, 91, 115].

6.8. Proofs

6.8.1 Proofs of Section 6.2

Proof of Lemma 6.1. In the independent case, we can write, by using the self-similarity property of Brownian motions, that

$$\sum_{i=1}^{N} \mathbb{E}\left[h_N(I_i - Q_i^{\beta_i,0}) + \beta_i\right] + (Nh_N + b_N)\mathbb{E}\left[\max_{j\le N}(Q_j^{\beta_j,0} - I_j)^+\right]$$
$$= \sum_{i=1}^{N} \mathbb{E}\left[h_N\left(I_i - \frac{1}{\beta_i}Q_i^{1,0}\right) + \beta_i\right] + (Nh_N + b_N)\mathbb{E}\left[\max_{j\le N}\left(\frac{1}{\beta_j}Q_j^{1,0} - I_j\right)^+\right].$$

We write $\eta_i = 1/\beta_i$. Thus,

$$\sum_{i=1}^{N} \mathbb{E}\left[h_N\left(I_i - \frac{1}{\beta_i}Q_i^{1,0}\right) + \beta_i\right] + (Nh_N + b_N)\mathbb{E}\left[\max_{j\le N}\left(\frac{1}{\beta_j}Q_j^{1,0} - I_j\right)^+\right]$$
$$= \sum_{i=1}^{N} \mathbb{E}\left[h_N(I_i - \eta_i Q_i^{1,0}) + \frac{1}{\eta_i}\right] + (Nh_N + b_N)\mathbb{E}\left[\max_{j\le N}(\eta_j Q_j^{1,0} - I_j)^+\right].$$

It is easy to see that $\sum_{i=1}^{N} \mathbb{E}\left[h_N(I_i - \eta_i Q_i^{1,0}) + 1/\eta_i\right]$ is strictly convex with respect to $(\eta_1, \ldots, \eta_N, I_1, \ldots, I_N)$, with $\eta_j, I_j > 0$. In order to examine whether $\mathbb{E}[\max_{j\le N}(\eta_j Q_j^{1,0} - I_j)^+]$ is convex, we should prove convexity of $\eta_j Q_j^{1,0} - I_j$, because taking the expectation of a convex function and taking maxima of convex functions preserve convexity. Since $\eta_j Q_j^{1,0} - I_j$

is linear in both η_j and I_j, convexity holds. Now, we write

$$C = \min_{(I_1,I_2,\ldots,I_N),(\beta_1,\beta_2,\ldots,\beta_N)} \sum_{i=1}^{N} \mathbb{E}\left[h_N(I_i - Q_i^{\beta_i,0}) + \beta_i\right] + (Nh_N + b_N)\mathbb{E}\left[\max_{j\leq N}(Q_j^{\beta_j,0} - I_j)^+\right]$$

with minimizers $(\beta_1^{(l)},\ldots,\beta_N^{(l)})$ and $(I_1^{(l)},\ldots,I_N^{(l)})$. Assume there exists i,j such that $\beta_i^{(l)} \neq \beta_j^{(l)}$ or $I_i^{(l)} \neq I_j^{(l)}$. Then, because of the symmetry of the problem with respect to the N servers, all the permutations of the minimizers give solutions. Assume there are k permutations, where the l-th permutation has minimizers $(\beta_1^{(l)},\ldots,\beta_N^{(l)})$ and $(I_1^{(l)},\ldots,I_N^{(l)})$. Now, define β_i and I_i such that they satisfy $1/\beta_i = \frac{1}{k}\sum_{l=1}^{k} 1/\beta_i^{(l)}$, and $I_i = \frac{1}{k}\sum_{l=1}^{k} I_i^{(l)}$. Because of the symmetry of the cost function around the N servers, we have that $\beta_i = \beta_j = \beta$, and $I_i = I_j = I$. Since we have a strictly convex function with respect to I_i and $1/\beta_i$,

$$C \geq \mathbb{E}\left[Nh_N(I - Q_i^{\beta,0})\right] + \beta N + (Nh_N + b_N)\mathbb{E}\left[\max_{j\leq N}(Q_j^{\beta,0} - I)^+\right].$$

Thus $I_i = I$, and $\beta_i = \beta$ are minimizers. An analogous derivation holds for the dependent case where we only minimize over one drift parameter. □

Remark 6.1. *In the dependent case where all servers choose a different drift parameter, we have that* $\sup_{s>0}(B_i(s) + B_A(s) - \beta_i s) = \sup_{s>0}(\hat{B}_i(s) + \hat{B}_A(s) - s)/\beta_i$ *where* $\hat{B}_i(s) = B_i(s/\beta_i^2)\beta_i$ *and* $\hat{B}_A(s) = B_A(s/\beta_i^2)\beta_i$. *However,* $\mathbb{E}\big[B_A(s/\beta_i^2)\beta_i B_A(s/\beta_j^2)\beta_j\big] = \sigma_A^2\beta_i\beta_j s/\max(\beta_i,\beta_j)^2 \neq \sigma_A^2 s$ *when* $\beta_i \neq \beta_j$. *Thus, when we have different drift parameters* β_i *and* β_j, *the joint distribution of* $\sup_{s>0}(B_i(s) + B_A(s) - \beta_i s)$ *and* $\sup_{s>0}(B_j(s) + B_A(s) - \beta_j s)$ *is not the same as the joint distribution of* $\sup_{s>0}(B_i(s) + B_A(s) - s)/\beta_i$ *and* $\sup_{s>0}(B_j(s) + B_A(s) - s)/\beta_j$. *So to prove Lemma 6.1 when the drifts are different, other techniques are needed.*

Remark 6.2. *In the case that* $\sigma_i \neq \sigma_j$, *in the independent case, the cost function given in Definition 6.3 can be simplified to*

$$\sum_{i=1}^{N} \mathbb{E}\left[h_N \frac{1}{\tilde{\beta}_i}\left(\tilde{I}_i - \frac{1}{2}\right) + \sigma_i^2\tilde{\beta}_i\right] + (Nh_N + b_N)\mathbb{E}\left[\max_{j\leq N}\frac{(Q_j^{1,0} - \tilde{I}_j)^+}{\tilde{\beta}_j}\right],$$

where $\tilde{I}_i = I_i\frac{\beta_i}{\sigma_i^2}$, $\tilde{\beta}_i = \frac{\beta_i}{\sigma_i^2}$, *and* $Q_j^{1,0} \overset{d}{=} Exp(2)$, *by again using the self-similarity property. Due to the term* $\sigma_i^2\tilde{\beta}_i$, *we cannot reduce the system to a minimization problem with two variables* $\tilde{I}$ *and* $\tilde{\beta}$, *because the problem is not symmetric anymore in each term. Thus, although the system is still strictly convex, for the minimizer we have that* $\tilde{\beta}_i$ *and* $\tilde{\beta}_j$ *are not necessarily the same. As this formula shows, the servers with larger standard deviations* σ_i *will cause more costs. What this formula also shows is that the problem with different standard deviations* σ_i *is equivalent to the problem with the same standard deviations* σ *but different capacity costs for each server.*

Remark 6.3. *In this chapter, we assume that capacity costs per supplier i equal the drift term* β_i. *One could extend this model to a system where the capacity costs per supplier is*

given by a function $k(\beta_i)$. In order to be able to conclude that we can reduce the complexity from $2N$ to 2 variables, and to minimize the cost function, we need to have that k is a strictly convex function with respect to $1/\beta_i$.

Proof of Lemma 6.2. We have that the function $F_N(I,\beta) > 0$, hence F_N has a global infimum, and since $\lim_{\beta\downarrow 0} F_N(I,\beta) = \infty$, $\lim_{\beta\to\infty} F_N(I,\beta) = \infty$ and $\lim_{I\to\infty} F_N(I,\beta) = \infty$, F_N has a global minimum. Now, assume $F_N(I_N,\beta_N) = \min_{(I,\beta)} F_N(I,\beta)$. Assume that there exists an $\hat{I}_N$ such that

$$\begin{aligned}\mathbb{E}[Nh_N(\hat{I}_N - Q_i^{1,\sigma_A} + (\bar{Q}_N^{1,\sigma_A} - \hat{I}_N)^+) + b_N(\bar{Q}_N^{1,\sigma_A} - \hat{I}_N)^+] \\ < \mathbb{E}[Nh_N(I_N - Q_i^{1,\sigma_A} + (\bar{Q}_N^{1,\sigma_A} - I_N)^+) + b_N(\bar{Q}_N^{1,\sigma_A} - I_N)^+].\end{aligned}$$

Then $F_N(\hat{I}_N,\beta_N) < F_N(I_N,\beta_N)$. This contradicts the statement that (I_N,β_N) gives the minimum of F_N. Hence, the optimal base-stock level minimizes $C_N(I)$. The proof that β_N minimizes $\frac{1}{\beta}C_N(I_N) + \beta N$ goes analogously.

