Scaling limits for critical queues with diminishing populations

Scaling limits
for critical queues
with diminishing populations

This work was financially supported by the European Research Council (ERC) and The Netherlands Organisation for Scientific Research (NWO).

Scaling limits for critical queues with diminishing populations,
by Gianmarco Bet

A catalogue record is available from the Eindhoven University of Technology library
ISBN: 978-90-386-4325-0

Cover design by Giulia Fongaro
Printed by Gildeprint, Enschede

SCALING LIMITS FOR CRITICAL QUEUES WITH DIMINISHING POPULATIONS

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
de Technische Universiteit Eindhoven, op gezag van
de rector magnificus, prof.dr.ir. F.P.T. Baaijens, voor
een commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen op
donderdag 11 september 2017 om 14:00 uur

door

Gianmarco Bet

geboren te Conegliano, Italië

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

voorzitter:	prof.dr.ir. B. Koren
1e promotor:	prof.dr. J.S.H. van Leeuwaarden
2e promotor:	prof.dr. R.W. van der Hofstad
leden:	prof.dr. R. Núñez-Queija (Universiteit van Amsterdam)
	dr. S. Bhamidi (University of North Carolina)
	prof.dr. E.A. Cator (Radboud Universiteit)
	dr. H. Honnappa (Purdue University)
	prof.dr. A.P. Zwart

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

Acknowledgments

A thesis is, in a very literal sense, a fruit of one's sweat and tears throughout the years. A thesis is also, in a broader sense, a collaborative effort. In fact, many people have directly or indirectly contributed with their own unique perspective to this particular piece of writing. I have but this small space to thank them all.

It is no hyperbole to state that the research contained here would not have been possible if not for the steadfast guidance of my supervisors, Johan van Leeuwaarden and Remco van der Hofstad. Johan, your joyful enthusiasm is contagious and your vast knowledge and interests are inspiring. I have benefited immensely from your experience and your passion for good writing. I like to think that, as a result of our time together, I have become a much better researcher. Remco, your extraordinary mathematical insight will never cease to amaze me. In the last four years, day after day you have surprised me with new perspectives and exciting ideas on topics that I (foolishly) thought I mastered. Your unending flow of ideas has been a constant inspiration, and was the original source of many of the results contained in this thesis. You have also introduced me to the difficult art of back-of-the-envelope computations. On top of all this, paraphrasing another student of yours, you are one of the nicest guys I know!

I am very thankful to Shankar Bhamidi, Eric Cator, Harsha Honnappa, Sindo Núñez-Queija and Bert Zwart for accepting to be in my defense committee and for taking the time to read this thesis.

Eurandom provides a warm and welcoming environment for researchers from all over the world to meet and exchange ideas. Thanks to the organizational work of Patty Koorn, I have been able to attend more conferences than I can count, all of them conveniently taking place right in front of my office. Thanks also to Chantal Reemers and Petra Rozema for all the help with administrative duties, you are some of the most friendly and helpful persons I know! I sincerely hope my ineptitude in bureaucratic matters has not caused too much stress to you.

One of the perks of working in such an open environment is the pleasure of discussing with many passionate and very smart researchers. Thanks to Francesca Nardi for our discussions on metastability, to Júlia Komjáthy for our discussions on explosive branching processes and to Bert Zwart for our discussions on large deviations. Many thanks to Shankar Bhamidi for being a most welcoming host during my visit to the University of North Carolina in Chapel Hill, and for all our discussions on critical digraphs and other topics related to my research.

During 2015 I have had the pleasure of supervising the bachelor project of Richard Post. Richard, your perseverance and enthusiasm made you the best student I could ever hope for. Thank you for all the insightful discussions. I have absolutely no doubt that you will be able to accomplish anything you will set your mind on. Nevertheless, I wish you the best of luck in your future endeavours. I have also been an instructor in numerous mathematics courses. I thank the one anonymous student who, showing unexpected compassion, wrote the only positive evaluation of my teaching.

Remaining sane during a PhD is sometimes not an easy task. For me, it would have been impossible without all the good friends in Eindhoven and Padova. I will try to acknowledge as many as these pages allow me to.

A heartfelt thank you goes to my 'artistic supervisors' Iris and Mathias for the countless hours of practice. Your unbridled passion for music is contagious and inspiring.

The Thursday nights spent rehearsing with my fellow actors Chris, Leo, Marina, Martijn and Raja of the Doppio theatre student association have been incredibly fun. Thanks also to our director Merja for successfully herding this unruly group of cats. Our hard work eventually paid off; as our most positive reviewer pointed out, our play was "not as bad as I expected".

When I set out to learn Dutch, I had no idea I would pick up a very good friend along the way. Thank you Jorn, I will miss our weekly bilingual chats.

Being part of the departmental PhD Council was a constant source of inspiration. Together, we organized many exciting events for the entire PhD community. A thank you goes to all my co-organizers for actually doing most of the work and to our former (but glorious) leader Christine for being always eager to lend a hand.

The occasional trip to distant, magical lands allowed me to maintain an healthy detachment from reality (so fundamental for every mathematician). Thank you Alberto, Antonio, Bea, Fabio and Tommaso for the (very) late-night sessions of Dungeons and Dragons and numerous other board games.

It is a well-known folk theorem that, wherever you will find yourself in the world, you will meet at least one Italian. This is particularly true for Eindhoven, where the Italians are secretly staging an invasion of the city. I wish to thank Alessandro Jr., Alessandro Sr., Carlo, Chiara, Enrico, Fabio, Lorenzo, Misa, Marta, Melania and Tommaso for making me feel at home even in this faraway land.

I have been very lucky to be part of such a social and open group of PhD students. Thank you all. I am particularly indebted to my fellow PhDs Britt, Fabio, Jori and Thomas, with whom I shared this remarkable journey. I wish you all the best. I was always looking forward to going to the office thanks to my lovely office mates Alessandro, Debankur, Kay, Marta and Jori. Together, we elevated the coffee break to an art form. A special thanks goes to Marta for (perhaps unwillingly) being a caring adopted mother to all of us.

Alessandro, we have worked, lived and often trained together every day for a remarkably long period of time. Somehow, this marriage-like arrangement worked exceptionally well. I will dearly miss our times together. A thank you goes to my house mates Fabio, Sándor and Stefano as well. Barbecue season will not be the same without you.

Jori, where do I even begin? In the last four years you have been a colleague, an office mate, a collaborator, a good friend and, lately, a rival in the conquest of the galaxy. We have worked, discussed and travelled together. You have a sharp mind and kind ways. You have taught me a lot and as such I owe much to you. Thank you.

Thank you to all the good friends in Padova who, every time I returned, made me feel as if I had never left.

Finally, my deepest thank you goes to mamma Luisa and papà Pietro for the unfaltering faith in me and the selfless support for my ambitions throughout the years.

A. and Z. are two endearing letters, the beginning and the end. Everything that is to come, all the subtleties of life, cannot but be experienced through them. Wherever the ride will go, it will be a good one.

Gianmarco Bet
July 2017

Contents

CHAPTER 1

Introduction

1.1 Queues with finite populations

Queueing phenomena occur whenever there is competition for a scarce resource. Often, this resource is time and the competition is over the attention of a server. As a result, abstract models of queues share some defining primitive features: *customers* arrive at a service station, where they require the attention of one or more *servers*. It is often not known precisely when a certain customer will require service, or for how long, and this assumption leads to the study of queues from a probabilistic point of view.

The queueing literature is to a large extent built on the assumption that customer arrivals are governed by some renewal process, an assumption that allows the use of powerful probabilistic techniques based on ergodic theory. In this thesis, however, we consider a *transitory* queueing model, known in the literature as the $\Delta_{(i)}/G/1$ queue, which operates only a finite amount of time and cannot be viewed as a standard regenerative process. The $\Delta_{(i)}/G/1$ queue assumes a finite population of customers entering the queue only once. As time progresses, more customers have joined the queue, and fewer customers can potentially join. This modelling assumption of a *diminishing population* of customers gives rise to a class of reflected stochastic processes that lack a stationary distribution, and instead display relevant behavior only during a finite time window. Therefore, only the time-dependent behavior is of interest.

Assume the arrival times of the customers are sampled independently from an identical distribution. The arrival times are then the order statistics of the sample, and the interarrival times are the differences of order statistics. Further assuming a single server, and generally distributed independent service times, this model was coined the $\Delta_{(i)}/G/1$ queue by Honnappa, Jain and Ward in [48], where they established fluid and diffusion limits for the queue-length process. The same authors introduced in [49] a wider class of transitory queues, with the $\Delta_{(i)}/G/1$ queue still as the prime example, and stochastic-process limits were established for large population sizes.

We will introduce a new heavy-traffic regime for the $\Delta_{(i)}/G/1$ queue, leading to stochastic-process limits and heavy-traffic approximations. Considering queueing processes in their critical regimes typically leads to a reduction in complexity, since the complicated processes can often be shown to converge to much simpler limiting stochastic processes. Stochastic-process limits have been studied for single-server queues that have a time-varying arrival rate. Newell [75, 76, 77, 78] pioneered this direction by deriving diffusion approximations, see also [58, 67]. Rigorous results in terms of stochastic-process limits were obtained by Mandelbaum and Massey in [71] (building on [73, 74]). Here, stochastic-process limits were established as refinements to deterministic ODE limits for the time-dependent $M/M/1$ queue, also known as the $M_t/M_t/1$ queue. See also [100] for a systematic treatment of the $M_t/G/1$ queue. The technique used in [71] to develop Functional Law of Large Numbers (FLLN) and Functional Central Limit Theorem (FCLT) results uses strong approximations and what is known as the *uniform accelaration* (UA) technique. UA relies on the assumption that the relevant time scale for changes in the queue-length process is of the order $O(1/\varepsilon)$ for some $\varepsilon > 0$. Accelerating the process in a uniform manner by scaling the arrival and service rates by ε then reveals the dominant model behavior as $\varepsilon \to 0$. While in [48, 49] the arrival and service rates are scaled in a similar manner, the time scale considered is of the order $O(1)$. In particular, the time of the process is not scaled. The UA technique has been extensively applied to non-stationary queueing systems with non-homogeneous Poisson input, but it remained unclear whether it is also useful for transitory queueing models as considered in [49]. We will show how the key idea behind the UA technique can be applied to these models. We shall now explain our approach in terms of the easiest setting, in which the identical distribution that generates the arrival times is exponential.

Let us now give the details of the basic $\Delta_{(i)}/G/1$ queue that plays a central role in this thesis. Assume a finite population of n customers, with

n very large, and where each customer has an independent exponential clock with mean $1/\lambda$. Customers join the queue when their clocks ring. The initial arrival process (close to time zero) is then roughly Poisson with rate $n\lambda$. However, as time progresses, the arrival intensity decreases, due to those customers that have left the system. Thus, the arrival process is a Poisson process that is thinned according to some time-dependent rule. Denote the mean service time by $\mathbb{E}[S] = 1/\mu$. In order to create heavy-traffic conditions we let the population size n grow to infinity, while at the same time making sure that the (initial) traffic intensity $\rho_n = n\lambda/\mu$ is close to one. The system can initially be underloaded (when $n\lambda < \mu$), overloaded (when $n\lambda > \mu$), or critically-loaded (when $n\lambda \approx \mu$). In case $n\lambda > \mu$, the queue initially shows a roughly linear increase and therefore the correct scaling of the queue length to obtain meaningful limits is n for a first order approximation (FLLN) and $n^{1/2}$ for second order approximations (FCLT). In particular, no time scaling is needed to obtain these approximations; these are the most relevant approximations obtained in [48].

We focus on the critically-loaded regime, and we combine this with UA through the population size n. In the spirit of UA, we let the arrival and service rates scale with n while also rescaling time so as to observe the queue-length process at a time scale of order $O(1/n^{\gamma})$ for some $\gamma > 0$. Denote the density of the arrival distribution as $f_T(t)$ (with $f_T(t) = \lambda \mathrm{e}^{-\lambda t}$, $t \geq 0$ as an important special case corresponding to exponential arrivals), and denote the i.i.d. service requirements of consecutive customers by $S_1, S_2, \ldots$ with generic random variable S. Assuming a service rate of n, the service times of consecutive customers are then $D_1 = S_1/n, D_2 = S_2/n, \ldots$ with generic random variable D. For the sake of clarity, we now first give a simplified version of the more general heavy-traffic condition (1.1.2) below. The heavy-traffic regime we consider is given by the condition

$$\rho_n := n f_T(0)\mathbb{E}[D] = f_T(0)\mathbb{E}[S] = 1, \quad \text{for } n \text{ large.} \tag{1.1.1}$$

For the exponential case, $f_T(0) = \lambda$, so that the condition reads $\rho_n = \lambda\mathbb{E}[S] = 1$ and can be interpreted as follows: For times close to zero, the expected number of newly arriving customer during one service time is roughly one. For general service times, the condition can be understood by interpreting $f_T(t)$ as the *instantaneous arrival rate* in t. Since we consider time scales of the order $O(1/n^{\gamma})$, only the mass in zero $f_T(0)$ matters for describing the new arrivals.

We shall actually consider a slightly more general definition of the random variables D_i and this leads to the more precise critical scaling

$$\rho_n = n f_T(0)\mathbb{E}[D] = 1 + \beta n^{-\eta}, \tag{1.1.2}$$

where $\eta > 0$ depends on the specific details of the model. The additional term $\beta n^{-\eta}$ arises from detailed calculations, but can be interpreted as the factor that describes the onset of the heavy-traffic period: when $\beta > 0$ (resp. < 0) the queue is initially slightly overloaded (resp. slightly underloaded).

The heavy-traffic regime (1.1.2) is defined by two features: The customer pool n grows to infinity and the initial (at time zero) rate of newly arriving customers is such that, on average, one new customer is expected to arrive during one service time. This gives rise to a large-scale system that (initially) operates close to full utilization, and is expected to utilize its resources efficiently. By this we mean that the server is typically busy, and that idle times are negligible. In fact, we will characterize the conditions under which sufficiently many customers will join the queue to guarantee that the system will have a substantial backlog of customers. We therefore will focus on the first busy period, and show how to set the initial number of customers already present in the queue at time $t = 0$, referred to as the *head start*, to create a considerable first busy period during which the server can work continuously.

It is clear that the $\Delta_{(i)}/G/1$ queue is strongly influenced by the service-time distribution. In particular, the heavy-traffic behavior is crucially different depending on whether the variance of the service-time distribution is finite or not. In the first part of this thesis we will assume that $\mathbb{E}[S^2] < \infty$. In this case, the queueing process is in the domain of attraction of Brownian motion. We will take the queue-length process and scale space and time. The resulting stochastic-process limit will turn out to be a (reflected) Brownian motion with quadratic drift. The latter process is defined as $\widehat{X}(t) = at + bt^2/2 + cW(t)$ with $(W(t))_{t\geq 0}$ a standard Brownian motion, and a, b, c constants. The constant b is negative, so that eventually the free process $(\widehat{X}(t))_{t\geq 0}$ drifts to minus infinity according to $bt^2/2$, causing the reflected process to be essentially stuck at zero. This is due to the *diminishing population* effect. One could interpret the quadratic term in the limit as the (cumulative) effect of the customers already served not being able to join the queue again. The stochastic-process limit provides insight into the macroscopic behavior (for n large) of the transitory queueing process, and the different phenomena occurring at different space-time scales. It also gives insight into the orders of the average queue lengths and the time scales of busy periods.

1.1.1 Comparison with known results

The first asymptotic results for the $\Delta_{(i)}/G/1$ queue were proven by Iglehart and Whitt in their seminal paper [51]. They prove that, when the arrival clocks are uniformly distributed, the fluctuations of the arrival counting process around its mean are given by a Brownian bridge. They do so through the theory of convergence of probability measures [17]. Building on the same framework, Honnappa, Jain and Ward [48, 49] perform an extensive analysis of the asymptotic behavior of the $\Delta_{(i)}/G/1$ queue. They show that the macroscopic asymptotic behavior of the $\Delta_{(i)}/G/1$ queue is given by

$$\frac{Q_n(t)}{n} \overset{\text{a.s.}}{\to} \bar{Q}(t) := \phi(F_T(t) - t/\mathbb{E}[S]), \tag{1.1.3}$$

where $t \mapsto F_T(t)$ is the cumulative density function of T and $\phi(f)(x) := f(x) - \inf_{y \leq x} f^-(y)$ is the reflection map; see Figure 1.1 for an example. The asymptotic approximation $Q_n(t) \approx n\bar{Q}(t)$ given by (1.1.3) can be further refined through an FCLT, describing the fluctuations of order $n^{1/2}$ around $n\bar{Q}(t)$. The resulting stochastic approximation [48] is given by

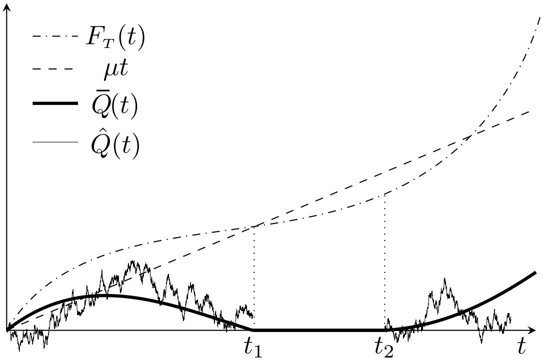

Figure 1.1: The fluid (thick line) and diffusion (thin line) limits for the $\Delta_{(i)}/G/1$ queue with arrival distribution $F_T(t)$, and service rate $1/\mathbb{E}[S] = \mu$. For $t \in [t_1, t_2]$, $\bar{Q}(t) = \widehat{Q}(t) = 0$.

$$\frac{Q_n(t) - n\bar{Q}(t)}{n^{1/2}} \overset{\text{d}}{\to} \widehat{Q}(t), \tag{1.1.4}$$

where $\widehat{Q}(\cdot)$ is a discontinuous process switching between three regimes: a free Brownian motion, a driftless, reflected Brownian motion, and the zero process; see Figure 1.1. The process $\widehat{Q}(\cdot)$ is obtained by applying a complicated functional to the sample paths of the sum of a Brownian bridge and a Brownian motion. In particular, the Brownian bridge arising in (1.1.4) represents the large-time, macroscopic effect of the finite population of customers.

Our work also has connections with the work of Mandelbaum and Massey [71] for the $M_t/M_t/1$ queue, who derive a fluid approximation through a FLLN and use this approximation to classify various operating regimes. In this setting, our results corresponds to the 'Onset of Critical Loading' regime [71, Theorem 3.4] and the results of [48] correspond to the FLLN and the FCLT [71, Theorems 2.1 and 2.2].

1.1.2 Other transitory models

We say that a queueing model is *transitory* if, denoting by $\mathcal{A}_n(t)$ the cumulative number of customers who have entered the queue by time t, almost surely [47, 49]

$$\lim_{t\to\infty} \mathcal{A}_n(t) < \infty. \tag{1.1.5}$$

Then, the $\Delta_{(i)}/G/1$ queue is a natural model for transitory queues. In fact, under mild assumptions all transitory queueing models satisfy the same FLLN (1.1.3) and FCLT (1.1.4) as the $\Delta_{(i)}/G/1$ queue [49]. Because of this, the $\Delta_{(i)}/G/1$ queue can be considered the standard model for transitory queues. Even the $M_t/M_t/1$ queue, which is time-inhomogeneous but not transitory, has the same asymptotic behavior as the $\Delta_{(i)}/G/1$ queue.

Let us elaborate on the relation between the two models by drawing a connection between the arrival process of the $\Delta_{(i)}/G/1$ queue and the arrival process of the $M_t/G/1$ queue. It is well known that, for a Poisson point process on the positive line with intensity function $f(t)$, conditioned on there being n points in an interval $[0, T]$, the points themselves are i.i.d. with distribution function $t \mapsto f(t)/\int_0^T f(s)\,\mathrm{d}s$. In particular, if the Poisson point process has finite total intensity $\int_0^\infty f(s)\,\mathrm{d}s < \infty$, conditioned on there being n points, they are independently and identically distributed over the positive line. Therefore, we can see the $\Delta_{(i)}/G/1$ model as a *conditioned* $M_t/G/1$ queue. More broadly, it is possible to model a transitory queue by considering a general, rather than Poisson, point process conditioned to have n points in a given time interval. This has been named the *conditional arrivals model* in [49]. Therefore, the conditional arrivals model corresponds to a $G_t/G/1$ queue [98] conditioned on n customers joining in total. However, if the underlying point process is not Poisson, the n points will not be i.i.d. under the conditioned measure and thus, in principle, the conditioned arrivals model and the $\Delta_{(i)}/G/1$ queue will behave differently. Nevertheless, as $n \to \infty$ and under additional assumptions, all conditioned arrival models satisfy (1.1.3) and (1.1.4) [49].

If the arrival times $t_1, t_2, \ldots, t_n$ of the n customers are fixed and prescheduled, and each customer arrives at a random time $t_i + \zeta_i$ close to its assigned time, the resulting queue is also transitory according to (1.1.5). This class of models is widely used in the study and optimization of airport operations [42, 94], and health clinics with appointment systems [43]. Unlike in the $\Delta_{(i)}/G/1$ queue, the arrival times are decidedly not i.i.d., even if the random unpunctualities of the customers are. However, if the number of customers is large, their assigned arrival times are evenly distributed in a finite interval $[0, T]$, and the support of the random noise ζ_i is small, then the cumulative arrivals process is well approximated by the cumulative distribution function of a uniform random variable on $[0, T]$. Intuitively, in the limit all the customers are statistically equivalent and the arrival time of a uniformly chosen customer will also be approximately uniform, if the support of ζ_i is sufficiently small. Therefore, for large n the $\Delta_{(i)}/G/1$ queue gives an asymptotically exact *mean-field* approximation of the pre-scheduled arrivals model, where customers are replaced by statistically equivalent entities.

1.1.3 The critical regime

In order to study the *critical* $\Delta_{(i)}/G/1$ queue, different time and space scalings than (1.1.4) are necessary. Let us sketch how these arise in the case of exponentially distributed arrival times with mean $1/\lambda$. Assuming that the system is never overloaded, that is $\sup_{t\geq 0} f_T(t)\mathbb{E}[S] \leq 1$, as well as (1.1.2), the macroscopic approximation of the queue given by (1.1.3) is identically zero. We then look for κ, γ such that

$$n^{-\kappa} Q_n(tn^{-\gamma}) \xrightarrow{\text{d}} \widehat{Q}(t). \tag{1.1.6}$$

As a result of (1.1.3), the leading order behavior of the cumulative number of arrivals $\mathcal{A}_n(t)$ is given by $nF_T(t)$. By Taylor expanding $F_T(tn^{-\gamma})$ as

$$F_T(tn^{-\gamma}) = 1 - \mathrm{e}^{\lambda tn^{-\gamma}} = \lambda tn^{-\gamma} - \lambda^2/2t^2n^{-2\gamma}, \tag{1.1.7}$$

and ignoring for the moment the first order term, we see that, if $-\kappa - 2\gamma + 1 = 0$, $Q_n(t)$ will have a deterministic drift given by $-\lambda^2/2t^2$. Moreover, since the service rate is rescaled to be n, in a time interval $[0, t]$, roughly

$tn^{1-\gamma}$ customers are served. The second order behavior of the number of services in $[0,t]$ will be, accordingly, of order $n^{(1-\gamma)/2}$. Therefore, in order to obtain a meaningful limit, we must have $n^{\kappa} = n^{(1-\gamma)/2}$, that is $2\kappa + \gamma - 1 = 0$. The two conditions above imply that $\kappa = \gamma = 1/3$. We will prove the following result:

Theorem 1 (Critically loaded $\Delta_{(i)}/G/1$ queue with exponential arrivals)**.** *Assume that the service times $(S_i)_{n=1}^{n}$ satisfy $\mathbb{E}[S^2] < \infty$ and that the heavy-traffic condition* (1.1.2) *holds with $\eta = 1/3$. Then*

$$n^{-1/3}Q_n(\cdot n^{-1/3}) \xrightarrow{\mathrm{d}} \phi(\widehat{X})(\cdot), \tag{1.1.8}$$

where $\widehat{X}(\cdot)$ is the diffusion process

$$\widehat{X}(t) = \beta\lambda t - \frac{\lambda^2}{2}t^2 + \sigma W(t), \tag{1.1.9}$$

with $\sigma^2 := \lambda^3\mathbb{E}[S^2]$ and $W(\cdot)$ a standard Brownian motion.

The limit (1.1.8) reveals important properties of the critical $\Delta_{(i)}/G/1$ queue. The negative quadratic drift is unique to time-inhomogeneous queueing models and encodes the transition of the system from heavy-traffic (critical load) to stability. This phenomenon is often referred to as *depletion-of-points effect*. The scaling limit result in (1.1.8) also implies that the depletion-of-points effect has a key impact on the performance of the system already at a short time scale of order $n^{-1/3}$. This should be contrasted with (1.1.4), where the effect of the finite population of customers is represented, in the limit, by a Brownian bridge. Refinements of the result (1.1.8) provide further insights on the performance of the queue, for example prescribing how to regulate the server speed and the initial number of customers in the queue in order to obtain a sizeable first busy period.

As an immediate consequence of Theorem 1 we get an asymptotic result for the length of the first busy period in the $\Delta_{(i)}/G/1$ queue, which we call BP_n. We assume that the queue length at time zero is deterministic and grows with n as

$$\lim_{n\to\infty} \frac{Q_n(0)}{n^{1/3}} = q > 0. \tag{1.1.10}$$

The lenght of the first busy period depends in a crucial way on both β and q since

$$n^{1/3}\mathrm{BP}_n \xrightarrow{\mathrm{d}} T^{\beta\lambda}_{\widehat{X}_q}(0), \tag{1.1.11}$$

where $T^{\beta\lambda}_{\widehat{X}_q}(0)$ is the time until the process $\widehat{X}_q(\cdot)$ crosses level 0, with

$$\widehat{X}_q(t) = q + \beta\lambda t - \frac{\lambda^2}{2}t^2 + \sigma W(t). \tag{1.1.12}$$

As above, $\sigma^2 = \lambda^3\mathbb{E}[S^2]$ and $W(\cdot)$ is a standard Brownian motion.

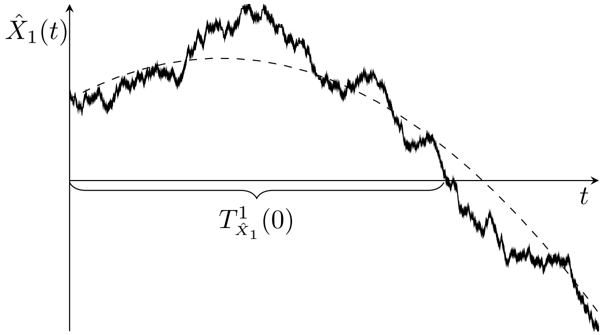

Figure 1.2: A sample path of $\widehat{X}_1(t)$ with $\beta = 1$ (solid) and the drift $1 + t - \frac{1}{2}t^2$ (dashed).

Equation (1.1.11) can be used to obtain numerical approximations of quantities related to the first busy period of the critically loaded $\Delta_{(i)}/G/1$ queue. In [72] an explicit expression for the first crossing time of zero of $\widehat{X}_q(\cdot)$ is given, and numerical simulations are carried out to display how the shape of the density crucially depends on the two parameters q and β. Let $\mathrm{Ai}(x)$ and $\mathrm{Bi}(x)$ denote the classical Airy functions [1]. When $\lambda = 1$, the first crossing time of zero of $\widehat{X}_q(\cdot)$ has probability density [72]

$$\begin{aligned} &f_q(t;\beta,\sigma) \\ &\quad = \mathrm{e}^{-((t-\beta)^3+\beta^3)/6\sigma^2-\beta a} \\ &\qquad \times \int_{-\infty}^{+\infty} \mathrm{e}^{tu}\frac{\mathrm{Bi}(cu)\mathrm{Ai}(c(u-a)) - \mathrm{Ai}(cu)\mathrm{Bi}(c(u-a))}{\pi(\mathrm{Ai}(cu)^2 + \mathrm{Bi}(cu)^2)}\mathrm{d}u, \end{aligned} \tag{1.1.13}$$

where $c = (2\sigma^2)^{1/3}$ and $a = q/\sigma^2 > 0$. Figure 1.3 shows the convergence of the empirical density function of the first excursion length of a $\Delta_{(i)}/G/1$ queue to the analytic expression (1.1.13) and also illustrates the influence of the parameters β and q.

Figure 1.3 suggests that to obtain a considerable first busy period, both parameters β and q must be chosen appropriately in order to avoid

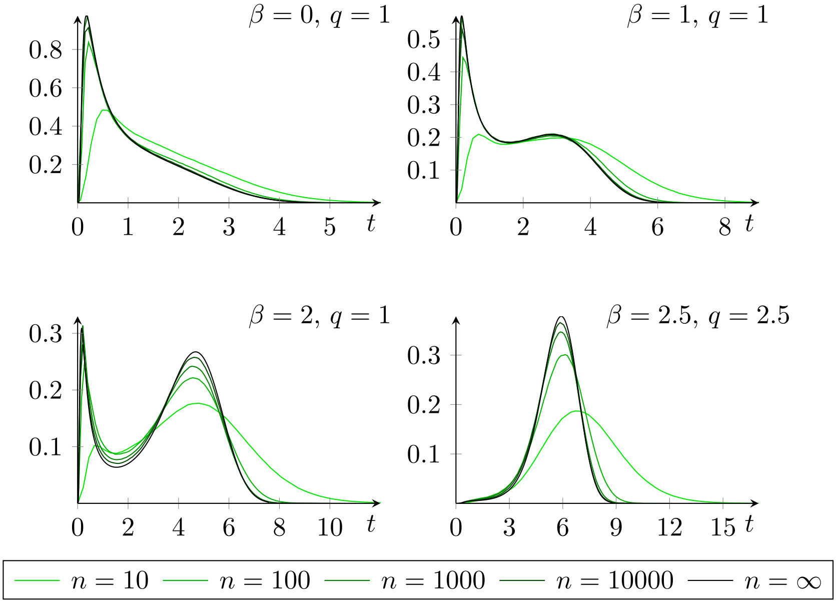

Figure 1.3: Density plot (black line) and Gaussian kernel density estimates (colored lines) of the first busy period length. The plots for finite n were obtained by averaging the results of 10^7 simulations.

a concentration of the probability mass close to zero. Indeed, in [72] it is conjectured that there exists a $\bar{q}$ such that for all $q > \bar{q}$ and every choice of β the distribution function is unimodal, while for $q < \bar{q}$ there exists a $\bar{\beta} = \bar{\beta}(q)$ such that for $\beta < \bar{\beta}$ the distribution is unimodal, and bimodal otherwise.

We conclude by showing numerical values for the mean busy period for exponential clock times with mean 1 and different values of q and β in Table 1.1. Observe that the approximation $\mathbb{E}[\mathrm{BP}_n] \approx n^{-1/3}\mathbb{E}[T^{\beta}_{\hat{X}}(0)]$ is accurate also for moderate values of n.

1.1.4 The embedded queue

We will provide an indirect proof and a direct proof of Theorem 1. In the indirect proof, we first prove that some *embedded* queueing process converges, and then argue that the difference between the embedded

	$q = 1, \beta = 1$		$q = 2, \beta = 1$	
n	$n^{1/3}\mathbb{E}[\mathrm{BP}_n]$	rel. error	$n^{1/3}\mathbb{E}[\mathrm{BP}_n]$	rel. error
10	3.0201	0.5072	4.0407	0.4079
100	2.2170	0.1062	3.2611	0.1362
1000	2.0341	0.0151	2.9813	0.0387
10000	2.0306	0.0133	2.9351	0.0226
100000	2.0295	0.0128	2.9145	0.0155
∞	2.0038	—	2.8701	—

Table 1.1: Mean busy period for the pre-limit queue with different population sizes and the exact expression for $n = \infty$ computed using (1.1.13). Each value for the pre-limit queue is the average of 10^4 simulations.

queue and the queue-length process is negligible. Let $Q_n^e(k)$ denote the number of customers in the queue just after the service completion of the k-th customer. The *embedded* queue-length process $Q_n^e(\cdot)$ is given by $Q_n^e(0) = q \geq 0$ and

$$Q_n^e(k) = (Q_n^e(k-1) + A_n(k) - 1)^+, \qquad k = 1, 2, \dots \tag{1.1.14}$$

with $x^+ = \max\{0, x\}$ and $A_n(k)$ the number of arrivals during the service time of the k-th customer. For the exponential case, $A_n(k)$ is given by

$$A_n(k) = \sum_{i \notin \nu_k} \mathbb{1}_{\{T_i \leq D_k\}} \tag{1.1.15}$$

where ν_k is the set of customers that is no longer in the population at the beginning of the service of the k-th customer. The system defined in (1.1.14) and (1.1.15) neglects idle times, which is a simplification of the $\Delta_{(i)}/G/1$ model. This in turn will greatly simplify the analysis since it allows for the representation of the process as (1.1.14) and (1.1.15).

It is possible to give an equivalent definition of the process $Q_n^e(\cdot)$ in (1.1.14) through the reflection map. Define the process $N_n(\cdot)$ by $N_n(0) = q$ and

$$N_n(k) = N_n(k-1) + A_n(k) - 1. \tag{1.1.16}$$

Then it is easy to see that

$$(Q_n^e(k))_{k \geq 0} = (\phi(N_n)(k))_{k \geq 0}. \tag{1.1.17}$$

We will first prove a limit theorem for $N_n(\cdot)$ and then apply the reflection map to obtain a limit for the queue-length process $Q^e_n(\cdot)$ defined in (1.1.14). For the case of exponential arrivals, assuming $N_n(0) = qn^{1/3}$, the free process $N_n(\cdot)$ converges to

$$n^{-1/3}N_n(\cdot n^{2/3}) \xrightarrow{\mathrm{d}} \widehat{N}(\cdot), \tag{1.1.18}$$

where

$$\widehat{N}(t) := q + \beta t - \frac{1}{2}t^2 + \sigma W(t) \tag{1.1.19}$$

with $\sigma^2 = \lambda^2\mathbb{E}[S^2]$ and $W(\cdot)$ a standard Brownian motion. Consequently, the reflected processes also converge as

$$n^{-1/3}Q^e_n(tn^{2/3}) \xrightarrow{\mathrm{d}} \phi(\widehat{N})(t). \tag{1.1.20}$$

The embedded queue neglects both idle times and the fluctuations of the queue-length process during one service. Therefore, in order to deduce Theorem 1 from (1.1.18), first we will show that the cumulative idle time in the critical $\Delta_{(i)}/G/1$ queue is negligible in the limit. Second, we will bound the fluctuations of the queue-length process during the service of one customer.

1.1.5 The queue-length process

The second proof of Theorem 1 is based on a direct representation of the $\Delta_{(i)}/G/1$ queue-length process. Denote the number of customers who arrive in the interval $[0, t]$ by

$$\mathcal{A}_n(t) = \sum_{i=1}^{n} \mathbb{1}_{\{T_i \le t\}}. \tag{1.1.21}$$

Let

$$\sigma_n(t) = \max\left\{k \ge 0 \mid \sum_{i=1}^{k} \frac{S_i}{n} \le t\right\} \tag{1.1.22}$$

be the renewal process associated with the rescaled service times. If $Q_n(0) \ge 0$ denotes the number of customers already in the queue at the beginning of the first service, then the queue-length process $Q_n(t)$ is given by

$$Q_n(t) = Q_n(0) + \mathcal{A}_n(t) - \sigma_n(B_n(t)), \tag{1.1.23}$$

where the time change $t \mapsto B_n(t)$ represents the *cumulative busy time process*, which is constant if and only if the server is idling, and increases linearly otherwise. $B_n(\cdot)$ depends both on $(T_i)_{i=1}^{\infty}$ and $(S_i)_{i=1}^{\infty}$ and as such makes the analysis of $Q_n(\cdot)$ challenging. An approach pioneered by Iglehart and Whitt [50] consists in studying a related queue in which the server never idles, but rather continues working according to the renewal process associated with $(S_i)_{i=1}^{\infty}$ even when the queue is empty. This is often referred to as the queue with autonomous service or the Borovkov modified system, see [96, Chapter 10.2]. It turns out that, under mild assumptions, the original queue and the Borovkov modified system are asymptotically equivalent in heavy traffic [96, Theorem 10.2.2], in the sense that the distance between the two queue-length process converges to zero. However, for this approach to work, the service time limit process needs to be continuous. Indeed, the distance between the two processes is bounded from above by the (scaled) maximum service time, or equivalently the maximum jump functional applied to the service time process. When the service time limit process is continuous, the maximum jump functional converges to zero. If, on the other hand, the service time limit process is discontinuous, then the distance between the two queues cannot be shown to converge to zero.