To prove that $C_N(I)$ is convex with respect to I, we observe that

$$\begin{aligned}\frac{d^2}{dI^2}C_N(I) &= (b_N + Nh_N)\,\frac{d^2}{dI^2}\,\mathbb{E}\left[(\bar{Q}_N^{1,\sigma_A} - I)^+\right] \\ &= (b_N + Nh_N)\,\frac{d^2}{dI^2}\int_I^\infty \mathbb{P}(\bar{Q}_N^{1,\sigma_A} > x)dx \\ &= (b_N + Nh_N)\,f(I) \geq 0,\end{aligned}$$

because f is the probability density function of $\bar{Q}_N^{1,\sigma_A}$. This density exists; see [41, Prop. 2a]. In conclusion, we have a convex minimization problem. Moreover, $\frac{d^2}{d\beta^2}\left(\frac{1}{\beta}C_N(I_N) + \beta N\right) = \frac{2}{\beta^3}C_N(I_N) > 0$. Thus $\frac{1}{\beta}C_N(I_N) + \beta N$ is also convex with respect to β. □

Proof of Lemma 6.3. $F_N(I,\beta)$ has the form $F_N(I,\beta) = \frac{1}{\beta}C_N(I) + \beta N$. Thus, in order to minimize $F_N(I_N^*,\beta)$, we know by Lemma 6.2 that we need to solve $\frac{d}{d\beta}F_N(I_N^*,\beta) = -\frac{1}{\beta^2}C_N(I_N^*) + N = 0$. Thus, $\beta_N^* = \frac{\sqrt{C_N(I_N^*)}}{\sqrt{N}}$, and $F_N(I_N^*,\beta_N^*) = 2\sqrt{NC_N(I_N^*)} = 2N\beta_N^*$. □

Proof of Lemma 6.4. To solve $\min_I C_N(I)$, we have to solve $\frac{d}{dI}C_N(I) = 0$. This gives for the optimal base-stock level I_N^* that

$$Nh_N - (Nh_N + b_N)\,\mathbb{P}(\bar{Q}_N^{1,\sigma_A} > I_N^*) = 0.$$

Hence, $I_N^* = P_N^{-1}\left(\frac{b_N}{Nh_N + b_N}\right)$, with P_N^{-1} the quantile function of $\bar{Q}_N^{1,\sigma_A}$. □

Proof of Lemma 6.6. Following Corollary 6.1, we have

$$\frac{F_N(I_N^*,\beta_N^*)}{F_N(\bar{I}_N,\bar{\beta}_N)} = \frac{2\sqrt{C_N(I_N^*)}\sqrt{\bar{C}_N(\bar{I}_N)}}{C_N(\bar{I}_N) + \bar{C}_N(\bar{I}_N)}.$$

Furthermore, observe that

$$\mathbb{E}[\bar{Q}_N^{1,\sigma_A}] \geq \mathbb{E}[\max_{i\leq N} \sup_{s>0}(B_i(s)-s)+B_A(\tau)] = \frac{\sigma^2}{2}\sum_{i=1}^{N}\frac{1}{i} \geq \frac{\sigma^2}{2}\log N,$$

where τ is the first hitting time of the supremum of $\max_{i\leq N}(B_i(t)-t)$. From this, it follows that for $I < \frac{\sigma^2}{2}\log N$, $\frac{\sigma^2}{2}\log N - I < \mathbb{E}[\bar{Q}_N^{1,\sigma_A} - I] < \mathbb{E}[(\bar{Q}_N^{1,\sigma_A} - I)^+]$. For $I > \frac{\sigma^2}{2}\log N$, $(\frac{\sigma^2}{2}\log N - I)^+ = 0 < \mathbb{E}[(\bar{Q}_N^{1,\sigma_A} - I)^+]$. In conclusion, $C_N(I) > \bar{C}_N(I)$. Therefore,

$$\frac{F_N(I_N^*,\beta_N^*)}{F_N(\bar{I}_N,\bar{\beta}_N)} = \frac{2\sqrt{C_N(I_N^*)}\sqrt{\bar{C}_N(\bar{I}_N)}}{C_N(\bar{I}_N)+\bar{C}_N(\bar{I}_N)} \geq \frac{\sqrt{C_N(I_N^*)}\sqrt{\bar{C}_N(\bar{I}_N)}}{C_N(\bar{I}_N)}.$$

We have $|C_N(I_N^*) - C_N(\bar{I}_N)| \leq (2Nh_N + b_N)|I_N^* - \bar{I}_N|$, and

$$|\bar{C}_N(\bar{I}_N) - C_N(\bar{I}_N)| \leq (Nh_N + b_N)\,\mathbb{E}\left[\left|\bar{Q}_N^{1,\sigma_A} - \frac{\sigma^2}{2}\log N\right|\right].$$

In the case that $\gamma_N = \gamma \in (0,1)$, we have by applying Lemma 6.5 that $|\bar{C}_N(\bar{I}_N) - C_N(\bar{I}_N)| = o((Nh_N + b_N)\log N)$. Furthermore, $C_N(\bar{I}_N) \sim Nh_N\frac{\sigma^2}{2}\log N$, and since $\bar{Q}_N^{1,\sigma_A}/\log N \xrightarrow{\mathbb{P}} \sigma^2/2$, as $N\to\infty$, we also have that $I_N^*/\log N \overset{N\to\infty}{\longrightarrow} \sigma^2/2$. Thus, $|C_N(I_N^*) - C_N(\bar{I}_N)| = o((Nh_N + b_N)\log N)$, and the lemma follows.

In the case that $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$, we first observe that $\bar{C}_N(\bar{I}_N) = Nh_N\left(\frac{\sigma^2}{2}\log N - \frac{\sigma^2+\sigma_A^2}{2}\right) \sim Nh_N\frac{\sigma^2}{2}\log N$. Furthermore,

$$\begin{aligned} C_N(\bar{I}_N) =& Nh_N\left(\frac{\sigma^2}{2}\log N - \frac{\sigma^2+\sigma_A^2}{2}\right) + (Nh_N + b_N)\,\mathbb{E}\left[\left(\bar{Q}_N^{1,\sigma_A} - \frac{\sigma^2}{2}\log N\right)^+\right] \\ \leq& Nh_N\left(\frac{\sigma^2}{2}\log N - \frac{\sigma^2+\sigma_A^2}{2}\right) + (Nh_N + b_N)\,\mathbb{E}\left[\left|\bar{Q}_N^{1,\sigma_A} - \frac{\sigma^2}{2}\log N\right|\right]. \end{aligned}$$

Thus,

$$\frac{C_N(\bar{I}_N)}{Nh_N\log N} \leq \frac{\sigma^2}{2} + o(1) + \frac{1}{\gamma_N}\frac{\mathbb{E}\left[\left|\bar{Q}_N^{1,\sigma_A} - \frac{\sigma^2}{2}\log N\right|\right]}{\log N}.$$

By Lemma 6.5, we know that $\mathbb{E}\left[\left|\bar{Q}_N^{1,\sigma_A} - \frac{\sigma^2}{2}\log N\right|\right]/\log N \overset{N\to\infty}{\longrightarrow} 0$. Thus

$$\limsup_{N\to\infty} C_N(\bar{I}_N)/(Nh_N\log N) \leq \sigma^2/2.$$

Finally,

$$\begin{aligned} C_N(I_N^*) =& Nh_N\left(I_N^* - \frac{\sigma^2+\sigma_A^2}{2}\right) + (Nh_N + b_N)\,\mathbb{E}\left[\left(\bar{Q}_N^{1,\sigma_A} - I_N^*\right)^+\right] \\ \geq& Nh_N\left(I_N^* - \frac{\sigma^2+\sigma_A^2}{2}\right) + (Nh_N + b_N)\,\mathbb{E}\left[\bar{Q}_N^{1,\sigma_A} - I_N^*\right] \\ \geq& -Nh_N\frac{\sigma^2+\sigma_A^2}{2} + (Nh_N + b_N)\frac{\sigma^2}{2}\log N - b_N I_N^*. \end{aligned}$$

The optimal base-stock level I_N^* satisfies $I_N^* = O(\log N)$, and $b_N/(Nh_N) \overset{N\to\infty}{\longrightarrow} 0$. Therefore, $\liminf_{N\to\infty} C_N(I_N^*)/(Nh_N \log N) \geq \sigma^2/2$. Combining these results gives that the lower bound $\sqrt{C_N(I_N^*)}\sqrt{\bar{C}_N(\bar{I}_N)}/\sqrt{C_N(\bar{I}_N)}$ for the fraction $F_N(I_N^*,\beta_N^*)/F_N(\bar{I}_N,\bar{\beta}_N)$ satisfies

$$\liminf_{N\to\infty}\frac{F_N(I_N^*,\beta_N^*)}{F_N(\bar{I}_N,\bar{\beta}_N)} \geq \liminf_{N\to\infty}\frac{\sqrt{C_N(I_N^*)}\sqrt{\bar{C}_N(\bar{I}_N)}}{C_N(\bar{I}_N)} = 1.$$

□

6.8.2 Proofs of Section 6.3

Proof of Lemma 6.7. In Lemma 6.4, it is shown that $I_N^* = P_N^{-1}(1-\gamma_N)$, with P_N^{-1} the quantile function of $\bar{Q}_N^{1,0}$. Because the random variables $(Q_i^{1,0}, i \leq N)$ are independent and exponentially distributed,

$$\mathbb{P}\big(\bar{Q}_N^{1,0} \leq P_N^{-1}(x)\big) = x = \left(1 - e^{-\frac{2}{\sigma^2}P_N^{-1}(x)}\right)^N.$$

From this, it follows that $P_N^{-1}(x) = \frac{\sigma^2}{2}\log\left(1\Big/\left(1-x^{\frac{1}{N}}\right)\right)$. □

Proof of Proposition 6.1. Minimizing $\hat{F}_N(\hat{I}_N,\hat{\beta}_N)$ goes analogously as minimizing $F_N(I_N,\beta_N)$ in Lemma 6.7. Hence, $\hat{I}_N = \hat{P}_N^{-1}(1-\gamma_N)$. Thus, we have to solve

$$\begin{aligned} \mathbb{P}\left(\frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log N \leq \hat{P}_N^{-1}(x)\right) = \mathbb{P}\left(G \leq \frac{2}{\sigma^2}\hat{P}_N^{-1}(x) - \log N\right) \\ = e^{-e^{-\left(\frac{2}{\sigma^2}\hat{P}_N^{-1}(x) - \log N\right)}} = x. \end{aligned}$$

Therefore, $\hat{P}_N^{-1}(x) = \frac{\sigma^2}{2}\log N - \frac{\sigma^2}{2}\log(-\log x)$. Hence, the optimal base-stock level is given in Equation (6.3.4). Furthermore,

$$\begin{aligned} \mathbb{E}\left[\left(\frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log N - \hat{I}_N\right)^+\right] =& \mathbb{E}\left[\left(\frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log\big(-\log(1-\gamma_N)\big)\right)^+\right] \\ =& \frac{\sigma^2}{2}\int_{-\log\left(-\log(1-\gamma_N)\right)}^{\infty} 1 - e^{-e^{-x}}\,dx. \end{aligned}$$

By using partial integration and substitution we can write

$$\frac{\sigma^2}{2}\int_{-\log\left(-\log(1-\gamma_N)\right)}^{\infty} 1-e^{-e^{-x}}\,dx = \frac{\sigma^2}{2}\left(\int_{-\log(1-\gamma_N)}^{\infty}\frac{e^{-t}}{t}dt+\Gamma+\log\left(-\log(1-\gamma_N)\right)\right).$$

Hence, this gives us the expression of $\hat{C}_N(\hat{I}_N)$ in (6.3.5). □

Lemma 6.11. *We define the random variable*

$$G_N := -\log\left(-\log\left(\left(1-\exp\left(-\frac{2}{\sigma^2}\bar{Q}_N^{1,0}\right)\right)^N\right)\right), \tag{6.8.1}$$

with $\bar{Q}_N^{1,0}$ given in Definition 6.1. The random variable G_N has the property that $\mathbb{P}(G_N < x) = e^{-e^{-x}}$, for all $N \geq 1$. Thus, G_N follows a Gumbel distribution. Moreover,

$$\bar{Q}_N^{1,0} > \frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N, \tag{6.8.2}$$

and $\bar{Q}_N^{1,0} - \frac{\sigma^2}{2}G_N - \frac{\sigma^2}{2}\log N$ strictly decreases as a function of $\bar{Q}_N^{1,0}$ with limit 0.

Proof. To prove that G_N follows a Gumbel distribution, we first observe that $\mathbb{P}(\bar{Q}_N^{1,0} < x) = \left(1-\exp\left(-\frac{2}{\sigma^2}x\right)\right)^N$. Therefore, $\left(1-\exp\left(-\frac{2}{\sigma^2}\bar{Q}_N^{1,0}\right)\right)^N \sim \text{Unif}[0,1]$. Then,

$$\begin{aligned}
\mathbb{P}(G_N < x) &= \mathbb{P}\left(-\log\left(-\log\left(\left(1-\exp\left(-\frac{2}{\sigma^2}\bar{Q}_N^{1,0}\right)\right)^N\right)\right) < x\right)\\
&= \mathbb{P}\left(-\log\left(\left(1-\exp\left(-\frac{2}{\sigma^2}\bar{Q}_N^{1,0}\right)\right)^N\right) > e^{-x}\right)\\
&= \mathbb{P}\left(\left(1-\exp\left(-\frac{2}{\sigma^2}\bar{Q}_N^{1,0}\right)\right)^N < e^{-e^{-x}}\right) = e^{-e^{-x}}.
\end{aligned}$$

To prove (6.8.2), we need to show that for all $x > 0$ and N

$$x > -\frac{\sigma^2}{2}\log\left(-\log\left(\left(1-\exp\left(-\frac{2}{\sigma^2}x\right)\right)^N\right)\right)+\frac{\sigma^2}{2}\log N.$$

This is equivalent to the inequality $x > -\frac{\sigma^2}{2}\log\left(-\log\left(1-\exp\left(-\frac{2}{\sigma^2}x\right)\right)\right)$, which is

equivalent to $1 - e^{-\frac{2}{\sigma^2}x} < e^{-e^{-\frac{2}{\sigma^2}x}}$, with $x > 0$. This is equivalent to $e^{-y} > 1 - y$ for $y \in (0, e^{-1}]$. Observe that for $y = 0$, we have equality, and we have for $y > 0$ that $(e^{-y})' > -1 = (1-y)'$. The statement follows. To prove that the larger $\bar{Q}_N^{1,0}$ becomes, the smaller the difference between $\bar{Q}_N^{1,0}$ and $\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N$ becomes, we first observe that

$$\begin{aligned}\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N &= -\frac{\sigma^2}{2}\log\left(-\log\left(\left(1-\exp\left(-\frac{2}{\sigma^2}\bar{Q}_N^{1,0}\right)\right)^N\right)\right) + \frac{\sigma^2}{2}\log N \\ &= -\frac{\sigma^2}{2}\log\left(-\log\left(1-\exp\left(-\frac{2}{\sigma^2}\bar{Q}_N^{1,0}\right)\right)\right).\end{aligned}$$

Thus, we need to obtain that $x + \frac{\sigma^2}{2}\log(-\log(1-e^{-\frac{2}{\sigma^2}x}))$ is strictly decreasing in x for $x > 0$. Taking the first derivative gives the inequality

$$\frac{e^{-\frac{2x}{\sigma^2}}}{\left(1-e^{-\frac{2x}{\sigma^2}}\right)\log\left(1-e^{-\frac{2x}{\sigma^2}}\right)} + 1 < 0.$$

This is equivalent to the inequality $-y/((1-y)\log(1-y)) > 1$ for $y \in (0,1)$, which can be rewritten to $\log y > 1 - 1/y$, which is a basic logarithm inequality. Finally, $\lim_{x\to\infty} x + \frac{\sigma^2}{2}\log(-\log(1-e^{-\frac{2}{\sigma^2}x})) = 0$. □

Lemma 6.12. *Let γ_N be given in* (6.2.4)*, then we have the following bounds on the cost functions C_N and $\hat{C}_N$ given in Definitions 6.3 and 6.4:*

$$\left|C_N(I_N^*) - C_N(\hat{I}_N)\right| \leq (I_N^* - \hat{I}_N)(Nh_N + b_N)\left(1 - \gamma_N - \left(1 + \frac{\log(1-\gamma_N)}{N}\right)^N\right), \quad (6.8.3)$$

$$\left|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)\right| \leq (I_N^* - \hat{I}_N)Nh_N\left(1 - \left(1 + \frac{\log(1-\gamma_N)}{N}\right)^N\right). \quad (6.8.4)$$

Proof. The optimal base-stock level I_N^* satisfies the equation given in (6.2.10). The approximation $\hat{I}_N$ satisfies a similar equation:

$$\mathbb{P}(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N \leq \hat{I}_N) = 1 - \gamma_N.$$

Due to the inequality in (6.8.2), it follows that $I_N^* > \hat{I}_N$, we have

$$\begin{aligned}&C_N(I_N^*) - C_N(\hat{I}_N) \\ &\quad = Nh_N(I_N^* - \hat{I}_N) + (Nh_N + b_N)\,\mathbb{E}[(\bar{Q}_N^{1,0} - I_N^*)^+ - (\bar{Q}_N^{1,0} - \hat{I}_N)^+] \\ &\quad = Nh_N(I_N^* - \hat{I}_N) + (Nh_N + b_N)\,\mathbb{E}[(\hat{I}_N - I_N^*)\mathbb{1}(\bar{Q}_N^{1,0} > I_N^*)]\end{aligned}$$

$$- (Nh_N + b_N)\,\mathbb{E}[(\bar{Q}_N^{1,0} - \hat{I}_N)^+ \mathbb{1}(\hat{I}_N < \bar{Q}_N^{1,0} < I_N^*)].$$

We have $\mathbb{P}(\bar{Q}_N^{1,0} > I_N^*) = \gamma_N = Nh_N/(Nh_N + b_N)$, thus

$$Nh_N(I_N^* - \hat{I}_N) + (Nh_N + b_N)\,\mathbb{E}[(\hat{I}_N - I_N^*)\mathbb{1}(\bar{Q}_N^{1,0} > I_N^*)] = 0.$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}[(\bar{Q}_N^{1,0} - \hat{I}_N)^+ \mathbb{1}(\hat{I}_N < \bar{Q}_N^{1,0} < I_N^*)] \\
&\quad \le (I_N^* - \hat{I}_N)\,\mathbb{P}(\hat{I}_N < \bar{Q}_N^{1,0} < I_N^*) \\
&\quad = (I_N^* - \hat{I}_N)\left(1 - \gamma_N - \left(1 + \frac{\log(1-\gamma_N)}{N}\right)^N\right).
\end{aligned}$$

Equation (6.8.3) follows. To prove Equation (6.8.4), we observe that

$$\begin{aligned}
&|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)| \\
&\quad = (Nh_N + b_N)\,\mathbb{E}\left[\left(\bar{Q}_N^{1,0} - \hat{I}_N\right)^+ - \left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N - \hat{I}_N\right)^+\right] \\
&\quad = (Nh_N + b_N)\,\mathbb{E}\left[\left(\bar{Q}_N^{1,0} - \frac{\sigma^2}{2}G_N - \frac{\sigma^2}{2}\log N\right)\mathbb{1}\left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N > \hat{I}_N\right)\right] \qquad (6.8.5) \\
&\qquad + (Nh_N + b_N)\,\mathbb{E}\left[\left(\bar{Q}_N^{1,0} - \hat{I}_N\right)\mathbb{1}\left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N < \hat{I}_N < \bar{Q}_N^{1,0}\right)\right]. \qquad (6.8.6)
\end{aligned}$$

Because G_N and $\bar{Q}_N^{1,0}$ are on the same probability space, we have $\mathbb{P}\Big(\bar{Q}_N^{1,0} = I_N^* \Big| \frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N = \hat{I}_N\Big) = 1$. Furthermore, $x + \frac{\sigma^2}{2}\log(-\log(1 - e^{-\frac{2}{\sigma^2}x}))$ is decreasing in x. Thus, we can bound

$$\begin{aligned}
&\mathbb{E}\left[\left(\bar{Q}_N^{1,0} - \frac{\sigma^2}{2}G_N - \frac{\sigma^2}{2}\log N\right)\mathbb{1}\left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N > \hat{I}_N\right)\right] \\
&\quad \le (I_N^* - \hat{I}_N)\,\mathbb{P}\left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N > \hat{I}_N\right) \\
&\quad = (I_N^* - \hat{I}_N)\gamma_N. \qquad (6.8.7)
\end{aligned}$$