Instead, we will adopt a different approach that will, among other things, allow us to deal with a discontinuous service time limit process. This consists in expressing $Q_n(\cdot)$ as the reflection of an appropriate free process $X_n(\cdot)$. Since, after rescaling, $X_n(\cdot)$ converges and the reflection mapping is continuous a.s. in the limit point, the process $Q_n(\cdot)$ also converges by the Continuous Mapping Theorem. The free process $X_n(\cdot)$ has the following interpretation: When the server is working, $X_n(\cdot)$ follows $Q_n(\cdot)$. When the queue is empty, $X_n(\cdot)$ decreases linearly at a rate proportional to the service rate. Therefore, while in the Borovkov modified system the server works continuously according to the service time renewal process, in the process $X_n(\cdot)$ the server provides instantaneous work with rate $1/\mathbb{E}[S]$ when there are no customers in the system. Consequently, the process $X_n(\cdot)$ can be seen as a fluid version of the Borovkov modified system. The process $Q_n(t)$ can then be represented as

$$Q_n(t) = \phi(X_n)(t), \qquad t \geq 0, \tag{1.1.24}$$

where $X_n(\cdot)$ is given by $X_n(0) = Q_n(0)$ and

$$X_n(t) = X_n(0) + \mathcal{A}_n(t) - \sigma_n(B_n(t)) - I_n(t)/\mathbb{E}[S]. \tag{1.1.25}$$

In Figure 1.4 we plot a sample path of the process $X_n(\cdot)$.

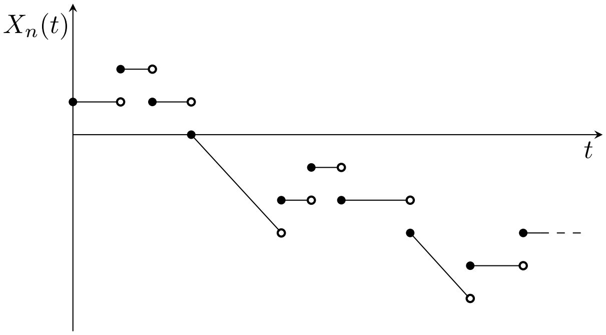

Figure 1.4: A sample path of the process $X_n(\cdot)$. When the queue is empty, $X_n(\cdot)$ has a constant negative slope of $-1/\mathbb{E}[S]$.

1.2 Heavy-tailed services

In some applications, the finite variance condition of Theorem 1 might not be realistic. When $\mathbb{E}[S^2] = \infty$, the random variable S is said to be *heavy tailed*. Heavy tails arise naturally in the areas of communication networks, insurance, and risk management. In the context of telecommunication traffic measurements, [27] collects empirical evidence that the distribution of available file sizes and transmission times is heavy tailed. See also [28, 99, 101] and references therein. Additionally, [27] shows that the file sizes and transmission times exhibit *power-law tails*, that is, when X denotes a generic file size or transmission time, the probability that X exceed a given threshold x is given by

$$\mathbb{P}(X > x) = cx^{-\gamma}, \qquad \gamma \in (0, 2), \tag{1.2.1}$$

for some $c > 0$, $x > c^{1/\gamma}$. Note that (1.2.1) implies that $\mathbb{E}[X^2] = \infty$. Several other characteristics of the World Wide Web exhibit power-law or heavy tails [60]. In the context of insurance companies, the time evolution of the monetary reserves is often governed by sizeable negative jumps with power-law tails, due to large but unpredictable claim sizes [7].

Later in this thesis we will drop the finite variance condition of Theorem 1 and study the queue-length process under the additional assumption that the service times are heavy tailed. More precisely, we assume that the service times follow a power-law distribution as (1.2.1) with power-law exponent $\gamma \in (1, 2)$. Under these assumptions, our model is the finite-pool analogue of the classical heavy-tailed $M/G/1$ queue; see below for a discussion. We will establish that in a similar heavy-traffic regime as in Theorem 1 the rescaled queue-length process converges to an γ-stable process with negative quadratic drift. Unlike Brownian motion, the γ-stable motion is discontinuous: it jumps an infinite number of times in every finite time interval. As a result, the γ-stable motion constitutes a suitable approximation for those processes that exhibit a large degree of burstiness and frequent large jumps. As in the finite variance case, the diminishing pool effect is still there in the form of the drift term, but the oscillations of the limiting queue length are much wilder. We will also show that, as a consequence of the larger fluctuations, the desired head start and canonical busy period should scale with n in a specific way that vitally depends on the exponent γ. Recall the definition of $X_n(\cdot)$ in (1.1.25). The behavior of the heavy-tailed $\Delta_{(i)}/G/1$ queue is contained in the following result:

Theorem 2 (The critically loaded heavy-tailed $\Delta_{(i)}/G/1$ queue)**.** *Assume that $(S_i)_{i=1}^n$ satisfy* (1.2.1) *and* (1.1.2)*, with $\eta = (\gamma-1)/(2\gamma-1)$. If $Q_n(0) = X_n(0) = qn^{1/(2\gamma-1)}$ for $q \geq 0$, then*

$$n^{-\frac{1}{2\gamma-1}} X_n(\cdot n^{-\frac{\gamma-1}{2\gamma-1}}) \xrightarrow{\text{d}} \widehat{X}(\cdot), \tag{1.2.2}$$

where

$$\widehat{X}(t) = q + \beta\lambda t - \frac{\lambda^2}{2}t^2 + s_\gamma \mathcal{S}(t), \tag{1.2.3}$$

$s_\gamma = \mathbb{E}[S]^{-1-1/\gamma}$ and $\mathcal{S}(\cdot)$ is a spectrally positive γ-stable process. Moreover,

$$n^{-\frac{1}{2\gamma-1}} Q_n(\cdot n^{\frac{1-\gamma}{2\gamma-1}}) \xrightarrow{\text{d}} \phi(\mathcal{N})(\cdot). \tag{1.2.4}$$

In Figure 1.5 we plot some sample paths of $\phi(\widehat{X})(\cdot)$ for different choices of γ for fixed q, β, λ, s_γ. We observe that as γ approaches 2, the reflected stable motion starts to resemble a reflected Brownian motion minus a quadratic drift. Figure 1.6 shows the first passage time as a function of the linear drift β, for fixed γ and q, and different values of the linear drift parameter β.

Similarly as in (1.1.11), we characterize the limiting distribution of the first busy period of the $\Delta_{(i)}/G/1$ queue $T_{Q_n}^{\beta\lambda}(0)$ as follows:

$$T_{Q_n}^{\beta\lambda}(0) \xrightarrow{\text{d}} T_{\widehat{X}}^{\beta\lambda}(0). \tag{1.2.5}$$

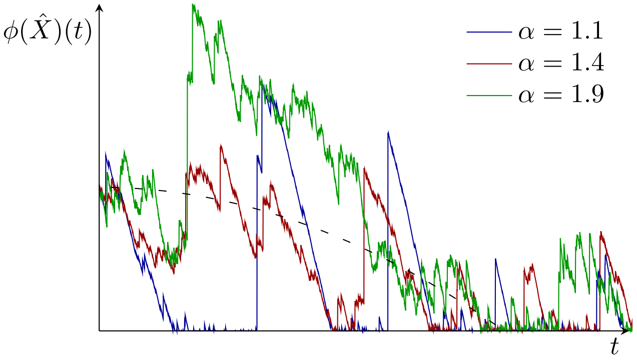

Figure 1.5: Sample paths of $\phi(\widehat{X})(\cdot)$ for different choices of the power-law exponent $\gamma \in (1,2)$. In all cases $q = 4$, $\beta = 0$, and $\lambda = s_\gamma = 1$. The dashed curve plots the function $t \mapsto 4 - t^2/2$.

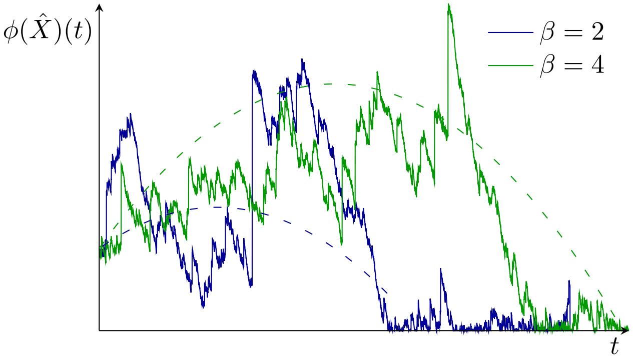

Figure 1.6: Sample paths of the process $\phi(\widehat{X})(t)$ for varying values of β. The dashed curves plot the functions $t \mapsto q + \beta\lambda t - \lambda^2/2t$. In all plots, $q = 4$, $\gamma = 1.8$, $\lambda = s_\gamma = 1$.

Figures 1.5 and 1.6 suggest that the hitting time $T_{d_{q,\beta}}(0)$ of the quadratic drift given by

$$d_{q_0,\beta}(t) := q + \beta\lambda t - \frac{\lambda^2}{2}t^2 \tag{1.2.6}$$

gives a first order approximation of $T_{\widehat{X}}^{\beta\lambda}(0)$. In particular, $T_{d_{q,\beta}}(0)$ is the solution of a quadratic equation, and is equal to

$$T_{d_{q,\beta}(\cdot)}(0) = \frac{-\beta + \sqrt{\beta^2 + 2q}}{\lambda}, \tag{1.2.7}$$

where we have assumed that $q \geq 0$. Note that the hitting time of zero of $\mathcal{S}(\cdot)$ is distributed as a $1/\gamma$ stable random variable by [86, Theorem 46.3]; see also [87]. The convergence result (1.2.5) allows us to estimate the tail probability for the length of the first busy period. In fact, we have the exact upper bound

$$\mathbb{P}(T^{\beta\lambda}_{\widehat{X}}(0) > t) \leq \mathbb{P}\Big(s_\gamma \mathcal{S}(t) > -q - \beta\lambda t + \frac{\lambda^2}{2} t^2\Big), \tag{1.2.8}$$

where we have used the trivial inclusion of events $\{T^{\beta\lambda}_{\widehat{X}}(0) > t\} \subseteq \{\widehat{X}(t) > 0\}$. By basic properties of stable laws, we have the asymptotic relation [85, pp. 16-17]

$$\begin{aligned} &\mathbb{P}\Big(Z_\gamma > \frac{1}{s_\gamma}(-qt^{-1/\gamma} - \beta\lambda t^{1-1/\gamma} + \frac{\lambda^2}{2} t^{2-1/\gamma})\Big) \\ &\quad \sim \frac{c_\gamma s_\gamma^\gamma}{(-qt^{-1/\gamma} - \beta\lambda t^{(\gamma-1)/\gamma} + \frac{\lambda^2}{2} t^{(2\gamma-1)/\gamma})^\gamma} \sim \frac{2^\gamma c_\gamma s_\gamma^\gamma}{\lambda^{2\gamma}} \frac{1}{t^{2\gamma-1}}, \end{aligned} \tag{1.2.9}$$

where Z_γ is distributed as a standard γ-stable law,

$$c_\gamma = \frac{1-\gamma}{\Gamma(2-\gamma)\cos(\pi\gamma/2)} \tag{1.2.10}$$

for $\gamma \neq 1$, and $t \mapsto \Gamma(t)$ is the standard Gamma function. On the other hand, due to the strong negative drift of $\widehat{X}(\cdot)$, it is natural to conjecture that the two events $\{T^{\beta\lambda}_{\widehat{X}}(0) > t\}$ and $\{\widehat{X}(t) > 0\}$ are of comparable measure when t is large. In Figure 1.7 we show that the tails of the empirical distribution of the first busy period behave like the upper bound (1.2.9); see [3, 46], where this is proven when $\mathcal{S}(\cdot)$ is replaced by a more complicated *thinned* Lévy process. However, the approximation becomes less effective as $\gamma \to 2$, and for $\gamma = 2$, (1.2.9) is not theoretically justified. In fact, for this finite variance case, Pittel [82] (see also [83, 84]) shows that the tail asymptotically behaves like

$$\mathbb{P}(T^{\beta\lambda}_{\widehat{X}}(0) > t) = \frac{1}{\sqrt{9\pi/8}t^{3/2}} \mathrm{e}^{-\frac{1}{8}t(t-2\beta)^2}(1 + o(1)), \qquad \text{as } t \to \infty. \tag{1.2.11}$$

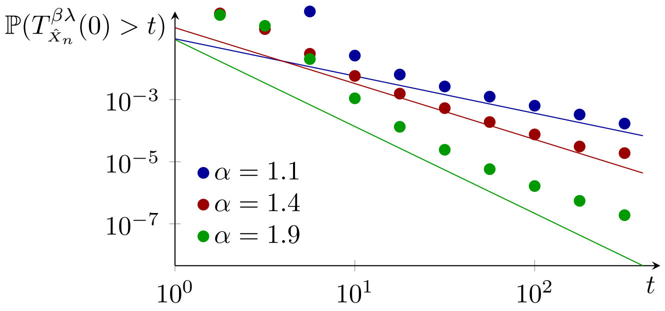

Figure 1.7: A log-log scale plot of the empirical tail distribution $\mathbb{P}(T^{\beta\lambda}_{\widehat{X}_n}(0) > t)$ of the first busy period of $Q_n(\cdot)$ for different values of $\gamma \in (1, 2)$. The solid lines represent the asymptotic approximation (1.2.9). In all plots, $n = 1000$ and $q = \beta = \lambda = s_\gamma = 1$.

1.3 Finite-pool queues and random graphs

We now introduce an extension of the $\Delta_{(i)}/G/1$ queue that we have called the $\Delta^{\alpha}_{(i)}/G/1$ queue. In this model, again n customers are triggered to join a queue after independent exponential times, but the rates of their exponential clocks depend on their service requirements. When a customer requires S units of service, its exponential clock rings after an exponential time with mean $\mathbb{E}[S^{-\alpha}]$ with $\alpha \in [0, 1]$. Depending on the value of the free parameter α, the arrival times are i.i.d. ($\alpha = 0$) or decrease with the service requirement ($\alpha \in (0, 1]$). For the case $\alpha = 0$, we retrieve the $\Delta_{(i)}/G/1$ queue with i.i.d. arrivals [48, 49], while the case $\alpha = 1$ is closely related to the critical inhomogeneous random graph studied in [15, 55].

The IRG is a generalization of the Erdős-Rényi random graph (ERRG) [36]. In the ERRG, each pair of different vertices chosen among n vertices is connected by an edge with a fixed probability p. The ERRG shows a very intricate behavior as the parameters n and p vary. In particular, the structure of the ERRG changes dramatically (it undergoes a *phase transition*) as $p = c/n$ crosses the critical threshold $c = 1$, when n is very large. More specifically, if $c > 1$, the largest connected component $\mathcal{C}_1$ of the ERRG will contain a positive fraction of all the n vertices (the so-called *giant component*), and all other connected components $\mathcal{C}_2, \mathcal{C}_3, \ldots$ will be negligible compared to $\mathcal{C}_1$. If, on the other hand $c < 1$, $\mathcal{C}_1$ will contain

$O(\log(n))$ vertices. The ERRG is said to be supercritical in the first case, and subcritical in the second. The behavior in the vicinity of $c = 1$ is more delicate and requires a finer analysis [5, 20, 68]. The most important result in this context was obtained by Aldous [5], who characterized the joint limit law of the sizes of the ordered components of the ERRG. The key insight of Aldous was that several characteristics of the ERRG can be encoded by a random walk representing the *exploration* of the random graph. This *exploration process* iteratively declares vertices as inactive, active and explored. All vertices are inactive at first. An arbitrary vertex $v(1)$ is declared active, and subsequently all its neighbours $\{v(2), v(3) \ldots\}$ are also declared active. Then, $v(1)$ is declared explored. The process then moves to $v(2)$ and repeats the steps; see Figure 1.8 and Figure 1.9.

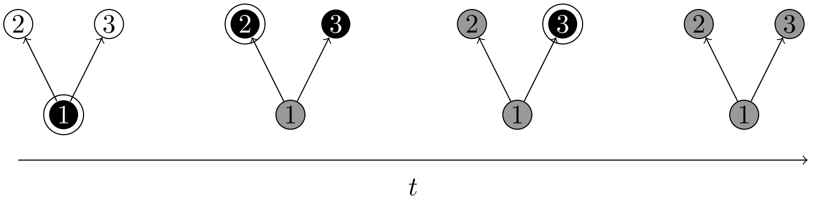

Figure 1.8: An example of an exploration process of a forest with one connected component (a tree). Active vertices are black, inactive vertices are white and explored vertices are grey. A circle around a node highlights which node is being explored. Vertices are numbered in order of appearence in the exploration.

As a consequence of the definition of the exploration process, the sizes of the ordered components are encoded by the time between successive minima of the process. In other words, this technique allowed Aldous to analyze the complicated random graph using powerful stochastic-processes tools. More specifically, Aldous proved that, if $c = 1 + \beta n^{-1/3}$ for $\beta \in \mathbb{R}$, then the exploration process converges, after appropriate rescaling, to the process $\widehat{N}(t) = \beta t - 1/2t^2 + \sigma W(t)$, where $W(t)$ is a standard Brownian motion. From this he concluded that the ordered component sizes of the ERRG, rescaled by $n^{2/3}$, converge to the times between successive minima of $\widehat{N}(t)$. The ERRG has received lots of attention in the past decades [21, 54], and continues to be a source of challenging problems [2].

The inhomogeneous random graph (IRG) generalizes the ERRG by assigning to each vertex i a (possibly random) weight $\mathcal{W}_i$. Vertices i and j

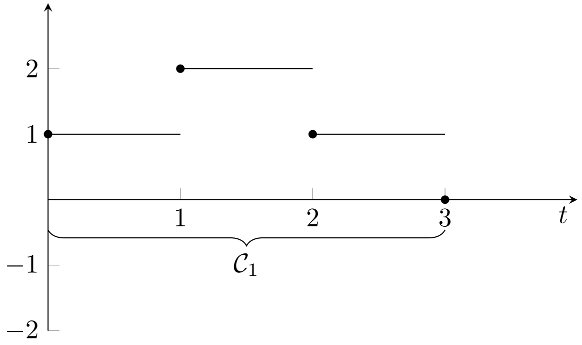

Figure 1.9: The plot of the exploration process associated with Figure 1.8. The component sizes are given by the difference between successive minima.

are connected (briefly $i \leftrightarrow j$) with probability

$$\mathbb{P}(i \leftrightarrow j \mid \mathcal{W}_i, \mathcal{W}_j) = 1 - \exp\Big(- \frac{\mathcal{W}_i \mathcal{W}_j}{\sum_{i=1}^n \mathcal{W}_i} \Big). \tag{1.3.1}$$

The graph thus constructed is also known as the Norros-Reittu random graph [79]. It is *inhomogeneous* because high weight vertices have a higher probability of having many neighbors. By choosing all weights equal to the same constant, we see that the ERRG is a special case of the IRG. Perhaps surprisingly, the phase transition behavior of the IRG is remarkably similar to the one of the ERRG [22]. In particular, if the weights $(\mathcal{W}_i)_{i=1}^n$ are i.i.d. with generic random variable w and $\mathbb{E}[\mathcal{W}^3] < \infty$, then in the critical regime $\mathbb{E}[\mathcal{W}^2]/\mathbb{E}[\mathcal{W}] = 1$, the distribution of the sizes of the connected components of the IRG converges to the distribution of the excursions above past minima of $\widehat{N}(t) = c_1 t - c_2 t^2 + c_3 W(t)$, for some $c_1 \in \mathbb{R}$, $c_2, c_3 > 0$ [15].

Now consider the embedded $\Delta_{(i)}^{\alpha}/G/1$ queue given by

$$Q_n^e = (Q_n^e(k-1) + A_n(k) - 1)^+, \tag{1.3.2}$$

where $A_n(k)$ denotes the number of arrivals during the k-th service. The probability that customer $i \in \{1, \dots, n\}$ joins during the service of customer j, conditioned on the rescaled service times S_i and S_j, is

$$\mathbb{P}(i \text{ joins during service of } j \mid S_j) = 1 - \exp\Big(- \frac{S_i^{\alpha} S_j}{n} \Big). \tag{1.3.3}$$

The similarity between (1.3.1) and (1.3.3) suggests that the two models are in fact closely related. More precisely, assume that initially in the queue there are q customers. Then, we construct a graph with vertex set $\{1, 2, \dots, n\}$ and in which two vertices i and j are joined by a *directed* edge (briefly $i \to j$) if and only if the i-th customer arrives during the service time of the j-th customer. Let us focus momentarily on the graph constructed from the first busy period of the queue. If $q = 1$, then the graph is a rooted tree with n labeled vertices, the root being labeled 1. If $q > 1$, then the graph is a forest consisting of q distinct rooted trees whose roots are labeled $1, \dots, q$ respectively. The total number of vertices in the forest is n. For this random graph model, we have

$$\mathbb{P}(i \to j \mid S_i, S_j) = 1 - \exp\Big(- \frac{S_i^\alpha S_j}{n} \Big), \tag{1.3.4}$$

corresponding to the situation where customer i joins the queue during the service of j. The service time S_i of customer i has the interpretation of the weight assigned to vertex i in the corresponding random graph. Moreover, the busy periods of the queue correspond to the connected components of the associated random graph. For $\alpha = 0$ we retrieve the standard $\Delta_{(i)}/G/1$ queue, while for $\alpha = 1$ the right-side expression is symmetric and we retrieve the IRG.

This random forest is exemplary for a deep relation between queues and random graphs, perhaps best explained by interpreting the embedded $\Delta_{(i)}/G/1$ queue as the exploration process of the corresponding graph. Let A_k denote the neutral neighbors of the k-th explored vertex. The exploration process then has increments $(A_k)_{k \geq 1}$ that each have a different distribution. The exploration process encodes useful information about the underlying random graph. For example, excursions above past minima are the sizes of the connected components. The critical behavior of random graphs connected with the emergence of a giant component has received tremendous attention [2, 13, 15, 16, 33, 55]. Interpreting active vertices as being in a queue, and vertices being explored as customers being served, we see that the exploration process and the (embedded) $\Delta_{(i)}^{\alpha}/G/1$ queue driven by $(A_n(k))_{k \geq 1}$ are identical.

The analysis of the $\Delta_{(i)}^{\alpha}/G/1$ queue and associated random forest is challenging because the random variables $(A_k)_{k \geq 1}$ are not i.i.d. In the case of i.i.d. $(A_k)_{k \geq 1}$, there exists an even deeper connection between queues and random graphs, established via branching processes instead of exploration processes [59]. To see this, declare the initial customers in the queue to be the 0-th generation. The customers (if any) arriving during the total service time of the initial i customers form the 1-st generation, and the customers (if any) arriving during the total service time of the customers in generation t form generation $t+1$ for $t \geq 1$. Through this connection, properties of branching processes can be carried over to the queueing processes and associated random graphs [34, 64, 66, 90, 91, 92]. Takács [90, 91, 92] proved several limit theorems for the case of i.i.d. $(A(k))_{k\geq 1}$, in which case the queue-length process and related processes such as the first busy period weakly converge to (functionals of) the Brownian excursion process. In that classical line, this thesis can be viewed as an extension to exploration processes with more complicated, non-Markovian, dependency structures in $(A_n(k))_{k\geq 1}$.

We will study the $\Delta_{(i)}^{\alpha}/G/1$ queue in heavy traffic, in a similar heavy-traffic regime as in (1.1.2). The initial traffic intensity ρ_n is kept close to one by imposing the relation

$$\rho = \lambda \mathbb{E}[S^{1+\alpha}](1+\beta n^{-1/3}) = 1+\beta n^{-1/3}. \tag{1.3.5}$$

In the $\Delta_{(i)}^{\alpha}/G/1$ queue the order of arrival of customers plays an important role. Accordingly, we define $c(i)$ as the i-th served customer, so that the arrival times of the customers are ordered as $T_{c(1)} \leq T_{c(2)} \leq \dots \leq T_{c(n)}$. The number of arrivals during the k-th service in (1.3.2) are given by

$$A_n(k) = \sum_{i\notin \nu_k} \mathbb{1}_{\{T_i \leq D_{c(k)}\}} \tag{1.3.6}$$

where $\nu_k \subseteq [n]$ is the set of customers no longer in the population at the beginning of the service of the k-th customer and D_i is the rescaled service time of customer i. Similarly as before, $Q_n^e(\cdot)$ in (1.3.2) can be alternatively represented as the reflected version of a process $N_n(\cdot)$, as

$$Q_n^e(k) = \phi(N_n)(k), \tag{1.3.7}$$

with $N_n(\cdot)$ given by $N_n(0) = Q_n^e(0) = q$ and by the recursion

$$N_n(k) = N_n(k-1) + A_n(k) - 1. \tag{1.3.8}$$

Whenever the server finishes processing one customer, and the queue is empty, the customer $c(i)$ to be placed into service is chosen according to the following size-biased distribution:

$$\mathbb{P}(c(i) = j \mid (S_i)_{i\in[n]}, \nu_{i-1}) = \frac{S_j^{\alpha}}{\sum_{l\notin \nu_{i-1}} S_l^{\alpha}}, \qquad j \notin \nu_{i-1}. \tag{1.3.9}$$

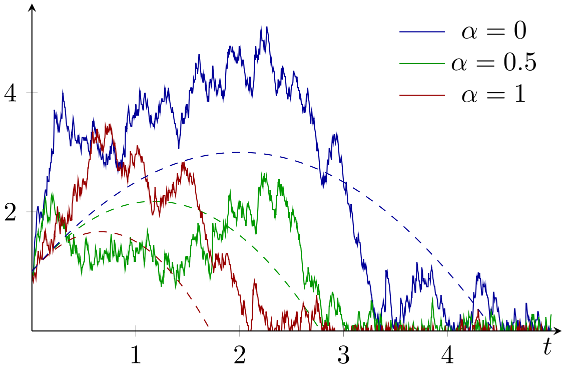

Figure 1.10: Sample paths of the process $n^{-1/3}Q^e_n(\cdot n^{2/3})$ for various values of α and $n = 10000$. The service times are taken unit-mean exponential. The dashed curves represent the drift $t \mapsto q + \beta t - \lambda\mathbb{E}[S^{1+2\alpha}]/(2\mathbb{E}[S^\alpha])t^2$. In all plots, $q = 1$, $\beta = 1$, $\lambda = 1/\mathbb{E}[S^{1+\alpha}]$.

We see that, with this choice, the process $N_n(\cdot)$ exactly describes the exploration process of the associated random graph and the total number of vertices in the tree (forest) is given by

$$T_{Q^e_n}(0) = \inf\{k \geq 0 : Q^e_n(k) = 0\}, \tag{1.3.10}$$

the hitting time of zero of the process $Q^e_n(\cdot)$. We will show that, for $\mathbb{E}[S^{2+\alpha}] < \infty$ and $N_n(0) = Q^e_n(0) = qn^{1/3}$,

$$n^{-1/3}N_n(tn^{2/3}) \xrightarrow{\mathrm{d}} \widehat{N}_q(t) = q + \beta t - \lambda\frac{\mathbb{E}[S^{1+2\alpha}]}{2\mathbb{E}[S^\alpha]}t^2 + \sigma W(t), \qquad n \to \infty \tag{1.3.11}$$

with $\sigma^2 = \lambda^2\mathbb{E}[S^\alpha]\mathbb{E}[S^{2+\alpha}]$ and $W(\cdot)$ a standard Brownian motion. By continuity arguments, this also implies that

$$n^{-1/3}Q^e_n(tn^{2/3}) \xrightarrow{\mathrm{d}} \phi(\widehat{N})(t), \tag{1.3.12}$$

see Figure 1.10.

As a straightforward consequence of (1.3.11) and (1.3.10)

$$|F_n| \xrightarrow{\mathrm{d}} T_{\widehat{N}_q}(0), \qquad n \to \infty, \tag{1.3.13}$$

where F_n denotes the cardinality of the tree constructed from the $\Delta^\alpha_{(i)}/G/1$ queue.

The result (1.3.11) extends (1.1.18) to the $\Delta_{(i)}^{\alpha}/G/1$ queue. A convergence result for the queue-length process of the $\Delta_{(i)}^{\alpha}/G/1$ queue follows from (1.3.11) by proving that, in the limit, the (cumulative) idle time is negligible and the embedded queue-length process is arbitrarily close to the queue-length process uniformly over compact intervals. We will develop this technique fully for the $\Delta_{(i)}/G/1$ queue, and will refrain from repeating it for the $\Delta_{(i)}^{\alpha}/G/1$ queue.

1.4 Outline

This thesis is organized as follows. In Chapter 2 we study the asymptotic behavior of the standard embedded $\Delta_{(i)}/G/1$ queue under the assumption that the variance of the service time is finite. First we treat the simpler case of exponentially distributed arrival times, and then we move to the more challenging general case. In both settings we show that, under a novel scaling regime and if the queue is critical in zero, the limit process is a Brownian motion with negative quadratic drift. We do this via discrete martingale techniques. Furthermore, we discuss a generalization of the results of Chapter 2 to arrival times whose first k derivatives are zero in zero. This assumption leads to a polynomial, rather than quadratic, drift in the limit process. Our results show that the limit process depends weakly on the arrival time distribution. Chapter 2 is based on Sections 4 and 5 of [10].

Chapter 3 expands on the ideas of Chapter 2 in various directions. We approach the asymptotic behavior of the $\Delta_{(i)}/G/1$ queue with two entirely different techniques. First, building on the results in Chapter 2, we show that the idle times are negligible and thus the embedded queue process and the $\Delta_{(i)}/G/1$ queue process have the same limit, up to rescaling of the coefficients. Second, we give a direct definition of the $\Delta_{(i)}/G/1$ queue process, in which the arrival process is given by an empirical distribution function, and prove directly its asymptotic behavior. We exploit this result to prove a sample path Little's Law, describing the relationship between the queue-length process and the virtual waiting time process. We conclude by proving that, if the $\Delta_{(i)}/G/1$ queue is subcritical in t, then $(Q_n(t))_{n\geq 1}$ is a tight family of random variables. The content of this chapter is based on the remaining sections of [10] and on [9].

In Chapter 4 we drop the finite variance assumption on the service times. More specifically, we assume that the service time is distributed as a power-law. We introduce a scaling regime that depends on the exponent of the power-law and that, under the appropriate heavy-traffic assumption, leads to a new scaling limit for the $\Delta_{(i)}/G/1$ queue. We show that in this setting the limit process is a pure jump process with a negative quadratic drift. The proof of this result makes use of a new representation of the arrival process as a thinned Poisson process (with time-dependent thinning). Chapter 4 is based on [12].

In Chapter 5 we introduce the $\Delta_{(i)}^{\alpha}/G/1$ queue and study its asymptotic behavior. We are able to prove a scaling limit for the $\Delta_{(i)}^{\alpha}/G/1$ queue by exploiting the martingale approach introduced in Chapter 2 and by carefully analysing the dependence structure of the arrival process. As a special case of this result, we retrieve the analogous result for the $\Delta_{(i)}/G/1$ queue in Chapter 2, and the well-known scaling limit for the critical inhomogeneous random graph. Chapter 5 is based on [11].

In Chapter 6 we depart from the stochastic-process limits framework. There, we study the $\Delta_{(i)}/G/1$ queue with a fixed and finite number of customers n. Assuming that both service times and arrival times are exponentially distributed, the resulting process jointly describing the queue length and the number of served customers is an absorbing Markov process. Exploiting the recursive structure of this process, we derive an explicit expression for the joint probability mass function of the number of customers in the first busy period and the maximum number of customers simultaneously in the queue during the first busy period.

Finally, in Chapter 7 we discuss our findings in the broader context of time-inhomogeneous queues and the random graph literature. We identify the shortcomings of our results and suggest ways to deal with them. Lastly, we discuss various interesting open problems that originated from the research conducted for this thesis.

CHAPTER 2

The embedded queue

In this chapter we study the asymptotic behavior of the embedded queue-length process of the heavy-traffic $\Delta_{(i)}/G/1$ queue by analyzing an approximating discrete-time process. We prove, first for exponentially distributed arrival times, and later for general arrival times, that when the second moment of the service time is finite, the approximating process converges to a reflected Brownian motion with parabolic drift. We do this by showing that the approximating process satisfies the conditions of a general Martingale Central Limit Theorem. When the arrival times are exponentials, we show that the approximating process in fact coincides with the embedded $\Delta_{(i)}/G/1$ queue. For general arrival times, this is true up to the end of the first busy period.

Lastly, we show that when the density $f_T(\cdot)$ of the arrival times in zero vanishes, the approximating process converges to a Brownian motion with polynomial drift, where the degree of the polynomial depends on the behavior of $f_T(\cdot)$ close to zero.

2.1 Model description

We now define in more detail the queueing model that we have introduced in Section 1.1.4. It will turn out that this model coincides with the embedded $\Delta_{(i)}/G/1$ queue up until the end of the first busy period. In fact, our model neglects idling, and this leads to a stochastic recursion driven by complicated, but tractable, increments.

We once more consider a population of n customers that all possess independent clocks $(T_i)_{i=1}^n$ with density function $t \mapsto f_T(t)$. Whenever a clock rings, that customer joins the queue. Customers are served in order of arrival. The service requirements of consecutive customers are given by the i.i.d. random variables $(S_i)_{i=1}^n$. We assume that $\mathbb{E}[S^2] < \infty$. We further assume that the service capacity per time unit scales as $c_n = n/(1+\beta n^{-1/3})$, so that the service times are given by

$$D_i := \frac{S_i}{c_n} = \frac{S_i}{n}(1+\beta n^{-1/3}), \qquad i = 1, \dots, n. \tag{2.1.1}$$

After her/his service is completed, a customer leaves the queue and is permanently removed from the system. We shall work under the heavy-traffic condition

$$\rho_n := n f_T(0)\mathbb{E}[D] = 1 + \beta n^{-1/3}. \tag{2.1.2}$$

Our crucial approximating assumption is the following: When, after a service completion, the system is empty, the customer with the smallest arrival time is drawn from the population and is immediately put into service.

As will become clear, considering the queue-length process embedded at service completions makes the process more amenable to mathematical analysis (e.g. allowing access to discrete-time martingale techniques). Let $Q_n^e(k)$ denote the number of customers that are waiting to be served just after the service completion of the k-th customer. Assume that the service of the first customer starts at time 0. The process $Q_n^e(\cdot)$ counting the number of customers waiting to be served and embedded at service completions, is given by $Q_n^e(0) = q \geq 0$ and

$$Q_n^e(k) = (Q_n^e(k-1) + A_n(k) - 1)^+, \qquad k = 1, 2, \dots \tag{2.1.3}$$

with $x^+ = \max\{0, x\}$ and $A_n(k)$ the number of arrivals during the service time of the k-th customer. Assuming $q > 0$ means that q customers are already waiting in the queue before the server starts working. For general arrival times, $A_n(k)$ is given by

$$A_n(k) = \sum_{i \notin \nu_k} \mathbb{1}_{\{\sum_{j=1}^{k-1} D_j \leq T_i \leq \sum_{j=1}^{k} D_j\}}, \tag{2.1.4}$$

where ν_k is the set of customers no longer in the population at the beginning of the service of the k-th customer. Note that (2.1.4) implies that the server works continuously, and thus does not idle. From (2.1.3) and (2.1.4) it is easy to see the following:

Lemma 1. *The process $Q^e_n(k)$ for $k \leq T_{Q^e_n}(0)$ is distributed as the $\Delta_{(i)}/G/1$ queue embedded at service completions.*

It turns out that the process $Q^e_n(\cdot)$ in (2.1.3) can be rewritten as the reflected version of a free process $N_n(\cdot)$. The process $N_n(\cdot)$ is defined as $N_n(0) = q \geq 0$ and

$$N_n(k) = N_n(k-1) + A_n(k) - 1, \tag{2.1.5}$$

with $A_n(k)$ given by (2.1.4). Then,

$$(Q^e_n(k))_{k\geq 0} = (\phi(N_n)(k))_{k\geq 0} \tag{2.1.6}$$

almost surely. We recall that the reflection mapping $\phi(\cdot)$ applied to a function $f(\cdot)$ is given by

$$\phi(f)(t) := f(t) - \inf_{s\leq t} f^-(s). \tag{2.1.7}$$

Note that the process defined in (2.1.5) may take negative values. The representation (2.1.5) allows us to write

$$N_n(k) = \sum_{i=1}^{k}(A_n(k) - 1). \tag{2.1.8}$$

The process $N_n(\cdot)$ is a random walk with increments given by $A_n(k) - 1$. However, due to the complicated dependence structure of the $A_n(k)$, $N_n(\cdot)$ is not a Markov process. Nevertheless, we will be able to prove a limit theorem for $N_n(\cdot)$, showing that $N_n(\cdot) \xrightarrow{\mathrm{d}} \widehat{N}(\cdot)$. Since the reflection map $\phi(\cdot)$ is continuous, this will allow us to conclude that $\phi(N_n) \xrightarrow{\mathrm{d}} \phi(\widehat{N})$.