Similarly, for (6.8.6), we observe that if $\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N < \hat{I}_N$, then $\bar{Q}_N^{1,0} < I_N^*$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\left(\bar{Q}_N^{1,0} - \hat{I}_N\right)\mathbb{1}\left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N < \hat{I}_N < \bar{Q}_N^{1,0}\right)\right] \\
&\quad \le (I_N^* - \hat{I}_N)\,\mathbb{P}\left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N < \hat{I}_N < \bar{Q}_N^{1,0}\right)
\end{aligned}$$

$$\leq (I_N^* - \hat{I}_N)\left(1 - \left(1 + \frac{\log(1-\gamma_N)}{N}\right)^N - \gamma_N\right). \tag{6.8.8}$$

Adding the bounds in (6.8.7) and (6.8.8) gives the result. □

Proof of Theorem 6.1. First, we assume that $\gamma_N = \gamma \in (0,1)$. Using Corollary 6.1, we have

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} = \frac{2\sqrt{C_N(I_N^*)}\sqrt{\hat{C}_N(\hat{I}_N)}}{C_N(\hat{I}_N) + \hat{C}_N(\hat{I}_N)}.$$

Because of the inequality in (6.8.2), we have for all I that $C_N(I) > \hat{C}_N(I)$. Thus

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} > \frac{2\sqrt{C_N(I_N^*)}\sqrt{\hat{C}_N(\hat{I}_N)}}{2C_N(\hat{I}_N)}.$$

By computing the Taylor series of the function $I_{1/x}^*$ around $x = 0$, we have

$$\begin{aligned} I_{1/x}^* &= \frac{\sigma^2}{2}\log\left(\frac{1}{1-(1-\gamma)^x}\right) \\ &= -\frac{\sigma^2}{2}\log x - \frac{\sigma^2}{2}\log(-\log(1-\gamma))) - \frac{\sigma^2}{4}x\log(1-\gamma) + O(x^2) \\ &= \hat{I}_{1/x} - \frac{\sigma^2}{4}x\log(1-\gamma) + O(x^2). \end{aligned}$$

Thus, $(I_N^* - \hat{I}_N) \sim -\sigma^2\log(1-\gamma)/(4N)$. Following (6.8.4), we can conclude that $|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)|/(Nh_N) = O(1/N)$. We can do the same for $\mathbb{P}(\hat{I}_N < \bar{Q}_N^{1,0} < I_N^*)$, and get

$$\left(1 - \gamma - \left(1 + \frac{\log(1-\gamma)}{N}\right)^N\right) \sim \frac{1}{2N}(1-\gamma)\log(1-\gamma)^2.$$

Thus, after applying the inequality in (6.8.3), we get $|C_N(I_N^*) - C_N(\hat{I}_N)|/(Nh_N + b_N) = O(1/N^2)$. We have that the approximation function $\hat{C}_N(\hat{I}_N)$ satisfies

$$\begin{aligned} &\hat{C}_N(\hat{I}_N) \\ &\quad = Nh_N\frac{\sigma^2}{2}(\log N - \log(-\log(1-\gamma)) - 1) + (Nh_N + b_N)\frac{\sigma^2}{2}\,\mathbb{E}\left[(G + \log(-\log(1-\gamma)))^+\right] \\ &\quad \sim Nh_N\frac{\sigma^2}{2}\log N, \end{aligned}$$

as $N \to \infty$, because $(Nh_N + b_N)/(Nh_N) = 1/\gamma$, and $-\log(-\log(1-\gamma))$ and $\mathbb{E}[(G_N + \log(-\log(1-\gamma)))^+]$ are of $O(1)$. In conclusion, we have

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} > \frac{\sqrt{C_N(I_N^*)}}{\sqrt{C_N(\hat{I}_N)}}\frac{\sqrt{\hat{C}_N(\hat{I}_N)}}{\sqrt{C_N(\hat{I}_N)}}$$

$$
\begin{aligned}
&= \frac{\sqrt{C_N(\hat{I}_N) - O((Nh_N + b_N)/N^2)}}{\sqrt{C_N(\hat{I}_N)}} \frac{\sqrt{C_N(\hat{I}_N) - O(Nh_N/N)}}{\sqrt{C_N(\hat{I}_N)}} \\
&= \sqrt{1 - O(1/(N^2 \log N))} \sqrt{1 - O(1/(N \log N))} \\
&= 1 - O(1/(N \log N)).
\end{aligned}
$$

Now, we assume that $\gamma_N \overset{N\to\infty}{\longrightarrow} 0$, then we have that $-\log(-\log(1-\gamma_N)) \sim -\log(\gamma_N)$, thus $\hat{I}_N \sim \frac{\sigma^2}{2}\log(N/\gamma_N)$. Also,

$$
\mathbb{E}\big[(G_N + \log(-\log(1-\gamma_N)))^+\big] \sim \mathbb{E}\big[(G_N + \log(\gamma_N))^+\big] \sim \gamma_N.
$$

From this, it follows that $\hat{C}_N(\hat{I}_N) \sim Nh_N \frac{\sigma^2}{2}\log(N/\gamma_N)$. Furthermore,

$$
\begin{aligned}
\mathbb{P}(\bar{Q}_N^{1,0} > \hat{I}_N) = 1 - \left(1 + \frac{\log(1-\gamma_N)}{N}\right)^N \\
\leq N\,\mathbb{P}(Q_i^{1,0} > \hat{I}_N) = -\log(1-\gamma_N) = \gamma_N(1 + O(\gamma_N/2)).
\end{aligned}
$$

Therefore, we have that

$$
\left(1 - \gamma_N - \left(1 + \frac{\log(1-\gamma_N)}{N}\right)^N\right) \leq -\log(1-\gamma_N) - \gamma_N = \frac{\gamma_N^2}{2}(1 + o(1)).
$$

Also

$$
\mathbb{P}(\bar{Q}_N^{1,0} < I_N^*) = \mathbb{P}\left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N < \hat{I}_N\right) = 1 - \gamma_N \overset{N\to\infty}{\longrightarrow} 1.
$$

In the first part of this proof, we showed that when $\gamma_N = \gamma$, $(I_N^* - \hat{I}_N) = O(1/N)$, now I_N^* is larger, because $\mathbb{P}(\bar{Q}_N^{1,0} < I_N^*) = 1 - \gamma_N \overset{N\to\infty}{\longrightarrow} 1$. Following the statement in Lemma 6.11 that the difference between $\bar{Q}_N^{1,0}$ and $\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N$ decreases as $\bar{Q}_N^{1,0}$ increases, we can conclude that $(I_N^* - \hat{I}_N) = O(1/N)$. Following the proof before, and by using the order bounds in (6.8.3) and (6.8.4), we have that

$$
\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} = 1 - O(\gamma_N/(N\log(N/\gamma_N))).
$$

Finally, we consider the case that $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$ and $\gamma_N \leq 1 - \exp(-N)$. Then, $\hat{I}_N \geq 0$. Furthermore, when $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$, we have $\log(-\log(1-\gamma_N)) \overset{N\to\infty}{\longrightarrow} \infty$. From this, it follows that

$$
\mathbb{E}\big[(G_N + \log(-\log(1-\gamma_N)))^+\big] \sim \log(-\log(1-\gamma_N)).
$$

Thus

$$
\hat{C}_N(\hat{I}_N) \sim \frac{\sigma^2}{2}Nh_N(\log N - \log(-\log(1-\gamma_N))) + \frac{\sigma^2}{2}(Nh_N + b_N)\log(-\log(1-\gamma_N))
$$

$$=\frac{\sigma^2}{2}Nh_N\log N+\frac{\sigma^2}{2}b_N\log(-\log(1-\gamma_N)).$$

Since we consider the efficiency-driven regime, we have $b_N/(Nh_N)\overset{N\to\infty}{\longrightarrow}0$. Also, it is easy to deduce that when $\gamma_N<1-\exp(-N)$, we have $\log(-\log(1-\gamma_N))<\log N$. Thus $\hat{C}_N(\hat{I}_N)\sim\frac{\sigma^2}{2}Nh_N\log N$. Furthermore, $I_N^*-\hat{I}_N=O(1)$; thus, the bounds in (6.8.3) and (6.8.4) are of $O(Nh_N)$. By using the same argument as in the proof for the balanced regime, we see that the approximate and optimal costs differ with order $1/\log N$:

$$\frac{F_N(I_N^*,\beta_N^*)}{F_N(\hat{I}_N,\hat{\beta}_N)}=1-O(1/\log N).$$

□

Proof of Lemma 6.8. Following Equations (6.8.3) and (6.8.4) and using the same arguments as in the proof of Theorem 6.1, we can find the same order bound for $F_N(I_N^*,\beta_N^*)/\hat{F}_N(\hat{I}_N,\hat{\beta}_N)=\sqrt{C_N(I_N^*)}/\sqrt{\hat{C}_N(\hat{I}_N)}$.

In the case that $\gamma_N=\gamma\in(0,1)$, we have

$$\begin{aligned}\hat{C}_N(\hat{I}_N)=&Nh_N\frac{\sigma^2}{2}\Big(\log N-\log(-\log(1-\gamma))-1\Big)\\&+(Nh_N+b_N)\frac{\sigma^2}{2}\,\mathbb{E}\Big[\Big(G+\log(-\log(1-\gamma))\Big)^+\Big].\end{aligned}$$

Thus $\hat{F}_N(\hat{I}_N,\hat{\beta}_N)/(N\log N)=2\sqrt{N}\sqrt{\hat{C}_N(\hat{I}_N)}/(N\log N)=O(\sqrt{h_N}/\sqrt{\log N})$.

When $\gamma_N\overset{N\to\infty}{\longrightarrow}0$, we have that $-\log(-\log(1-\gamma_N))\sim-\log(\gamma_N)$, thus $\hat{I}_N\sim\frac{\sigma^2}{2}\log(N/\gamma_N)$. Also,

$$\mathbb{E}\big[(G_N+\log(-\log(1-\gamma_N)))^+\big]\sim\mathbb{E}\big[(G_N+\log(\gamma_N))^+\big]\sim\gamma_N.$$

From this, it follows that

$$\hat{C}_N(\hat{I}_N)\sim Nh_N\frac{\sigma^2}{2}\Big(\log(N/\gamma_N)-1\Big)+(Nh_N+b_N)\frac{\sigma^2}{2}\gamma_N.$$

Therefore, $2\sqrt{N}\sqrt{\hat{C}_N(\hat{I}_N)}\gamma_N/(N\log(N/\gamma_N))=O(\gamma_N\sqrt{h_N}/\sqrt{\log(N/\gamma_N)})$.

When $\gamma_N\overset{N\to\infty}{\longrightarrow}1$, we have

$$\begin{aligned}\hat{C}_N(\hat{I}_N)\sim&\frac{\sigma^2}{2}Nh_N(\log N-\log(-\log(1-\gamma_N)))+\frac{\sigma^2}{2}(Nh_N+b_N)\log(-\log(1-\gamma_N))\\=&\frac{\sigma^2}{2}Nh_N\log N+\frac{\sigma^2}{2}b_N\log(-\log(1-\gamma_N)).\end{aligned}$$

Thus, $2\sqrt{N}\sqrt{\hat{C}_N(\hat{I}_N)}/\log N=O(N\sqrt{h_N}/\sqrt{\log N})$. □

6.8.3 Proofs of Section 6.4

Proof of Lemma 6.9. From Lemma 6.2, we know that the optimal inventory I_N^A satisfies

$$\frac{d}{dI}\mathbb{E}[Nh_N(I_N^A - Q_i^{1,\sigma_A} + (\bar{Q}_N^{1,\sigma_A} - I_N^A)^+) + b_N(\bar{Q}_N^{1,\sigma_A} - I_N^A)^+] = 0.$$

We have

$$\begin{aligned}
&\frac{d}{dI}\mathbb{E}[Nh_N(I_N^A - Q_i^{1,\sigma_A} + (\bar{Q}_N^{1,\sigma_A} - I_N^A)^+) + b_N(\bar{Q}_N^{1,\sigma_A} - I_N^A)^+] \\
&\quad = Nh_N - (Nh_N + b_N)\,\mathbb{P}\Big(\bar{Q}_N^{1,\sigma_A} > I_N^A\Big) \\
&\quad = Nh_N - (Nh_N + b_N)\,\mathbb{P}\left(\frac{\sqrt{2}}{\sigma\sigma_A}\frac{\bar{Q}_N^{1,\sigma_A} - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} > \frac{\sqrt{2}}{\sigma\sigma_A}\frac{I_N^A - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right).
\end{aligned}$$

Therefore, I_N^A satisfies $\frac{\sqrt{2}}{\sigma\sigma_A}(I_N^A - \frac{\sigma^2}{2}\log N)/\sqrt{\log N} = {P_N^A}^{-1}(1-\gamma_N)$. □

Proof of Proposition 6.2. We have to find I and β such that $F_N(I,\beta)$ is minimized. As before, we know that the optimal $\hat{I}_N^A$ should satisfy

$$Nh_N - (Nh_N + b_N)\,\mathbb{P}\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X > \hat{I}_N^A\right) = 0.$$

Thus, $\hat{I}_N^A$ as given in (6.4.2) minimizes $\hat{C}_N^A(I)$. We know that

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X - \hat{I}_N^A\right)^+\right] \\
&\quad = \int_{\frac{\hat{I}_N^A - \frac{\sigma^2}{2}\log N}{\frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}}}^{\infty}\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}x - \hat{I}_N^A\right)\phi(x)dx \\
&\quad = \left(\frac{\sigma^2}{2}\log N - \hat{I}_N^A\right)\mathbb{P}\left(\frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X \geq \hat{I}_N^A - \frac{\sigma^2}{2}\log N\right) \\
&\qquad + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(\sigma^2\log N - 2\hat{I}_N^A\right)^2}{4\sigma^2\sigma_A^2\log N}\right) \\
&\quad = -\frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}\Phi^{-1}(1-\gamma_N)\gamma_N \\
&\qquad + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\Phi^{-1}(1-\gamma_N)^2\Big).
\end{aligned}$$

The expression in Equation (6.4.3) follows. □

Proof of Theorem 6.2. Using Corollary 6.1, we have

$$\frac{F_N(I_N^A,\beta_N^A)}{F_N(\hat{I}_N^A,\hat{\beta}_N^A)} = \frac{2\sqrt{C_N(I_N^A)}\sqrt{\hat{C}_N^A(\hat{I}_N^A)}}{C_N(\hat{I}_N^A)+\hat{C}_N^A(\hat{I}_N^A)}.$$

First, assume $\hat{C}_N^A(\hat{I}_N^A) > C_N(\hat{I}_N^A)$. Then, $F_N(I_N^A,\beta_N^A)/F_N(\hat{I}_N^A,\hat{\beta}_N^A) > \sqrt{C_N(I_N^A)/\hat{C}_N^A(\hat{I}_N^A)}$. We have

$$|\hat{C}_N^A(\hat{I}_N^A) - C_N(I_N^A)| \leq (2Nh_N+b_N)|I_N^A-\hat{I}_N^A| + (Nh_N+b_N)\,\mathbb{E}\left[\left|\bar{Q}_N^{1,\sigma_A} - \frac{\sigma^2}{2}\log N - \frac{\sigma\sigma_A}{\sqrt{2}}X\right|\right].$$

We know by [148, Lem. 21.2, p. 305], that $(I_N^A-\hat{I}_N^A)/\sqrt{\log N} \overset{N\to\infty}{\longrightarrow} 0$. Furthermore, we have shown in Lemma 3.3 that $\mathbb{E}\left[\left|\bar{Q}_N^{1,\sigma_A} - \frac{\sigma^2}{2}\log N - \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X\right|\Big/\sqrt{\log N}\right] \overset{N\to\infty}{\longrightarrow} 0$. From this, it follows that $|\hat{C}_N^A(\hat{I}_N^A) - C_N(I_N^A)| = o((Nh_N+b_N)\sqrt{\log N})$. Since $\hat{C}_N^A(\hat{I}_N^A) \sim \frac{\sigma^2}{2}Nh_N\log N$, we have $\frac{\sqrt{C_N(I_N^A)}}{\sqrt{\hat{C}_N^A(\hat{I}_N^A)}} = 1 - o\left((Nh_N+b_N)\sqrt{\log N}/(Nh_N\log N)\right) = 1 - o\left(1/\sqrt{\log N}\right)$.

Second, assume $\hat{C}_N^A(\hat{I}_N^A) < C_N(\hat{I}_N^A)$, then

$$\frac{F_N(I_N^A,\beta_N^A)}{F_N(\hat{I}_N^A,\hat{\beta}_N^A)} > \frac{\sqrt{C_N(I_N^A)\hat{C}_N^A(\hat{I}_N^A)}}{C_N(\hat{I}_N^A)} = \frac{\sqrt{C_N(I_N^A)}}{\sqrt{C_N(\hat{I}_N^A)}}\frac{\sqrt{\hat{C}_N^A(\hat{I}_N^A)}}{\sqrt{C_N(\hat{I}_N^A)}}.$$

With an analogous derivation, we obtain the same order bound. □

Proof of Lemma 6.10. We have $\hat{I}_N^A = \frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}\Phi^{-1}(1-\gamma)$. Furthermore, $|I_N^A - \hat{I}_N^A| = o(\sqrt{\log N})$, thus (6.4.4) follows. Furthermore, by using the same argument as in Lemma 6.8, (6.4.5) follows. □

6.9. Mixed-behavior approximations

Though we have a symbolic expression for β_N^M in (6.5.3), it is not completely clear how to compute the part

$$\mathbb{E}\left[\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X + \frac{\sigma^2}{2}G - I_N^M\right)^+\right] = \int_{I_N^M}^{\infty}\mathbb{P}\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X + \frac{\sigma^2}{2}G > x\right)dx$$

in β_N^M. First, observe that we can write

$$\begin{aligned}
&\mathbb{P}\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X + \frac{\sigma^2}{2}G > x\right) \\
&\quad = \mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X + G > \frac{2}{\sigma^2}x - \log N\right) \\
&\quad = \int_{-\infty}^{\infty} \mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X > \frac{2}{\sigma^2}x - \log N - z\right)\exp(-\exp(-z) - z)dz.
\end{aligned}$$

Now, we write $z = -\log s$. Then,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X > \frac{2}{\sigma^2}x - \log N - z\right)\exp(-\exp(-z) - z)dz \\
&\qquad\qquad = \int_0^{\infty} \mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X > \frac{2}{\sigma^2}x - \log N + \log s\right)\exp(-s)ds.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X + \frac{\sigma^2}{2}G - I_N^M\right)^+\right] \\
&\quad = \int_{I_N^M}^{\infty}\int_0^{\infty} \mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X > \frac{2}{\sigma^2}x - \log N + \log s\right)\exp(-s)dsdx \\
&\quad = \int_0^{\infty}\int_{I_N^M}^{\infty} \mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X > \frac{2}{\sigma^2}x - \log N + \log s\right)\exp(-s)dxds.
\end{aligned}$$

It turns out that

$$\int_{I_N^M}^{\infty} \mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X > \frac{2}{\sigma^2}x - \log N + \log s\right)\exp(-s)dx$$

can be expressed in terms of error functions. Thus, since I_N^M can be numerically found by solving Equation (6.5.2), $\mathbb{E}\left[\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X + \frac{\sigma^2}{2}G - I_N^M\right)^+\right]$ can be computed numerically as well. Observe that the procedure to obtain I_N^M and β_N^M is efficient and that its running time is independent of the system size N.