All the processes that we consider are elements of the space $\mathcal{D} := \mathcal{D}([0,\infty))$ of càdlàg functions, which admit left limits and are continuous from the right. To simplify notation, for a discrete-time process $X(\cdot) : \mathbb{N} \to \mathbb{R}$, we write $X(t)$, with $t \in [0,\infty)$, instead of $X(\lfloor t \rfloor)$. In particular, a process defined in this way always admits càdlàg paths. The space $\mathcal{D}$ is endowed with the usual Skorokhod J_1 topology. We then say that a process converges in distribution in $(\mathcal{D}, J_1)$ when it converges as a random measure on the space $\mathcal{D}$, when this is endowed with the J_1 topology.

When the arrival times are exponentially distributed, the critical behavior of our approximating model is determined in the following theorem:

Theorem 3 (Convergence of the approximating process for exponential arrivals)**.** *Assume that $(T_i)_{i=1}^n$ are i.i.d. rate λ exponential arrival random* *variables, and that the service times* $(S_i)_{i=1}^n$ *are such that* $\mathbb{E}[S^2] < \infty$*. Then, as* $n \to \infty$*,*

$$n^{-1/3}N_n(\cdot n^{2/3}) \xrightarrow{\mathrm{d}} \widehat{N}(\cdot), \qquad \text{in } (\mathcal{D}, J_1), \tag{2.1.9}$$

where $\widehat{N}(\cdot)$ *is the diffusion process*

$$\widehat{N}(t) := \beta t - \frac{1}{2}t^2 + \sigma W(t), \tag{2.1.10}$$

with $\sigma^2 := \lambda^2\mathbb{E}[S^2]$ *and* $W(\cdot)$ *a standard Brownian motion. Consequently, as* $n \to \infty$*,*

$$n^{-1/3}Q_n^e(\cdot n^{2/3}) \xrightarrow{\mathrm{d}} \phi(\widehat{N})(\cdot), \qquad \text{in } (\mathcal{D}, J_1). \tag{2.1.11}$$

Note that, since each service $D_i = S_i(1 + \beta n^{-1/3})/n$ is of order $D_i = O_{\mathbb{P}}(1/n)$, $n^{2/3}$ services will take roughly $\sum_{i=1}^{n^{2/3}} S_i/n \approx \mathbb{E}[S]n^{-1/3}$ time units.

2.1.1 General arrivals

When the arrival times T_i are drawn from a general distribution, the cumulative distribution function $F_T(\cdot)$ and the density function $f_T(\cdot)$ of T_i must satisfy some technical regularity properties, which we now describe. First, we assume that $f_T(\cdot)$ is continuous, with $f_T(0) \in (0, \infty)$. We also assume that the sublinear terms of the distribution function $F_T(\cdot)$ decay as quickly as

$$F_T(x) - F_T(\bar{x}) = f_T(\bar{x})(x - \bar{x}) + o(|x - \bar{x}|^{4/3}), \qquad \forall \bar{x} \in (0, \infty). \tag{2.1.12}$$

This is, for example, the case when $F_T(\cdot) \in \mathcal{C}^2([0, \infty))$. Furthermore, we assume that $f_T'(\cdot)$ exists and is continuous in a neighborhood of zero. This implies that

$$\sup_{\bar{x} \leq cy^{1/3}} |F_T(\bar{x} + y) - f_T(\bar{x}) - f_T(\bar{x})y| \leq \sup_{\substack{\bar{x} \leq cy^{1/3} \\ \zeta \in (\bar{x}, \bar{x}+y)}} \left|\frac{f_T'(\zeta)}{2}y^2\right| \leq \frac{M}{2}y^2, \tag{2.1.13}$$

where $M > 0$ is the supremum of $f_T'(\cdot)$ in a neighborhood of zero. Equation (2.1.13) is a technical condition that will be useful later on. Our assumptions on $f_T(\cdot)$ imply also that

$$f_T(x) = f_T(0) + f_T'(0)x + o(x). \tag{2.1.14}$$

Since $\lim_{x\to\infty} f_T(x) = 0$ and $f_T(\cdot)$ is continuous on $[0,\infty)$, it admits a maximum in $[0,\infty)$. Our analysis will rely crucially on the assumption

$$f_T(0) = \sup_{x\geq 0} f_T(x). \tag{2.1.15}$$

When the arrival times follow a general distribution, the heavy-traffic behavior of the approximating model is given by the following theorem:

Theorem 4 (Convergence of the approximating process for general arrivals). *Assume that the arrival times $(T_i)_{i=1}^n$ satisfy* (2.1.12)–(2.1.15), *and that the service times $(S_i)_{i=1}^n$ are such that $\mathbb{E}[S^2] < \infty$. Then, as $n \to \infty$,*

$$n^{-1/3} N_n(\cdot n^{2/3}) \xrightarrow{\text{d}} \widehat{N}(\cdot), \tag{2.1.16}$$

where $\widehat{N}(\cdot)$ is the diffusion process

$$\widehat{N}(t) := \beta t + \frac{f_T'(0)}{2f_T(0)^2} t^2 + \sigma W(t), \tag{2.1.17}$$

with $\sigma^2 := f_T(0)^2 \mathbb{E}[S^2]$ and $W(\cdot)$ a standard Brownian motion. Moreover, as $n \to \infty$,

$$n^{-1/3} Q_n^e(\cdot n^{2/3}) \xrightarrow{\text{d}} \phi(\widehat{N})(\cdot). \tag{2.1.18}$$

We carry out the involved proof of Theorem 4 in Section 2.3. Note that when T is exponentially distributed, $f_T'(0)/(2f_T(0)^2) = -1/2$, so that Theorem 4 is in fact a generalization of Theorem 3.

We now provide a heuristic argument that explains the scaling exponents in Theorem 4. Setting $\Sigma_i := \sum_{l=1}^i D_l/n$, we estimate $N_n(\cdot)$ at time tn^p as

$$\begin{aligned} N_n(tn^p) &= \sum_{i=1}^{tn^p} \Big(\sum_{j\notin \nu_i} \mathbb{1}_{\{\sum_{l=1}^{i-1} D_l \leq T_j \leq \sum_{l=1}^{i} D_l\}} - 1 \Big) \\ &\approx \sum_{i=1}^{tn^p} (n(F_T(\Sigma_i) - F_T(\Sigma_{i-1})) - 1) \\ &\approx \sum_{i=1}^{tn^p} (f_T(\Sigma_{i-1})\mathbb{E}[S] - 1) \approx \sum_{i=1}^{tn^p} \sum_{l=1}^{i-1} \frac{S_l}{n} f_T'(0)\mathbb{E}[S], \end{aligned} \tag{2.1.19}$$

where in the last approximation we used our heavy-traffic assumption (2.1.2). This computation gives us the leading order term of the process

$N_n(\cdot)$ up to a multiplicative constant as

$$N_n(tn^p) \approx \sum_{i=1}^{tn^p} \frac{i}{n} \approx \frac{t^2}{2} n^{2p-1}. \tag{2.1.20}$$

The process $N_n(tn^p)$ is the sum of tn^p contributions, thus (ignoring dependencies) the correct spatial scaling in order to obtain Gaussian fluctuations is $n^{p/2}$. Equating the order of magnitude of the first order approximation (2.1.19) and $n^{p/2}$ gives $2p - 1 = p/2$, so that p should be $2/3$.

For general arrival times, the coupling between the approximating model and the $\Delta_{(i)}/G/1$ queue established in Lemma 1 breaks down after the end of the first busy period since the clocks T_i are no longer memoryless. However, Lemma 1 still allows us to prove results for the first busy period of the $\Delta_{(i)}/G/1$ queue with general arrivals. The functional $f \mapsto T_f(0)$ denotes the first hitting time of 0 of a function $f(\cdot)$. We have the following:

Theorem 5 (Number of customers in the first busy period)**.** *The number of customers served in the first busy period of the $\Delta_{(i)}/G/1$ queue is given by $T_{Q_n^e}(0)$. Furthermore, assuming* (2.1.12)-(2.1.15) *and that $Q_n^e(0) = qn^{1/3}$,*

$$n^{-2/3} T_{Q_n^e}(0) \xrightarrow{\text{d}} T^{\beta}_{\widehat{N}_q}(0), \tag{2.1.21}$$

where $T^{\beta}_{\widehat{N}_q}(0)$ is the first hitting time of zero of the process

$$\widehat{N}_q(t) := q + \beta t + \frac{f_T'(0)}{2 f_T(0)^2} t^2 + \sigma W(t), \tag{2.1.22}$$

and $\sigma^2 := f_T(0)^2 \mathbb{E}[S^2]$.

Proof. The functional $T_f : \mathcal{D} \to \mathbb{R}$, $f \mapsto T_f(0)$ is a.s. continuous in $\widehat{N}_q(\cdot)$ by [52, Chapter VI, Proposition 2.11], when $\mathcal{D}$ is endowed with the Skorokhod J_1 topology. Moreover,

$$\begin{aligned} n^{-2/3} T_{Q_n^e}(0) &= n^{-2/3} \inf\{t > 0 : Q_n^e(t) \le 0\} \\ &= \inf\{t > 0 : n^{-1/3} Q_n^e(tn^{2/3}) \le 0\}. \end{aligned} \tag{2.1.23}$$

Since $n^{-2/3} Q_n^e(\cdot n^{-1/3}) \xrightarrow{\text{d}} \phi(\widehat{N}_q)(\cdot)$ by Theorem 4, the conclusion follows from the Continuous-Mapping Theorem. □

It does not seem possible to extend Theorem 4 directly to the $\Delta_{(i)}/G/1$ queue beyond the first busy period. However, the limiting process (2.1.17) only depends on the distribution of T through $f_T(0)$ (λ for exponential clocks), suggesting that the result is insensitive to the arrival clocks distribution, as long as $f_T(0) > 0$. In the next chapter we will show that the *queue-length process* of the $\Delta_{(i)}/G/1$ queue with general arrivals converges to (2.1.17), after a suitable scaling of time.

2.1.2 Preliminaries

We will first prove Theorem 3 and then move to the technically more demanding Theorem 4.

Let us now set some notation and present some useful results. All random variables that we consider are defined on some complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Elements of Ω will always be denoted by ω. Given two real-valued random variables X, Y we say that X *stochastically dominates* Y, and we denote it by $Y \preceq X$, if

$$\mathbb{P}(X \leq x) \leq \mathbb{P}(Y \leq x), \qquad \forall x \in \mathbb{R}, \tag{2.1.24}$$

so that for every non-decreasing function $f(\cdot) : \mathbb{R} \to \mathbb{R}$

$$\mathbb{E}[f(Y)] \leq \mathbb{E}[f(X)]. \tag{2.1.25}$$

If X and Y are defined on the same probability space Ω, and $X(\omega) \leq Y(\omega)$ for almost every $\omega \in \Omega$, then we write $X \overset{\text{a.s.}}{\leq} Y$. We write $f(n) = O(g(n))$ for functions $f(\cdot)$, $g(\cdot) \geq 0$ and $n \to \infty$ if there exists a constant $c > 0$ such that $\lim_{n\to\infty} f(n)/g(n) \leq c$. We write $f(n) = o(g(n))$ if $\lim_{n\to\infty} f(n)/g(n) = 0$. Furthermore, we write $O_{\mathbb{P}}(a_n)$ for a sequence of real-valued random variables X_n for which $|X_n|/a_n$ is tight as $n \to \infty$. Moreover, we write $o_{\mathbb{P}}(a_n)$ for a sequence of random variables X_n for which $|X_n|/a_n \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$. We say that a sequence of events $(E_n)_{n=1}^{\infty}$ holds with high probability (briefly, w.h.p.) if $\mathbb{P}(E_n) \to 1$ as $n \to \infty$. We denote by $|A|$ the cardinality of a set A.

Following [17], we say that X_n *converges in distribution* (or *converges weakly*) to X (and denote it by $X_n \xrightarrow{\text{d}} X$) if $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ as $n \to \infty$ for every $f(\cdot)$ that is real-valued, bounded and continuous. In particular, if X is $\mathcal{D}$-valued, $f(\cdot)$ can be any continuous function from $\mathcal{D}$ to $\mathbb{R}$. Thus, to formally establish convergence in distribution in the space of $\mathcal{D}$-valued random variables, a metric, or a topology, on $\mathcal{D}$ is needed (in order to define continuity of the functions $f(\cdot)$). Several topologies on the space

$\mathcal{D}$ have been defined (all by Skorokhod in his celebrated paper [88]). For our purposes we will consider the J_1 topology, which can be described as being generated by some metric d_∞ on $\mathcal{D}([0,\infty),\mathbb{R})$ defined as an extension of some metric d_t on $\mathcal{D}([0,t],\mathbb{R})$. The latter is defined as follows. Let $\|\cdot\|$ indicate the supremum norm, $\mathrm{id}(\cdot)$ the identity function on $[0,t]$ and Λ_t the space of non-decreasing homeomorphisms on $[0,t]$. Define, for any $x_1, x_2 \in \mathcal{D}$,

$$d_t(x_1,x_2) := \inf_{\lambda\in\Lambda_t} \{\max\{\|\lambda(\cdot)-\mathrm{id}(\cdot)\|, \|x_1(\cdot)-x_2(\lambda(\cdot))\|\}\} \qquad (2.1.26)$$

and

$$d_\infty(x_1,x_2) := \int_0^\infty \mathrm{e}^{-t}[d_t(x_1,x_2)\wedge 1]\mathrm{d}t. \qquad (2.1.27)$$

[95] shows that (2.1.27) is the correct way of extending the metric, and thus the topology, from $\mathcal{D}([0,t],\mathbb{R})$ to $\mathcal{D}([0,\infty),\mathbb{R})$, since convergence with respect to d_∞ is equivalent to convergence with respect to d_t on any compact subset $[0,t]$. When dealing with vectors of functions we make use of the *weak* J_1 topology JW_1. This coincides with the product topology on $\mathcal{D}\times\mathcal{D}\times\cdots\times\mathcal{D}=\mathcal{D}^k$.

In order to prove Theorems 3 and 4 we first show that the (rescaled) process $N_n(\cdot)$ converges weakly to $\widehat{N}(\cdot)$, and then we deduce the convergence of the reflected process $\phi(N_n)(\cdot)$ exploiting the Continuous-Mapping Theorem below. In fact, this procedure follows a general technique known as the *Continuous-Mapping approach* (see [95, 96] for a detailed description).

Theorem 6 (Continuous-Mapping Theorem). *If $X_n \xrightarrow{\mathrm{d}} X$ and f is continuous almost surely with respect to the distribution of X, then $f(X_n) \xrightarrow{\mathrm{d}} f(X)$.*

Through the Continuous-Mapping approach one reduces the problem of establishing convergence of random objects to one of continuity of suitable functions. Suppose we have shown that $n^{-1/3}N_n(\cdot n^{2/3}) \xrightarrow{\mathrm{d}} \widehat{N}(\cdot)$. To prove Theorem 4 we are left to prove that the reflection map (2.1.7) is continuous almost surely with respect to the distribution of $\widehat{N}(\cdot)$. For this, note that $\mathbb{P}(\widehat{N}(\cdot)\in\mathcal{C})=1$, where $\mathcal{C}=\mathcal{C}([0,\infty),\mathbb{R})\subset\mathcal{D}$ denotes the space of continuous functions from $[0,\infty)$ to $\mathbb{R}$. Then by [95, Theorem 4.1] and [95, Theorem 6.1], $\phi(\cdot)$ is continuous almost surely with respect to the distribution of $\widehat{N}(\cdot)$.

To prove the convergence of $n^{-1/3}N_n(\cdot n^{2/3})$ we make use of a general Martingale Functional Central Limit Theorem [38, Section 7] (MFCLT) in the special case where the limit process is a standard Brownian motion; for a thorough overview, see [97]. For a discrete-time process $k \mapsto X(k)$, we consider its continuous-time version obtained by piece-wise constant interpolation. We denote the continuous-time version of $X(k)$ again as $X(t) = X(\lfloor t \rfloor)$ with a slight abuse of notation.

We shall present the MFCLT below, as it is stated in [97]. Recall that, when $M(t)$ is a square-integrable martingale with respect to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$, the *predictable quadratic variation process* associated with $M(\cdot)$ is the unique non-decreasing, non-negative, predictable, integrable process $V(\cdot)$ such that $M^2(t) - V(t)$ is a martingale with respect to $\{\mathcal{F}_t\}_{t\geq 0}$.

Theorem 7 (Martingale Functional Central Limit Theorem). *Assume that $\{\mathcal{F}^n_t\}_{t\geq 0, n\in\mathbb{N}}$ is a family of increasing filtrations and let $\{\bar{M}_n(\cdot)\}_{n=1}^{\infty}$ be a sequence of continuous-time, real-valued, square-integrable martingales, each with respect to $(\mathcal{F}^n_t)_{t\geq 0}$, such that $\bar{M}_n(0) = 0$. Assume that $\bar{V}_n(\cdot)$, the predictable quadratic variation process associated with $\bar{M}_n(\cdot)$, and $\bar{M}_n(\cdot)$ satisfy the following conditions:*

(i) $\bar{V}_n(t) \xrightarrow{\mathbb{P}} \sigma^2 t, \qquad \forall t \in \mathbb{R}^+,$

(ii) $\lim_{n\to\infty} \mathbb{E}[\sup_{t\leq \bar{t}} |\bar{V}_n(t) - \bar{V}_n(t^-)|] = 0, \qquad \forall \bar{t} \in \mathbb{R}^+,$

(iii) $\lim_{n\to\infty} \mathbb{E}[\sup_{t\leq \bar{t}} |\bar{M}_n(t) - \bar{M}_n(t^-)|^2] = 0, \qquad \forall \bar{t} \in \mathbb{R}^+.$

Then, as $n \to \infty$, $\bar{M}_n(\cdot)$ converges in distribution in $\mathcal{D}([0,\infty))$ to a centered Brownian motion with variance $\sigma^2 t$.

Before applying the MFCLT, we recall the Doob decomposition of the process $N_n(\cdot)$, writing it as the sum of a martingale—which will converge to the Brownian motion—and an appropriate drift term, as follows:

$$\begin{aligned} N_n(k) &= \sum_{i=1}^{k}(A_n(i) - \mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}]) + \sum_{i=1}^{k}(\mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}] - 1) \\ &=: M_n(k) + C_n(k), \end{aligned} \tag{2.1.28}$$

with $\{\mathcal{F}_i\}_{i\geq 1}$ the filtration generated by $(A_n(k))_{k\geq 1}$, i.e. $\mathcal{F}_i = \sigma(\{A_n\}_{k=1}^{i})$. Another Doob decomposition of interest is

$$M_n^2(k) = Z_n(k) + V_n(k) \tag{2.1.29}$$

with $Z_n(\cdot)$ a martingale and $V_n(\cdot)$ the discrete-time predictable quadratic variation of the process $M_n(\cdot)$. Note that for every fixed n and k, $|M_n(k)|$ is bounded and thus its second moment is finite. Therefore $V_n(k)$ exists and is given by

$$\begin{aligned} V_n(k) &= \sum_{i=1}^{k} \mathbb{E}[(A_n(i) - \mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}])^2 \mid \mathcal{F}_{i-1}] \\ &= \sum_{i=1}^{k} (\mathbb{E}[A_n(i)^2 \mid \mathcal{F}_{i-1}] - \mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}]^2). \end{aligned} \tag{2.1.30}$$

To see this, we rewrite

$$\begin{aligned} M_n^2(k) &= \sum_{i=1}^{k} (A_n(i) - \mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}])^2 \\ &\quad + \sum_{\substack{i,j \le k \\ i \ne j}} (A_n(i) - \mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}])(A_n(j) - \mathbb{E}[A_n(j) \mid \mathcal{F}_{j-1}]) \\ &=: \sum_{i=1}^{k} (A_n(i) - \mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}])^2 + L_n(k). \end{aligned} \tag{2.1.31}$$

It is easy to see that $L_n(\cdot)$ is also a martingale. The decomposition (2.1.29) follows from

$$\begin{aligned} Z_n(k) &:= \sum_{i=1}^{k} (A_n(i) - \mathbb{E}[A(i) \mid \mathcal{F}_{i-1}])^2 \\ &\qquad - \sum_{i=1}^{k} \mathbb{E}[(A_n(i) - \mathbb{E}[A(i) \mid \mathcal{F}_{i-1}])^2 \mid \mathcal{F}_{i-1}] + L_n(k), \\ V_n(k) &:= \sum_{i=1}^{k} \mathbb{E}[(A_n(i) - \mathbb{E}[A(i) \mid \mathcal{F}_{i-1}])^2 \mid \mathcal{F}_{i-1}]. \end{aligned} \tag{2.1.32}$$

Note that $Z_n(\cdot)$ is the sum of two martingales and thus is also a martingale.

2.2 Proof of Theorem 3

When the arrival clocks are exponentially distributed, the number of customers that join the queue during one service has a simple expression by virtue of the memoryless property. Conditioned on ν_k, (2.1.4) is distributed as

$$\sum_{i=1}^{P_n(k)} \mathbb{1}_{\{T_{i,k} \le D_k\}}, \tag{2.2.1}$$

where $T_{i,k} \overset{\mathrm{d}}{=} T_i$, which means that the clocks are re-drawn after each service, and $P_n(k) := |[n] \setminus \nu_k| = n - Q_n^e(k-1) - k$ is the number of customers still in the population. Since the dependence of $T_{i,k}$ on k does not play a role in our analysis, we will write T_i instead of $T_{i,k}$. The exponential distribution satisfies the assumptions (2.1.12)–(2.1.15). However, we will prove Theorem 3 under a weaker assumption. In particular we will assume that $F_T(\cdot)$ and $f_T(\cdot)$ satisfy

$$F_T(x) = f_T(0)x + o(x^{4/3}). \tag{2.2.2}$$

2.2.1 Supporting lemmas

In this section we prove various lemmas that we will make use of during the proof. For Lemma 2, we restrict ourselves to the case $\beta = 0$ for simplicity.

Lemma 2. *Let S, T be positive random variables. Let $D := S/n$. If* (2.2.2) *holds for T and $\mathbb{E}[S^2] < \infty$, then*

$$\mathbb{E}[|\mathbb{P}(T \le D \mid D) - f_T(0)D|] = o(n^{-4/3}), \tag{2.2.3}$$

$$\mathbb{E}[|D(\mathbb{P}(T \le D \mid D) - f_T(0)D)|] = o(n^{-2}), \tag{2.2.4}$$

$$\mathbb{E}[|\mathbb{P}(T \le D \mid D) - f_T(0)D|^2] = o(n^{-2}). \tag{2.2.5}$$

Proof. Since

$$\mathbb{P}(T \le D|D) = F_T(D) = f_T(0)D + o(S^{4/3}n^{-4/3}), \tag{2.2.6}$$

pointwise convergence trivially holds. As $n \to \infty$,

$$n^{4/3}|F_T(D) - f_T(0)D| \overset{\text{a.s.}}{\to} 0. \tag{2.2.7}$$

By assumption (2.2.2) there exists a constant $c > 0$ such that

$$|F_T(x) - f_T(0)x| \le cx^{4/3}. \tag{2.2.8}$$

Consequently,

$$n^{4/3}|F_T(D) - f_T(0)D| \le cn^{4/3}S^{4/3}n^{-4/3} \tag{2.2.9}$$

almost surely. Since $\mathbb{E}[S^{4/3}] < \infty$, the random variable $n^{4/3}|F_T(D) - f_T D|$ is bounded by an integrable random variable not depending on n. The Dominated Convergence Theorem then gives us (2.2.3). Equations (2.2.4) and (2.2.5) are proven similarly. Pointwise convergence is again trivial. Next, note that there exist constant $c_1, c_2 > 0$ such that

$$x|F_T(x) - f_T(0)x| \leq c_1 x^2, \qquad |F_T(x) - f_T(0)x|^2 \leq c_2 x^2. \tag{2.2.10}$$

Indeed, for $x \ll 1$, $|F_T(x) - f_T(0)x|^2 \leq c_2 x^{8/3} \leq c_2 x^2$ for some $c_2 > 0$, and for $x \gg 1$ it is enough to notice that $F_T(x)$ is bounded. The first bound in (2.2.10) is obtained in the same way. Since $\mathbb{E}[S^2] < \infty$ by assumption, (2.2.4) and (2.2.5) again follow by the Dominated Convergence Theorem. □

Roughly speaking, Lemma 2 states that the small-o term in the Taylor expansion satisfies

$$\mathbb{E}[o_{\mathbb{P}}(x)] = o(x). \tag{2.2.11}$$

Assuming that, as customers join the queue, the customer population does not deplete gives a stochastic upper bound on $A_n(k)$. The random variable that describes the number of arriving customers is then given by

$$A'_n := \sum_{i=1}^{n} \mathbb{1}_{\{T_i \leq D\}}. \tag{2.2.12}$$

The upper bound (2.2.12) allows us to circumvent the difficulties of dealing with the complicated set ν_k. Note that

$$A_n(k) \preceq A'_n \tag{2.2.13}$$

for *all* $k = 1, 2, \ldots$ The next lemmas shed light into the behavior of the process $N_n(\cdot)$:

Lemma 3. *For $k = O(n^{2/3})$,*

$$n^{-2/3} N_n(k) \preceq G_n(k), \tag{2.2.14}$$

where $G_n(k)$ is a random variable such that $G_n(k) \xrightarrow{\mathbb{P}} 0$.

Proof. First note that

$$N_n(k-1) \preceq \sum_{j=1}^{k-1} \Big(\sum_{l=1}^{n} \mathbb{1}_{\{T_l \leq D_j\}} - 1 \Big) = \sum_{j=1}^{k-1} (A'_n - 1). \tag{2.2.15}$$

By the Weak LLN for uncorrelated random variables (see e.g. [63]) it is enough to show that $\sup_{n\in\mathbb{N}} \mathrm{Var}(A'_n) < \infty$. Write

$$\begin{aligned}\mathrm{Var}(A'_n)^2 &= \mathbb{E}[(A'_n)^2] - \mathbb{E}[A'_n]^2 \\ &= \mathbb{E}[A'_n] + \mathbb{E}[\sum_{\substack{i,j\le n\\ i\neq j}} \mathbb{1}_{\{T_i\le D\}}\mathbb{1}_{\{T_j\le D\}}] - \mathbb{E}[A'_n]^2. \end{aligned} \tag{2.2.16}$$

The terms $\mathbb{E}[A'_n]$ and $\mathbb{E}[A'_n]^2$ are uniformly bounded since

$$\mathbb{E}[A'_n] = 1 + \beta n^{-1/3} + o(n^{-1/3}). \tag{2.2.17}$$

Moreover,

$$\mathbb{E}[\sum_{\substack{i,j\le n\\ i\neq j}} \mathbb{1}_{\{T_i\le D\}}\mathbb{1}_{\{T_j\le D\}}] = \sum_{\substack{i,j\le n\\ i\neq j}} \mathbb{E}[F_T(D)^2] \le c + o(1), \tag{2.2.18}$$

for some $c > 1$, where we have performed a Taylor expansion of $F_T(\cdot)$ and have used Lemma 2 to bound the lower order terms. Both error terms in (2.2.17) and (2.2.18) can be bounded from above by a constant independent of n. Therefore $\sup_{n\in\mathbb{N}} \mathrm{Var}(A'_n) < \infty$ and the Weak LLN allows us to conclude the statement in the lemma. □

Note that the convergence established in Lemma 3 is not uniform in $j \le k$, with $k = O(n^{2/3})$. We now work towards this result.

We will make use of a well-known property of the order statistics of exponential random variables. Recall that $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the order statistics of random variables $X_1, \ldots, X_n$.

Lemma 4. *Let $E_1, \ldots, E_n$ be independent exponentially distributed random variables with mean one. Then,*

$$(E_{(j)})_{j=1}^n \overset{\mathrm{d}}{=} \Big(\sum_{s=1}^{j} \frac{E_s}{n-s+1}\Big)_{j=1}^n. \tag{2.2.19}$$

In particular there exists a coupling between $(E_{(j)})_{j=1}^n$ and $(E_j)_{j=1}^n$ such that

$$\frac{E_1 + \cdots + E_j}{n} \le E_{(j)} \tag{2.2.20}$$

almost surely for all $j \le n$.

See [31, Section 2.5] for a proof of Lemma 4. Next, we investigate the random variable A'_n. The following lemma states that on average, in the limit, the contribution to the queue length of arrivals of order $n^{1/3}$ or greater is negligible:

Lemma 5. *A'^2_n is stochastically dominated by a family of uniformly integrable (with respect to n) random variables. In particular,*

$$\mathbb{E}[(A'_n)^2 \mathbb{1}_{\{(A'_n)^2 > \varepsilon n^{2/3}\}}] \to 0, \tag{2.2.21}$$

as $n \to \infty$.

Proof. Recall that, when E_i is a mean one exponential random variable, $U_i = 1 - \exp(-E_i)$ is a uniform random variable on $[0,1]$. The same result implies that

$$T_{(i)} \overset{\mathrm{d}}{=} F_T^{-1}(1 - \exp(-E_{(i)})), \tag{2.2.22}$$

so that

$$A'_n = \sum_{i=1}^{n} \mathbb{1}_{\{T_i \leq D\}} = \sum_{i=1}^{n} \mathbb{1}_{\{T_{(i)} \leq D\}} \overset{\mathrm{d}}{=} \sum_{i=1}^{n} \mathbb{1}_{\{F_T^{-1}(1-\exp(-E_{(i)})) \leq D\}}. \tag{2.2.23}$$

Since the function $x \mapsto F_T^{-1}(1 - \exp(-x))$ is monotone, by Lemma 4 we get

$$\begin{aligned} \sum_{i=1}^{n} \mathbb{1}_{\{F_T^{-1}(1-\exp(-E_{(i)})) \leq D\}} &\preceq \sum_{i=1}^{n} \mathbb{1}_{\{F_T^{-1}(1-\exp(-\sum_{j=1}^{i} E_j/n)) \leq D\}} \\ &\overset{\text{a.s.}}{=} \sum_{i=1}^{n} \mathbb{1}_{\{\sum_{j=1}^{i} E_j \leq -n\log(1-F_T(D))\}}. \end{aligned} \tag{2.2.24}$$

By (2.2.2), $F_T(x)/x$ is bounded from above by a positive constant $K \in \mathbb{R}^+$, so that

$$\sum_{i=1}^{n} \mathbb{1}_{\{\sum_{j=1}^{i} E_j \leq -n\log(1-F_T(D))\}} \preceq \sum_{i=1}^{n} \mathbb{1}_{\{\sum_{j=1}^{i} E_j \leq -n\log(1-KD)\}}. \tag{2.2.25}$$

Fix ε and let c be such that $-\log(1-x) \leq cx$ for all $0 \leq x \leq 1-\varepsilon$. We do this in order to remove the dependencies from n. We then obtain that

$$\begin{aligned} A'_n &\preceq \Big(\sum_{i=1}^{\infty} \mathbb{1}_{\{\sum_{j=1}^{i} E_j \leq cnKD\}} \Big) \mathbb{1}_{\{KD \leq 1-\varepsilon\}} + A'_n \mathbb{1}_{\{KD > 1-\varepsilon\}} \\ &\preceq N(cnKD) + A'_n \mathbb{1}_{\{KD > 1-\varepsilon\}}, \end{aligned} \tag{2.2.26}$$

where

$$N(t,\omega) := \sum_{i=1}^{\infty} \mathbb{1}_{\{\sum_{j=1}^{i} E_j \le t\}}(\omega) \tag{2.2.27}$$

is a Poisson process having rate one. We now prove that each of the two terms in (2.2.26) is a family of uniformly integrable random variables, and thus also their sum is. Since by assumption S_k has a finite second moment, also $N(cnKD)$ has. Since the latter does not depend on n, it is uniformly integrable with respect to n. Moving to the second term, note that, since $A'_n \le n$ almost surely,

$$\begin{aligned}\mathbb{E}[A'^2_n \mathbb{1}_{\{K\cdot D>1-\varepsilon\}}] &\le n^2 \mathbb{P}\Big(S_k > \frac{(1-\varepsilon)n}{K(1+\beta n^{-1/3})}\Big) \\ &\le \frac{n^2K^2(1+\beta n^{-1/3})^2}{(1-\varepsilon)^2 n^2} \mathbb{E}[S_k^2 \mathbb{1}_{\{S_k > \frac{(1-\varepsilon)n}{K(1+\beta n^{-1/3})}\}}]. \end{aligned} \tag{2.2.28}$$

Since $\mathbb{E}[S^2] < \infty$, as $n \to \infty$

$$\mathbb{E}[S_k^2 \mathbb{1}_{\{S_k > \frac{(1-\varepsilon)n}{K(1+\beta n^{-1/3})}\}}] \to 0. \tag{2.2.29}$$

The second moments of the second term in (2.2.26) converge to zero as n tends to infinity, and thus $\{A'^2_n \mathbb{1}_{\{KD>1-\varepsilon\}}\}_{n\ge1}$ is a uniformly integrable family. Therefore, $(N(cnKD) + A'_n \mathbb{1}_{\{KD>1-\varepsilon\}})^2$ is uniformly integrable. We have then shown that $(A'^2_n)_{n\ge1}$ is stochastically dominated by a random variable with uniformly integrable second moments. The second claim then follows by the stochastic domination result in (2.1.25). □

2.2.2 Proof of Theorem 3

Recall that $N_n(k)$ can be decomposed as $N_n(k) = M_n(k) + C_n(k)$, where $M_n(k)$ is a martingale and $C_n(k)$ is a drift term. Moreover, we also wrote $M_n^2(k)$ as $M_n^2(k) = Z_n(k) + V_n(k)$ with $Z_n(k)$ the Doob martingale and $V_n(k)$ its drift. The proof then consists of verifying the following conditions: For every $\bar{t} \in \mathbb{R}^+$,

(i) $\sup_{t\le\bar{t}} |n^{-1/3}C_n(tn^{2/3}) - \beta t + \frac{1}{2}t^2| \xrightarrow{\mathbb{P}} 0$,

(ii) $n^{-2/3}V_n(\bar{t}n^{2/3}) \xrightarrow{\mathbb{P}} \sigma^2\bar{t}$,

(iii) $\lim_{n\to\infty} n^{-2/3}\mathbb{E}[\sup_{t\le\bar{t}} |V_n(tn^{2/3}) - V_n(tn^{2/3}-)|] = 0$,

(iv) $\lim_{n\to\infty} n^{-2/3}\mathbb{E}[\sup_{t\leq\bar{t}} |M_n(tn^{2/3}) - M_n(tn^{2/3}-)|^2] = 0$.

Recall that $\sigma^2 = f_T(0)^2\mathbb{E}[S^2]$.

Condition (i) implies the convergence of the drift term, while conditions (ii)-(iv) imply the convergence of the (rescaled) process $M_n(k)$ to a centered Brownian motion, by Theorem 7. By standard convergence arguments, we can then conclude that the rescaled version of the sum $C_n(k) + M_n(k)$ converges in distribution to the sum of the respective limits.

Proof of (i)

We first prove (i) and to that end, we expand the term $\mathbb{E}[A_n(i)|\mathcal{F}_{i-1}]$. Recall that ν_i denotes the set of the customers that are no longer in the population at the beginning of the service of the i-th customer. Then,

$$\begin{aligned}\mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}] &= \sum_{s\notin\nu_i} \mathbb{E}[\mathbb{1}_{\{T_s\leq D_i\}} \mid \mathcal{F}_{i-1}] \\ &= \sum_{s\notin\nu_i} \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{T_s\leq D_i\}} \mid \mathcal{F}_{i-1}, D_i] \mid \mathcal{F}_{i-1}] \\ &= \sum_{s\notin\nu_i} (\mathbb{E}[f_T(0)D_i \mid \mathcal{F}_{i-1}] + o(n^{-4/3})), \qquad (2.2.30)\end{aligned}$$

where, in the last equality, we have used Lemma 2 and the error term is independent of s. Since D_i is independent from $\mathcal{F}_{i-1}$, we obtain

$$\mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}] = \sum_{s\notin\nu_i} (\mathbb{E}[f_T(0)D_i] + o(n^{-4/3})). \qquad (2.2.31)$$

The summation can then be simplified to

$$\begin{aligned}\mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}] &= (n - |\nu_i|)(f_T(0)\mathbb{E}[D_i] + o(n^{-4/3})) \\ &= f_T(0)\mathbb{E}[S_i](1+\beta n^{-1/3}) - f_T(0)\mathbb{E}[S_i](1+\beta n^{-1/3})\frac{|\nu_i|}{n} \\ &\quad + (n - |\nu_i|)o(n^{-4/3}), \qquad (2.2.32)\end{aligned}$$

since the error terms in (2.2.30) are uniform in s. Then,

$$\begin{aligned}&\mathbb{E}[A_n(i) - 1 \mid \mathcal{F}_{i-1}] \\ &= f_T(0)(1+\beta n^{-1/3})\mathbb{E}[S_i] - 1 - f_T(0)(1+\beta n^{-1/3})\mathbb{E}[S_i]\frac{|\nu_i|}{n} + o(n^{-1/3}). \qquad (2.2.33)\end{aligned}$$

By (1.1.2),

$$\mathbb{E}[A_n(i) - 1 \mid \mathcal{F}_{i-1}] = \beta n^{-1/3} - \frac{|\nu_i|}{n}(1 + O(n^{-1/3})) + o(n^{-1/3}). \quad (2.2.34)$$

Note that, since the service times are independent from the history of the system, conditioning on $\mathcal{F}_{i-1}$ has no effect. Since $|\nu_i| = i + N_n(i-1)$, the drift term in the decomposition of $N_n(k)$ can be written as

$$\begin{aligned} C_n(k) = k\beta n^{-1/3} - \Big(\frac{k^2+k}{2} + \sum_{i=1}^{k} N_n(i-1)\Big)(n^{-1} + O(n^{-4/3})) \\ + ko(n^{-1/3}). \quad (2.2.35) \end{aligned}$$

The term $-\sum_{i=1}^{k} N_n(i-1)$ in (2.2.35) accounts for the fact that the customers already in the queue cannot rejoin it. This term converges to zero as n tends to infinity (after appropriate scaling) by the following result:

Lemma 6. *As $n \to \infty$,*

$$n^{-2/3} \sup_{j \le an^{2/3}} |N_n(j)| \xrightarrow{\mathbb{P}} 0. \quad (2.2.36)$$

Proof. Recall that $N_n(j) = M_n(j) + C_n(j)$. Then,

$$\begin{aligned} \mathbb{P}(n^{-2/3} \sup_{j \le an^{2/3}} |N_n(j)| \ge 2\varepsilon) \le \mathbb{P}(n^{-2/3} \sup_{j \le an^{2/3}} |M_n(j)| \ge \varepsilon) \\ + \mathbb{P}(n^{-2/3} \sup_{j \le an^{2/3}} |C_n(j)| \ge \varepsilon). \quad (2.2.37) \end{aligned}$$

We will bound the first and second terms separately. Applying Doob's inequality to the martingale $M_n(\cdot)$ gives

$$\mathbb{P}(n^{-2/3} \sup_{j \le an^{2/3}} |M_n(j)| \ge \varepsilon) \le \frac{\mathbb{E}[M_n^2(an^{2/3})]}{(\varepsilon n^{2/3})^2}. \quad (2.2.38)$$

By (2.1.29), $\mathbb{E}[M_n^2(k)] = \mathbb{E}[V_n(k)]$. Expanding this term gives

$$\begin{aligned} \mathbb{E}[V_n(k)] &= \mathbb{E}[\sum_{i=1}^{k} (\mathbb{E}[A_n(i)^2 \mid \mathcal{F}_{i-1}] - \mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}]^2)] \\ &\le \mathbb{E}[\sum_{i=1}^{k} \mathbb{E}[A_n(i)^2 \mid \mathcal{F}_{i-1}]] \le k\mathbb{E}[A_n'^2], \quad (2.2.39) \end{aligned}$$

where A'_n is defined as in (2.2.12). By rescaling we get

$$n^{-4/3}\mathbb{E}[V_n(an^{2/3})] \leq an^{-2/3}\mathbb{E}[A'^2_n]. \tag{2.2.40}$$

A'_n has a finite second moment uniformly in n by (2.2.16)–(2.2.18), thus the right-most term in (2.2.40) tends to zero as n tends to infinity.

For the second term in (2.2.37) we make use of the decomposition of the drift term in (2.2.35). From there we obtain

$$-(n^{-1}+O(n^{-4/3}))\sum_{i=1}^{k}(i+N_n(i-1))+ko(n^{-1/3}) \leq C_n(k), \tag{2.2.41}$$

and

$$C_n(k) \leq k\beta n^{-1/3}+ko(n^{-1/3}), \tag{2.2.42}$$

and thus, almost surely,

$$\begin{aligned}\sup_{j\leq an^{2/3}}|C_n(j)| \leq{}& (an^{2/3}\beta n^{-1/3}+an^{2/3}o(n^{-1/3})) \\ &\vee((n^{-1}+O(n^{-4/3}))\sum_{i=1}^{an^{2/3}}(i+N_n(i-1))+an^{2/3}o(n^{-1/3})),\end{aligned} \tag{2.2.43}$$

since the two bounds (2.2.41) and (2.2.42) are monotone functions of k. By rescaling the first term by $n^{-2/3}$ we obtain $a\beta n^{-1/3}+ao(n^{-1/3})$, which tends to zero almost surely as n goes to infinity. The second term in (2.2.43) needs more attention. Notice that the function $i \mapsto i+N_n(i)$ is non-negative and non-decreasing. Thus, we bound all the terms in the sum by the final term:

$$\sum_{i=1}^{an^{2/3}}(i+N_n(i-1)) \overset{\text{a.s.}}{\leq} an^{2/3}(an^{2/3}+N_n(an^{2/3}-1)). \tag{2.2.44}$$

Rescaling by $n^{-2/3}$ we get for the second term in (2.2.43) that, almost surely,

$$\begin{aligned}n^{-2/3}\Big((n^{-1}+O(n^{-4/3}))\sum_{i=1}^{an^{2/3}}(i+N_n(i-1))+an^{2/3}o(n^{-1/3})\Big)& \\ \leq a(n^{-1}+O(n^{-4/3}))(an^{2/3}+N_n(an^{2/3}-1))+ao(n^{-1/3}).&\end{aligned} \tag{2.2.45}$$

The $O(n^{-4/3})$ term is of lower order than n^{-1}, and thus can be ignored. By Lemma 3, the right-hand side of (2.2.45) tends to zero in probability as n tends to infinity and this concludes the proof that the second term in (2.2.37) converges in probability to zero. □

Lemma 6 proves that the process $n^{-2/3}N_n(\cdot)$ tends to zero in probability, *uniformly* in $j \leq an^{2/3}$, which is stronger than the conclusion of Lemma 3. Substituting $k = tn^{2/3}$ into (2.2.35) and multiplying by $n^{-1/3}$ yields

$$\begin{aligned} & n^{-1/3}C_n(tn^{2/3}) \qquad (2.2.46) \\ & \quad = \beta t - \Big(\frac{t^2 + tn^{-2/3}}{2} + \sum_{i=1}^{tn^{2/3}} N_n(i-1)\Big)(1 + O(n^{-1/3})) + o(1). \end{aligned}$$

Both the small-o and the big-O terms in (2.2.46) are independent of t. Indeed, the small-o term originates from Lemma 2 (and is therefore independent of k) and the big-O term was introduced in (2.2.34) and depends only on n and β. Therefore, the convergence of $n^{-1/3}C_n(tn^{2/3})$ is uniform in $t \leq \bar{t}$ for fixed $\bar{t}$ as required, and this concludes the proof of Lemma 6 and thus of (i). □

Proof of (ii)

In order to prove (ii) we first compute

$$\begin{aligned} & \mathbb{E}[A_n(i)^2 \mid \mathcal{F}_{i-1}] \\ & \quad = \mathbb{E}[\sum_{j \notin \mathcal{V}_i} \mathbb{1}^2_{\{T_j \leq D_i\}} + \sum_{\substack{l \neq m \\ l,m \notin \mathcal{V}_i}} \mathbb{1}_{\{T_l \leq D_i\}} \mathbb{1}_{\{T_m \leq D_i\}} \mid \mathcal{F}_{i-1}] \\ & \quad = \mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}] + \mathbb{E}[\sum_{\substack{l \neq m \\ l,m \notin \mathcal{V}_i}} \mathbb{1}_{\{T_l \leq D_i\}} \mathbb{1}_{\{T_m \leq D_i\}} \mid \mathcal{F}_{i-1}], \qquad (2.2.47) \end{aligned}$$

which yields

$$\begin{aligned} \mathbb{E}[A_n(i)^2 \mid \mathcal{F}_{i-1}] - \mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}] & = \sum_{\substack{l \neq m \\ l,m \notin \mathcal{V}_i}} \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{T_l \leq D_i\}} \mathbb{1}_{\{T_m \leq D_i\}} \mid \mathcal{F}_{i-1}, D_i] \mid \mathcal{F}_{i-1}] \\ & = \sum_{\substack{l \neq m \\ l,m \notin \mathcal{V}_i}} \mathbb{E}[\mathbb{P}(T_l \leq D_i \mid D_i)\mathbb{P}(T_m \leq D_i \mid D_i) \mid \mathcal{F}_{i-1}] \\ & = \sum_{\substack{l \neq m \\ l,m \notin \mathcal{V}_i}} \mathbb{E}[F_T(D_i)^2 \mid \mathcal{F}_{i-1}]. \qquad (2.2.48) \end{aligned}$$

Exploiting (2.2.2) we rewrite the summation term as

$$\begin{aligned}\mathbb{E}[F_T(D_i)^2 \mid \mathcal{F}_{i-1}] = f_T(0)^2\mathbb{E}[D_i^2] + 2f_T(0)\mathbb{E}[D_i(F_T(D_i) - f_T(0)D_i)]\\ + \mathbb{E}[(F_T(D_i) - f_T(0)D_i)^2]. \qquad (2.2.49)\end{aligned}$$

The second and third terms are $o(n^{-2})$ by Lemma 2. We rewrite (2.2.48) as

$$\begin{aligned}&\sum_{\substack{l\neq m\\ l,m\notin v_i}} \Big(\frac{f_T(0)^2}{n^2}(1+\beta n^{-1/3})^2\mathbb{E}[S^2] + o(n^{-2})\Big)\\ &\qquad = \frac{|\Xi_i| f_T(0)^2}{n^2}(1+\beta n^{-1/3})^2\mathbb{E}[S^2] + o(1). \qquad (2.2.50)\end{aligned}$$

where $\Xi_i := \{(l,m) :\ l \neq m,\ l,m \notin v_i\}$. Note that the cardinality of Ξ_i is

$$|\Xi_i| = (n - N_n(i-1) - i)^2 - (n - N_n(i-1) - i), \qquad (2.2.51)$$

thus of the order n^2. Then, for $k = O(n^{2/3})$,

$$\begin{aligned}C_n(k) &= \sum_{i=1}^{k}(\mathbb{E}[A_n(i)^2 \mid \mathcal{F}_{i-1}] - \mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}]^2)\\ &= \sum_{i=1}^{k}\Big(\frac{|\Xi_i| f_T(0)^2}{n^2}(1+\beta n^{-1/3})^2\mathbb{E}[S^2] + o(1)\Big)\\ &\quad + \sum_{i=1}^{k}(\mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}] - \mathbb{E}[A_n(i) \mid \mathcal{F}_{i-1}]^2). \qquad (2.2.52)\end{aligned}$$

Using (2.2.34), together with the fact that $|v_i|/n = O_{\mathbb{P}}(n^{-1/3})$ uniformly for $i = O(n^{2/3})$, we get

$$\begin{aligned}C_n(k) &= \sum_{i=1}^{k}\Big(\frac{|\Xi_i| f_T(0)^2}{n^2}(1+\beta n^{-1/3})^2\mathbb{E}[S^2] + O_{\mathbb{P}}(n^{-1/3}) + o(1)\Big)\\ &= \sum_{i=1}^{k}\Big(\frac{(n - N_n(i-1) - i)^2 - (n - N_n(i-1) - i)}{n^2}\\ &\qquad \times f_T(0)^2(1+\beta n^{-1/3})^2\mathbb{E}[S^2]\Big) + O_{\mathbb{P}}(kn^{-1/3}) + o(k). \qquad (2.2.53)\end{aligned}$$

We then split the term inside the summation to isolate the contribution of the process $N_n(\cdot)$, and write

$$\begin{aligned}V_n(k) &= \sum_{i=1}^{k} \frac{(n-i)^2-(n-i)}{n^2} f_T(0)^2(1+\beta n^{-1/3})^2\mathbb{E}[S^2]\\ &+ \sum_{i=1}^{k} \frac{N_n(i-1)(N_n(i-1)-2(n-i)+1)}{n^2} f_T(0)^2(1+\beta n^{-1/3})^2\mathbb{E}[S^2]\\ &+ O_{\mathbb{P}}(kn^{-1/3}) + o(k). \qquad (2.2.54)\end{aligned}$$

The second term accounts for the process history. By Lemma 6, this term tends to zero in probability after rescaling. A computation shows that

$$\begin{aligned}\frac{f_T(0)^2\mathbb{E}[S^2]}{n^2}\sum_{l=n-k}^{n-1}(l^2-l) &= \frac{\sigma^2}{n^2}\left(\frac{2}{3}k+k^2-\frac{1}{3}k^3-2kn-k^2n+kn^2\right)\\ &= \sigma^2 k + O(k^2n^{-1}). \qquad (2.2.55)\end{aligned}$$

The remaining terms were omitted because they are of order smaller than $O(k^2n^{-1})$ when $k = sn^{2/3}$. When rescaling space and time appropriately in (2.2.54) we finally obtain that

$$n^{-2/3}V_n(tn^{2/3}) \xrightarrow{\mathbb{P}} \sigma^2 t, \qquad (2.2.56)$$

as required, completing the proof of (ii). □

Proof of (iii)

The process $V_n(\cdot)$ in (2.1.29) is almost surely increasing. To prove (iii), we will estimate the largest possible jump

$$\begin{aligned}&n^{-2/3}|V_n(k+1)-V_n(k)|\\ &\quad = n^{-2/3}|\mathbb{E}[A_n(k+1)^2 \mid \mathcal{F}_k] - \mathbb{E}[A_n(k+1) \mid \mathcal{F}_k]^2|, \qquad (2.2.57)\end{aligned}$$

with $k = O(n^{2/3})$. We will apply the Dominated Convergence Theorem. The jump (2.2.57) has already been implicitly computed as the term in the summation in (2.2.54) and it takes the form

$$\begin{aligned}&n^{-2/3}|\mathbb{E}[A_n(k+1)^2 \mid \mathcal{F}_k] - \mathbb{E}[A_n(k+1) \mid \mathcal{F}_k]^2|\\ &= n^{-2/3}\Big|f_T(0)^2\mathbb{E}[S^2]\Big(\frac{(n-k-1)^2-(n-k-1)}{n^2}\\ &\quad + \frac{N_n(k)(N_n(k)-2(n-k-1)+1)}{n^2}\Big) + O_{\mathbb{P}}(n^{-1/3})\Big| \qquad (2.2.58)\end{aligned}$$

The term $O_{\mathbb{P}}(n^{-1/3})$ is a byproduct of $\mathbb{E}[A_n(k)|\mathcal{F}_{k-1}] - \mathbb{E}[A_n(k)|\mathcal{F}_{k-1}]^2$, as computed in (2.2.52) and following calculations. We now compute it precisely by using the exact expression for $\mathbb{E}[A_n(k)|\mathcal{F}_{k-1}]$ found in (2.2.34),. We obtain the almost sure bound

$$\begin{aligned}|\mathbb{E}[A_n(k) \mid \mathcal{F}_{k-1}] - \mathbb{E}[A_n(k) \mid \mathcal{F}_{k-1}]^2| \\ \leq \beta n^{-1/3} + 2|\nu_k|n^{-1} + 2\beta|\nu_k|n^{-4/3} + o(n^{-1/3}).\end{aligned} \tag{2.2.59}$$

The right-hand side of (2.2.59) is bounded by 3 for all sufficiently large values of n, uniformly in $k \leq tn^{2/3}$, since $|\nu_k| \leq n$. Plugging this into (2.2.58) we get that, almost surely,

$$\begin{aligned}n^{-2/3}|\mathbb{E}[A_n(k+1)^2 \mid \mathcal{F}_k] - \mathbb{E}[A_n(k+1) \mid \mathcal{F}_k]^2| \\ \leq n^{-2/3}\Big|\sigma^2\Big(\frac{(n-k-1)^2-(n-k-1)}{n^2} \\ + \frac{N_n(k)(N_n(k)-2(n-k-1)+1)}{n^2}\Big) + 3\Big|.\end{aligned} \tag{2.2.60}$$

Since $|N_n(k)| \leq n$, there exists a constant C such that, uniformly in $k \leq tn^{2/3}$,

$$\begin{aligned}n^{-2/3}|\mathbb{E}[A_n(k+1)^2 \mid \mathcal{F}_k] - \mathbb{E}[A_n(k+1) \mid \mathcal{F}_k]^2 \\ \leq n^{-2/3}(3+\sigma^2 C),\end{aligned} \tag{2.2.61}$$

almost surely. Therefore, both assumptions of the Dominated Convergence Theorem hold and this concludes the proof of (iii). □

Proof of (iv)

We prove (iv) through a coupling argument. First, note that

$$\begin{aligned}n^{-2/3}\mathbb{E}[\sup_{t\leq \bar{t}} |M_n(tn^{2/3}) - M_n(tn^{2/3}-)|^2] \\ = n^{-2/3}\mathbb{E}[\sup_{k\leq \bar{t}n^{2/3}} |A_n(k) - \mathbb{E}[A_n(k) \mid \mathcal{F}_{k-1}]|^2] \\ \leq n^{-2/3}\mathbb{E}[\sup_{k\leq \bar{t}n^{2/3}} |A_n(k)|^2] + n^{-2/3}\mathbb{E}[\sup_{k\leq \bar{t}n^{2/3}} \mathbb{E}[A_n(k) \mid \mathcal{F}_{k-1}]^2].\end{aligned} \tag{2.2.62}$$

We start with the second term in (2.2.62). Using the computations in (2.2.34), and that the second term there is negative, yields

$$0 \overset{\text{a.s.}}{\leq} \mathbb{E}[A_n(k) \mid \mathcal{F}_{k-1}] \overset{\text{a.s.}}{\leq} 1 + O(n^{-1/3}). \tag{2.2.63}$$

For the first term we use a coupling argument. For $\varepsilon > 0$, we split

$$\begin{aligned}\mathbb{E}[\sup_{k\leq \bar{t}n^{2/3}} |A_n(k)|^2] = \mathbb{E}[&\sup_{k\leq \bar{t}n^{2/3}} A_n(k)^2 \mathbb{1}_{\{\sup_{k\leq \bar{t}n^{2/3}} A_n(k)^2 \leq \varepsilon n^{2/3}\}}] \\ &+ \mathbb{E}[\sup_{k\leq \bar{t}n^{2/3}} A_n(k)^2 \mathbb{1}_{\{\sup_{k\leq \bar{t}n^{2/3}} A(k)^2 > \varepsilon n^{2/3}\}}]. \end{aligned} \tag{2.2.64}$$

Multiplying (2.2.64) by $n^{-2/3}$, the first term is bounded by ε. For the second term we estimate

$$\begin{aligned}\mathbb{E}[\sup_{k\leq \bar{t}n^{2/3}} & A_n(k)^2 \mathbb{1}_{\{\sup_{k\leq \bar{t}n^{2/3}} A_n(k)^2 > \varepsilon n^{2/3}\}}] \\ &\leq \sum_{k=1}^{\bar{t}n^{2/3}} \mathbb{E}[A_n(k)^2 \mathbb{1}_{\{A_n(k)^2 > \varepsilon n^{2/3}\}}] \\ &\leq \sum_{k=1}^{\bar{t}n^{2/3}} \mathbb{E}[{A'_n}^2 \mathbb{1}_{\{{A'_n}^2 > \varepsilon n^{2/3}\}}] \\ &= \bar{t}n^{2/3}\mathbb{E}[{A'_n}^2 \mathbb{1}_{\{{A'_n}^2 > \varepsilon n^{2/3}\}}], \end{aligned} \tag{2.2.65}$$

where we have used the stochastic domination in (2.2.12). By Lemma 5, $\mathbb{E}[{A'_n}^2 \mathbb{1}_{\{{A'_n}^2 > \varepsilon n^{2/3}\}}] \to 0$ and thus, as $n \to \infty$

$$n^{-2/3}\mathbb{E}[\sup_{k\leq \bar{t}n^{2/3}} A_n(k)^2 \mathbb{1}_{\{\sup_{k\leq \bar{t}n^{2/3}} A_n(k)^2 > \varepsilon n^{2/3}\}}] \to 0. \tag{2.2.66}$$

This concludes the proof of (iv). □

2.3 Proof of Theorem 4

In this section we prove Theorem 4. We assume that the arrival times $(T_i)_{i=1}^n$ follow a general distribution with cumulative distribution function $F_T(\cdot)$ and density function $f_T(\cdot)$ satisfying (2.1.12)–(2.1.15). The number of arrivals during the k-th service is given by

$$A_n(k) = \sum_{i\notin \nu_k} \mathbb{1}_{\{\sum_{j=1}^{k-1} D_j < T_i \leq \sum_{j=1}^{k} D_j\}}, \tag{2.3.1}$$

where ν_k is the set of customers no longer in the population at the beginning of the service of the k-th customer. We assume that $F_T(\cdot)$ can be Taylor expanded in a neighborhood of every point, as in (2.1.12), and that the density $f_T(\cdot)$ can be Taylor expanded in a neighborhood of zero, as in (2.1.14).

2.3.1 Supporting lemmas

For readability, throughout this section we will denote $\Sigma_j := \sum_{i=1}^{j} D_i$. In this section we present results that generalize various lemmas in Section 2.2.1. For each result, we provide the proof only when it is substantially different from the proof of the analogous result in Section 2.2.1. The following lemma estimates the error term in the Taylor expansion of $F_T(\cdot)$.

Lemma 7. *If $k = O(n^{2/3})$, then*

$$\mathbb{E}[|F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1})D_k| \mid \Sigma_{k-1}] = o_{\mathbb{P}}(n^{-4/3}), \quad (2.3.2)$$

$$\mathbb{E}[|D_k(F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1})D_k)| \mid \Sigma_{k-1}] = o_{\mathbb{P}}(n^{-2}), \quad (2.3.3)$$

$$\mathbb{E}[|F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1})D_k|^2 \mid \Sigma_{k-1}] = o_{\mathbb{P}}(n^{-2}). \quad (2.3.4)$$

Moreover, all the statements of convergence hold uniformly for $k = O(n^{2/3})$.

Proof. We give the proof for (2.3.3), the rest are shown in an analogous way. Note that, by our assumptions on $F_T(\cdot)$,

$$\begin{aligned}\mathbb{E}[n^2|D_k(F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1})D_k)| \mid \Sigma_{k-1}]\\ \leq \sup_{y \leq cn^{-1/3}} n^2\mathbb{E}[|F_T(y + D_k) - F_T(y) - f_T(y)D_k \mid D_k]\\ \leq n^2\mathbb{E}[\sup_{y \leq cn^{-1/3}} |F_T(y + D_k) - F_T(y) - f_T(y)D_k \mid D_k] \quad (2.3.5)\end{aligned}$$

since with high probability $\Sigma_{k-1} \leq cn^{-1/3}$ for some $c > 0$. The right term tends to zero by the Dominated Convergence Theorem and (2.1.13), and this immediately implies (2.3.3). □

The stochastic upper bound (2.2.12) is generalized as

$$A'_n(k) := \sum_{i=1}^{n} \mathbb{1}_{\{\Sigma_{k-1} \leq T_i \leq \Sigma_k\}}. \quad (2.3.6)$$

In the exponential case the process $N_n(\cdot)$ is roughly of the order $n^{1/3}$ around time $tn^{2/3}$. This is also the case in this more general setting, as the following lemma shows:

Lemma 8. *For $k = 1, 2, \ldots$ we have that*

$$n^{-2/3}N_n(k) \preceq G_n(k), \quad (2.3.7)$$

where $G_n(k)$ *is a random variable such that* $G_n(k) \xrightarrow{\mathbb{P}} 0$ *uniformly in* $k = O(n^{2/3})$.

Proof. Let us set $k = tn^{2/3}$. Fix an arbitrary $\varepsilon > 0$. Then,

$$\begin{aligned}
&\mathbb{P}\Big(n^{-2/3}\sum_{j=1}^{tn^{2/3}} A_n(j) - n^{-2/3}t \geq \varepsilon\Big) \qquad (2.3.8)\\
&\leq \mathbb{P}\Big(n^{-2/3}\sum_{j=1}^{tn^{2/3}} A'_n(j) - n^{-2/3}t \geq \varepsilon\Big)\\
&\leq \mathbb{P}\Big(n^{-2/3}|\sum_{j=1}^{tn^{2/3}} A'_n(j) - n^{-2/3}t| \geq \varepsilon\Big)\\
&= \mathbb{P}\Big(n^{-2/3}|\sum_{l=1}^{n} \mathbb{1}_{\{T_l \leq \Sigma_{tn^{2/3}}\}} - n^{-2/3}t| \geq \varepsilon\Big)\\
&\leq n^{-4/3}\varepsilon^{-2}\mathbb{E}\Big[\Big(\sum_{l=1}^{n} \mathbb{1}_{\{T_l \leq \Sigma_{tn^{2/3}}\}}\Big)^2 \mathbb{1}_{\{|\sum_{l=1}^{n} \mathbb{1}_{\{T_l \leq \Sigma_{tn^{2/3}}\}} - n^{-2/3}t| \geq \varepsilon n^{2/3}\}}\Big].
\end{aligned}$$

Next, we bound the expected value on the right-hand side of (2.3.8). To this end we define the event

$$\mathcal{E}_n := \Big\{\Big|\sum_{l=1}^{n} \mathbb{1}_{\{T_l \leq \Sigma_{tn^{2/3}}\}} - n^{-2/3}t\Big| \geq \varepsilon n^{2/3}\Big\}, \qquad (2.3.9)$$

and write

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\sum_{l=1}^{n}\mathbb{1}_{\{T_l \leq \Sigma_{tn^{2/3}}\}}\Big)^2 \mathbb{1}_{\mathcal{E}_n}\Big]\\
&\qquad\leq n + \mathbb{E}\Big[\sum_{h \neq k} \mathbb{1}_{\{T_h \leq \Sigma_{tn^{2/3}}\}}\mathbb{1}_{\{T_k \leq \Sigma_{tn^{2/3}}\}}\mathbb{1}_{\mathcal{E}_n}\Big]\\
&\qquad\leq n + n^2\mathbb{E}[F_T(\Sigma_{tn^{2/3}})^2\mathbb{1}_{\mathcal{E}_n}] \leq n + cn^2\mathbb{E}[(\Sigma_{tn^{2/3}})^2\mathbb{1}_{\mathcal{E}_n}]\\
&\qquad\leq n + cn^{4/3}\mathbb{E}[S^2\mathbb{1}_{\mathcal{E}_n}], \qquad (2.3.10)
\end{aligned}$$

for a large constant $c > 0$. In the last step in (2.3.10) we have used the Cauchy-Schwarz inequality. Since $\mathbb{P}(\mathcal{E}_n) \to 0$ and $\mathbb{E}[S^2] < \infty$, by plugging (2.3.10) into (2.3.8) and by the Dominated Convergence Theorem, we get the desired convergence. □

The next lemma will be a crucial ingredient of the proof of Lemma 10, below, which is the equivalent in this setting of Lemma 5.

Lemma 9. *Let $(E_{(i)})_{i=1}^{n-v}$ $(v \leq n)$ be the order statistics of n exponential unit mean random variables. Define $|\mathrm{Y}_{(0,c)}^{(n-v)}|$ as the cardinality of the set*

$$\mathrm{Y}_{(0,c)}^{(n-v)} := \{j \in [n-v] : E_{(j)} \in (0,c)/n\}. \tag{2.3.11}$$

Then,

$$|\mathrm{Y}_{(0,c)}^{(n-v)}| \preceq N\Big(\frac{n-v}{n}c\Big), \tag{2.3.12}$$

where $N(t)$ is a Poisson process with unit rate.

Proof. The statement is a consequence of Lemma 4. Fix $j \in \{1, \dots, n-v\}$. By definition of stochastic domination

$$\mathbb{P}\Big(E_{(j)} \leq \frac{c}{n}\Big) \leq \mathbb{P}\Big(\frac{\sum_{i=1}^{j} E_i}{n-v} \leq \frac{c}{n}\Big) \leq \mathbb{P}\Big(\Pi_j \leq \frac{n-v}{n}c\Big), \tag{2.3.13}$$

where Π_j is the j-th point of a Poisson process with rate one. The computation in (2.3.13) intuitively means that there are more Poisson points in an interval of length $(n-v)c/n$ than order statistics in an interval of length c/n. This gives (2.3.12). □

Since

$$N\Big(\frac{n-v}{n}c\Big) \preceq N(c), \qquad \forall v \leq n, \tag{2.3.14}$$

it follows from (2.3.12) that

$$|\mathrm{Y}_{(0,c)}^{(n-v)}| \preceq N(c), \qquad \forall v \leq n. \tag{2.3.15}$$

Corollary 1. *Under the same assumptions as in* Lemma 9,

$$\mathrm{Y}_{(a,b)}^{(n)} \preceq N(b-a). \tag{2.3.16}$$

Proof. By Lemma 9,

$$\mathbb{P}(N(b-a) \leq x) \leq \mathbb{P}\big(|\mathrm{Y}_{(0,b-a)}^{(n-v)}| \leq x\big). \tag{2.3.17}$$

Note that, by the memoryless property,

$$\mathbb{P}(|\mathrm{Y}_{(0,b-a)}^{(n-v)}| \leq x) = \mathbb{P}(|\mathrm{Y}_{(a,b)}^{(n)}| \leq x \mid |\mathrm{Y}_{(0,a)}^{(n)}| = v), \tag{2.3.18}$$

almost surely. Since the left side of (2.3.17) does not depend on v, by combining (2.3.17) and (2.3.18) and taking expectations on both sides in order to remove the conditioning, we get

$$\mathbb{P}(N(b-a) \leq x) \leq \mathbb{P}(|\mathrm{Y}^{(n)}_{(a,b)}| \leq x), \tag{2.3.19}$$

allowing us to conclude the claim. □

Lemma 9 simplifies the task of estimating quantities involving order statistics of exponentials by replacing them with a Poisson process.

One of the cornerstones of the analysis in Section 2.2 was the uniform integrability of $(A'^2_n)_{n\geq 1}$ in Lemma 5. An analogous version holds in this general setting:

Lemma 10. *$A_n(k)$ is stochastically bounded by a random variable with uniformly integrable (with respect to n) second moment, uniformly in $k \leq tn^{2/3}$.*

Proof. Note that $T_{(i)} \overset{\mathrm{d}}{=} F_T^{-1}(1-\exp(-E_{(i)}))$, where $(E_{(i)})_{i=1}^n$ are the order statistics of unit mean exponential random variables. Then,

$$\begin{aligned} A_n(k) \overset{\text{a.s.}}{\leq} A'_n(k) &\overset{\mathrm{d}}{=} \sum_{i=1}^n \mathbb{1}_{\{\Sigma_{k-1}\leq F_T^{-1}(1-\exp(-E_{(i)}))\leq \Sigma_k\}} \\ &\overset{\mathrm{d}}{=} \sum_{i=1}^n \mathbb{1}_{\{F_T(\Sigma_{k-1})-1\leq -\exp(-E_{(i)})\leq F_T(\Sigma_k)-1\}} \\ &\overset{\mathrm{d}}{=} \sum_{i=1}^n \mathbb{1}_{\{-\log(1-F_T(\Sigma_{k-1}))\leq E_{(i)}\leq -\log(1-F_T(\Sigma_k))\}}. \end{aligned} \tag{2.3.20}$$

By Corollary 1,

$$A_n(k) \preceq N\Big(n \log\Big(\frac{1-F_T(\Sigma_{k-1})}{1-F_T(\Sigma_k)}\Big)\Big). \tag{2.3.21}$$

By splitting the event space Ω we write (2.3.21) as

$$A_n(k) \preceq N\Big(n \log\Big(\frac{1-F_T(\Sigma_{k-1})}{1-F_T(\Sigma_k)}\Big)\Big)\mathbb{1}_{\{\Sigma_k\leq \bar{x}\}} + A_n(k)\mathbb{1}_{\{\Sigma_k>\bar{x}\}}, \tag{2.3.22}$$

where $\bar{x}$ is independent of n and will be determined later on. We now show that the first term in (2.3.22) is bounded by a random variable independent of n and with finite second moment. Moreover, the second moment of the second term converges to zero as n tends to infinity. These two facts together imply that the right-hand side of (2.3.22) has uniformly integrable second moments.

The first term is bounded as follows. Choose $\bar{x}$ so that $1 - F_T(\bar{x}) > 0$. By Taylor expanding the function

$$x \mapsto \log\Big(\frac{1 - F_T(\Sigma_{k-1})}{1 - F_T(\Sigma_{k-1} + x)}\Big), \tag{2.3.23}$$

for some $x* \in (\Sigma_{k-1}, \Sigma_k)$ we get

$$\begin{aligned} N\Big(n \log\Big(\frac{1 - F_T(\Sigma_{k-1})}{1 - F_T(\Sigma_k)}\Big)\Big)\mathbb{1}_{\{\Sigma_k \le \bar{x}\}} &= N\Big(n\frac{f_T(x^*)}{1 - F_T(x^*)}\frac{S_k}{n}\Big)\mathbb{1}_{\{\Sigma_k \le \bar{x}\}} \\ &\preceq N\Big(\frac{f_T(0)}{1 - F_T(\bar{x})}S_k\Big), \end{aligned} \tag{2.3.24}$$

where we have used that the density $f_T(\cdot)$ has finite maximum value $f_T(0)$. The right-most term in (2.3.24) has finite second moment, since $\mathbb{E}[S^2] < \infty$. For the second term we proceed as follows:

$$A_n(k)^2\mathbb{1}_{\{\Sigma_k \ge \bar{x}\}} \overset{\text{a.s.}}{\le} A_n(k)^2\mathbb{1}_{\{D_k \ge \bar{x}/2\}} + A_n(k)^2\mathbb{1}_{\{\Sigma_k \ge \bar{x}, D_k < \bar{x}/2\}}, \tag{2.3.25}$$

The mean of the first term is be bounded by

$$\begin{aligned} \mathbb{E}[A_n(k)^2\mathbb{1}_{\{D_k \ge \bar{x}/2\}}] &\le n^2\mathbb{P}(S_k \ge n\bar{x}/2) \\ &\le 4n^2\frac{\mathbb{E}[S_k^2\mathbb{1}_{\{S_k \ge n\bar{x}/2\}}]}{(n\bar{x})^2}, \end{aligned} \tag{2.3.26}$$

and the right-hand side tends to zero as n tends to infinity since $\mathbb{E}[S^2] < \infty$. For the second term, more work is needed. First observe that

$$\mathbb{1}_{\{\Sigma_k \ge \bar{x}, D_k < \bar{x}/2\}} \le \mathbb{1}_{\{\Sigma_{k-1} \ge \bar{x}/2\}}. \tag{2.3.27}$$

Bounding $A_n(k)^2$ by $nA_n'(k)$, we get

$$\begin{aligned} &\mathbb{E}[A_n(k)^2\mathbb{1}_{\{\Sigma_k \ge \bar{x}, D_k < \bar{x}/2\}}] \\ &\qquad \le n^2\mathbb{E}[\mathbb{E}[\mathbb{1}_{\{\Sigma_{k-1} \le T_i \le \Sigma_k\}}\mathbb{1}_{\{\Sigma_{k-1} \ge \bar{x}/2\}} \mid \Sigma_k]] \\ &\qquad = n^2\mathbb{E}[\mathbb{1}_{\{\Sigma_{k-1} \ge \bar{x}/2\}}\mathbb{E}[\mathbb{1}_{\{\Sigma_{k-1} \le T_i \le \Sigma_k\}} \mid \Sigma_k]] \\ &\qquad = n^2\mathbb{E}[\mathbb{1}_{\{\Sigma_{k-1} \ge \bar{x}/2\}}(F_T(\Sigma_k) - F_T(\Sigma_{k-1}))]. \end{aligned} \tag{2.3.28}$$

By applying the Mean Value Theorem to $F_T(\cdot)$, we obtain

$$|F_T(\Sigma_k) - F_T(\Sigma_{k-1})| \overset{\text{a.s.}}{\leq} f_T(0) D_k, \tag{2.3.29}$$

since $f_T(0) = \max_{t \in \mathbb{R}^+} f_T(t)$. Plugging this into (2.3.28),

$$\begin{aligned} \mathbb{E}[A_n(k)^2 \mathbb{1}_{\{\Sigma_k \geq \bar{x}, D_k < \bar{x}/2\}}] &\leq n f_T(0) \mathbb{E}[S_k \mathbb{1}_{\{\Sigma_{k-1} \geq \bar{x}/2\}}] \\ &= n f_T(0) \mathbb{E}[S_k] \mathbb{P}(\Sigma_{k-1} \geq \bar{x}/2), \end{aligned} \tag{2.3.30}$$

the equality following from independence of S_k and Σ_{k-1}.