Bibliography

[1] Joseph Abate and Ward Whitt. Transient behavior of regulated Brownian motion, I: starting at the origin. *Advances in Applied Probability*, 19(3):560–598, 1987.

[2] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, tenth edition, 1972.

[3] Victor S Adamchik and HM Srivastava. Some series of the zeta and related functions. *Analysis*, 18(2):131–144, 1998.

[4] Robert J Adler and Jonathan E Taylor. *Random fields and geometry*, volume 80. Springer, 2007.

[5] Najam Ahmad, Albert Gordon Greenberg, Parantap Lahiri, Dave Maltz, Parveen K Patel, Sudipta Sengupta, and Kushagra V Vaid. Distributed load balancer, February 11 2010. US Patent App. 12/189,438.

[6] Yalçin Akçay and Susan H Xu. Joint inventory replenishment and component allocation optimization in an assemble-to-order system. *Management Science*, 50(1):99–116, 2004.

[7] Klaus Altendorfer and Stefan Minner. Simultaneous optimization of capacity and planned lead time in a two-stage production system with different customer due dates. *European Journal of Operational Research*, 213(1):134–146, 2011.

[8] Clive W Anderson. Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *Journal of Applied Probability*, 7(1):99–113, 1970.

[9] Clive W Anderson. Local limit theorems for the maxima of discrete random variables. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 88, pages 161–165. Cambridge University Press, 1980.

[10] Clive W Anderson, Stuart G Coles, and Jürg Hüsler. Maxima of Poisson-like variables and related triangular arrays. *The Annals of Applied Probability*, 7(4):953–971, 1997.

[11] Theodore W Anderson and Stephen M Samuels. Some inequalities among binomial and Poisson probabilities. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 1–12, 1965.

[12] ASML Holding N.V. ASML annual report 2020. `https://www.asml.com/en/investors/annual-report/2020`, 2021.

[13] Søren Asmussen. *Applied probability and queues*, volume 2. Springer, 2003.

[14] Søren Asmussen, Peter W Glynn, and Jim Pitman. Discretization error in simulation of one-dimensional reflecting Brownian motion. *The Annals of Applied Probability*, 5(4):875–896, 1995.

[15] Zümbül Atan and Martine Rousseau. Inventory optimization for perishables subject to supply disruptions. *Optimization Letters*, 10(1):89–108, 2016.

[16] Rami Atar, Avishai Mandelbaum, and Asaf Zviran. Control of fork-join networks in heavy traffic. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 823–830. IEEE, 2012.

[17] François Baccelli. Two parallel queues created by arrivals with two demands: The $M/G/2$ symmetrical case. *Technical report RR–0426, INRIA*, 1985.

[18] François Baccelli, Armand M Makowski, and Adam Shwartz. The fork-join queue and related systems with synchronization constraints: Stochastic ordering and computable bounds. *Advances in Applied Probability*, 21(3):629–660, 1989.

[19] Raghu Raj Bahadur and R Ranga Rao. On deviations of the sample mean. *The Annals of Mathematical Statistics*, 31(4):1015–1027, 1960.

[20] August A Balkema, Paul Embrechts, and Natalia Nolde. Meta densities and the shape of their sample clouds. *Journal of Multivariate Analysis*, 101(7):1738–1754, 2010.

[21] Rocco Ballerini and Sidney I Resnick. Records from improving populations. *Journal of Applied probability*, 22(3):487–502, 1985.

[22] Rocco Ballerini and Sidney I Resnick. Records in the presence of a linear trend. *Advances in Applied Probability*, 19(4):801–828, 1987.

[23] Ole Barndorff-Nielsen. On the limit behaviour of extreme order statistics. *The Annals of Mathematical Statistics*, 34(3):992–1002, 1963.

[24] Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef L Teugels. *Statistics of extremes: theory and applications*, volume 558. John Wiley & Sons, 2004.

[25] George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[26] Narayan Bhat. *An introduction to queueing theory: modeling and analysis in applications*, volume 36. Springer, 2008.

[27] Marco Bijvank, Woonghee Tim Huh, Ganesh Janakiraman, and Wanmo Kang. Robustness of order-up-to policies in lost-sales inventory systems. *Operations Research*, 62(5):1040–1047, 2014.

[28] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[29] Nicholas H Bingham, Charles M Goldie, and Jozef L Teugels. *Regular variation*. Number 27. Cambridge University Press, 1989.

[30] Ramesh Bollapragada, Uday S Rao, and Jun Zhang. Managing two-stage serial inventory systems under demand and supply uncertainty and customer service level requirements. *IIE transactions*, 36(1):73–85, 2004.

[31] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[32] Andrei N Borodin and Paavo Salminen. *Handbook of Brownian motion-facts and formulae*. Springer Science & Business Media, 2015.

[33] Sem Borst, Avi Mandelbaum, and Martin I Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, 2004.

[34] James R Bradley and Peter W Glynn. Managing capacity and inventory jointly in manufacturing systems. *Management Science*, 48(2):273–288, 2002.

[35] Fabien Brosset, Thierry Klein, Agnès Lagnoux, and Pierre Petit. Large deviations at the transition for sums of Weibull-like random variables. In *Séminaire de Probabilités LI*, pages 239–257. Springer, 2022.

[36] Aadhaar Chaturvedi and Victor Martínez-de Albéniz. Safety stock, excess capacity or diversification: Trade-offs under supply and demand uncertainty. *Production andOperations Management*, 25(1):77–95, 2016.

[37] Hong Chen and David D Yao. *Fundamentals of queueing networks: Performance, asymptotics, and optimization*, volume 4. Springer, 2001.

[38] Jacob W Cohen. Some results on regular variation for distributions in queueing and fluctuation theory. *Journal of applied probability*, 10(2):343–353, 1973.

[39] Delta Commission. Beschouwingen over stormvloeden en getijbeweging, 1960.

[40] Harald Cramér. Sur un nouveau théorème-limite de la théorie des probabilités. *Actualités scientifiques et industrielles*, 736:5–23, 1938.

[41] Jim G Dai and J Michael Harrison. Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *The Annals of Applied Probability*, 2(1):65–86, 1992.

[42] Yves Dallery, Zhen Liu, and Don Towsley. Equivalence, reversibility, symmetry and concavity properties in fork-join queuing networks with blocking. *Journal of the ACM (JACM)*, 41(5):903–942, 1994.

[43] Yves Dallery, Zhen Liu, and Don Towsley. Properties of fork/join queueing networks with blocking under various operating mechanisms. *IEEE Transactions on Robotics and Automation*, 13(4):503–518, 1997.

[44] Richard A Davis, Edward Mulrow, and Sidney I Resnick. Almost sure limit sets of random samples in $\mathbb{R}^d$. *Advances in Applied Probability*, 20(3):573–599, 1988.

[45] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38. Springer Science & Business Media, 2009.

[46] Krzysztof Dębicki, Enkelejd Hashorva, Lanpeng Ji, and Kamil Tabiś. Extremes of vector-valued Gaussian processes: Exact asymptotics. *Stochastic Processes and their Applications*, 125(11):4039–4065, 2015.

[47] Krzysztof Dębicki, Lanpeng Ji, and Tomasz Rolski. Exact asymptotics of component-wise extrema of two-dimensional Brownian motion. *Extremes*, 23(4):569–602, 2020.

[48] Ronald A Doney. Stochastic bounds for Lévy processes. *The Annals of Probability*, 32(2):1545–1552, 2004.

[49] Monroe D Donsker. *An invariance principle for certain probability limit theorems*, volume 6. Memoirs of the American Mathematical Society, 1951.

[50] Andrzej Duda and Tadeusz Czachórski. Performance evaluation of fork and join synchronization primitives. *Acta Informatica*, 24(5):525–553, 1987.

[51] Meyer Dwass. Extremal processes. *The Annals of Mathematical Statistics*, 35(4):1718–1725, 1964.

[52] Meyer Dwass. Extremal processes, II. *Illinois Journal of Mathematics*, 10(3):381–391, 1966.

[53] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

[54] Carl-Gustav Esséen. On the Liapunov limit error in the theory of probability. *Arkiv för Matematik, Astronomi och Fysik*, 28:1–19, 1942.

[55] Lloyd Fisher. Limiting sets and convex hulls of samples from product measures. *The Annals of Mathematical Statistics*, 40(5):1824–1832, 1969.

[56] Ronald A Fisher and Leonard HC Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, volume 24, pages 180–190. Cambridge University Press, 1928.

[57] Leopold Flatto. Two parallel queues created by arrivals with two demands II. *SIAM Journal on Applied Mathematics*, 45(5):861–878, 1985.

[58] Leopold Flatto and Sann Hahn. Two parallel queues created by arrivals with two demands I. *SIAM Journal on Applied Mathematics*, 44(5):1041–1053, 1984.

[59] Dao H Fuk and Sergey V Nagaev. Probability inequalities for sums of independent random variables. *Theory of Probability & Its Applications*, 16(4):643–660, 1971.

[60] Ayalvadi J Ganesh. *Big queues*. Springer Science & Business Media, 2004.

[61] Noah Gans, Ger Koole, and Avishai Mandelbaum. Telephone call centers: tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

[62] Jean Geffroy. Contributions à la théorie des valeurs extrêmes. *Publications de l'Institut de statistique de l'Université de Paris*, 7:37–185, 1958.

[63] Paul Glasserman. Bounds and asymptotics for planning critical safety stocks. *Operations Research*, 45(2):244–257, 1997.

[64] Boris Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, pages 423–453, 1943.

[65] Claude Godreche, Satya N Majumdar, and Gregory Schehr. Record statistics of a strongly correlated time series: random walks and Lévy flights. *Journal of Physics A: Mathematical and Theoretical*, 50(33):333001, 2017.

[66] Ragavendran Gopalakrishnan, Sherwin Doroudi, Amy R Ward, and Adam Wierman. Routing and staffing when servers are strategic. *Operations Research*, 64(4):1033–1050, 2016.