It is easy to see that the right-hand side converges to zero by using Chebyshev's inequality. Indeed, taking n so large that $n^{2/3}\mathbb{E}[S] \leq n\bar{x}/4$,

$$\begin{aligned} \mathbb{P}(\Sigma_{k-1} \geq \bar{x}/2) &\leq \mathbb{P}\Big(\Big| \sum_{i=1}^{tn^{2/3}} S_i - n^{2/3}\mathbb{E}[S_i] \Big| \geq n\bar{x}/4 \Big) \\ &\leq 16 \frac{tn^{2/3} \mathrm{Var}(S_i)}{n^2 \bar{x}} = o(n^{-1}). \end{aligned} \tag{2.3.31}$$

This concludes the proof that the second moment of the second term in (2.3.22) tends to zero as n tends to infinity. □

We conclude with a useful application of Doob's inequality, summarized in the following lemma:

Lemma 11. *Assume $(S_i)_{i=}^{\infty}$ is a sequence of* i.i.d. *random variables with finite second moment. Then, for any $\alpha, \beta > 0$ such that $\alpha < 2\beta$,*

$$\frac{\sup_{k \leq tn^{\alpha}} |\sum_{i=1}^{k} S_i - k\mathbb{E}[S]|}{n^{\beta}} \xrightarrow{\mathbb{P}} 0. \tag{2.3.32}$$

Proof. Define $M_k := \sum_{j=1}^{k}(S_j - \mathbb{E}[S])$. Then $k \mapsto M_k$ is a martingale. Therefore, by Doob's inequality applied to the sub-martingale $k \mapsto |M_k|$, we have

$$\mathbb{P}\Big(\frac{\sup_{k \leq tn^{\alpha}} |M_k|}{n^{\beta}} > \varepsilon \Big) \leq \frac{\mathbb{E}[M_{tn^{\alpha}}^2]}{\varepsilon^2 n^{2\beta}} = \frac{tn^{\alpha}\mathbb{E}[(S - \mathbb{E}[S])^2]}{\varepsilon^2 n^{2\beta}}. \tag{2.3.33}$$

This converges to zero since $\alpha < 2\beta$. Note that ε can depend on n, for example by defining $\varepsilon := n^{-\delta}$ and choosing δ such that $\delta < \beta - \alpha/2$. □

2.3.2 Proof of Theorem 4

The proof consists, again, in verifying the three conditions of Theorem 7, and establishing the convergence of the drift $C_n(\cdot)$, as

(i) $\sup_{t\leq\bar{t}} |n^{-1/3}C_n(tn^{2/3}) - \beta t - f_T'(0)/(2f_T(0)^2)t^2| \xrightarrow{\mathbb{P}} 0, \quad \forall \bar{t} \in \mathbb{R}^+,$

The filtration we consider henceforth is defined as

$$\mathcal{F}_i := \sigma(\{A_n(j), D_j\}_{j\leq i}). \tag{2.3.34}$$

Proof of (i)

We obtain the asymptotic drift by computing

$$\begin{aligned}\mathbb{E}[A_n(k)|\mathcal{F}_{k-1}] &= \sum_{i\notin\nu_k} \mathbb{E}[\mathbb{1}_{\{\Sigma_{k-1}\leq T_i\leq\Sigma_k\}} \mid \mathcal{F}_{k-1}]\\ &= \sum_{i\notin\nu_k} \mathbb{E}[\mathbb{1}_{\{T_i\leq\Sigma_k\}} \mid \Sigma_{k-1}, \{T_i \geq \Sigma_{k-1}\}],\end{aligned} \tag{2.3.35}$$

where, as above, ν_i denotes the set of the customers that no longer remain in the population at the beginning of the service of the i-th customer.

Adding the conditioning on $\{T_i \geq \Sigma_{k-1}\}$ does not influence the conditional expectation, since T_i is such that $i \notin \nu_{k-1}$. Indeed, note that $i \notin \nu_k$ implies $T_i \geq \Sigma_{k-1}$. Then, defining for simplicity $\mathcal{E}_{k-1} := \{\Sigma_{k-1}, \{T_i \geq \Sigma_{k-1}\}\}$, we compute

$$\begin{aligned}\mathbb{E}[A_n(k)|\mathcal{F}_{k-1}] &= \sum_{i\notin\nu_k} \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{\Sigma_{k-1}\leq T_i\leq\Sigma_k\}} \mid D_k, \mathcal{E}_{k-1}] \mid \mathcal{E}_{k-1}]\\ &= (n-|\nu_k|)\mathbb{E}\Big[\frac{F_T(\Sigma_k) - F_T(\Sigma_{k-1})}{1-F_T(\Sigma_{k-1})} \mid \mathcal{E}_{k-1}\Big]\\ &= \frac{(n-|\nu_k|)}{1-F_T(\Sigma_{k-1})}\mathbb{E}[F_T(\Sigma_k) - F_T(\Sigma_{k-1}) \mid \mathcal{E}_{k-1}].\end{aligned} \tag{2.3.36}$$

We now rearrange the terms in order to distinguish between the ones contributing to the limit and those vanishing, as follows

$$\begin{aligned}
&\mathbb{E}[A_n(k) \mid \mathcal{F}_{k-1}] - 1 \\
&\quad= \frac{(n-|\nu_k|)}{1-F_T(\Sigma_{k-1})}\mathbb{E}[F_T(\Sigma_k) - F_T(\Sigma_{k-1}) \mid \mathcal{E}_{k-1}] - \frac{1-F_T(\Sigma_{k-1})}{1-F_T(\Sigma_{k-1})} \\
&\quad= \frac{n}{1-F_T(\Sigma_{k-1})}\mathbb{E}[F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - n^{-1} \mid \mathcal{E}_{k-1}] \\
&\qquad - \frac{1}{1-F_T(\Sigma_{k-1})}\mathbb{E}[|\nu_k|(F_T(\Sigma_k) - F_T(\Sigma_{k-1})) - F_T(\Sigma_{k-1}) \mid \mathcal{E}_{k-1}] \\
&\quad=: A_n^{(1)}(k) - A_n^{(2)}(k). \qquad (2.3.37)
\end{aligned}$$

$A_n^{(1)}(k)$ groups all the terms appearing in the limit, while $A_n^{(2)}(k)$ groups all the terms of lower order, that vanish in the limit. We treat them separately, starting with $A_n^{(1)}(k)$. The term $F_T(\Sigma_k) - F_T(\Sigma_{k-1})$ is simplified through our assumptions. By (2.1.12) and Lemma 7,

$$\begin{aligned}
&A_n^{(1)}(k) \\
&= \frac{n}{1-F_T(\Sigma_{k-1})}\mathbb{E}[f_T(\Sigma_{k-1})D_k - n^{-1} + o_{\mathbb{P}}(n^{-4/3}) \mid \mathcal{E}_{k-1}], \qquad (2.3.38)
\end{aligned}$$

and by (2.1.14),

$$\begin{aligned}
\mathbb{E}[f_T(\Sigma_{k-1})D_k \mid \mathcal{E}_{k-1}] &= (f_T(0) + f_T'(0)\Sigma_{k-1} + o(\Sigma_{k-1}))\mathbb{E}[D] \qquad (2.3.39) \\
&= f_T(0)\mathbb{E}[D_k] + f_T'(0)\Sigma_{k-1}\mathbb{E}[D_k] + o(\Sigma_{k-1})\mathbb{E}[D_k],
\end{aligned}$$

where, with an abuse of notation, we denoted the term $|f_T(\Sigma_{k-1}) - f_T(0) - f_T'(0)\Sigma_{k-1}|$ as $o(\Sigma_{k-1})$. Since by the strong Law of Large Numbers, for $k = O(n^{2/3})$,

$$n^{1/3}|f_T(\Sigma_{k-1}) - f_T(0) - f_T'(0)\Sigma_{k-1}| \overset{\text{a.s.}}{\to} 0, \qquad (2.3.40)$$

also convergence in probability holds, that is, $o(\Sigma_{k-1}) = o_{\mathbb{P}}(n^{-1/3})$. In particular we see that $o(\Sigma_{k-1})\mathbb{E}[D_k] = o_{\mathbb{P}}(n^{-4/3})$ uniformly in $k \leq tn^{2/3}$.

Plugging (2.3.39) into (2.3.38) yields

$$\begin{aligned}A_n^{(1)}(k) &= \frac{n}{1-F_T(\Sigma_{k-1})}\\ &\quad\times\Big(f_T(0)\mathbb{E}[D_k]-\frac{1}{n}+f_T'(0)\mathbb{E}[D_k]\Sigma_{k-1}+o_{\mathbb{P}}(n^{-4/3})\Big)\\ &= \frac{1}{1-F_T(\Sigma_{k-1})}\\ &\quad\times\Big(f_T(0)\mathbb{E}[S_k]-1+\frac{f_T'(0)}{n}\sum_{j=1}^{k-1}S_j\mathbb{E}[S_k]+o_{\mathbb{P}}(n^{-1/3})\Big). \qquad (2.3.41)\end{aligned}$$

The criticality assumption $f_T(0)\mathbb{E}[S_k]=1+\beta n^{-1/3}+o(n^{-1/3})$ then leads to

$$A_n^{(1)}(k)=\frac{1}{1-F_T(\Sigma_{k-1})}\Big(\beta+f_T'(0)\mathbb{E}[S_k]\frac{\sum_{j=1}^{k-1}S_j}{n^{2/3}}+o_{\mathbb{P}}(1)\Big)n^{-1/3}. \qquad (2.3.42)$$

Since the drift term is defined as

$$\begin{aligned}C_n(s) &= \sum_{k=1}^{s}(\mathbb{E}[A_n(k)\mid\mathcal{F}_{k-1}]-1)\\ &= \sum_{k=1}^{s}(A_n^{(1)}(k)-A_n^{(2)}(k)) =: C_n^{(1)}(s)+C_n^{(2)}(s), \qquad (2.3.43)\end{aligned}$$

we sum (2.3.42) over k, obtaining

$$\begin{aligned}C_n^{(1)}(s) &= \sum_{k=1}^{s}\frac{\beta n^{-1/3}}{1-F_T(\Sigma_{k-1})}+\frac{f_T'(0)\mathbb{E}[S_1]}{n}\sum_{k=1}^{s}\frac{\sum_{j=1}^{k-1}S_j}{1-F_T(\Sigma_{k-1})}\\ &\quad+\sum_{k=1}^{s}\frac{1}{1-F_T(\Sigma_{k-1})}o_{\mathbb{P}}(n^{-1/3}). \qquad (2.3.44)\end{aligned}$$

Scaling time as $s=tn^{2/3}$ and multiplying the drift by $n^{-1/3}$, we obtain

$$\begin{aligned}n^{-1/3}C_n^{(1)}(tn^{2/3}) &= \sum_{k=1}^{tn^{2/3}}\frac{\beta n^{-2/3}}{1-F_T(\Sigma_{k-1})} \qquad (2.3.45)\\ &\quad+\frac{f_T'(0)\mathbb{E}[S_1]}{n^{4/3}}\sum_{k=1}^{tn^{2/3}}\frac{\sum_{j=1}^{k-1}S_j}{1-F_T(\Sigma_{k-1})}+\sum_{k=1}^{tn^{2/3}}\frac{n^{-1/3}o_{\mathbb{P}}(n^{-1/3})}{1-F_T(\Sigma_{k-1})}.\end{aligned}$$

The following lemma will be useful in analysing (2.3.45):

Lemma 12. *Let* $(S_i)_{i=1}^{\infty}$ *be a sequence of* i.i.d. *random variables such that* $\mathbb{E}[S_1^2] < \infty$. *Then*

$$\left| \frac{\sum_{n=1}^{N} \sum_{i=1}^{n} S_i}{N^2} - \frac{\mathbb{E}[S_1]}{2} \right| \xrightarrow{\mathbb{P}} 0, \qquad \text{as } N \to \infty. \tag{2.3.46}$$

Moreover,

$$\left| \frac{\sum_{n=1}^{N} (\sum_{i=1}^{n} S_i)^2}{N^3} - \frac{\mathbb{E}[S_1]^2}{3} \right| \xrightarrow{\mathbb{P}} 0, \qquad \text{as } N \to \infty. \tag{2.3.47}$$

Proof. Both claims are proved through Lemma 11. We omit the details. □

Another useful fact is the following Taylor expansion:

$$\begin{aligned} \frac{1}{1 - F_T(\Sigma_{k-1})} &= 1 + F_T(\Sigma_{k-1}) + \Big(\frac{1}{1 - F_T(\Sigma_{k-1})} - 1 - F_T(\Sigma_{k-1}) \Big) \\ &= 1 + O_{\mathbb{P}}(\Sigma_{k-1}). \end{aligned} \tag{2.3.48}$$

In what follows, we compute the limits (in probability) for each term in (2.3.45).

First term in (2.3.45). For the first term, by (2.3.48) and Lemma 12,

$$\begin{aligned} &\sup_{t \le \bar{t}} \Big| n^{-2/3} \sum_{k=1}^{tn^{2/3}} \frac{1}{1 - F_T(\Sigma_{k-1})} - t \Big| \\ &= \sup_{t \le \bar{t}} \Big| n^{-2/3} \sum_{k=1}^{tn^{2/3}} \Big(F_T(\Sigma_{k-1}) + \Big(\frac{1}{1 - F_T(\Sigma_{k-1})} - 1 - F_T(\Sigma_{k-1}) \Big) \Big) \Big| \\ &\le 2n^{-2/3} \sum_{k=1}^{\bar{t}n^{2/3}} F_T(\Sigma_{k-1}), \end{aligned} \tag{2.3.49}$$

where we dominated the error term in the Taylor expansion (2.3.48) by $F_T(\Sigma_{k-1})$ and used the fact that, as a function of t, the summation is an increasing function. We dominate the sum in the right-hand side of (2.3.49) uniformly as follows:

$$\begin{aligned} n^{-2/3} \sum_{k=1}^{\bar{t}n^{2/3}} 2F_T(\Sigma_{k-1}) &\le n^{-2/3} \sum_{k=1}^{\bar{t}n^{2/3}} 2 \sup_{k \le \bar{t}n^{2/3}} F_T(\Sigma_{k-1}) \\ &= 2\bar{t} F_T(\Sigma_{\bar{t}n^{2/3}}). \end{aligned} \tag{2.3.50}$$

The right-hand side of (2.3.50) tends to zero almost surely, and thus also in probability.

Second term in (2.3.45)**.** Again, by (2.3.48), the second term simplifies to

$$\begin{aligned} &\frac{f_T'(0)\mathbb{E}[S]}{n^{4/3}}\sum_{k=1}^{tn^{2/3}}\frac{\sum_{j=1}^{k-1}S_j}{1-F_T(\Sigma_{k-1})} \qquad (2.3.51)\\ &= \frac{f_T'(0)\mathbb{E}[S]}{n^{4/3}}\sum_{k=1}^{tn^{2/3}}\sum_{j=1}^{k-1}S_j + \frac{f_T'(0)\mathbb{E}[S]f_T(0)}{n^{4/3}}\sum_{k=1}^{tn^{2/3}}\sum_{j=1}^{k-1}S_j O_{\mathbb{P}}(\Sigma_{k-1}). \end{aligned}$$

By Lemma 12, the first term converges to $t^2 f_T'(0)\mathbb{E}[S]^2/2$ uniformly in t. Indeed,

$$\begin{aligned} &\sup_{t\le\bar{t}}\Big|n^{-4/3}\sum_{k=1}^{tn^{2/3}}\sum_{j=1}^{k-1}S_j - \frac{t^2}{2}\mathbb{E}[S]\Big|\\ &\quad= \sup_{t\le\bar{t}}\Big|\frac{\sum_{n=1}^{tn^{2/3}} n\mathbb{E}[S]}{n^{4/3}} - \frac{t^2}{2}\mathbb{E}[S] + \frac{\sum_{k=1}^{tn^{2/3}}(\sum_{i=1}^{k}S_i - k\mathbb{E}[S])}{n^{4/3}}\Big|\\ &\quad= \sup_{t\le\bar{t}}\Big|\frac{t\mathbb{E}[S]}{2n^{2/3}} + \frac{\sum_{k=1}^{tn^{2/3}}(\sum_{i=1}^{k}S_i - k\mathbb{E}[S])}{n^{4/3}}\Big|\\ &\quad\le \bar{t}\frac{\mathbb{E}[S]}{2n^{2/3}} + \bar{t}\frac{\sup_{k\le tn^{2/3}}|\sum_{i=1}^{k}S_i - k\mathbb{E}[S]|}{n^{2/3}}, \qquad (2.3.52) \end{aligned}$$

almost surely. The second term converges to zero in probability by Lemma 12, since

$$\begin{aligned} &\sup_{t\le\bar{t}}\Big|\frac{f_T'(0)\mathbb{E}[S]f_T(0)}{n^{4/3}}\sum_{k=1}^{tn^{2/3}}\Big(\sum_{j=1}^{k-1}S_j\Big)O_{\mathbb{P}}(\Sigma_{k-1})\Big|\\ &\qquad\le \frac{cf_T'(0)\mathbb{E}[S]f_T(0)}{n^{7/3}}\sum_{k=1}^{\bar{t}n^{2/3}}\Big|\Big(\sum_{j=1}^{k-1}S_j\Big)^2\Big|, \qquad (2.3.53) \end{aligned}$$

where we used the domination $O_{\mathbb{P}}(\Sigma_{k-1}) \le c\Sigma_{k-1}$. The right-most term in (2.3.53) then converges to zero in probability by Lemma 12.

Third term in (2.3.45)**.** The error term originates from the Taylor expansion of $n(F_T(\Sigma_k) - F_T(\Sigma_{k-1}))$ done in (2.3.38). To see that it is uniform in $k \le \bar{k}$, we write $F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1})D_k = \epsilon_k$ and, since by assumption $f_T'(\cdot)$ exists and is continuous in a neighborhood of 0, we bound it as

$$|n\epsilon_k| \leq n \sup_{x \leq cn^{-1/3}} |f_T'(x)| \frac{D_k^2}{2}, \tag{2.3.54}$$

similarly as in Lemma 7, when $k = O(n^{2/3})$. In particular,

$$\sup_{k \leq tn^{2/3}} \mathbb{E}[n\epsilon_k | \mathcal{E}_{k-1}] \leq \sup_{x \leq cn^{-1/3}} |f_T'(x)| \frac{\mathbb{E}[S^2]}{2n} = o(n^{-1/3}), \tag{2.3.55}$$

and the right-hand side is independent of k. This concludes the bound on the third term in (2.3.45) and thus the proof that

$$\sup_{t \leq \bar{t}} |n^{-1/3} C_n^{(1)}(tn^{2/3}) - \beta t - f_T'(0)/(2 f_T(0)^2) t^2| \xrightarrow{\mathbb{P}} 0. \tag{2.3.56}$$

□

To conclude, we prove that $\sup_{t \leq \bar{t}} n^{-1/3} |C_n^{(2)}(tn^{2/3})|$ vanishes in the limit. We develop the terms of $A_n^{(2)}(k)$ similarly as before, obtaining

$$\begin{aligned} A_n^{(2)}(k) &= \frac{1}{1 - F_T(\Sigma_{k-1})} \\ &\quad \times \mathbb{E}[|\nu_k|(f_T(\Sigma_{k-1}) D_k + o_{\mathbb{P}}(n^{-1})) - f_T(0)\Sigma_{k-1} + o_{\mathbb{P}}(n^{-1/3}) \mid \mathcal{E}_{k-1}] \\ &= \frac{1}{1 - F_T(\Sigma_{k-1})} \\ &\quad \times \mathbb{E}[|\nu_k| f_T(\Sigma_{k-1}) D_k - f_T(0)\Sigma_{k-1} + |\nu_k| o_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-1/3}) \mid \mathcal{E}_{k-1}] \\ &= \frac{\mathbb{E}[|\nu_k| f_T(0) D_k - f_T(0)\Sigma_{k-1} + f_T'(0)|\nu_k|\Sigma_{k-1} D_k}{1 - F_T(\Sigma_{k-1})} \\ &\quad + \frac{|\nu_k| o_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-1/3}) \mid \mathcal{E}_{k-1}]}{1 - F_T(\Sigma_{k-1})}, \end{aligned} \tag{2.3.57}$$

where $o_{\mathbb{P}}(n^{-1})$ is a convenient notation for the term $F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1}) D_k$ and $o_{\mathbb{P}}(n^{-1/3})$ for $F_T(\Sigma_{k-1}) - f_T(0)\Sigma_{k-1}$. Next, we sum (2.3.57) over k to obtain

$$\begin{aligned} C_n^{(2)}(s) = \sum_{k=1}^{s} & \frac{1}{1 - F_T(\Sigma_{k-1})} \\ & \times \mathbb{E}[|\nu_k| f_T(0) D_k - f_T(0)\Sigma_{k-1} + f_T'(0)|\nu_{k-1}|\Sigma_{k-1} D_k \mid \mathcal{E}_{k-1}] \\ & + \sum_{k=1}^{s} \frac{1}{1 - F_T(\Sigma_{k-1})} \mathbb{E}[|\nu_k| o_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-1/3}) \mid \mathcal{E}_{k-1}]. \end{aligned} \tag{2.3.58}$$

Recall that $|\nu_k| = k + N_n(k-1)$. Intuitively, k is of a much larger order than $N_n(k-1)$, therefore at first approximation we ignore $N_n(k-1)$ and we will later prove convergence of the terms containing it. We rescale, and split $C_n^{(2)}(k)$ into

$$\begin{aligned} n^{-1/3}C_n^{(2)}(tn^{2/3}) = n^{-1/3} &\sum_{k=1}^{tn^{2/3}} \frac{1}{1-F_T(\Sigma_{k-1})} \\ &\times \mathbb{E}[kf_T(0)D_k - f_T(0)\Sigma_{k-1} \mid \mathcal{E}_{k-1}] \\ &+ n^{-1/3}\sum_{k=1}^{tn^{2/3}} \frac{1}{1-F_T(\Sigma_{k-1})} \\ &\times \mathbb{E}[f_T'(0)k\Sigma_{k-1}D_k + ko_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-1/3}) \mid \mathcal{E}_{k-1}] \\ &+ \varepsilon_n, \end{aligned} \tag{2.3.59}$$

where ε_n represents the terms containing $N_n(k-1)$. Again we rescale and study each term separately.

First term in (2.3.59)**.** Expanding $(1 - F_T(\Sigma_{k-1}))^{-1}$ gives

$$\begin{aligned} &\frac{f_T(0)}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} \mathbb{E}\Big[kS_k - \sum_{j=1}^{k-1} S_j \mid \mathcal{E}_{k-1}\Big] \\ &\qquad + \frac{f_T(0)^2}{n^{1/3}} \sum_{k=1}^{tn^{2/3}} O_{\mathbb{P}}(\Sigma_{k-1})\mathbb{E}[kD_k - \Sigma_{k-1} \mid \mathcal{E}_{k-1}], \end{aligned} \tag{2.3.60}$$

The second term is almost surely dominated by the first for n sufficiently large, so that it is enough to show (uniform) convergence of the first term. By Lemma 12,

$$-\frac{1}{n^{4/3}}\sum_{k=1}^{tn^{2/3}}\sum_{j=1}^{k-1} S_j \xrightarrow{\mathbb{P}} -\frac{t^2}{2}\mathbb{E}[S], \qquad \frac{1}{n^{4/3}}\sum_{k=1}^{tn^{2/3}} k\mathbb{E}[S_k] \xrightarrow{\mathbb{P}} \frac{t^2}{2}\mathbb{E}[S]. \tag{2.3.61}$$

Therefore, (2.3.60) converges to zero in probability. Moreover, the convergence is uniform in $t \leq \bar{t}$ by Lemma 11.

Second term in (2.3.59)**.** Expanding $(1 - F_T(\Sigma_{k-1}))^{-1}$ and ignoring all but the highest order term, which can be almost surely dominated, we get for the second term

$$f_T'(0)\mathbb{E}[S]n^{-7/3}\sum_{k=1}^{tn^{2/3}} k \sum_{j=1}^{k-1} S_j + n^{-1/3}\sum_{k=1}^{tn^{2/3}} ko_{\mathbb{P}}(n^{-1}) + tn^{1/3}o_{\mathbb{P}}(n^{-1/3}). \tag{2.3.62}$$

One can check, similarly as in Lemma 12, that $N^{-3}\sum_{k=1}^{N} k \sum_{j=1}^{k} S_j$ converges in probability to a non-trivial limit. Since $7/3 > (2/3)3$, the first term converges to zero in probability. In addition, it converges uniformly in $t \leq \bar{t}$ because of the monotonicity of the sum. The small-o terms are dominated uniformly as has already been done in (2.3.54).

Third term in (2.3.59)**.** The remaining term is

$$\varepsilon_n = \sum_{k=1}^{tn^{2/3}} \frac{n^{-1/3}}{1 - F_T(\Sigma_{k-1})} \times \mathbb{E}[N_n(k-1)(f_T(0)D_k + f'_T(0)\Sigma_{k-1}D_k + o_{\mathbb{P}}(n^{-1})) \mid \mathcal{E}_{k-1}]. \tag{2.3.63}$$

Again it is sufficient to show that the first term in the Taylor expansion of $(1 - F_T(\Sigma_{k-1}))^{-1}$ converges uniformly. This simplifies the previous expression to

$$\frac{f_T(0)\mathbb{E}[S]}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} N_n(k-1) + \frac{f'_T(0)\mathbb{E}[S]}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} N_n(k-1)\Sigma_{k-1} + n^{-4/3} \sum_{k=1}^{tn^{2/3}} N_n(k-1) o_{\mathbb{P}}(n^{-1}). \tag{2.3.64}$$

The second and third terms are again almost surely dominated by the first for n large. Moreover, the first converges to zero uniformly in probability by the following lemma:

Lemma 13. *As $n \to \infty$,*

$$n^{-2/3} \sup_{j \leq an^{2/3}} |N_n(j)| \xrightarrow{\mathbb{P}} 0. \tag{2.3.65}$$

Proof. The proof follows the ideas of the proof of Lemma 6. We split $N_n(j)$ as the sum of a martingale and a predictable process, $N_n(j) = M_n(j) + C_n(j)$, and bound each one separately. The probability that $|M_n(j)|$ has large jumps $\mathbb{P}(n^{-2/3} \sup_{j \leq tn^{2/3}} |M_n(j)| \geq \varepsilon n^{2/3})$ is bounded through Doob's inequality, giving the upper bound $\mathbb{E}[M_n^2(an^{2/3})]/(\varepsilon n^{2/3})^2$. As was noted in Lemma 6, $\mathbb{E}[M_n(k)^2] = \mathbb{E}[V_n(k)]$, where $V_n(k)$ is the predictable quadratic variation of $M_n(k)$, and its expectation is given by

$$\begin{aligned}\mathbb{E}[V_n(k)] &= \mathbb{E}\Big[\sum_{i=1}^{k}(\mathbb{E}[A_n(i)^2 \mid \mathcal{F}_{i-1}] - \mathbb{E}^2[A_n(i) \mid \mathcal{F}_{i-1}])\Big] \\ &\leq \sum_{i=1}^{k} \mathbb{E}[A_n(i)^2]. \end{aligned} \tag{2.3.66}$$

This term is bounded exploiting Lemma 10. We have

$$\frac{1}{(\varepsilon n^{2/3})^2}\mathbb{E}[V_n(an^{2/3})] \leq \frac{1}{(\varepsilon n^{2/3})^2}\sum_{i=1}^{an^{2/3}} \mathbb{E}[(A'_n(i))^2], \tag{2.3.67}$$

which tends to zero because $\mathbb{E}[A'^2_n(i)] < \infty$ uniformly in $i = O(n^{2/3})$ by Lemma 10. The $C_n(j)$ term computed in (2.3.37) and (2.3.43) is the difference of two increasing processes. Therefore, it can be bounded from above and below as was done in Lemma 5. We omit the details. □

This concludes the proof that

$$\sup_{t\leq \bar{t}} |n^{-1/3}C_n^{(2)}(tn^{2/3})| \xrightarrow{\mathbb{P}} 0, \tag{2.3.68}$$

and thus we have proven that

$$\sup_{t\leq \bar{t}} |n^{-1/3}C_n(n^{2/3}t) - \beta t - t^2 f'_T(0)\mathbb{E}[S]^2/2| \xrightarrow{\mathbb{P}} 0. \tag{2.3.69}$$

This completes the proof of (i). □

Proof of (ii)

First we compute $\mathbb{E}[A_n(k)^2 \mid \mathcal{F}_{k-1}]$. By proceeding as in (2.2.48) we obtain

$$\begin{aligned}&\mathbb{E}[A_n(k)^2 \mid\mathcal{F}_{k-1}] - \mathbb{E}[A_n(k) \mid \mathcal{F}_{k-1}] \\ &\quad= \mathbb{E}\Big[\sum_{\substack{l\neq m\\ l,m\notin \nu_k}} \mathbb{1}_{\{\Sigma_{k-1}\leq T_m\leq \Sigma_k\}}\mathbb{1}_{\{\Sigma_{k-1}\leq T_l\leq \Sigma_k\}} \mid \mathcal{F}_{k-1}\Big] \\ &\quad= \sum_{\substack{l\neq m\\ l,m\notin \nu_k}} \mathbb{E}\Big[\frac{F_T(\Sigma_k)-F_T(\Sigma_{k-1})}{1-F_T(\Sigma_{k-1})}\frac{F_T(\Sigma_k)-F_T(\Sigma_{k-1})}{1-F_T(\Sigma_{k-1})} \mid \mathcal{F}_{k-1}\Big] \\ &\quad= \frac{\sum_{l\neq m}\mathbb{E}[(f_T(\Sigma_{k-1})D_k + o_{\mathbb{P}}(D_k^{-4/3}))^2 \mid \mathcal{F}_{k-1}]}{(1-F_T(\Sigma_{k-1}))^2}, \end{aligned} \tag{2.3.70}$$

where the sum is over the set $\{l, m \leq n : l \neq m, l, m \notin \nu_k\}$ when not specified. We also denoted, for convenience, $F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1})D_k$ as $o_{\mathbb{P}}(D_k^{4/3})$. We proceed as in (2.2.49) and (2.2.50). By Lemma 7,

$$\begin{aligned}
&\mathbb{E}[A_n(k)^2 \mid \mathcal{F}_{k-1}] - \mathbb{E}[A_n(k) \mid \mathcal{F}_{k-1}] \\
&\quad = \frac{1}{(1 - F_T(\Sigma_{k-1}))^2} \\
&\qquad \times \sum_{l \neq m} \mathbb{E}[f_T(\Sigma_{k-1})^2 D_k^2 + 2 f_T(\Sigma_{k-1}) D_k o_{\mathbb{P}}(D_k^{4/3}) + o_{\mathbb{P}}(D_k^2) \mid \mathcal{F}_{k-1}] \\
&\quad = \frac{1}{(1 - F_T(\Sigma_{k-1}))^2} \\
&\qquad \times \sum_{l \neq m} ((f_T(0) + f_T'(0)\Sigma_{k-1} + o_{\mathbb{P}}(\Sigma_{k-1}))^2 \mathbb{E}[D_k^2] + o_{\mathbb{P}}(n^{-2})). \quad (2.3.71)
\end{aligned}$$

Here $o_{\mathbb{P}}(\Sigma_{k-1})$ is a shorthand notation for $f_T(\Sigma_{k-1}) - f_T(0) - f_T'(0)\Sigma_{k-1}$. Developing the coefficient of $\mathbb{E}[D_k^2]$ reveals that it has the form $f_T(0)^2 + \alpha_k n^{-1/3} + \beta_k o(n^{-1/3})$, with α_k and β_k converging in probability to a constant, for $k = O(n^{2/3})$. We can ignore all the terms except the one with the leading order, $f_T(0)^2$. From this point onwards the computations are identical to (2.2.52), concluding the proof of (ii). □

Proof of (iii) and (iv)

The proof of (iii) in the exponentials arrivals case can be carried over to the general arrivals case without any significant changes, since it relies only on (ii) and Lemma 13. For (iv), we split the quantity according to

$$\begin{aligned}
&n^{-2/3}\mathbb{E}[\sup_{t \leq \bar{t}} |M_n(tn^{2/3}) - M_n(tn^{2/3}-)|^2] \qquad (2.3.72) \\
&\qquad \leq n^{-2/3}\mathbb{E}[\sup_{k \leq \bar{t}n^{2/3}} |A_n(k)|^2] + n^{-2/3}\mathbb{E}[\sup_{k \leq \bar{t}n^{2/3}} |\mathbb{E}[A_n(k) \mid \mathcal{F}_{k-1}]|^2].
\end{aligned}$$

The second term is straightforward. Indeed, (2.3.37) and (2.3.42) give the crude bound

$$\mathbb{E}[A_n(k)|\mathcal{F}_{k-1}] \overset{\text{a.s.}}{\leq} A_n^{(1)}(k) \overset{\text{a.s.}}{\leq} c + o_{\mathbb{P}}(1), \qquad (2.3.73)$$

for $k = O(n^{2/3})$ and some positive constant $c > 1$, uniform over $k \leq \bar{t}n^{2/3}$. The first term can also be estimated imitating (2.2.64). Indeed, fix $\varepsilon > 0$

and split it as

$$\mathbb{E}[\sup_{k\leq \bar{t}n^{2/3}} |A_n(k)|^2] = \mathbb{E}[\sup_{k\leq \bar{t}n^{2/3}} A_n(k)^2 \mathbb{1}_{\{\sup_{k\leq \bar{t}n^{2/3}} A_n(k)^2\leq \varepsilon n^{2/3}\}}]$$
$$+ \mathbb{E}[\sup_{k\leq \bar{t}n^{2/3}} A_n(k)^2 \mathbb{1}_{\{\sup_{k\leq \bar{t}n^{2/3}} A(k)^2 > \varepsilon n^{2/3}\}}]. \quad (2.3.74)$$

The first term is bounded by ε. We bound the second term as in (2.2.65) and using Lemma 10. We omit the details. □

2.4 Arrivals with ℓ-th order contact

The goal of this section is to drop the assumption $f'_T(0) \neq 0$ of Section 2.3. In fact, we will prove a limit theorem for the more general case where the function $t \mapsto f_T(t) - 1/\mathbb{E}[S]$ has ℓ-th order contact in zero, defined as follows:

Definition 1 (ℓ-th order contact point)**.** Given a smooth, real-valued, function $f(\cdot)$, and $\ell \in \mathbb{N}$, we say that $f(\cdot)$ has ℓ-th order contact in $\bar{t}$ if $f(\bar{t}) = 0$, $f^{(l)}(\bar{t}) = 0$ for $l = 1, \ldots, \ell - 1$ and $f^{(\ell)}(\bar{t}) \neq 0$.

If, in Definition 1, $f(\bar{t}) = o(1)$, we still say that $f(\cdot)$ has an ℓ-th order contact in $\bar{t}$. Indeed, our criticality assumption is $f_T(0) - 1/\mathbb{E}[S] = o(1)$, where the error term is specified later. The assumption that the argmax of $f_T(\cdot)$ is zero allows us to consider both odd and even order contacts. We will assume again that the service times are given by $D_i := S_i/c_n$, where $(S_i)_{i=1}^n$ is a sequence of i.i.d. random variables such that $\mathbb{E}[S^2] < \infty$ and c_n is the rate at which the server processes the customers. In this case, we have the following result for the embedded queue $Q^e_n(\cdot)$ introduced in Chapter 2:

Theorem 8 (Asymptotics for the critical ℓ-th order embedded queue)**.** *Assume that the function $f_T(t) - 1/\mathbb{E}[S]$ has ℓ-th order contact in 0, where $\ell \geq 1$. Define*

$$\tau = \frac{\ell}{\ell + 1/2}. \quad (2.4.1)$$

Assume that the service times $(S_i)_{i=1}^n$ are such that $\mathbb{E}[S^2] < \infty$ and that the service speed is given by $c_n = n/(1 + \beta n^{-\tau/2})$. Assume further that the heavy-traffic condition $nf_T(0)\mathbb{E}[D] = 1 + \beta n^{-\gamma/2}$ holds. Then, as $n \to \infty$,

$$n^{-\tau/2} Q^e_n(\cdot n^\tau) \xrightarrow{\mathrm{d}} \phi(\widehat{N})(\cdot), \qquad \text{in } (\mathcal{D}, J_1), \quad (2.4.2)$$

where $\widehat{N}(\cdot)$ is given by

$$\widehat{N}(t) := \beta t - ct^{\ell+1} + \sigma W(t), \tag{2.4.3}$$

for constants $c, \sigma \in \mathbb{R}^+$, and $W(\cdot)$ a standard Brownian motion.

Note that $\ell = 1$ gives the same scaling as in Theorem 4. Moreover, the case $\ell = 2$ has already been known in the literature for quite some time, at least at a heuristic level. Newell [77] derived the correct exponents ($\tau = 4/5$) through an argument using the Fokker-Planck equation associated with the queue-length process. Note also that

$$\lim_{\ell\to\infty} \frac{\tau}{2} = \lim_{\ell\to\infty} \frac{\ell}{2\ell+1} = \frac{1}{2}, \tag{2.4.4}$$

suggesting that the correct scaling for the uniform arrivals case (∞-order contact) is the diffusive one.

The scaling constants. We will again express $Q_n^e(\cdot) = \phi(N_n)(\cdot)$ and, generalizing the heuristics in (2.1.19), we get

$$\begin{aligned} N_n(tn^\tau) &\approx \sum_{i=1}^{tn^\tau} \Big(\sum_{j\notin v_i} \mathbb{1}_{\{\sum_{l=1}^{i-1} D_l \le T_j \le \sum_{l=1}^{i} D_l\}} - 1\Big) \\ &\approx \sum_{i=1}^{tn^\tau} (n(F_T(\Sigma_i) - F_T(\Sigma_{i-1})) - 1) \\ &\approx \sum_{i=1}^{tn^\tau} (f_T(\Sigma_{i-1})\mathbb{E}[S] - 1) \approx \sum_{i=1}^{tn^\tau} \Big(\sum_{l=1}^{i-1} \frac{S_l}{n} \Big)^{\ell} f_T(0)^{(\ell)} \mathbb{E}[S], \end{aligned} \tag{2.4.5}$$

where $\Sigma_i := \sum_{j=1}^{i} D_i$. Thus the leading order term (up to a multiplicative constant) of the queue-length process is

$$N_n(tn^\tau) \approx \sum_{i=1}^{tn^\tau} \frac{i^\ell}{n^\ell} \approx t^{\ell+1} n^{(\ell+1)\tau - \ell}. \tag{2.4.6}$$

The Brownian fluctuations of the random sum in (2.4.5) are of order $n^{\tau/2}$. Equating the order of magnitude of the fluctuations and that of (2.4.6) gives

$$(\ell+1)\tau - \ell = \tau/2 \;\Rightarrow\; \tau = \frac{\ell}{\ell + 1/2}. \tag{2.4.7}$$

The proof of Theorem 8 follows from the proof for the case $\ell = 1$. Remarkably, the higher moments of S (higher than two) are not required to be finite. In Section 2.4.1, we perform the key steps in the analysis in order to show how to proceed in this general case.

There are no explicit formulas for the distribution of the first hitting time of zero of a Brownian motion with parabolic drift and thus we resort to numerical simulations. In Table 2.1 we provide numerical simulations for the case of truncated normal arrival times, so that $\ell = 2$. Recall that if Z is normally distributed with mean 0 and variance σ^2, then $|Z|$ is equivalent to a zero mean normal distribution that is conditioned to be positive, which is a particular case of the so-called truncated normal distribution; see [80]. In particular, $f_T'(0) = 0$ and $f_T''(0) < 0$.

	$q = 1, \beta = 1$	$q = 2, \beta = 1$
n	$n^{1/5}\mathbb{E}[\mathrm{BP}_n]$	$n^{1/5}\mathbb{E}[\mathrm{BP}_n]$
10	3.8340	5.3984
100	3.0997	4.1232
1000	2.8378	3.8772
10000	2.7801	3.7721
100000	2.7942	3.7548

Table 2.1: Mean busy period for the pre-limit queue with truncated normal arrivals and different population sizes. The truncated normal distribution has scale parameter $\sigma = \sqrt{\pi}/\sqrt{2}$, and $f_T'(0) = 0$, $f_T''(0) < 0$. Each value for the pre-limit queue is the average of 10^4 simulations.

2.4.1 Proof of Theorem 8

Note that τ in (2.4.1) is such that $\tau < 1$. Some simple relations hold between τ and ℓ and we will use these throughout this section. These are given by

$$\ell - \frac{\tau}{2} = \tau\ell, \qquad \ell + \frac{\tau}{2} = \tau(\ell + 1). \tag{2.4.8}$$

We now explicitly state our assumptions for the ℓ-th order contact case. We assume that the distribution function of the arrival times satisfies

$$F_T(x) - F_T(\bar{x}) = f_T(\bar{x})(x - \bar{x}) + o(|x - \bar{x}|^{1+\tau/2}). \tag{2.4.9}$$

This is, for example, the case when $F_T(\cdot) \in \mathcal{C}^2([0, \infty))$. Further, we assume that the maximum of the density $f_T(\cdot)$ is obtained in zero. The heavy-traffic condition is then given by

$$n f_T(0)\mathbb{E}[D] = 1 + \beta n^{-\tau/2}. \tag{2.4.10}$$

We proceed by verifying conditions (i)- (iv), where (i) is given by

(i) $\sup_{t \leq \bar{t}} |n^{-1/3} C_n(tn^{2/3}) - \beta t - ct^{\ell}| \xrightarrow{\mathbb{P}} 0, \qquad c > 0, \ \forall \bar{t} \in \mathbb{R}^+,$

and conditions (ii)-(iv) correspond to the conditions in Theorem 7. The drift $C_n(\cdot)$ in (i) is defined as in (2.3.43). We will treat condition (i) in great detail as this changes profoundly, since the limiting drift is significantly different. We will then discuss how (ii)-(iv) follow from the computations in Section 2.3.2.

Proof of (i). Recall that $\mathcal{E}_{k-1} := \{\Sigma_{k-1}, \{T_i \geq \Sigma_{k-1}\}\}$ and $\Sigma_j := \sum_{i=1}^{j} D_i$. The conditioned number of arrivals during one service are given by

$$\begin{aligned}
&\mathbb{E}[A_n(k) \mid \mathcal{F}_{k-1}] - 1 \\
&\quad = \frac{n}{1 - F_T(\Sigma_{k-1})} \mathbb{E}[(F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - 1/n) \mid \mathcal{E}_{k-1}] \\
&\qquad - \frac{1}{1 - F_T(\Sigma_{k-1})} \mathbb{E}[(|\nu_k|(F_T(\Sigma_k) - F_T(\Sigma_{k-1})) - F_T(\Sigma_{k-1})) \mid \mathcal{E}_{k-1}] \\
&\quad =: A_n^{(1)}(k) - A_n^{(2)}(k).
\end{aligned} \tag{2.4.11}$$

Correspondingly, the drift is decomposed as

$$C_n(s) = \sum_{k=1}^{s} (A_n^{(1)}(k) - A_n^{(2)}(k)) =: C_n^{(1)}(s) + C_n^{(2)}(s). \tag{2.4.12}$$

The process $A_n^{(1)}(\cdot)$ represents the terms appearing in the limit, while $A_n^{(2)}(\cdot)$ represents the terms that vanish. Performing a Taylor expansion gives us

$$\begin{aligned}
&A_n^{(1)}(k) \\
&\quad = \frac{1}{1 - F_T(\Sigma_{k-1})} \Big(\beta + \frac{f_T(0)^{(\ell)}}{\ell!} \mathbb{E}[S] \frac{(\sum_{j=1}^{k-1} S_j)^{\ell}}{n^{\ell\tau}} + o_{\mathbb{P}}(1) \Big) n^{-\tau/2}.
\end{aligned} \tag{2.4.13}$$

It is easy to check that both the linear part of the drift and the error term converge uniformly by proceeding as in (2.3.50) and the following computations. Therefore, we focus on the second term of $A_n^{(1)}(\cdot)$, that is, on

$$\Big(\mathbb{E}[S]\frac{f_T(0)^{(\ell)}}{\ell!}\Big)n^{-\tau/2}\sum_{i=1}^{tn^\tau}\frac{1}{1-F_T(\Sigma_{i-1})}\Big(\frac{\sum_{j=1}^{i-1}S_j}{n}\Big)^{\ell}, \tag{2.4.14}$$

for which we prove (uniform) convergence in probability. We begin by computing

$$\begin{aligned}
&\Big|\frac{1}{n^{\tau(\ell+1)}}\sum_{i=1}^{tn^\tau}\Big(\sum_{j=1}^{i-1}S_j\Big)^{\ell}-\frac{t^{\ell+1}}{\ell+1}\mathbb{E}[S]^{\ell}\Big| \\
&\qquad=\frac{1}{n^{\tau(\ell+1)}}\Big|\sum_{i=1}^{tn^\tau}\Big(\sum_{j=1}^{i-1}S_j\Big)^{\ell}-\sum_{i=1}^{tn^\tau}((i-1)\mathbb{E}[S])^{\ell}+o(n^{\tau(\ell+1)})\Big| \\
&\qquad\le\frac{1}{n^{\tau(\ell+1)}}\sum_{i=1}^{tn^\tau}\Big|\Big(\sum_{j=1}^{i-1}S_j\Big)^{\ell}-(i-1)^{\ell}\mathbb{E}[S]^{\ell}\Big|+o(1) \qquad (2.4.15)
\end{aligned}$$

We will make use of the following lemma:

Lemma 14. *Assume $(S_i)_{i\ge 0}$ is a sequence of* i.i.d. *random variables such that $\mathbb{E}[S^2]<\infty$. Then for any $\tau>0,\beta\in\mathbb{R}$ such that $-\tau<2\beta$,*

$$\frac{\sup_{k\le tn^\tau}|(\sum_{i=1}^{k}S_i)^{\ell}-k^{\ell}\mathbb{E}[S]^{\ell}|}{n^{\tau\ell+\beta}}\xrightarrow{\mathbb{P}}0. \tag{2.4.16}$$

Proof. The proof is an application of Lemma 11, hence we only sketch it. We have

$$\begin{aligned}
&\sup_{k\le tn^\tau}\Big|\Big(\sum_{i=1}^{k}S_i\Big)^{\ell}-k^{\ell}\mathbb{E}[S]^{\ell}\Big| \\
&\qquad\le\sup_{k\le tn^\tau}\Big|\Big(k\mathbb{E}[S]+\sup_{k\le tn^\tau}\Big|\sum_{i=1}^{k}(S_i-\mathbb{E}[S])\Big|\Big)^{\ell}-k^{\ell}\mathbb{E}[S]^{\ell}\Big| \qquad (2.4.17)
\end{aligned}$$

The term on the right can be shown to converge to zero when appropriately rescaled. In fact, the leading order term is given by

$$\begin{aligned}
&\sup_{k\le tn^\tau}k^{\ell-1}\mathbb{E}[S]^{\ell-1}\sup_{k\le tn^\tau}|\sum_{i=1}^{k}(S_i-\mathbb{E}[S])| \\
&\qquad=(tn)^{\tau(\ell-1)}\mathbb{E}[S]^{\ell-1}\sup_{k\le tn^\tau}|\sum_{i=1}^{k}(S_i-\mathbb{E}[S])|, \qquad (2.4.18)
\end{aligned}$$

which converges to zero when divided by $n^{\tau\ell+\beta}$, with $\beta > -\tau/2$ by Lemma 11. □

Note that when $\ell = 1$ we recover Lemma 11, since in that case $\tau + \beta > \tau - \tau/2 = \tau/2$. By Lemma 14 the right side of (2.4.15) converges to zero. The convergence is uniform in $t \leq \bar{t}$ by monotonicity in t. We can similarly analyse $A_n^{(2)}(\cdot)$ and $C_n^{(2)}(\cdot)$. Equation (2.3.58) in this case is

$$\begin{aligned} C_n^{(2)}(s) = \sum_{k=1}^{s} & \frac{1}{1 - F_T(\Sigma_{k-1})} \\ & \times \mathbb{E}\Big[|\nu_k| f_T(0) D_k - f_T(0)\Sigma_{k-1} + \frac{f_T(0)^{(\ell)}}{\ell!}|\nu_k|(\Sigma_{k-1})^{\ell} D_k\Big], \\ + \sum_{k=1}^{s} & \frac{1}{1 - F_T(\Sigma_{k-1})} \\ & \times \mathbb{E}\Big[|\nu_k| o_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-\tau/2}) \mid \mathcal{E}_{k-1}\Big]. \end{aligned} \tag{2.4.19}$$

where $o_{\mathbb{P}}(n^{-1}) =: |F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1})D_k|$ and $o_{\mathbb{P}}(n^{-\tau/2}) =: |F_T(\Sigma_{k-1}) - f_T(0)\Sigma_{k-1}|$. This gives a decomposition of the drift $C_n^{(2)}(\cdot)$ similar to (2.3.59) as

$$\begin{aligned} n^{-\tau/2} C_n^{(2)}(tn^{\tau}) = n^{-\tau/2} & \sum_{k=1}^{tn^{\tau}} \frac{1}{1 - F_T(\Sigma_{k-1})} \mathbb{E}[k f_T(0) D_k - f_T(0)\Sigma_{k-1} \mid \mathcal{E}_{k-1}] \\ + \sum_{k=1}^{tn^{\tau}} & \frac{n^{-\tau/2}}{1 - F_T(\Sigma_{k-1})} \\ & \times \mathbb{E}[f_T(0)^{(\ell)} k (\Sigma_{k-1})^{\ell} D_k + k o_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-\tau/2}) \mid \mathcal{E}_{k-1}] \\ + \varepsilon_n, & \end{aligned} \tag{2.4.20}$$

where ε_n groups all the terms containing $N_n(k)$. For the first term, we compute (again ignoring the higher order terms in the expansion of $(1 - F_T(\Sigma_{k-1}))^{-1}$)

$$\begin{aligned} n^{-\tau/2} \Big| \sum_{k=1}^{tn^{\tau}} & \big(k f_T(0)\mathbb{E}[D_1] - f_T(0)\Sigma_{k-1}\big)\Big| \\ & = f_T(0) n^{-1-\tau/2} \Big| \sum_{k=1}^{tn^{\tau}} \Big(\sum_{j=1}^{k-1} (S_j - \mathbb{E}[S]) \Big) \Big| \\ & \leq f_T(0) n^{-1-\tau/2} \sum_{k=1}^{tn^{\tau}} \Big| \sum_{j=1}^{k-1} (S_j - \mathbb{E}[S]) \Big|. \end{aligned} \tag{2.4.21}$$

Applying Lemma 11 with $\beta = 1 - \tau/2$ (note that $2\beta < \tau$) shows that the first term in (2.4.20) tends to zero in probability, uniformly in $t \leq \bar{t}$. The remaining terms are treated, without additional complications, similarly to the analogous terms in Section 2.3.

Proof of (ii). The proof of (ii) was based on an analysis of the leading-order term of the quadratic variation, in which the ℓ-th derivative of the density played no role.

Proof of (iii). The proof of (iii) relies on Lemma 13 which, in turn, relies on the analysis of the order statistics done in Lemma 10. Since the latter does not depend on the derivatives of the density (but rather on its continuity), the proof carries over.

Proof of (iv). Again, the proof relies on the analysis carried out in Lemma 10.

Having proved conditions (i)–(iv), the proof of Theorem 8 is complete. □

2.5 Conclusions

In this chapter we have shown that, in the limit for $n \to \infty$, the depletion-of-points effect gives rise to a negative quadratic drift of the embedded queue process. This implies, in particular, that, even for very large n, the finite pool of customers has a sizeable impact on the performance of the system after only $n^{2/3} \ll n$ services. We have also shown that, under mild assumptions, the *form* of the limiting process does not depend on the arrival time or service-time distribution. In fact, the density in zero of the arrival time distribution, its derivative in zero and the second moment of the service time completely determine the limiting process. Next we will show that the $\Delta_{(i)}/G/1$ *queue-length process* satisfies a similar scaling limit as the embedded queue, as long as the scaling exponents are modified accordingly.

CHAPTER 3

The queue-length process

In this chapter, we build on the results of Chapter 2 to show that, when the arrival times are exponentially distributed, the *queue-length process* of the $\Delta_{(i)}/G/1$ queue, when appropriately rescaled, converges to a Brownian motion with parabolic drift. To this end, we construct a time-change and we show that the supremum distance between the time-changed embedded queue and the queue-length process is given by the maximum number of customers arriving during one service. We prove that the time-change converges to a constant times the identity function, and that the rescaled maximum number of arrivals during one service converges to zero. This implies that the difference between the embedded queue of Chapter 2 and the queue-length process is negligible in the limit.

Next we adopt a different perspective, and prove the convergence of the queue-length process directly. We give an explicit representation of the queue-length process in terms of the empirical distribution function of the arrival times, and we prove a functional Central Limit Theorem under our special scaling. Next, building on this result we prove a sample-path Little's Law for the critical $\Delta_{(i)}/G/1$ queue.

We conclude the chapter by analysing the queue-length process of the *subcritical* $\Delta_{(i)}/G/1$ queue. We prove that the queue-length process converges pointwise to the stationary distribution of a $M/G/1$ queue with the same service times of the original $\Delta_{(i)}/G/1$ queue and arrival rate given by the density of the arrival times.

3.1 Model description

In this section we give an explicit construction of the $\Delta_{(i)}/G/1$ queue-length process. All the functions that we consider are elements of $\mathcal{D}$. We endow $\mathcal{D}$ with the Skorokhod J_1 topology. Define the total number of customers who arrive in the interval $[0,t]$ as

$$\mathcal{A}_n(t) := \sum_{i=1}^{n} \mathbb{1}_{\{T_i \le t\}}. \tag{3.1.1}$$

Let

$$\sigma(t) := \max\Big\{k \ge 0 : \sum_{i=1}^{k} S_i \le t\Big\} \tag{3.1.2}$$

be the renewal process associated with the job sizes $S_1, S_2, \ldots, S_n$. We rescale the server speed to be equal to $c_n = n/(1+\beta n^{-1/3})$, and let $\sigma_n(t) := \sigma(c_n t)$ denote the renewal process associated to the service times $D_i = S_i/c_n$. Define the net-put process as

$$P_n(t) := \sum_{i=1}^{\mathcal{A}_n(t)} S_i - c_n t. \tag{3.1.3}$$

The process $P_n(\cdot)$ is used in defining the rescaled *cumulative busy time process* as

$$B_n(t) := t - \frac{I_n(t)}{c_n} = t - \inf_{0\le s\le t}\Big(\frac{P_n(s)}{c_n}\Big)^{-}, \tag{3.1.4}$$

where $f(x)^- = \min\{0, f(x)\}$ (resp. $f(x)^+ = \max\{0, f(x)\}$), and $I_n(\cdot)$ represents the cumulative idle time. With definition (3.1.4), the total time that the server has spent working up to time t is given by $c_n B_n(t)$.

Finally, the queue-length process $Q_n(\cdot)$ is given by

$$Q_n(t) := X_n(0) + \mathcal{A}_n(t) - \sigma_n(B_n(t)), \tag{3.1.5}$$

where $X_n(0)$ denotes the number of customers already in the queue at the beginning of the first service.

Note that the time change $t \mapsto B_n(t)$ depends, through $P_n(\cdot)$, both on $(T_i)_{i=1}^n$ and $(S_i)_{i=1}^n$. Because of this, a direct analysis of the asymptotic behavior of $Q_n(\cdot)$ is often challenging, and various techniques have been developed to overcome this difficulty; see the discussion in Section 1.1.5. In fact, the advantage of the embedded process approach of Chapter 2 is that one does not have to deal with the process $B_n(t)$.

The main theorem of this section is the following:

Theorem 9 (Critically loaded $\Delta_{(i)}/G/1$ queue with exponential arrivals)**.** *Let $(T_i)_{i=1}^n$ be i.i.d. rate λ exponential random variables and let the i.i.d. service times $(S_i)_{i=1}^n$ be such that $\mathbb{E}[S^2] < \infty$. Assume that the heavy-traffic condition* (2.1.2) *holds. Then*

$$n^{-1/3}Q_n(\cdot n^{-1/3}) \xrightarrow{\mathrm{d}} \phi(\widehat{X})(\cdot), \tag{3.1.6}$$

where $\widehat{X}(\cdot)$ is the diffusion process

$$\widehat{X}(t) := \beta\lambda t - \frac{\lambda^2}{2}t^2 + \sigma W(t), \tag{3.1.7}$$

with $\sigma^2 := \lambda^3\mathbb{E}[S^2]$ and $W(\cdot)$ a standard Brownian motion.

The diffusion process (3.1.7) should be compared to (2.1.10). The former is obtained from the latter by the linear time change $t \mapsto \lambda t = t/\mathbb{E}[S]$. For a critical system, in the limit a single service is instantaneous, suggesting that the embedded queue and the queue-length process coincide. However, the (rescaled) *cumulative service time* converges to a deterministic process, leading to the time change described above.

As an immediate consequence of our approach we get an asymptotic result for BP_n, the length of the first busy period in the $\Delta_{(i)}/G/1$ queue. This result will be valid for general arrival times. In order to obtain a sizeable first busy period, we assume that the queue length at time zero grows with n as

$$\lim_{n\to\infty}\frac{Q_n(0)}{n^{1/3}} = q > 0. \tag{3.1.8}$$

The next theorem is the continuous-time analogue of Theorem 5, and it will show that the size of the first busy period depends crucially both on β and q.

Theorem 10 (First busy period of the critical $\Delta_{(i)}/G/1$ queue with general arrivals)**.** *Let $Q_n(\cdot)$ be the queue-length process of the $\Delta_{(i)}/G/1$ queue. Assume that the heavy-traffic condition* (1.1.2) *holds. Assume further that* (3.1.8) *holds and that $f_T'(0) < 0$.. Let BP_n denote the first busy period of $Q_n(\cdot)$. Then*

$$n^{1/3}\mathrm{BP}_n \xrightarrow{\mathrm{d}} T_{\widehat{X}_q}^{\beta f_T}(0), \tag{3.1.9}$$

where $T_{W_q}^{\beta\lambda}(0)$ is the time until the process $\widehat{X}_q(\cdot)$ crosses level 0*, with*

$$\widehat{X}_q(t) = q + \beta f_T(0)t + \frac{f_T'(0)}{2}t^2 + \sigma W(t). \tag{3.1.10}$$

Here we have $\sigma^2 := f_T^3(0)\mathbb{E}[S^2]$ and $W(\cdot)$ is a standard Brownian motion.

	$q=1, \beta=1$		$q=2, \beta=1$	
n	$n^{1/3}\mathbb{E}[\text{BP}_n]$	rel. error	$n^{1/3}\mathbb{E}[\text{BP}_n]$	rel. error
10	2.8630	0.6581	3.8646	0.5620
100	1.9862	0.1503	2.9665	0.1991
1000	1.8103	0.0484	2.6486	0.0706
10000	1.7725	0.0265	2.5596	0.0346
100000	1.7440	0.0100	2.5050	0.0125
∞	1.7267	—	2.4740	—

Table 3.1: Mean busy period for the pre-limit queue with hyperexponential arrivals and different population sizes and the exact expression for $n=\infty$ computed using (1.1.13). The hyperexponential distribution is distributed as a rate $\lambda_1 = 2$ exponential random variable with probability $p_1 = 0.2$ and as a rate $\lambda_1 = 3/4$ exponential random variable with probability $p_2 = 0.8$. Each value for the pre-limit queue is the average of 10^4 simulations.

In Table 3.1 we show numerically that the (rescaled) average busy period of the $\Delta_{(i)}/G/1$ queue with hyperexponential arrivals converges to the exact value obtained with the explicit expression 1.1.13. The arrival random variable is exponentially distributed with rate $\lambda_1 = 2$ with probability $p_1 = 0.2$ and with rate $\lambda_2 = 3/4$ with probability $p_2 = 0.8$. The relative error is computed as $|n^{1/3}\mathbb{E}[\text{BP}_n] - \mathbb{E}[T^{\beta f_T}_{\widehat{X}_q}(0)]|/\mathbb{E}[T^{\beta f_T}_{\widehat{X}_q}(0)]$. Note that formula (1.1.13) holds for a parabolic drift of the form $-t^2/2$. However, this can be extended to more general coefficients of the parabolic term by some simple scaling properties. In particular, the first hitting time of zero of $\widehat{X}_q(\cdot)$ is distributed as

$$T^{\beta f_T}_{\widehat{X}_q}(0) \overset{\text{d}}{=} k^{-2/3} T^{\beta f_T(0)k^{-1/3}}_{\widehat{X}_{qk^{1/3}}}(0), \tag{3.1.11}$$

where $k = f_T'(0)$. Relation (3.1.11) follows from a more general scaling relation, see e.g. [45, Section 4.1].

We give two proofs of Theorem 9. In Section 3.2 we extend the techniques of Chapter 2 for the embedded process to deal with the continuous-time case. The second approach is presented in Section 3.3 and relies on the explicit expression (3.1.5).

3.2 An indirect approach to Theorem 9

In this section we show how Theorem 9 is deduced from Theorem 3 of Chapter 2 through a time change argument. We denote by $k \mapsto \bar{Q}_n(k)$ denoting the $\Delta_{(i)}/G/1$ queue embedded at service completions, we first argue that $\bar{Q}_n(\cdot)$ is closely related to Q^e_n:

Lemma 15 (Distribution of the embedded $\Delta_{(i)}/G/1$ queue). *For all $k \geq 1$,*

$$Q^e_n(k) \stackrel{\mathrm{d}}{=} \bar{Q}_n(k). \tag{3.2.1}$$

Proof. We couple the two queues as follows. The sequence of service times $(S_i)_{i=1}^n$ are taken to be the same for the two queues while the arrival clocks coincide until the end of the first busy period. After that, assign new clocks to the customers still in the population. The first customer after the idle period of $\bar{Q}_n(\cdot)$ is also the customer placed into service in $Q^e_n(\cdot)$. At the beginning of the busy period, assign new clocks to the customers still in the population. The coupling then proceeds in this manner until the population in both queues is depleted. By the memoryless property of exponential random variables, these new processes (with the clocks drawn multiple times) coincide in distribution with the original ones (with the clocks drawn at the start of the system), since

$$\begin{aligned}\mathbb{P}(T_i \geq B + I + x \mid T_i \geq B + I) &= \mathbb{P}(T_i \geq x) \\ &= \mathbb{P}(T_i \geq B + x \mid T_i \geq B),\end{aligned} \tag{3.2.2}$$

where $B = B(t)$ is the busy time process and $I = I(t)$ the idle time process at the instant t in which a new busy period starts. The number of arrivals during one service time is then the same in the two coupled queues because the arrival times are equal. In particular, the queues sampled at the end of a service time have the same distribution. □

The next step is to prove that the supremum distance between $Q^e_n(\cdot)$ and $Q_n(\cdot)$ (when suitably rescaled in space and time) converges to zero. Let $\|f(\cdot)\|_T := \sup_{t \leq T} |f(t)|$ denote the supremum norm. The claim is contained in the following proposition:

Proposition 1 (Asymptotic equivalence of the approximating model). *For each $T > 0$, as $n \to \infty$,*

$$n^{-1/3} \| Q^e_n(\cdot n^{2/3}/\mathbb{E}[S]) - Q_n(\cdot n^{-1/3}) \|_T \xrightarrow{\mathbb{P}} 0. \tag{3.2.3}$$

By Slutsky's theorem, Proposition 1 and Theorem 3 of Chapter 2 imply that

$$n^{-1/3}Q_n(\cdot n^{-1/3}) \xrightarrow{\text{d}} \phi(\widehat{X})(\cdot/\mathbb{E}[S]) = \phi(\widehat{X})(\lambda\cdot). \tag{3.2.4}$$

Without loss of generality, we will assume from now on that $\mathbb{E}[S] = 1/\lambda = 1$. To prove (3.2.3) we split

$$\begin{aligned} &\|Q^e_n(\cdot/\mathbb{E}[S]) - Q_n(\cdot)\|_T \\ &\qquad \leq \|Q^e_n(\cdot/\mathbb{E}[S]) - Q^e_n(\varphi_n(\cdot))\|_T + \|Q^e_n(\varphi_n(\cdot)) - Q_n(\cdot)\|_T, \end{aligned} \tag{3.2.5}$$

for an appropriate yet still unspecified time change $\varphi_n(\cdot)$. Thus, we are left to prove that $n^{-1/3}\|Q^e_n(\cdot n^{2/3}/\mathbb{E}[S]) - Q^e_n(\varphi_n(\cdot)n^{2/3})\|_T \xrightarrow{\mathbb{P}} 0$ and $n^{-1/3}\|Q^e_n(\varphi_n(\cdot)n^{2/3}) - Q_n(\cdot n^{-1/3})\|_T \xrightarrow{\mathbb{P}} 0$. The idea behind introducing $\varphi_n(\cdot)$ is to rescale time so that each time-step (corresponding to one service) is replaced by the actual length of the service time. In this way, the interval $[0,1]$ is replaced by $[0, D_1]$, the interval $[1,2]$ is replaced by $[D_1, D_1 + D_2]$ (if a customer has arrived during the first service), and so on. The time change $\varphi_n(\cdot)$ must also take into account idle times. The precise expression of $\varphi_n(\cdot)$ is given in Section 3.2.2 below. In the following lemma we prove that the time change $\varphi_n(\cdot)$ is, in the limit, a constant times the identity function:

Lemma 16. *As $n \to \infty$,*

$$\sup_{t\leq T} |t/\mathbb{E}[S] - \varphi_n(t)| \xrightarrow{\mathbb{P}} 0. \tag{3.2.6}$$

Consequently,

$$n^{-1/3}\|Q^e_n(\cdot n^{2/3}/\mathbb{E}[S]) - Q^e_n(\varphi_n(\cdot)n^{2/3})\|_T \xrightarrow{\mathbb{P}} 0. \tag{3.2.7}$$

The proof of (3.2.6) crucially relies on the fact that, under our heavy-traffic assumption, the idle time process of the $\Delta_{(i)}/G/1$ queue is negligible in the limit. We postpone the proof of this and of Lemma 16 to Section 3.2.2.

After this time change the two queues $Q^e_n(\varphi_n(\cdot))$ and $Q_n(\cdot)$ are synchronized in time. It still remains to be proven that their supremum distance converges to zero. However, in their coupling $Q^e_n(\varphi_n(\cdot))$ is constructed by sampling $Q_n(\cdot)$ at service completions, so that the two coincide at the time of each service completion. In other words, the maximum distance between $Q^e_n(\varphi_n(\cdot))$ and $Q_n(\cdot)$ is the maximum number of arrivals during a single service time until time T, that is

$$\|Q_n^e(\varphi_n(\cdot)n^{2/3}) - Q_n(\cdot n^{-1/3})\|_T = \max_{k\leq Tn^{2/3}} A_n(k), \tag{3.2.8}$$

where $A_n(k)$ is the number of arrivals during the k-th service time, as defined in (2.2.1). In the following lemma we will prove that the quantity on the right of (3.2.8) negligible, thus concluding the proof of Proposition 1.

Lemma 17. *As $n \to \infty$,*

$$n^{-1/3} \max_{k\leq Tn^{2/3}} A_n(k) \xrightarrow{\mathbb{P}} 0. \tag{3.2.9}$$

Consequently,

$$n^{-1/3}\|Q_n^e(\varphi_n(\cdot)n^{2/3}) - Q_n(\cdot n^{-1/3})\|_T \xrightarrow{\mathbb{P}} 0. \tag{3.2.10}$$

The following section is dedicated to discussing the idle times in the $\Delta_{(i)}/G/1$ queue, and how they relate to the $Q_n^e(\cdot)$ process. Next, we will prove Lemmas 16 and 17.

3.2.1 Idle times

Whenever the queue $Q_n^e(\cdot)$ is empty at the end of a service (say, the k-th service), the customer with the smallest arrival time is drawn from the pool and placed into service. Denote this customer by $c(k)$. Since the minimum of n rate one exponential random variables is again an exponential random variable with rate n, conditioned on $|\nu_k|$, $I_k := T_{c(k)}$ is distributed as

$$I_k \overset{\mathrm{d}}{=} \frac{\mathrm{Exp}(1)}{n - |\nu_k|}, \tag{3.2.11}$$

where $\mathrm{Exp}(1)$ is an exponential random variable with rate one. The random variable I_k represents *the time the server would have idled if customer $c(k)$ was not immediately placed into service.* Alternatively, I_k is the idle period after the k-th service in the coupled $\Delta_{(i)}/G/1$ queue. Thus we name I_k a *virtual* idle time. In particular $|\nu_k| = O(n^{2/3})$ for $k = O(n^{2/3})$, and therefore its contribution in (3.2.11) is negligible and we can think of the virtual idle periods up to time $k = O(n^{2/3})$ as being independent exponential random variables with rate n.

Let $\beta_n(k)$ be the number of customer who have been taken from the population for immediate service before the completion of the $(k+1)$-th service. Equivalently, $\beta_n(k)$ is the number of idle periods in the coupled $\Delta_{(i)}/G/1$ queue before the completion of the $(k+1)$-th service. The *cumulative* virtual idle time up to the k-th service completion takes the form

$$\mathcal{I}(k) = \sum_{i=1}^{\beta_n(k)} I_i. \tag{3.2.12}$$

For exponential arrivals, an explicit expression for the virtual idle periods is available in (3.2.11), thus we can estimate the average cumulative virtual idle time at step $k = O(n^{2/3})$ to be

$$\sum_{i=1}^{\beta_n(k)} I_i \approx \frac{\beta_n(k)}{n}. \tag{3.2.13}$$

We now aim at making (3.2.13) rigorous by proving that $\mathcal{I}(k)$ is asymptotically negligible, uniformly in $k = O(n^{2/3})$. First we show that $\beta_n(k) = O_{\mathbb{P}}(n^{1/3})$ and to this end we prove the following representation:

Lemma 18. *For every $k = 1, 2, \dots$*

$$\beta_n(k) = -\inf_{j\leq k}(N_n(j) \wedge 0) \tag{3.2.14}$$

almost surely.

Proof. Equation (3.2.14) holds for $k = 0$ because in this case both $\beta_n(0) = 0$ and $N_n(0) = 0$. Assume (3.2.14) holds for $k \geq 1$. Without loss of generality we can also assume that

$$\beta_n(k) = -\inf_{j\leq k}(N_n(j) \wedge 0) = -N_n(k). \tag{3.2.15}$$

Let $\bar{k}$ be the minimum index such that $\bar{k} > k$, $N_n(\bar{k}-1) = N_n(k)$ and $A_n(\bar{k}) = 0$. Equivalently, at the end of the $\bar{k}$-th service time there are no customers in the queue (and it is the first time after the k-th service that this happens). By the definition of $\beta_n(\cdot)$, we have that $\beta_n(\bar{k}-1) = \beta_n(k)$ and $\beta_n(\bar{k}) = \beta_n(k)+1$. On the other hand, we have that $N_n(\bar{k}) = N_n(\bar{k}-1) + A_n(\bar{k}) - 1 = N_n(k) - 1$. This gives

$$\begin{aligned}\beta_n(\bar{k}) &= \beta_n(k) + 1\\ &= -N_n(k) + 1 = -N_n(\bar{k}) = -\inf_{j\leq \bar{k}}(N_n(j) \wedge 0).\end{aligned} \tag{3.2.16}$$

Moreover, for every q such that $k < q < \bar{k}$,

$$-\inf_{j\le q}(N_n(j)\wedge 0) = -N_n(k), \tag{3.2.17}$$

and $\beta_n(q) = \beta_n(k)$, by definition of $\bar{k}$. □

Next we show that $n^{-1/3}\beta_n(tn^{2/3})$ converges in distribution to a non-trivial random variable, hence (3.2.13) is negligible in the limit when $k = O(n^{2/3})$. Recall that $\beta_n(k)$ denotes the number of customers that have been removed from the population and directly put into service before the end of the $(k+1)$-st service.

Lemma 19 (Convergence of the number of idle periods). *Fix $t \in (0,\infty)$. As $n \to \infty$,*

$$n^{-1/3}\beta_n(tn^{2/3}) \xrightarrow{\mathrm{d}} -\inf_{s\le t}(\widehat{X}(s)\wedge 0), \tag{3.2.18}$$

where $\widehat{X}(t) = \beta t - 1/2t^2 + \sigma W(t)$.