[67] Laurens de Haan and Ana Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2006.

[68] Rasoul Haji and Gordon F Newell. A relation between stationary queue and waiting time distributions. *Journal of Applied Probability*, 8(3):617–620, 1971.

[69] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[70] Mor Harchol-Balter. Open problems in queueing theory inspired by datacenter computing. *Queueing Systems*, 97(1):3–37, 2021.

[71] J Michael Harrison. *Brownian motion and stochastic flow systems*. Wiley New York, 1985.

[72] J Michael Harrison. *Brownian models of performance and control*. Cambridge University Press, 2013.

[73] Woonghee Tim Huh, Ganesh Janakiraman, John A Muckstadt, and Paat Rusmevichientong. Asymptotic optimality of order-up-to policies in lost sales inventory systems. *Management Science*, 55(3):404–420, 2009.

[74] Donald L Iglehart and Ward Whitt. Multiple channel queues in heavy traffic. I. *Advances in Applied Probability*, 2(1):150–177, 1970.

[75] Donald L Iglehart and Ward Whitt. Multiple channel queues in heavy traffic. II: Sequences, networks, and batches. *Advances in Applied Probability*, 2(2):355–369, 1970.

[76] Shigeru Kanemitsu, Hiroshi Kumagai, Hari M Srivastava, and Masami Yoshimoto. Some integral and asymptotic formulas associated with the Hurwitz zeta function. *Applied Mathematics and Computation*, 154(3):641–664, 2004.

[77] Frank Karsten, Marco Slikker, and Geert-Jan van Houtum. Inventory pooling games for expensive, low-demand spare parts. *Naval Research Logistics (NRL)*, 59(5):311–324, 2012.

[78] David G Kendall. Some problems in the theory of dams. *Journal of the Royal Statistical Society. Series B (Methodological)*, 19(2):207–233, 1957.

[79] Harry Kesten. Convergence in distribution of lightly trimmed and untrimmed sums are equivalent. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 113, pages 615–638. Cambridge University Press, 1993.

[80] Cheeha Kim and Ashok K Agrawala. Analysis of the fork-join queue. *IEEE Transactions on Computers*, 38(2):250–255, 1989.

[81] Alan C Kimber. A note on Poisson maxima. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 63(4):551–552, 1983.

[82] John FC Kingman. The single server queue in heavy traffic. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 57, pages 902–904. Cambridge University Press, 1961.

[83] John FC Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society: Series B (Methodological)*, 24(2):383–392, 1962.

[84] Stephanus J de Klein. *Fredholm integral equations in queueing analysis*. PhD thesis, Rijksuniversiteit Utrecht, 1988.

[85] Steffen T Klosterhalfen, Stefan Minner, and Sean P Willems. Strategic safety stock placement in supply networks with static dual supply. *Manufacturing and Service Operations Management*, 16(2):204–219, 2014.

[86] Charles Knessl. On the diffusion approximation to a fork and join queueing model. *SIAM Journal on Applied Mathematics*, 51(1):160–171, 1991.

[87] Sung-Seok Ko and Richard F. Serfozo. Response times in $M/M/s$ fork-join networks. *Advances in Applied Probability*, 36(3):854–871, 2004.

[88] Sung-Seok Ko and Richard F Serfozo. Sojourn times in $G/M/1$ fork-join networks. *Naval Research Logistics (NRL)*, 55(5):432–443, 2008.

[89] Dmitry Korshunov. On distribution tail of the maximum of a random walk. *Stochastic Processes and their Applications*, 72(1):97–103, 1997.

[90] Steven Kou and Haowen Zhong. First-passage times of two-dimensional Brownian motion. *Advances in Applied Probability*, 48(4):1045–1060, 2016.

[91] Sunil Kumar and Ramandeep S Randhawa. Exploiting market size in service systems. *Manufacturing & Service Operations Management*, 12(3):511–526, 2010.

[92] Tze L Lai and Herbert Robbins. A class of dependent random variables and their maxima. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 42(2):89–111, 1978.

[93] Lucien Le Cam. An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181–1197, 1960.

[94] Malcolm R Leadbetter, Georg Lindgren, and Holger Rootzén. *Extremes and related properties of random sequences and processes*. Springer Science & Business Media, 1983.

[95] Johan SH van Leeuwaarden, Britt WJ Mathijsen, and Bert Zwart. Economies-of-scale in many-server queueing systems: tutorial and partial review of the QED Halfin-Whitt heavy-traffic regime. *SIAM Review*, 61(3):403–440, 2019.

[96] David V Lindley. The theory of queues with a single server. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 277–289. Cambridge University Press, 1952.

[97] John DC Little. A proof for the queuing formula: $L = \lambda W$. *Operations research*, 9(3):383–387, 1961.

[98] Hongyuan Lu and Guodong Pang. Gaussian limits for a fork-join network with nonexchangeable synchronization in heavy traffic. *Mathematics of Operations Research*, 41(2):560–595, 2015.

[99] Hongyuan Lu and Guodong Pang. Heavy-traffic limits for a fork-join network in the Halfin-Whitt regime. *Stochastic Systems*, 6(2):519–600, 2017.

[100] Hongyuan Lu and Guodong Pang. Heavy-traffic limits for an infinite-server fork-join queueing system with dependent and disruptive services. *Queueing Systems*, 85(1-2):67–115, 2017.

[101] Yi Lu, Qiaomin Xie, Gabriel Kliot, Alan Geller, James R Larus, and Albert Greenberg. Join-Idle-Queue: a novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation*, 68(11):1056–1071, 2011.

[102] Michel Mandjes. *Large deviations for Gaussian queues: modelling communication networks*. John Wiley & Sons, 2007.

[103] Andrea Marin and Sabina Rossi. Dynamic control of the join-queue lengths in saturated fork-join stations. In *International Conference on Quantitative Evaluation of Systems*, pages 123–138. Springer, 2016.

[104] Andrea Marin and Sabina Rossi. Power control in saturated fork-join queueing systems. *Performance Evaluation*, 116:101–118, 2017.

[105] Andrea Marin, Sabina Rossi, and Matteo Sottana. Biased processor sharing in fork-join queues. In *International Conference on Quantitative Evaluation of Systems*, pages 273–288. Springer, 2018.

[106] Maria E Mayorga and Hyun-Soo Ahn. Joint management of capacity and inventory in make-to-stock production systems with multi-class demand. *European Journal of Operational Research*, 212(2):312–324, 2011.

[107] Mirjam Meijer, Dennis Schol, Willem van Jaarsveld, Maria Vlasiou, and Bert Zwart. Extreme-value theory for large fork-join queues, with applications to high-tech supply chains. `https://arxiv.org/abs/2105.09189`, 2021.

[108] Carlos Mena, Andrew Humphries, and Thomas Y Choi. Toward a theory of multi-tier supply chain management. *Journal of Supply Chain Management*, 49(2):58–77, 2013.

[109] Reinhard Michel. On the constant in the nonuniform version of the Berry-Esséen theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55(1):109–117, 1981.

[110] Thomas Mikosch and Aleksandr V Nagaev. Large deviations of heavy-tailed sums with applications in insurance. *Extremes*, 1(1):81–110, 1998.

[111] Aleksandr V Nagaev. Integral limit theorems taking large deviations into account when Cramér's condition does not hold. I. *Theory of Probability & Its Applications*, 14(1):51–64, 1969.

[112] Aleksandr V Nagaev. Integral limit theorems taking large deviations into account when Cramér's condition does not hold. II. *Theory of Probability & Its Applications*, 14(2):193–208, 1969.

[113] Sergey V Nagaev. Some limit theorems for large deviations. *Theory of Probability & Its Applications*, 10(2):214–235, 1965.

[114] Sergey V Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, pages 745–789, 1979.

[115] Jayakrishnan Nair, Adam Wierman, and Bert Zwart. Provisioning of large-scale systems: the interplay between network effects and strategic behavior in the user base. *Management Science*, 62(6):1830–1841, 2016.

[116] Jayakrishnan Nair, Adam Wierman, and Bert Zwart. *The fundamentals of heavy tails: Properties, emergence, and estimation*, volume 53. Cambridge University Press, 2022.

[117] Randolph Nelson and Asser N Tantawi. *Approximating task response times in fork/join queues.* IBM Thomas J. Watson Research Division, 1987.

[118] Randolph Nelson and Asser N Tantawi. Approximate analysis of fork/join synchronization in parallel queues. *IEEE Transactions on Computers*, 37(6):739–743, 1988.

[119] Viên Nguyen. Processing networks with parallel and sequential tasks: Heavy traffic analysis and Brownian limits. *The Annals of Applied Probability*, pages 28–55, 1993.

[120] Viên Nguyen. The trouble with diversity: fork-join networks with heterogeneous customer population. *The Annals of Applied Probability*, pages 1–25, 1994.

[121] Wenting Pan and Kut C So. Component procurement strategies in decentralized assembly systems under supply uncertainty. *IIE Transactions*, 48(3):267–282, 2016.

[122] Valentin V Petrov. *Sums of independent random variables.* Springer, Berlin, 1975.

[123] James Pickands III. Moment convergence of sample extremes. *The Annals of Mathematical Statistics*, 39(3):881–889, 1968.

[124] James Pickands III. Asymptotic properties of the maximum in a stationary Gaussian process. *Transactions of the American Mathematical Society*, 145:75–86, 1969.

[125] James Pickands III. Upcrossing probabilities for stationary Gaussian processes. *Transactions of the American Mathematical Society*, 145:51–73, 1969.

[126] Vladimir I Piterbarg. *Asymptotic methods in the theory of Gaussian processes and fields*, volume 148. American Mathematical Society, 1996.

[127] Yuri Prohorov. Transient phenomena in processes of mass service. *Litovski Matematicheski Sbornik*, 3:199–205, 1963.