Proof. The operator $\psi : f \mapsto \psi(f)(t) = -\inf_{s\le t}(f(s)\wedge 0)$ acting from $\mathcal{D}$ to itself is Lipschitz continuous with respect to the Skorokhod J_1 topology by [95, Theorem 6.1]. Note that

$$n^{-1/3}\beta_n(tn^{2/3}) = \psi(n^{-1/3}N_n(\cdot n^{2/3}))(t). \tag{3.2.19}$$

Then, since $n^{-1/3}N_n(\cdot n^{2/3}) \xrightarrow{\mathrm{d}} W$, the Continuous-Mapping Theorem gives

$$\psi(n^{-1/3}N_n(\cdot\, n^{2/3})) \xrightarrow{\mathrm{d}} \psi(W), \tag{3.2.20}$$

and this is (3.2.18). □

A consequence of Lemma 19, is that the cumulative virtual idle time is asymptotically negligible, as shown in the following lemma:

Lemma 20 (Convergence of the cumulative idle time). *Conditioned on $\{Q_n^e(s), s \in (0, tn^{2/3})\}$, as $n \to \infty$,*

$$\frac{n}{\beta_n(tn^{2/3})}\sum_{i=1}^{\beta_n(tn^{2/3})} I_i \xrightarrow{\mathbb{P}} 1. \tag{3.2.21}$$

Proof. As was noted in (3.2.11), I_i is distributed as an exponential random variable with rate $n - |\nu_{k_i}|$, where k_i is the time step corresponding to the i-th customer being placed directly into service. We rewrite the sum as

$$\begin{aligned}
&\frac{n}{\beta_n(tn^{2/3})} \sum_{i=1}^{\beta_n(tn^{2/3})} \frac{E_i}{n - |\nu_{k_i}|} \\
&\quad = \frac{1}{\beta_n(tn^{2/3})} \sum_{i=1}^{\beta_n(tn^{2/3})} \frac{E_i}{1 - \frac{|\nu_{k_i}|}{n}} \\
&\quad = \frac{1}{\beta_n(tn^{2/3})} \sum_{i=1}^{\beta_n(tn^{2/3})} E_i + \frac{1}{\beta_n(tn^{2/3})} \sum_{i=1}^{\beta_n(tn^{2/3})} E_i \frac{|\nu_{k_i}|}{n} + \varepsilon_n, \qquad (3.2.22)
\end{aligned}$$

where $\varepsilon_n = o_{\mathbb{P}}(1)$ and $(E_i)_{n=1}^{\infty}$ are i.i.d. exponential random variables with rate 1.

By Lemma 19, $\beta_n(tn^{2/3}) \geq cn^{\alpha}$ w.h.p. for a fixed $c > 0$ and $\alpha \in (0, 1/3)$. By the LLN, the first term in (3.2.22) converges in probability to 1, and by Lemma 6 the second term, and consequently the error term, converges to zero. □

Lemma 20 intuitively says that the total virtual idle time up to time $tn^{2/3}$ is of the same order of magnitude of the number of virtual idle periods up to time $tn^{2/3}$ times the average interarrival time. In particular, as we prove below, $\mathcal{I}(tn^{2/3}) = o_{\mathbb{P}}(1)$, that is, the cumulative virtual idle time up to times of the order $n^{2/3}$ is negligible:

Corollary 2 (Cumulative idle time is negligible)**.** *Fix $T > 0$. Then,*

$$\sup_{t \leq T} \mathcal{I}(tn^{2/3}) \xrightarrow{\mathbb{P}} 0. \qquad (3.2.23)$$

Proof. By monotonicity, $\sup_{t \leq T} \mathcal{I}(tn^{2/3}) = \mathcal{I}(T)$. Fix $\varepsilon > 0$. Then, by Lemma 20,

$$\mathcal{I}(Tn^{2/3}) \leq (1 + \varepsilon)\beta_n(tn^{2/3})/n \qquad (3.2.24)$$

with high probability. By Lemma 19, $\beta_n(tn^{2/3})/n$ converges in probability to zero, so that, as $n \to \infty$,

$$\mathcal{I}(Tn^{2/3}) \xrightarrow{\mathbb{P}} 0, \qquad (3.2.25)$$

concluding the proof. □

3.2.2 Proof of Theorem 9

We begin by constructing the time change $\varphi_n(\cdot)$. Fix a realization of the arrival and service times $(S_i)_{i=1}^n = (S_i(\omega))_{i=1}^n$ and $(T_i)_{i=1}^n = (T_i(\omega))_{i=1}^n$. For simplicity we assume $\mathbb{E}[S] = 1$ and $\beta = 0$, the generalization to a different choice of parameters being straightforward. We define $\varphi_n(t)$ piece-wise, depending whether a customer is in service at time t or the server is idling at time t.

Because the time scaling in $Q_n(\cdot n^{-1/3})$ is $n^{-1/3}$, for the two processes $Q_n(\cdot n^{-1/3})$ and $Q_n^e(\varphi_n(\cdot)n^{2/3})$ to be on a comparable time scale, the time change $\varphi_n(\cdot)$ must be such that

$$\varphi_n(\cdot) = n^{-2/3}\phi_n(\cdot n^{-1/3}), \tag{3.2.26}$$

for a suitable $\phi_n(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}^+$. We now provide a precise expression of $\phi_n(\cdot)$. We remark that many other choices for $\phi_n(\cdot)$ would work for proving Lemma 16 and Lemma 17, and the one we present here is only the simplest one.

To increase readability we define

$$\vartheta_{k_2}(k_1) := \sum_{i=1}^{k_1} D_i + \mathcal{I}(k_2) = \frac{\sum_{i=1}^{k_1} S_i}{n} + \mathcal{I}(k_2). \tag{3.2.27}$$

First assume that at time t a service (say, the k-th service) is taking place, then

$$\phi_n(t) = (k-1) + \frac{1}{S_k/n}(t - \nu_k(k-1)) \qquad \text{for } t \in [\vartheta_k(k-1), \vartheta_k(k)]. \tag{3.2.28}$$

Note that, since the queue is serving at time t, $\mathcal{I}(k-1) = \mathcal{I}(k)$. In other words $\phi_n(t)$ is the line joining the points $(\vartheta_k(k-1), k-1)$ and $(\vartheta_k(k), k)$, where the term $\mathcal{I}(k)$ in $\vartheta_k(k)$ takes into account previous idle periods which might have occurred before t. Assume now that an idle period (say, the $(\beta(k)+1)$-th idle period) is under way at time t. Then $\phi_n(t)$ takes the form

$$\phi_n(t) = k \qquad \text{for } t \in [\vartheta_k(k), \vartheta_k(k) + I_{\beta(k)+1}]. \tag{3.2.29}$$

In other words, $\phi_n(t)$ is constant during an idle period. This is because $\phi_n(t)$ represents the *number of completed services*. Again, here $\sum_{i=1}^k S_i/n$ takes previous services that have occurred before t into account. Summarizing, $\phi_n(\cdot)$ takes the form

$$\phi_n(t) = \Big(k - 1 + \frac{n}{S_k}(t - \vartheta_k(k-1))\Big), \tag{3.2.30}$$

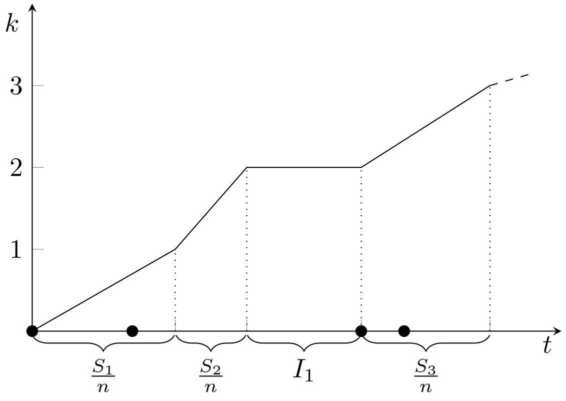

Figure 3.1: An example of the time change $\phi_n(\cdot)$. The •'s indicate arrival times of customers.

for $t \in [\vartheta_k(k-1), \vartheta_k(k)]$ for some k, and

$$\phi_n(t) = k, \tag{3.2.31}$$

for $t \in [\vartheta_k(k), \vartheta_k(k) + I_{\beta(k)+1}]$ for some k. In particular $\phi_n(\cdot)$ (and therefore also $\varphi_n(\cdot)$) is a piecewise linear continuous function. See Figure 3.1 for one possible sample path of $\phi_n(\cdot)$.

We now focus on $\varphi_n(\cdot)$ in (3.2.26) and show that it converges to the identity function. The time change $\varphi_n(t)$ takes the form

$$\varphi_n(t) = n^{-2/3}\Big(k-1+\frac{n}{S_k}\Big(\frac{t}{n^{1/3}} - \vartheta_k(k-1)\Big)\Big), \tag{3.2.32}$$

for $t/n^{1/3} \in [\vartheta_k(k-1), \vartheta_k(k)]$ for some k, and

$$\varphi_n(t) = n^{-2/3}k. \tag{3.2.33}$$

for $t/n^{1/3} \in [\vartheta_k(k), \vartheta_k(k) + I_{\beta(k)+1}]$ for some k. Note that the only values of k for which $\phi_n(t)$ has a meaningful limit as $n \to \infty$ are $k = O(n^{2/3})$. This observation is consistent with the fact that the original scaling for $Q_n^e(\cdot)$ is $Q_n^e(\cdot n^{2/3})$, that is, in order to obtain a meaningful limit we observe the queue length at times when $O(n^{2/3})$ services have been completed. As a consequence, we assume that $k = sn^{2/3}$ for some $s \in \mathbb{R}^+$. Summarizing, the final form of $\varphi_n(t)$ is

$$\varphi_n(t) = s - \frac{1}{n^{2/3}} + \frac{1}{S_{sn^{2/3}}}(t - \vartheta_{sn^{2/3}}(sn^{2/3}-1)), \tag{3.2.34}$$

for $t \in n^{1/3}[\vartheta_{sn^{2/3}}(sn^{2/3}-1), \vartheta_{sn^{2/3}}(sn^{2/3})]$, and

$$\varphi_n(t) = s, \tag{3.2.35}$$

for $t \in n^{1/3}[\vartheta_{sn^{2/3}}(sn^{\frac{2}{3}}), \vartheta_{sn^{2/3}}(sn^{2/3}) + I_{\beta(sn^{2/3})}]$. We now turn to proving Lemma 16.

Proof of Lemma 16. First note that

$$\sup_{t\le T}\Big| -\frac{1}{n^{2/3}} + \frac{1}{S_{sn^{2/3}}}(t - \vartheta_{sn^{2/3}}(sn^{2/3}-1))\Big| \le \frac{2}{n^{2/3}} \to 0, \qquad n\to\infty, \tag{3.2.36}$$

implying that we can treat $\varphi_n(t)$ as piece-wise constant, $\varphi_n(t) \equiv s$ on intervals of the form $[l(s), u(s)]$. We now prove (3.2.6). Since the function $t \mapsto t - \varphi_n(t) = t - s$ (for some fixed s) defined on an interval $[l, u] = [l(s), u(s)]$ is linear in t, it obtains its maximum either in l or u. This implies

$$\begin{aligned}&\lim_{n\to\infty} \|t - \varphi_n(t)\|_T \\ &\le \lim_{n\to\infty}(\sup_{s\le S}|\vartheta_{sn^{2/3}}(sn^{2/3}-1) - s| \vee \sup_{s\le S} \quad |\vartheta_{sn^{2/3}}(sn^{2/3}) + I_{\beta(sn^{2/3})} - s| \\ &\quad \vee \sup_{s\le S}|\vartheta_{sn^{2/3}}(sn^{2/3}) - s|),\end{aligned} \tag{3.2.37}$$

where we recall that $x \vee y := \max\{x, y\}$ and $S = S(T) = T(1+\varepsilon)/\mathbb{E}[S] > 0$ for some $\varepsilon > 0$. The inequality is implied by the fact that the three suprema are taken over a larger set, by the definition of $S(T)$. We prove convergence to zero of one of the three terms on the right in (3.2.37), the others being analogous. The triangle inequality to the last term yields

$$\sup_{s\le S}\Big|\frac{\sum_{i=1}^{sn^{2/3}} S_i}{n^{2/3}} + n^{1/3}\mathcal{I}(sn^{2/3}) - s\Big| \le \sup_{s\le S}\Big|\frac{\sum_{i=1}^{sn^{2/3}} S_i}{n^{2/3}} - s\Big| + \sup_{s\le S}|n^{1/3}\mathcal{I}(sn^{2/3})|. \tag{3.2.38}$$

The first term converges to zero in probability by the FLLN. The second term converges to zero in probability by Corollary 2. Indeed, the proof of Lemma 20 and Corollary 2 show that $\mathcal{I}(sn^{2/3}) = O_{\mathbb{P}}(n^{-2/3})$. This concludes the proof of (3.2.6).

We now turn to proving (3.2.7). Note that $\varphi_n(\cdot) \xrightarrow{\mathbb{P}} \mathrm{id}$, the identity map on $[0, T]$. In particular, id is deterministic, so that by [96, Theorem 11.4.5],

$$(n^{-1/3} Q_n^e(\cdot n^{2/3}), \varphi_n) \xrightarrow{\mathrm{d}} (\phi(\widehat{X})(\cdot)\mathrm{id}(\cdot)). \tag{3.2.39}$$

By Skorokhod's representation theorem and (3.2.39) there exist $\overline{Q_n^e}(\cdot)$, $\overline{\varphi_n}(\cdot)$ and $\overline{\phi}(\widehat{X})(\cdot)$ defined on the same probability space $\overline{\Omega}$ such that

$$(\overline{Q_n^e}(\cdot n^{2/3}), \overline{\varphi_n}) \overset{\mathrm{d}}{=} (Q_n^e(\cdot n^{2/3}), \varphi_n) \tag{3.2.40}$$

and

$$(n^{-1/3}\overline{Q_n^e}(\cdot n^{2/3}), \overline{\varphi_n}) \xrightarrow{\text{a.s.}} (\overline{\phi}(\widehat{X})(\cdot), \mathrm{id}(\cdot)). \tag{3.2.41}$$

We now dominate (3.2.7) using the random variables provided by the representation theorem, as follows:

$$\begin{aligned} &n^{-1/3}\|\overline{Q_n^e}(\cdot n^{2/3}) - \overline{Q_n^e}(\varphi_n(\cdot)n^{2/3})\|_T \\ &\quad\le n^{-1/3}(\|\overline{Q_n^e}(\cdot n^{2/3}) - \overline{\phi}(\widehat{X})(\cdot)\|_T + \|\overline{\phi}(\widehat{X})(\cdot) - \overline{Q_n^e}(\varphi_n(\cdot)n^{2/3})\|_T). \end{aligned} \tag{3.2.42}$$

Since the limiting process $\overline{\phi}(\widehat{X})(\cdot)$ is almost surely continuous, by standard arguments both the convergence $n^{-1/3}\overline{Q_n^e}(\cdot n^{2/3}) \xrightarrow{\text{a.s.}} \overline{\phi}(\widehat{X})(\cdot)$ and $n^{-1/3}\overline{Q_n^e}(\varphi_n(\cdot)n^{2/3}) \xrightarrow{\text{a.s.}} \overline{\phi}(\widehat{X})(\cdot)$ hold with respect to the uniform topology. In particular, both terms on the right in (3.2.42) converge to zero in probability. Moreover, since $(\overline{Q}(\cdot n^{2/3}), \overline{\varphi_n}(\cdot)) \overset{\mathrm{d}}{=} (Q_n^e(\cdot n^{2/3}), \varphi_n(\cdot))$,

$$\begin{aligned} &n^{-1/3}\|\overline{Q_n^e}(\cdot n^{2/3}) - \overline{Q_n^e}(\varphi_n(\cdot)n^{2/3})\|_T \\ &\qquad\overset{\mathrm{d}}{=} n^{-1/3}\|Q_n^e(\cdot n^{2/3}) - Q_n^e(\varphi_n(\cdot)n^{2/3})\|_T, \end{aligned} \tag{3.2.43}$$

so that

$$n^{-1/3}\|Q_n^e(\cdot n^{2/3}) - Q_n^e(\varphi_n(\cdot)n^{2/3})\|_T \xrightarrow{\mathbb{P}} 0, \tag{3.2.44}$$

as desired. This concludes the proof of Lemma 16. □

The fact that during one service the number of arrivals is asymptotically small is crucial in proving that $Q_n^e(\cdot)$ and $Q_n(\cdot)$ are close in the supremum norm. We prove this fact in the following lemma:

Proof of Lemma 17. Note that if $t/n^{1/3} = \sum_{i=1}^k S_i/n + \mathcal{I}(k)$ for $k \in \mathbb{N}$ (i.e. if a service is completed in t, or an idle period is undergoing in t), then

$$Q_n^e(\phi_n(t)) = Q_n(t). \tag{3.2.45}$$

This follows from the definition of the time change $\phi_n(\cdot)$, as discussed above. This implies that during any service time $Q_n^e(\phi_n(t))$ and $Q_n(t)$ differ by the number of arrivals that have occurred *during that* service time. Moreover, during any idle time $Q_n^e(\phi_n(t))$ and $Q_n(t)$ are both equal to zero. Therefore,

$$\|Q_n^e(\varphi_n(\cdot)n^{2/3}) - Q_n(\cdot n^{-1/3})\|_T = \max_{k \le Tn^{2/3}} A_n(k). \tag{3.2.46}$$

Hence (3.2.9) implies (3.2.10). Let now $\varepsilon > 0$ be arbitrary. Then

$$\begin{aligned} &\mathbb{P}(n^{-1/3} \max_{k \le Tn^{2/3}} A_n(k) \ge \varepsilon) \\ &\qquad = \mathbb{P}(n^{-1/3} \max_{k \le Tn^{2/3}} A_n(k) \mathbb{1}_{\{A_n(k) > \varepsilon n^{1/3}\}} \ge \varepsilon). \end{aligned} \tag{3.2.47}$$

In other words, only the very large values of $A_n(\cdot)$ contribute to the probability being computed. By Markov's inequality,

$$\begin{aligned} \mathbb{P}(n^{-1/3} \max_{k \le Tn^{2/3}} A_n(k) \ge \varepsilon) &\le \frac{\mathbb{E}[\max_{k \le Tn^{2/3}} A_n(k)^2 \mathbb{1}_{\{A_n(k) > \varepsilon n^{1/3}\}}]}{(n^{1/3}\varepsilon)^2} \\ &\le \sum_{k=1}^{Tn^{2/3}} \frac{\mathbb{E}[A_n(k)^2 \mathbb{1}_{\{A_n(k) > \varepsilon n^{1/3}\}}]}{n^{2/3}\varepsilon^2}. \end{aligned} \tag{3.2.48}$$

The almost sure domination $A_n(k) \le A_n'(k) = \sum_{i=1}^n \mathbb{1}_{\{T_i \le S_k/n\}}$, valid for all $k \le Tn^{2/3}$ simultaneously, gives

$$\begin{aligned} \mathbb{P}(n^{-1/3} \max_{t \le T} A_n(tn^{2/3}) \ge \varepsilon) &\le \sum_{k=1}^{Tn^{2/3}} \frac{\mathbb{E}[A_n'(k)^2 \mathbb{1}_{\{A_n'(k)(k) > \varepsilon n^{1/3}\}}]}{n^{2/3}\varepsilon^2} \\ &= T\varepsilon^{-2} \mathbb{E}[A_n'(1)^2 \mathbb{1}_{\{A_n'(1) > \varepsilon n^{1/3}\}}]. \end{aligned} \tag{3.2.49}$$

The right-most term in (3.2.49) tends to zero because $A_n'(\cdot)^2$ is stochastically dominated by a uniformly integrable random variable, as was proven in Lemma 5. □

3.3 A direct approach to Theorem 9

Both the pre-limit (3.1.5) and the limit queue-length process in (3.1.7) are easily characterized through explicit formulas. This suggests that it is possible to prove Theorem 9 by using the elementary approach to stochastic process convergence, as detailed e.g. in [17]. Assume that a sequence of processes $(\mathcal{S}_n(\cdot))_{n=1}^{\infty}$ and a candidate limit $\mathcal{S}(\cdot)$ are given. This method consists in proving separately the tightness of the family $(\mathcal{S}_n(\cdot))_{n=1}^{\infty}$, seen as measures on a certain function space, and the convergence of the finite-dimensional distributions, that is, as $n \to \infty$,

$$\mathbb{P}(S_n(t_1) \in A_1, \ldots, S_n(t_k) \in A_k) \to \mathbb{P}(S(t_1) \in A_1, \ldots, S(t_k) \in A_k), \tag{3.3.1}$$

for each $k \geq 1$ and $t_1, \ldots, t_k$. Condition (3.3.1) characterize the limit process uniquely. By exploiting this method, we prove that the queue-length process of the $\Delta_{(i)}/G/1$ queue converges in distribution to a Brownian motion with negative quadratic drift, reflected at zero. In particular, the proof we give is substantially simpler than the one in Section 3.2, requiring only the standard notions of stochastic process convergence theory [17]. This approach has two advantages. First, we impose mild assumptions on the arrival time distribution, thus generalizing [10], where the arrival times were assumed to be exponentially distributed. Second, as a consequence of our main theorem, several results relating quantities of interest other than the queue length can be deduced. As an example of this, we prove a sample path Little's Law.

The techniques of this section allow us to extend Theorem 9 to general arrival times $(T_i)_{i=1}^n$. Furthermore, the stochastic component of the limit process is defined more precisely in terms of the random fluctuations of the arrival process and of the service process. For simplicity, we will restrict to the setting $\beta = 0$, and we will prove the following:

Theorem 11 (Scaling limit of the critical $\Delta_{(i)}/G/1$ queue with general arrivals)**.** *Let $(T_i)_{i=1}^n$ be such that $f_T(0) > 0$ and let the service times $(S_i)_{i=1}^n$ be such that $\mathbb{E}[S^2] < \infty$. Assume that the heavy-traffic condition* (2.1.2) *holds with $\beta = 0$. Then*

$$n^{-1/3}Q_n(\cdot n^{-1/3}) \xrightarrow{\text{d}} \phi(\widehat{X})(\cdot), \qquad \text{in } (\mathcal{D}, J_1), \tag{3.3.2}$$

where

$$\widehat{X}(t) = W_1(f_T(0)t) - \frac{\sigma}{\mathbb{E}[S]^{3/2}} W_2(t) + \frac{f_T'(0)}{2} t^2, \tag{3.3.3}$$

and $W_1(\cdot), W_2(\cdot)$ are two independent standard Brownian motions.

We shall compare Theorem 11 and Theorem 9. The latter result shows that, when $\beta = 0$ the queue-length process converges to $\phi(\widehat{X})(t)$, where

$\widehat{X}(t) = \sigma W(t) - t^2/2$, and $\sigma^2 = \mathbb{E}[S^2]/\mathbb{E}[S]^3$. The random process consisting of the sum of two Brownian motions in (3.3.3) is distributionally equivalent to a single Brownian motion with variance equal to

$$f_T(0) + \frac{\mathbb{E}[S^2] - \mathbb{E}[S]^2}{\mathbb{E}[S]^3}. \tag{3.3.4}$$

By the heavy-traffic condition (2.1.2) this simplifies to

$$\frac{\mathbb{E}[S]^2 + \mathbb{E}[S^2] - \mathbb{E}[S]^2}{\mathbb{E}[S]^3} = \frac{\mathbb{E}[S^2]}{\mathbb{E}[S]^3}. \tag{3.3.5}$$

Therefore, the two limits are equal in distribution.

The cumulative busy time process

We now give an explicit analytical characterization of $B_n(\cdot)$. To this end, we need to introduce several auxiliary processes. The total amount of work that has entered the queue by time t (briefly, the *cumulative input*) is given by

$$C_n(t) := \sum_{i=1}^{\mathcal{A}_n(t)} S_i. \tag{3.3.6}$$

Recall that, when the server works with speed c_n, the *net-put process* $P_n(\cdot)$ of the queue is given by

$$P_n(t) := C_n(t) - c_n t = c_n \Big(\sum_{i=1}^{\mathcal{A}_n(t)} \frac{S_i}{c_n} - t \Big). \tag{3.3.7}$$

The *workload* process is then defined as

$$L_n(t) := \phi(P_n)(t) = P_n(t) - \inf_{s \le t} (P_n(s))^-. \tag{3.3.8}$$

Note that $L_n(t)$ is positive if and only if

$$C_n(t) \ge c_n t + \inf_{s \le t} (P_n(s))^- = c_n t - \psi(P_n)(t). \tag{3.3.9}$$

By construction, $\psi(P_n)(t)$ increases (linearly) if and only if the server is idling, and is constant otherwise. In other words, $I_n(t) := \psi(P_n)(t)$ has the interpretation of *cumulative idle time*. Consequently the term on the right-hand side of (3.3.9) is the *cumulative busy time* process, and we define its rescaled version as

$$B_n(t) := t - \psi\Big(\frac{P_n}{c_n}\Big)(t) = t - \frac{I_n(t)}{c_n}, \tag{3.3.10}$$

where for notational convenience we have rescaled $B_n(\cdot)$ by the server speed c_n. Note that $B_n(\cdot)$ increases only if the server is working, and is constant otherwise. With this notation, the total amount of time units the server has worked until time t is given by $c_n B_n(t)$. Then, (3.3.9) reads

$$C_n(t) \geq c_n B_n(t), \tag{3.3.11}$$

so that the workload is positive if and only if the cumulative input up to time t is larger than the total time the server has spent processing jobs, and in that case it decreases linearly in time.

The queue-length process

It is more convenient to express $Q_n(\cdot)$ as a reflection of a simpler process $X_n(\cdot)$. We will refer to $X_n(\cdot)$ as the *free process*. We rewrite (3.1.5) as

$$\begin{aligned} Q_n(t) &= (\mathcal{A}_n(t) - \sigma_n(B_n(t)) - f_T(0)I_n(t)) + f_T(0)I_n(t) \qquad (3.3.12)\\ &= \Big(\mathcal{A}_n(t) - \sigma_n(B_n(t)) + \frac{c_n B_n(t)}{\mathbb{E}[S]} - f_T(0)c_n t\Big) + f_T(0)I_n(t), \end{aligned}$$

where we have used (2.1.2) and (3.3.10) in the second equality. Recall also the definitions of $\mathcal{A}_n(\cdot)$ and $\sigma_n(\cdot)$ in (3.1.1) and (3.1.2). We define

$$X_n(t) = \mathcal{A}_n(t) - \sigma_n(B_n(t)) + \frac{c_n B_n(t)}{\mathbb{E}[S]} - f_T(0)c_n t. \tag{3.3.13}$$

For a given process $X_n(t)$, the *Skorokhod problem* associated to $X_n(t)$ consists in finding two processes $P(t)$ and $R(t)$ such that $P(t) = X_n(t) + R(t) \geq 0$, $R(t)$ is increasing, and $\int_0^\infty X_n(t)\mathrm{d}R(t) = 0$. Note that $I_n(\cdot)$ is increasing and, by definition of $Q_n(t)$ and $I_n(t)$,

$$\int_0^\infty Q_n(t)\mathrm{d}I_n(t) = 0. \tag{3.3.14}$$

Then $Q_n(t)$ and $I_n(t)$ are a solution to the Skorokhod problem associated with $X_n(t)$ and, by applying [6, Proposition 2.2, p.251] we have the representation

$$Q_n(t) = X_n + \psi(X_n)(t) = \phi(X_n)(t), \tag{3.3.15}$$

where

$$\psi(X_n)(t) = f_T(0)I_n(t) = -\Big(\frac{c_n B_n(t)}{\mathbb{E}[S]} - f_T(0)c_n t\Big). \tag{3.3.16}$$

The fluid and diffusive scaling regimes

The *fluid-scaled heavy-traffic* queue-length process is defined as

$$\bar{Q}_n(t) := \frac{Q_n(tn^{-1/3})}{n^{2/3}} = n^{1/3}\Big(\frac{\mathcal{A}_n(tn^{-1/3})}{n} - \frac{\sigma_n(B_n(tn^{-1/3}))}{n}\Big). \tag{3.3.17}$$

Correspondingly, since we assume $c_n = n$, $\bar{X}_n(\cdot)$ is defined as

$$\begin{aligned}\bar{X}_n(t) :=& \, n^{1/3}\Big(\frac{\mathcal{A}_n(tn^{-1/3})}{n} - \frac{\sigma_n(B_n(tn^{-1/3}))}{n}\Big) \\ &+ n^{1/3}\frac{B_n(tn^{-1/3})}{\mathbb{E}[S]} - f_T(0)t \\ =& \, n^{1/3}\Big(\frac{\mathcal{A}_n(n^{-1/3}t)}{n} - F_T(tn^{-1/3})\Big) \\ &- n^{1/3}\Big(\frac{\sigma_n(B_n(tn^{-1/3}))}{n} - \frac{B_n(tn^{-1/3})}{\mathbb{E}[S]}\Big) \\ &+ (n^{1/3}F_T(tn^{-1/3}) - f_T(0)t).\end{aligned} \tag{3.3.18}$$

where in the second equality we have added and subtracted $F_T(t)$ in order to rewrite $\bar{X}_n(t)$. It can be shown through an application of the functional Law of Large Numbers that, as $n \to \infty$, the fluid-scaled process $\bar{Q}_n(\cdot)$ converges to a deterministic process $\bar{Q}(\cdot)$. However, under our heavy-traffic assumption the process $\bar{Q}(\cdot)$ is identically zero. Because of this, the diffusion-scaled queue-length process can be rewritten as

$$\widehat{Q}_n(t) = n^{1/3}(\bar{Q}_n(t) - \bar{Q}(t)) = n^{1/3}\bar{Q}_n(t). \tag{3.3.19}$$

Accordingly, $\widehat{X}_n(t)$ is defined as

$$\begin{aligned}\widehat{X}_n(t) :=& \, n^{1/3}\bar{X}_n(t) \\ =& \, n^{2/3}\Big(\frac{\mathcal{A}_n(tn^{-1/3})}{n} - F_T(tn^{-1/3})\Big) \\ &- n^{2/3}\Big(\frac{\sigma_n(B_n(tn^{-1/3}))}{n} - \frac{B_n(tn^{-1/3})}{\mathbb{E}[S]}\Big) \\ &+ n^{2/3}(F_T(tn^{-1/3}) - f_T(0)tn^{-1/3}).\end{aligned} \tag{3.3.20}$$

In order to prove Theorem 11 we will rely on an analogous result for $\widehat{X}_n(\cdot)$. In fact, Theorem 11 is a straightforward consequence of the following theorem:

Theorem 12 (Scaling limit of the free process). *As $n \to \infty$,*

$$\widehat{X}_n(t) \xrightarrow{\mathrm{d}} \widehat{X}(t), \qquad \text{in } (\mathcal{D}, J_1), \tag{3.3.21}$$

where $\widehat{X}(\cdot)$ is as in (3.3.3).

The scaling exponents

Let us now give a heuristic motivation for the scaling exponents in (3.3.20). Define the general time scaling exponent as $-\alpha$ and the spatial scaling exponent as β, for some $\alpha, \beta > 0$ to be determined, so that $\widehat{X}_n$ is given by

$$\begin{aligned}\widehat{X}_n &= n^{\beta}\Big(\frac{\mathcal{A}_n(tn^{-\alpha})}{n} - F_T(tn^{-\alpha})\Big) + n^{\beta}\Big(\frac{\sigma_n(B_n(tn^{-\alpha}))}{n} - \frac{B_n(tn^{-\alpha})}{\mathbb{E}[S]}\Big)\\ &\quad + n^{\beta}(F_T(tn^{-\alpha}) - f_T(0)tn^{-\alpha}).\end{aligned} \tag{3.3.22}$$

For the deterministic drift to converge to a non-trivial limit it is necessary that α, β be such that $2\alpha = \beta$. Indeed, replacing $F_T(tn^{-\alpha})$ with its Taylor expansion up to the second term, we get

$$n^{\beta}(F_T(tn^{-\alpha}) - f_T(0)tn^{-\alpha}) = n^{\beta}\Big(\frac{f_T'(0)}{2}t^2 n^{-2\alpha} + o(n^{-2\alpha})\Big). \tag{3.3.23}$$

Moreover, a necessary condition for the first term in (3.3.22) to converge to a non-trivial random process is that, for fixed time $t > 0$, its variance is of order $O(1)$. This is given by

$$\begin{aligned}\mathrm{Var}\Big(n^{\beta}\frac{\mathcal{A}_n(tn^{-\alpha})}{n}\Big) &= \frac{n^{2\beta}}{n}\mathrm{Var}(\mathbb{1}_{\{T \le tn^{-\alpha}\}})\\ &= \frac{n^{2\beta}}{n}\mathbb{P}(T \le tn^{-\alpha})(1 - \mathbb{P}(T \le tn^{-\alpha}))\\ &= \frac{n^{2\beta}}{n}(f_T(0)tn^{-\alpha} + o(n^{-\alpha})).\end{aligned} \tag{3.3.24}$$

Then, α and β should be such that

$$\frac{n^{2\beta-\alpha}}{n} = O(1), \tag{3.3.25}$$

which, together with $\beta = 2\alpha$, imply that $\alpha = 1/3$ and $\beta = 2/3$.

3.3.1 Proof of Theorem 9

The proof of Theorem 11 proceeds in several steps. These consist in proving convergence of the three terms in (3.3.22) to the respective terms in (3.3.3) separately. The first term in (3.3.22) is the centred and rescaled empirical distribution function of the sequence $(T_i)_{i=1}^n$. Therefore, its convergence to $W_1(f_T(0)t)$ can be seen as a 'local Donsker's Theorem', in which the limiting Brownian Bridge is replaced by a Brownian motion. The second term in (3.3.22) is a time-changed, centred and rescaled renewal process and thus converges by a random time-change theorem and the FCLT for renewal processes. The third term also converges trivially to the limiting quadratic drift. Then, the convergence (3.3.2) follows immediately from (3.3.21) by the continuity of the Skorokhod reflection $\phi(x)$ in all $x \in \mathcal{C}$, the space of real-valued continuous functions, see [96, Theorem 13.5.1].

A local Donsker's Theorem

For sake of simplicity, let us define

$$\widehat{A}_n(t) := n^{2/3}\Big(\frac{\mathcal{A}_n(tn^{-1/3})}{n} - F_T(tn^{-1/3})\Big) \tag{3.3.26}$$

and

$$\widehat{A}(t) := W_1(f_T(0)t). \tag{3.3.27}$$

The goal of this section is to prove the following:

Lemma 21 (Convergence of the arrival process)**.** *As $n \to \infty$,*

$$\widehat{A}_n(\cdot) \xrightarrow{\mathrm{d}} \widehat{A}(\cdot), \qquad \text{in } (\mathcal{D}, J_1). \tag{3.3.28}$$

Proof. The proof proceeds in two steps. First, we prove convergence of the finite-dimensional distributions. This characterizes the limit uniquely. Second, we prove tightness of the family $(\widehat{A}_n)_{n=1}^\infty$, seen as elements of $\mathcal{P}(\mathcal{D})$, the space of measures on the Polish space $\mathcal{D}$ of càdlàg functions. By definition, we say that the finite-dimensional distributions of $\widehat{A}_n(\cdot)$ converge to the finite-dimensional distributions of $\widehat{A}(\cdot)$ if, for every $n \in \mathbb{N}$ and for each choice of $(t_i)_{i=1}^n$ such that $0 < t_1 < t_2 < \ldots < t_n < \infty$ it holds that, as $n \to \infty$,

$$(\widehat{A}_n(t_1), \ldots, \widehat{A}_n(t_n)) \xrightarrow{\mathrm{d}} (\widehat{A}(t_1), \ldots, \widehat{A}(t_n)). \tag{3.3.29}$$

For simplicity we shall prove (3.3.29) for $t_1 < t_2$, the generalization to an arbitrary choice of $(t_i)_{i=1}^n$ being straightforward. We then aim to show that, as $n \to \infty$,

$$(\widehat{A}_n(t_1), \widehat{A}_n(t_2)) \xrightarrow{\mathrm{d}} (\widehat{A}(t_1), \widehat{A}(t_2)). \tag{3.3.30}$$

Let $\mathcal{N}(m, v)$ denote a normally distributed random variable with mean m and covariance matrix v. Then $(\widehat{A}(t_1), \widehat{A}(t_2)) \sim \mathcal{N}(m, V_{t_1,t_2})$, with mean $m = (0,0)$ and covariance matrix V_{t_1,t_2} given by

$$V_{t_1,t_2} = f_T(0) \begin{pmatrix} t_1 & t_1 \wedge t_2 \\ t_1 \wedge t_2 & t_2 \end{pmatrix}, \tag{3.3.31}$$

where $a \wedge b = \min\{a, b\}$. To show joint convergence, we apply the Cramér-Wold device. Given an arbitrary vector $\gamma = (\gamma_1, \gamma_2) \in \mathbb{R}^2$, we aim to show that, as $n \to \infty$,

$$\gamma_1 \widehat{A}_n(t_1) + \gamma_2 \widehat{A}_n(t_2) \xrightarrow{\mathrm{d}} \gamma_1 \widehat{A}(t_1) + \gamma_2 \widehat{A}(t_2). \tag{3.3.32}$$

This is done through the following straightforward generalization of the Lindeberg-Feller CLT:

Theorem 13 (Lindeberg-Feller CLT [63]). *Let $(X_{n,l})_{l=1}^n$ be an array of random variables such that $\mathbb{E}[X_{n,l}] = 0$ for all $n \geq 1$ and $l \leq n$ and $\sum_{l=1}^n \mathrm{Var}(X_{n,l}) \to 1$. Define*

$$S_n := X_{n,1} + \ldots + X_{n,n}. \tag{3.3.33}$$

Assume that the Lindeberg condition *holds, i.e. for $\varepsilon > 0$,*

$$\frac{1}{\mathrm{Var}(S_n)} \sum_{l=1}^n \mathbb{E}[X_{n,l}^2 \mathbb{1}_{\{X_{n,l}^2 > \varepsilon^2 \mathrm{Var}(S_n)\}}] \to 0, \tag{3.3.34}$$

as $n \to \infty$. Then S_n converges in distribution to a standard normal random variable.