[128] Zhan Qiu, Juan F Pérez, and Peter G Harrison. Beyond the mean in fork-join queues: Efficient approximation for response-time tails. *Performance Evaluation*, 91:99–116, 2015.

[129] Youri Raaijmakers, Sem Borst, and Onno Boxma. Fork–join and redundancy systems with heavy-tailed job sizes. *Queueing Systems*, pages 1–29, 2022.

[130] Kondreddy N Reddy and Akhilesh Kumar. Capacity investment and inventory planning for a hybrid manufacturing-remanufacturing system in the circular economy. *International Journal of Production Research*, 59(8):2450–2478, 2021.

[131] Josh Reed and Bo Zhang. Managing capacity and inventory jointly for multi-server make-to-stock queues. *Queueing Systems*, 86(1–2):61–94, 2017.

[132] Sidney I Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

[133] Sidney I Resnick and Michael Rubinovitch. The structure of extremal processes. *Advances in Applied Probability*, 5(2):287–307, 1973.

[134] Dennis Schol, Maria Vlasiou, and Bert Zwart. Large fork-join queues with nearly deterministic arrival and service times. *Mathematics of Operations Research*, 47(2):1335–1364, 2021.

[135] Dennis Schol, Maria Vlasiou, and Bert Zwart. Extreme values for the waiting time in large fork-join queues. *in preparation*, 2022.

[136] Dennis Schol, Maria Vlasiou, and Bert Zwart. Maximum waiting time in heavy-tailed fork-join queues. `https://arxiv.org/abs/2211.02313`, 2022.

[137] Dennis Schol, Maria Vlasiou, and Bert Zwart. Tail asymptotics for the delay in a Brownian fork-join queue. `https://arxiv.org/abs/2208.04796`, 2022.

[138] Adam Shwartz and Alan Weiss. Induced rare events: analysis via large deviations and time reversal. *Advances in Applied Probability*, 25(3):667–689, 1993.

[139] Karl Sigman and Ward Whitt. Heavy-traffic limits for nearly deterministic queues. *Journal of Applied Probability*, 48(3):657–678, 2011.

[140] Karl Sigman and Ward Whitt. Heavy-traffic limits for nearly deterministic queues: stationary distributions. *Queueing Systems*, 69(2):145, 2011.

[141] Gordon Simons and Norman L Johnson. On the convergence of binomial to Poisson distributions. *The Annals of Mathematical Statistics*, 42(5):1735–1736, 1971.

[142] David Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41(2):463–501, 1962.

[143] Andrei Sleptchenko, Matthijs C van der Heijden, and Aart van Harten. Trade-off between inventory and repair capacity in spare part networks. *Journal of the Operational Research Society*, 54(3):263–272, 2003.

[144] Toby Sterling. Intel orders ASML system for well over $340 mln in quest for chipmaking edge. `https://www.reuters.com/technology/intel-orders-asml-machine-still-drawing-board-chipmakers-look-an-edge-2022-01-19/`. Accessed: 2023-01-18.

[145] Xiaoming Tan and Charles Knessl. A fork-join queueing model: Diffusion approximation, integral representations and asymptotics. *Queueing Systems*, 22(3):287–322, 1996.

[146] Alexander Thomasian and Asser N Tantawi. Approximate solutions for $M/G/1$ fork/join synchronization. In *Proceedings of Winter Simulation Conference*, pages 361–368. IEEE, 1994.

[147] Muhammad Tirmazi, Adam Barker, Nan Deng, Muhammad E Haque, Zhijing Gene Qin, Steven Hand, Mor Harchol-Balter, and John Wilkes. Borg: the next generation. In *Proceedings of the fifteenth European conference on computer systems*, pages 1–14, 2020.

[148] Aad W van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[149] Elizabeth Varki. Mean value technique for closed fork-join networks. *ACM SIGMETRICS Performance Evaluation Review*, 27(1):103–112, 1999.

[150] Elizabeth Varki, Arif Merchant, and Hui Chen. The $M/M/1$ fork-join queue with variable sub-tasks. `https://www.cs.unh.edu/~varki/publication/2002-nov-open.pdf`, 2002.

[151] Subir Varma. *Heavy and light traffic approximations for queues with synchronization constraints*. PhD thesis, University of Maryland, 1990.

[152] Subir Varma and Armand M Makowski. Interpolation approximations for symmetric fork-join queues. *Performance Evaluation*, 20(1-3):245–265, 1994.

[153] Weina Wang, Mor Harchol-Balter, Haotian Jiang, Alan Scheller-Wolf, and Rayadurgam Srikant. Delay asymptotics and bounds for multitask parallel jobs. *Queueing Systems*, 91(3):207–239, 2019.

[154] Ward Whitt. Some useful functions for functional limit theorems. *Mathematics of Operations Research*, 5(1):67–85, 1980.

[155] Paul E Wright. Two parallel processors with coupled inputs. *Advances in Applied Probability*, 24(4):986–1007, 1992.

[156] Cathy H Xia, Zhen Liu, Don Towsley, and Marc Lelarge. Scalability of fork/join queueing networks with blocking. *ACM SIGMETRICS Performance Evaluation Review*, 35(1):133–144, 2007.

[157] Yun Zeng, Jian Tan, and Cathy H Xia. Fork and join queueing networks with heavy tails: Scaling dimension and throughput limit. *Journal of the ACM (JACM)*, 68(3):1–30, 2021.

[158] Songfeng Zheng. An improved Bennett's inequality. *Communications in Statistics-Theory and Methods*, 47(17):4152–4159, 2018.

[159] Ryszard Zielinski. Optimal nonparametric quantile estimators. Towards a general theory. A survey. *Communications in Statistics-Theory and Methods*, 38(7):980–992, 2009.

[160] Xuxia Zou, Shaligram Pokharel, and Rajesh Piplani. Channel coordination in an assembly system facing uncertain demand with synchronized processing time and delivery quantity. *International Journal of Production Research*, 42(22):4673–4689, 2004.

Summary

The work in this thesis is inspired by modeling delays in supply chains for high-tech manufacturers, such as ASML, Philips Healthcare, and Boeing; these supply chains are large. A typical property is that many high-tech suppliers specialize in producing and delivering a specific component of the final product. In this system, the slowest supplier determines the delay of the manufacturer.

To model this delay, we consider the N-server fork-join queueing network, in which each server represents a unique supplier, and the arrival stream denotes orders from the manufacturer. The literature on this network for large N is scarce. First, we investigate the behavior of the longest queue and the longest waiting time by proving limiting results as the number of servers N converges to infinity. Second, we propose centralized base-stock and capacity policies to minimize costs incurred by delays. To achieve these objectives, we use results from extreme-value theory, diffusion approximations, large deviations principles, theory on heavy-tailed random variables, and newsvendor problems.

In Chapter 2, we assume the arrivals and services to be nearly deterministic. The aim of this study is to approximate the length of the largest of the N queues in the network. We present a fluid limit and a steady-state result for the maximum queue length, as N goes to infinity. In order to achieve this fluid limit, we have to scale time and space appropriately, where these scalings depend on the number of servers N. We extend these results to a fork-join queue with non-zero initial queue lengths with few assumptions on the distribution of these initial queue lengths. To prove the fluid limit for this case, we obtain an extreme-value result on the sum of two independent random variables, which is of independent interest.

In Chapter 3, the main result we derive is a second-order convergence result on the longest waiting time using a scaling that is more refined than the one in Chapter 2. The rescaled longest waiting time converges in distribution to a Gaussian random variable. By applying distributional Little's law, we show that a similar convergence result holds for the maximum queue length. Finally, we present a similar convergence result for the Brownian fork-join queue in steady-state.

We focus on the Brownian fork-join queue in Chapter 4 as well, here we derive large deviations results for the maximum queue length. We show that there are two regimes of large deviations, one in which the dependence structure between queue lengths is recognized in the limit, and one where we see asymptotic independence between the queue lengths.

In Chapter 5, we model the longest waiting time in an N-server fork-join queue with

heavy-tailed services, which is motivated by applications in parallel cloud computing. We derive a fluid limit and a steady-state result, where the limiting process is a supremum of a drifted process with Fréchet marginals.

In Chapter 6, we apply earlier obtained results in an industrial setting. Namely, we wish to optimize the performance of the supply chain of a multi-component assembly system involving original equipment manufacturers. Specifically, we model the supply chain with an N-server Brownian fork-join queue. We argue the validity of the Brownian fork-join queue and look at the system in steady state. Thus, we model the queue lengths as all-time suprema of Brownian motions with drift. As each server faces the same arrivals, these queue lengths are again mutually dependent. We define a newsvendor problem where costs are caused by (i) the maximum queue length, which is a measure of the delay in the system, (ii) the base-stock level, and (iii) the average speed of service. Now, as the cumulative distribution function cannot be written down explicitly, it is impossible to solve the resulting minimization problem directly. However, we use the central limit result derived in Chapter 3 to obtain an explicit approximate solution of the minimization problem. We prove asymptotic optimality of this approximate solution and test its performance using numerical experiments.

About the author

Dennis Schol was born in Dirksland, the Netherlands, on May 20, 1995. He finished his secondary education at CSG Prins Maurits in Middelharnis in 2013. In 2016, Dennis obtained a bachelor's degree in Technische Wiskunde at Eindhoven University of Technology (TU/e), and in 2018, he received a master's degree in Industrial and Applied Mathematics at TU/e.

In October 2018, Dennis started as a PhD candidate in the Department of Mathematics and Computer Science at TU/e. Under the supervision of Prof. Dr. Maria Vlasiou and Prof. Dr. Bert Zwart, he studied the extreme values of large fork-join queueing systems. During his PhD project, Dennis collaborated with Dr. Mirjam Meijer and Dr. Willem van Jaarsveld from the department of Industrial Engineering & Innovation Sciences at TU/e. Dennis attended several national and international conferences to present his work. The most important results of this PhD research are described in this thesis.

Dennis will defend his PhD thesis on May 16, 2023.