In the usual formulation of the Lindeberg-Feller CLT it is assumed that $\sum_{l=1}^n \mathrm{Var}(X_{n,l}) = 1$. The proof of the theorem, as presented e.g. in [63] can be directly generalized to accommodate for the assumption that $\sum_{l=1}^n \mathrm{Var}(X_{n,l}) \to 1$. We now take $X_{n,l}$ to be

$$\begin{aligned} X_{n,l} = \gamma_1 &\frac{\mathbb{1}_{\{T_l \leq t_1 n^{-1/3}\}} - F_T(t_1 n^{-1/3})}{n^{1/3} v_{t_1,t_2}} \\ &+ \gamma_2 \frac{\mathbb{1}_{\{T_l \leq t_2 n^{-1/3}\}} - F_T(t_2 n^{-1/3})}{n^{1/3} v_{t_1,t_2}}, \end{aligned} \tag{3.3.35}$$

where v_{t_1,t_2} is a normalizing constant and is given by

$$v_{t_1,t_2} = \frac{1}{\sqrt{f_T(0)(\gamma_1^2 t_1 + \gamma_2^2 t_2 + 2\gamma_1\gamma_2 t_1)}}. \tag{3.3.36}$$

Recall that $t_1 < t_2$ by assumption. In order to deduce the desired convergence in (3.3.32) we are left to check the conditions of Theorem 13. Trivially, $\mathbb{E}[X_{n,l}] = 0$. We compute $\mathrm{Var}(X_{n,l})$ explicitly as follows:

$$\begin{aligned}\mathrm{Var}(X_{n,l}) &= \frac{\gamma_1^2}{n^{2/3}v_{t_1,t_2}^2}(F_T(t_1 n^{-1/3}) - F_T(t_1 n^{-1/3})^2) \\ &\quad + \frac{\gamma_2^2}{n^{2/3}v_{t_1,t_2}^2}(F_T(t_2 n^{-1/3}) - F_T(t_2 n^{-1/3})^2) \\ &\quad + \frac{2\gamma_1\gamma_2}{n^{2/3}v_{t_1,t_2}^2}(F_T(t_1 n^{-1/3}) - F_T(t_1 n^{-1/3})F_T(t_2 n^{-1/3})) \\ &= \frac{f_T(0)}{v_{t_1,t_2}^2}\Big(\frac{\gamma_1^2}{n}t_1 + \frac{\gamma_2^2}{n}t_2 + 2\frac{\gamma_1\gamma_2}{n}t_1\Big) + O(n^{-4/3}),\end{aligned} \tag{3.3.37}$$

where in the second equality we Taylor expanded the distribution function $F_T(\cdot)$. In particular,

$$\sum_{l=1}^{n} \mathrm{Var}(X_{n,l}) = 1 + O(n^{-1/3}). \tag{3.3.38}$$

The *Lindeberg condition* is also satisfied, since

$$\begin{aligned}&\sum_{l=1}^{n} \frac{1}{n^{2/3}v_{t_1,t_2}} \\ &\times \mathbb{E}[(\mathbb{1}_{\{T_i \le t_1 n^{-1/3}\}} - F_T(t_1 n^{-1/3}))^2 \mathbb{1}_{\{(\mathbb{1}_{\{T_i \le t_1 n^{-1/3}\}} - F_T(t_1 n^{-1/3})) \ge \varepsilon n^{1/3}\}}] = 0,\end{aligned} \tag{3.3.39}$$

since $\mathbb{1}_{\{(\mathbb{1}_{\{T_i \le t_1 n^{-1/3}\}} - F_T(t_1 n^{-1/3})) \ge \varepsilon n^{1/3}\}} = 0$ almost surely. The first term is of the order $O(n^{-1/3})$, while the second is identically zero for n large enough.

By Theorem 13,

$$\frac{1}{v_{t_1,t_2}}(\gamma_1, \gamma_2) \cdot (\widehat{A}_n(t_1), \widehat{A}_n(t_2)) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1), \tag{3.3.40}$$

where $\cdot$ denotes the usual scalar product. However, since

$$(\gamma_1, \gamma_2)^{\mathrm{t}} \cdot V_{t_1,t_2} \cdot (\gamma_1, \gamma_2) = v^2_{t_1,t_2}, \tag{3.3.41}$$

where q^{t} denotes the transpose of a vector q, so that

$$\mathcal{N}(0,1) \overset{\mathrm{d}}{=} \frac{1}{v_{t_1,t_2}}(\gamma_1, \gamma_2) \cdot \mathcal{N}((0,0), V_{t_1,t_2}). \tag{3.3.42}$$

This together with (3.3.40) implies (3.3.32). By an application of the Cramér-Wold device, joint convergence follows.

The last step of the proof is to show that $(\widehat{A}_n(\cdot))_{n=1}^{\infty}$ is a tight family of random variables on $\mathcal{D}$. By [17, Theorem 13.5], in particular equation (13.14), it is enough for $(\widehat{A}_n(\cdot))_{n=1}^{\infty}$ to satisfy the following condition. For every $T > 0$,

$$\mathbb{E}[|\widehat{A}_n(t) - \widehat{A}_n(t_1)|^2 |\widehat{A}_n(t_2) - \widehat{A}_n(t)|^2] \leq (f_{\mathrm{inc}}(t_2) - f_{\mathrm{inc}}(t_1))^2, \tag{3.3.43}$$

for $0 \leq t_1 \leq t \leq t_2 \leq T$ and $f_{\mathrm{inc}}(\cdot)$ is a non-decreasing function. Checking (3.3.43) amounts to computing the mean appearing on the left side of the equation. Define

$$\begin{aligned} p_1 &:= F_T(tn^{-1/3}) - F_T(t_1 n^{-1/3}), \\ p_2 &:= F_T(t_2 n^{-1/3}) - F_T(tn^{-1/3}). \end{aligned} \tag{3.3.44}$$

Define also

$$\alpha_i := \begin{cases} 1 - p_1, & \text{if } T_i n^{-1/3} \in (t_1, t], \\ -p_1, & \text{if } T_i n^{-1/3} \notin (t_1, t], \end{cases} \tag{3.3.45}$$

and

$$\beta_i := \begin{cases} 1 - p_2, & \text{if } T_i n^{-1/3} \in (t, t_2], \\ -p_2, & \text{if } T_i n^{-1/3} \notin (t, t_2], \end{cases} \tag{3.3.46}$$

where we have omitted dependencies on n to avoid cumbersome notation. Note that $\mathbb{E}[\alpha_1] = \mathbb{E}[\beta_1] = 0$. With the help of these definitions, (3.3.43) can be immediately rewritten in the following form:

$$\mathbb{E}\Big[\Big(\sum_{i=1}^{n} \alpha_i\Big)^2 \Big(\sum_{i=1}^{n} \beta_i\Big)^2\Big] \leq n^{4/3}(f_{\mathrm{inc}}(t_2) - f_{\mathrm{inc}}(t_1))^2. \tag{3.3.47}$$

We will take $f_{\mathrm{inc}}(t) = \sqrt{c}t$ for a certain constant $c > 0$. By definition α_i (resp. β_i) is independent from α_j and β_j for $j \neq i$, so that the left side of (3.3.47) can be simplified as

$$n\mathbb{E}[\alpha_1^2 \beta_1^2] + n(n-1)\mathbb{E}[\alpha_1^2]\mathbb{E}[\beta_2^2] + 2n(n-1)\mathbb{E}[\alpha_1\beta_1]\mathbb{E}[\alpha_2\beta_2]. \tag{3.3.48}$$

The first term $n\mathbb{E}[\alpha_1^2\beta_1^2]$ is of lower order, so we focus on the remaining two. A simple computation gives

$$\begin{aligned}\mathbb{E}[\alpha_1^2] &= p_1(1-p_1) \le p_1,\\ \mathbb{E}[\beta_1^2] &= p_2(1-p_2) \le p_2,\\ \mathbb{E}[\alpha_1\beta_2] &= -p_1p_2,\end{aligned} \tag{3.3.49}$$

so that, since $p_1 \le (p_1+p_2)$ and $p_2 \le (p_1+p_2)$,

$$\begin{aligned}\mathbb{E}\Big[\Big(\sum_{i=1}^n \alpha_i\Big)^2\Big(\sum_{i=1}^n \beta_i\Big)^2\Big] &\le c_0n^2p_1p_2 \le c_0n^2(p_2+p_1)^2\\ &= c_0n^2(F_T(t_2n^{-1/3}) - F_T(t_1n^{-1/3}))^2\\ &\le c_1n^{4/3}f_T(0)(t_2-t_1)^2,\end{aligned} \tag{3.3.50}$$

for a sufficiently large $c_1 > 0$. Therefore, we have verified (3.3.47) with $f_{\text{inc}}(t) = \sqrt{c_1f_T(0)}t,$. □

A functional CLT for renewal processes

We define

$$\widehat{\sigma}_n(t) := n^{2/3}\Big(\frac{\sigma_n(tn^{-1/3})}{n} - \frac{1}{\mathbb{E}[S]}tn^{-1/3}\Big) \tag{3.3.51}$$

and

$$\widehat{\sigma}(t) := \frac{\sigma}{\mathbb{E}[S]^{3/2}}W_2(t), \tag{3.3.52}$$

where $\sigma^2 = \text{Var}(S)$. In this section we prove the following lemma:

Lemma 22 (Convergence of the service process)**.** *As $n \to \infty$,*

$$\widehat{\sigma}_n(\cdot) \xrightarrow{\text{d}} \widehat{\sigma}(\cdot), \qquad \text{in } (\mathcal{D}, J_1). \tag{3.3.53}$$

Proof. Note that $\sigma_n(tn^{-1/3}) = \sigma_{n^{2/3}}(t)$. Moreover,

$$n^{2/3}\Big(\frac{\sigma_n(tn^{-1/3})}{n} - \frac{1}{\mathbb{E}[S]}tn^{-1/3}\Big) = \frac{\sigma_{n^{2/3}}(t) - \mathbb{E}[S]^{-1}tn^{2/3}}{n^{1/3}}. \tag{3.3.54}$$

Therefore, (3.3.53) is a consequence of the FCLT for renewal processes, see e.g. [17, Theorem 14.6]. □

Convergence of the cumulative busy time

In this section we exploit Lemma 22 and the random time change theorem to prove that the rescaled service process in (3.3.20) converges. First, we prove some scaling limits for the arrival process. Define the fluid-scaled arrival process as

$$\bar{A}_n(t) := \frac{\mathcal{A}_n(tn^{-1/3})}{n^{2/3}}. \tag{3.3.55}$$

The following generalized Markov inequality is useful when proving the strong Law of Large Numbers.

Lemma 23 (Generalized Markov inequality). *For any $p = 1, 2, \dots$ and any random variable X such that $\mathbb{E}[|X|^p] < \infty$,*

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}[|X|^p]}{\varepsilon^p}. \tag{3.3.56}$$

Lemma 23 together with the Borel-Cantelli lemma allow us to prove the following:

Lemma 24 (LLN for the arrival process). *For any fixed $t \geq 0$, as $n \to \infty$,*

$$|\bar{A}_n(t) - f_T(0)t| \overset{\text{a.s.}}{\to} 0. \tag{3.3.57}$$

Proof. First, we rewrite

$$\bar{A}_n(t) - f_T(0)t = \frac{1}{n}\sum_{i=1}^{n}(n^{1/3}\mathbb{1}_{\{T_i \leq tn^{-1/3}\}} - n^{1/3}F_T(tn^{-1/3})). \tag{3.3.58}$$

We define

$$Y_i := n^{1/3}\mathbb{1}_{\{T_i \leq tn^{-1/3}\}} - n^{1/3}F_T(tn^{-1/3}). \tag{3.3.59}$$

In order to apply the Borel-Cantelli lemma, we compute

$$\begin{aligned}\mathbb{P}(|\sum_{i=1}^{n} Y_i| \geq \varepsilon n) &\leq \frac{\mathbb{E}[|\sum_{i=1}^{n} Y_i|^4]}{n^4\varepsilon^4} \\ &= \frac{n\mathbb{E}[|Y_1|^4] + 3n(n-1)\mathbb{E}[|Y_1|^2]^2}{n^4\varepsilon^4}.\end{aligned} \tag{3.3.60}$$

It is immediate to see that the leading orders of the moments are

$$\begin{aligned}\mathbb{E}[|Y_1|^4] &= O(n^{4/3}\mathbb{P}(T_i \leq tn^{-1/3})) = O(tn), \\ \mathbb{E}[|Y_1|^2] &= O(n^{2/3}\mathbb{P}(T_i \leq tn^{-1/3})) = O(tn^{1/3}).\end{aligned} \tag{3.3.61}$$

We conclude that, for a large constant $c_1 > 0$,

$$\mathbb{P}(|\sum_{i=1}^{n} Y_i| \geq \varepsilon n) \leq c_1 \frac{tn^2 + 3tn^{8/3}}{n^4 \varepsilon^4}. \tag{3.3.62}$$

Define the event $\mathcal{A} := \{|\sum_{i=1}^{n} Y_i| \geq \varepsilon n \text{ for infinitely many } n\}$. Since

$$\sum_{n=1}^{\infty} \mathbb{P}(|\sum_{i=1}^{n} Y_i| \geq \varepsilon n) \leq c_1 \sum_{n=1}^{\infty} \frac{tn^2 + 3tn^{8/3}}{n^4 \varepsilon^4} \leq c_2 \sum_{n=1}^{\infty} \frac{1}{n^{4/3} \varepsilon^4} < \infty, \tag{3.3.63}$$

for some large constant $c_2 > 0$, by the Borel-Cantelli lemma,

$$\mathbb{P}(\mathcal{A}) = 0. \tag{3.3.64}$$

Since $\varepsilon > 0$ is arbitrary, this concludes the proof of (3.3.57). □

We will now extend the convergence (3.3.57) to uniform convergence over compact subsets of the positive half-line. Our result can be interpreted as a special Glivenko-Cantelli theorem. This is summarized in the following lemma.

Lemma 25 (Glivenko-Cantelli Theorem for the arrival process). *As $n \to \infty$,*

$$\bar{A}_n(t) \overset{\text{a.s.}}{\to} f_T(0)t, \qquad \text{in } (\mathcal{D}, U). \tag{3.3.65}$$

Consequently, as $n \to \infty$,

$$n^{1/3} C_n(tn^{-1/3}) \overset{\text{a.s.}}{\to} t \qquad \text{in } (\mathcal{D}, U). \tag{3.3.66}$$

Proof. Let $T > 0$ be arbitrary. The claim (3.3.65) is then equivalent to

$$\lim_{n\to\infty} \sup_{t \leq T} |\bar{A}_n(t) - f_T(0)t| = 0, \tag{3.3.67}$$

almost surely as $n \to \infty$. Let N be a large arbitrary natural number and define

$$t_j := \frac{1}{f_T(0)} \frac{j}{N} T, \qquad j = 1, \dots, N, \tag{3.3.68}$$

so that $f_T(0)t_j = jT/N$. The idea is that both $\mathcal{A}_n(t)$ and $f_T(0)t$ are increasing, so for $t \in (t_{j-1}, t_j)$ the difference of the two can be bounded by their values in t_{j-1} and t_j. Then, we have convergence because of Lemma 24 and because N is fixed. Formally, define the error as

$$\begin{aligned} E_{n,N} := \max_{j=1,\dots,N} (&|\mathcal{A}_n(t_j n^{-1/3})/n^{2/3} - f_T(0)t_j| \\ &+ |\mathcal{A}_n((t_j n^{-1/3})^-)/n^{2/3} - f_T(0)t_j^-|). \end{aligned} \tag{3.3.69}$$

where $f(t^-) := \lim_{s \nearrow t} f(s)$. For $t \in (t_{j-1}, t_j)$ we upper bound $\bar{A}_n(t)$ as follows

$$\bar{A}_n(t) \leq \bar{A}_n(t_j^-) \leq f_T(0)t_j^- + E_{n,N} \leq f_T(0)t + E_{n,N} + \frac{T}{N}, \tag{3.3.70}$$

where in the last inequality we have used that $|f_T(0)t_j - f_T(0)t_{j-1}| \leq T/N$. Analogously, for the lower bound

$$\bar{A}_n(t) \geq \bar{A}_n(t_{j-1}) \geq f_T(0)t_{j-1} - E_{n,N} \geq f_T(0)t - E_{n,N} - \frac{T}{N}. \tag{3.3.71}$$

Summarizing the two bounds, since $E_{n,N}$ and T/N do not depend on t,

$$\sup_{t \leq T} |\bar{A}_n(t) - f_T(0)t| \leq E_{n,N} + \frac{T}{N}. \tag{3.3.72}$$

Since N is fixed, almost surely

$$\lim_{n \to \infty} E_{n,N} = 0, \tag{3.3.73}$$

by Lemma 24. Letting $N \to \infty$, we obtain (3.3.65).

The convergence (3.3.66) follows from (3.3.65). Indeed, by the functional strong Law of Large Numbers [24, Theorem 5.10]

$$\sum_{i=1}^{tn^{2/3}} \frac{S_i}{n^{2/3}} \overset{\text{a.s.}}{\to} \mathbb{E}[S]t \qquad \text{in } (\mathcal{D}, U). \tag{3.3.74}$$

Since $\bar{A}_n(t)$ converges to a deterministic limit, we also have the joint convergence

$$\Big(\sum_{i=1}^{tn^{2/3}} \frac{S_i}{n^{2/3}}, \bar{A}_n(t) \Big) \overset{\text{a.s.}}{\to} (\mathbb{E}[S]t, f_T(0)t), \qquad \text{in } (\mathcal{D}^2, WJ_1). \tag{3.3.75}$$

Recall that WJ_1 denotes the product J_1 topology on $\mathcal{D} \times \mathcal{D} \times \cdots \times \mathcal{D} = \mathcal{D}^k$. Note that $\mathcal{A}_n(\cdot)$ is non-decreasing. Then, by a time-change theorem [17, Lemma p.151],

$$\sum_{i=1}^{\bar{A}_n(t)n^{2/3}} \frac{S_i}{n^{2/3}} \overset{\text{a.s.}}{\to} \mathbb{E}[S]f_T(0)t \qquad \text{in } (\mathcal{D}, U). \tag{3.3.76}$$

Recall that convergence in $(\mathcal{D}, J_1)$ to a continuous function implies convergence in $(\mathcal{D}, U)$. Moreover, $\mathbb{E}[S]f_T(0) = 1$ by the heavy-traffic condition (2.1.2), and this concludes the proof of (3.3.66). □

Since $t \mapsto f_T(0)t$ is not a proper distribution function, Theorem 25 is a *local* version of the usual Glivenko-Cantelli Theorem. Let us now define the fluid-scaled cumulative busy time process as

$$\bar{B}_n(t) := n^{1/3} B_n(tn^{-1/3}). \tag{3.3.77}$$

We are able to prove the following lemma:

Lemma 26 (Convergence of the time-changed service process). *As $n \to \infty$,*

$$\bar{B}_n(\cdot) \overset{\text{a.s.}}{\to} \text{id}(\cdot), \qquad \text{in } (\mathcal{D}, U), \tag{3.3.78}$$

Proof. $B_n(t)$ can be rewritten as

$$B_n(t) = t + \psi(P_n)(t) = t + \inf_{s \le t}(C_n(s) - s)^-. \tag{3.3.79}$$

By Lemma 25, $n^{1/3}(C_n(tn^{-1/3}) - tn^{-1/3}) \overset{\text{a.s.}}{\to} 0$ in $(\mathcal{D}, U)$. Moreover, the null function is a continuity point of $\psi(\cdot)$ with probability one [96, Lemma 13.4.1]. The claim then follows from the Continuous-Mapping Theorem [96, Theorem 3.4.3]. □

Proof of Theorem 11

Since $\bar{B}_n(\cdot)$ converges to a deterministic limit,

$$(\widehat{A}_n(\cdot), \widehat{\sigma}_n(\cdot), \bar{B}_n(\cdot)) \overset{\text{d}}{\to} (\widehat{A}(\cdot), \widehat{\sigma}(\cdot), \text{id}(\cdot)), \qquad \text{in } (\mathcal{D}^3, WJ_1). \tag{3.3.80}$$

Note also that $\widehat{A}_n(\cdot)$ and $\widehat{\sigma}_n(\cdot)$ are independent processes, so that $\widehat{A}(\cdot)$ and $\widehat{\sigma}(\cdot)$ are also independent. Applying the random time-change theorem [17, Lemma p.151], we get

$$(\widehat{A}_n(\cdot), \widehat{\sigma}_n(\bar{B}_n(\cdot))) \overset{\text{d}}{\to} (\widehat{A}(\cdot), \widehat{\sigma}(\cdot)), \qquad \text{in } (\mathcal{D}^2, WJ_1). \tag{3.3.81}$$

Since the limit points are continuous, by [95, Theorem 4.1] addition is also continuous, so that, in $(\mathcal{D}, J_1)$,

$$\widehat{A}_n(\cdot) - \widehat{\sigma}_n(\cdot) + n^{2/3}(F_T(\cdot n^{-1/3}) - f_T(0)\text{id}(\cdot n^{-1/3})) \overset{\text{d}}{\to} \widehat{X}(\cdot), \tag{3.3.82}$$

where

$$\widehat{X}(t) = \widehat{A}(t) - \widehat{\sigma}(t) - \frac{f_T'(0)}{2} t^2, \tag{3.3.83}$$

concluding the proof of (3.3.21). By [96, Theorem 13.5.1], the reflection map $\phi(\cdot)$ is continuous when $\mathcal{D}$ is endowed with the J_1 topology, from which (3.3.2) follows. □

3.3.2 Sample path Little's Law

In this section we apply the ideas and results from the previous sections to derive a 'sample path Little's Law' for the $\Delta_{(i)}/G/1$ queue. The standard formulation of Little's Law relates the expected waiting time $\mathbb{E}[W]$, to the expected queue length $\mathbb{E}[L_q]$ as $\mathbb{E}[L_q] = \lambda\mathbb{E}[W]$, where λ is the rate at which customers arrive in the system. We will work instead with the *virtual* waiting time $W_n(t)$, defined as

$$W_n(t) := C_n(t) - B_n(t). \tag{3.3.84}$$

Accordingly, we define the diffusion-scaled virtual waiting time as

$$\begin{aligned}\hat{W}_n(t) &:= n^{2/3}(C_n(tn^{-1/3}) - B_n(tn^{-1/3}))\\ &= n^{1/3}\Big(\sum_{i=1}^{\mathcal{A}_n(tn^{-1/3})}\frac{S_i}{n^{2/3}} - \bar{B}_n(t)\Big).\end{aligned} \tag{3.3.85}$$

First, we rewrite the expression for $\hat{W}_n(t)$ as

$$\begin{aligned}\hat{W}_n(t) = n^{1/3}&\Big(\sum_{i=1}^{\bar{A}_n(t)n^{2/3}}\frac{S_i}{n^{2/3}} - \mathbb{E}[S]\bar{A}_n(t)\Big)\\ &+ n^{1/3}\mathbb{E}[S]\Big(\bar{A}_n(t) - n^{1/3}F_T(tn^{-1/3})\Big)\\ &+ n^{1/3}\mathbb{E}[S]\Big(F_T(tn^{-1/3}) - f_T(0)t\Big)\\ &+ n^{1/3}\mathbb{E}[S]\Big(f_T(0)t - \bar{B}_n(t)/\mathbb{E}[S]\Big).\end{aligned} \tag{3.3.86}$$

By (3.3.16), $n^{1/3}(f_T(0)t - \bar{B}_n(t)/\mathbb{E}[S]) = \psi(\widehat{X}_n)(t)$, so that (3.3.86) can be further simplified as

$$\begin{aligned}\hat{W}_n(t) &= \mathbb{E}[S]\widehat{Q}_n(t)\\ &+ n^{1/3}\Big(\sum_{i=1}^{\bar{A}_n(t)n^{2/3}}\frac{S_i}{n^{2/3}} - \mathbb{E}[S]\bar{A}_n(t)\Big) + \mathbb{E}[S]\widehat{\sigma}_n(\bar{B}_n(t)).\end{aligned} \tag{3.3.87}$$

We now focus on the second and third terms in (3.3.87). Let us ignore the time change $t \mapsto \bar{A}_n(t)$ and $t \mapsto \bar{B}_n(t)$ for the moment. Then, the second line in (3.3.87) is the difference between the diffusion-scaled partial sums and the diffusion-scaled counting process associated with the sequence of random variables $(S_i)_{n=1}^{\infty}$. These converge to the same limiting Brownian motion, so that their contribution to $\hat{W}_n(t)$ vanishes in the limit. We now aim make this reasoning rigorous:

Theorem 14 (Diffusion sample path Little's Law)**.** *As $n \to \infty$,*

$$\hat{W}_n(\cdot) \xrightarrow{\mathrm{d}} \hat{W}(\cdot), \qquad \text{in } (\mathcal{D}, J_1), \tag{3.3.88}$$

where

$$\hat{W}(t) := \mathbb{E}[S]\widehat{Q}(t). \tag{3.3.89}$$

Proof. Define the diffusion-scaled partial sum process as

$$\hat{\mathcal{P}}_n(t) = n^{1/3}\Big(\sum_{i=1}^{tn^{2/3}} \frac{S_i}{n^{2/3}} - \mathbb{E}[S]t\Big). \tag{3.3.90}$$

By [96, Theorem 7.3.2], $\hat{\mathcal{P}}_n(\cdot)$ and $\widehat{\sigma}_n(\cdot)$ jointly converge as

$$(\hat{\mathcal{P}}_n(\cdot), \widehat{\sigma}_n(\cdot)) \xrightarrow{\mathrm{d}} (-\mathbb{E}[S]\widehat{\sigma}(\mathbb{E}[S]\cdot), \widehat{\sigma}(\cdot)), \qquad \text{in } (\mathcal{D}^2, WJ_1), \tag{3.3.91}$$

where $\widehat{\sigma}_n(\cdot)$ and $\widehat{\sigma}(\cdot)$ are the same as in (3.3.53). Since $\widehat{A}_n(\cdot)$ is independent from $\hat{\mathcal{P}}_n(\cdot)$ and $\widehat{\sigma}_n(\cdot)$, in $(\mathcal{D}^3, WJ_1)$,

$$(\widehat{A}_n(\cdot), \hat{\mathcal{P}}_n(\cdot), \widehat{\sigma}_n(\cdot)) \xrightarrow{\mathrm{d}} (\widehat{A}(\cdot), -\mathbb{E}[S]\widehat{\sigma}(\mathbb{E}[S]\cdot), \widehat{\sigma}(\cdot)). \tag{3.3.92}$$

Moreover, since $\bar{A}_n(\cdot)$ and $\bar{B}_n(\cdot)$ converge to deterministic limits, by [96, Theorem 11.4.5] the above convergence can be strengthened to

$$\begin{aligned}&(\widehat{A}_n(\cdot), \hat{\mathcal{P}}_n(\cdot), \widehat{\sigma}_n(\cdot), \bar{A}_n(\cdot), \bar{B}_n(\cdot))\\ &\qquad \xrightarrow{\mathrm{d}} (\widehat{A}(\cdot), -\mathbb{E}[S]\widehat{\sigma}(\mathbb{E}[S]\cdot), \widehat{\sigma}(\cdot), f_T(0)\mathrm{id}(\cdot), \mathrm{id}(\cdot)),\end{aligned} \tag{3.3.93}$$

in $(\mathcal{D}^4, WJ_1)$. It follows that

$$\begin{aligned}&(\widehat{A}_n(\cdot), \hat{\mathcal{P}}_n(\bar{A}_n(\cdot)), \mathbb{E}[S]\widehat{\sigma}_n(\bar{B}_n(\cdot)))\\ &\qquad \xrightarrow{\mathrm{d}} (\widehat{A}(\cdot), -\mathbb{E}[S]\widehat{\sigma}(\cdot), \mathbb{E}[S]\widehat{\sigma}(\cdot)),\end{aligned} \tag{3.3.94}$$

in $(\mathcal{D}^3, WJ_1)$ by the heavy-traffic assumption (2.1.2). The limit processes are continuous with probability one, and thus their sum converges to the sum of the limits. This observation, together with the Continuous-Mapping Theorem and (3.3.94) imply that, as $n \to \infty$,

$$\mathbb{E}[S]\widehat{Q}_n(t) + \hat{\mathcal{P}}_n(\bar{A}_n(\cdot)) + \mathbb{E}[S]\widehat{\sigma}_n(\bar{B}_n(\cdot)) \xrightarrow{\mathrm{d}} \widehat{Q}(\cdot), \quad \text{in } (\mathcal{D}, J_1), \tag{3.3.95}$$

concluding the proof. □

Thanks to the heavy-traffic condition $\mathbb{E}[S] = 1/f_T(0)$, we retrieve the usual form of Little's Law as

$$\widehat{Q}(t) = f_T(0)\hat{W}(t). \tag{3.3.96}$$

Note that $f_T(0) = \lambda$ when T is exponentially distributed with mean $1/\lambda$.

Theorem 14 should be contrasted with the analogous result [48, Proposition 4]. There, an extra diffusion term in the expression of $\hat{W}(t)$ appears. This term is a function of the fluid limit of the queue-length process. However, in our setting, this limit is the zero process, as can be seen in (3.3.19), where no centering is needed.

3.4 The subcritical regime

This chapter is dedicated to proving the following lemma:

Lemma 27. *Assume that $f_T(s) = \mathrm{e}^{-s}$ and*

$$\sup_{t\geq 0} f_T(t)\mathbb{E}[S] = \mathbb{E}[S] < 1. \tag{3.4.1}$$

Then, for any fixed $t > 0$ there exists a non-trivial integer random variable $\mathcal{Q}(t)$ such that, as $n \to \infty$

$$Q_n(t) \xrightarrow{\mathrm{d}} \mathcal{Q}(t). \tag{3.4.2}$$

Moreover, $\mathcal{Q}(t)$ is the stationary distribution of the queue length of a $M/G/1$ queue with constant arrival rate $f_T(t)$ and service distributions given by $(S_i)_{i=1}^{\infty}$.

We start by describing the idea of the proof. First, we show that without loss of generality we can assume that the queue is empty at $t - \delta$, for small $\delta > 0$. Then we stochastically bound $Q_n(t)$ at time t from below and from above by two $M/G/1$ queues at time δn, with arrival intensity approximately equal to $F_T(t) - F_T(t-\delta)$. By letting $n \to \infty$, the $M/G/1$ queues acting as lower and upper bound converge to their stationary distributions, and subsequently letting $\delta \to 0$, they converge to the stationary distribution of an $M/G/1$ queue with arrival rate $f_T(t)$. Let us now give the details.

Proof. Fix $0 < \delta \ll 1$. We define $Q_{n,\delta}(t)$ the queue length of the $\Delta_{(i)}/G/1$ queue conditioned on starting in 0 at $t - \delta$. Then,

$$Q_n(t) = \max\{Q_{n,\delta}(t), Q_n(t-\delta) + \Delta N_n\}, \tag{3.4.3}$$

where ΔN_n is defined as $\Delta N_n := N_n(t) - N_n(t-\delta)$. In [48] the authors show that $\lim_{n\to\infty} N_n(t)/n = F_T(t) - \mu t$ almost surely. Therefore $\lim_{n\to\infty} \Delta N_n/n = F_T(t) - F_T(t-\delta) - \mu\delta$. By assumption (3.4.1), δ can be chosen sufficiently small so that $F_T(t) - F_T(t-\delta) - \mu\delta < 0$. Moreover, by [48] $\lim_{n\to\infty} Q_n(s)/n = 0$ almost surely for any s. Therefore, $\lim_{n\to\infty} Q_n(t-\delta) + \Delta N_n = -\infty$ and therefore $Q_n(t) - Q_{n,\delta}(t) \searrow 0$ as $n \to \infty$ almost surely.

The arrival process $\mathcal{A}_n(t)$, consisting of order statistics of n i.i.d. exponential random variables, can be cast as a thinned Poisson process, with time-dependent thinning. This construction will play an important role in the next chapter. We briefly introduce it here. Consider a rate n Poisson process $\Pi(\cdot)$ and associate to each point PP_i a 'mark' M_i chosen uniformly at random from the set $\{1,2,\ldots,n\}$. If $M_i \notin \{M_1, M_2, \ldots, M_{i-1}\}$, the point is accepted and otherwise it is rejected. The probability of acceptance P_i is then a random variable and, given $M_1, M_2, \ldots, M_{i-1}$, is given by $P_i := 1 - \frac{|\{M_1,M_2,\ldots,M_{i-1}\}|}{n}$. The process $A_n^m(\cdot)$ constructed in this way jumps almost surely exactly n times. In fact, we will show that this process is distributionally equivalent to the cumulative arrival process of the $\Delta_{(i)}/G/1$ queue. In order to lower bound (resp. upper bound) $\mathcal{A}_n(t)$ with a homogeneous Poisson process we compute the largest (resp. smallest) number of different marks (arrivals) in the time interval $(t-\delta, t)$ and use this as a constant thinning parameter. By [48, Proposition 1],

$$\frac{\mathcal{A}_n(t)}{n} \overset{\text{a.s.}}{\to} F_T(t). \tag{3.4.4}$$

Therefore, for every fixed $\varepsilon > 0$, with high probability,

$$\begin{aligned} \frac{\mathcal{A}_n(t)}{n} &\in ((1-\varepsilon)F_T(t), (1+\varepsilon)F_T(t)), \\ \frac{\mathcal{A}_n(t-\delta)}{n} &\in ((1-\varepsilon)F_T(t-\delta), (1+\varepsilon)F_T(t-\delta)). \end{aligned} \tag{3.4.5}$$

The number of cumulative arrivals $\mathcal{A}_n(t)$ coincides with the number of different 'marks' seen up to time t in our equivalent description of the arrival process. Then, on the event given by (3.4.5), the largest and smallest number of arrivals in the interval $(t-\delta, t)$ are, respectively,

$$\begin{aligned} &n((1+\varepsilon)F_T(t) - (1-\varepsilon)F_T(t-\delta)) \\ &\qquad = n\delta(f_T(t) + \varepsilon(F_T(t) + F_T(t-\delta))/\delta + o_\delta(1)) =: n\overline{p}. \end{aligned} \tag{3.4.6}$$

and

$$n((1-\varepsilon)F_T(t)-(1+\varepsilon)F_T(t-\delta))$$
$$= n\delta(f_T(t)-\varepsilon(F_T(t)+F_T(t-\delta))/\delta+o_\delta(1)) =: n\underline{p}, \tag{3.4.7}$$

where $o_\delta(1)$ denotes a quantity such that $\lim_{\delta\to 0} o_\delta(1)=0$.

We now couple the process $A_n^m(\cdot)$ with two homogeneous thinned Poisson processes that act as upper and lower bound on $A_n^m(\cdot)$. Recall that $A_n^m(\cdot)$ is defined by a *time-dependant thinning* of a rate n Poisson process $\Pi(\cdot)$. We interpret this thinning as assigning to each point i of $\Pi(\cdot)$ a Bernoulli random variable $\mathrm{Be}_i(P_i)$ that accepts the point with the time-dependent probability P_i. We simultaneously couple each of these Bernoulli random variables in the interval $(t-\delta,t)$ with other two Bernoulli random variables such that, almost surely,

$$\mathrm{Be}_i(\underline{p})\le\mathrm{Be}_i(P_i)\le\mathrm{Be}_i(\overline{p}), \tag{3.4.8}$$

where $\overline{p}$ and $\underline{p}$ are defined in (3.4.6)–(3.4.7). This implies the almost sure stochastic domination

$$\underline{N}_{\underline{p}/\delta}(n\delta^2)\le A_n^m(t)\le\overline{N}_{\overline{p}/\delta}(n\delta^2), \tag{3.4.9}$$

where $\underline{N}_{\underline{p}/\delta}(n\delta^2)$ (resp. $\overline{N}_{\overline{p}/\delta}(n\delta^2)$) represents a Poisson process with rate $\underline{p}/\delta$ (resp. $\overline{p}/\delta$) at time $n\delta^2$. The coupling (3.4.9) is constructed as follows: $A_n^m(t)$ starts at zero in $t-\delta$ and for each Poisson point, the test to accept it is performed with the three coupled Bernoulli random variables. We obtain (3.4.9) by observing that $\underline{N}_{\underline{p}n}(\delta) \overset{\mathrm{d}}{=} \underline{N}_{\underline{p}/\delta}(n\delta^2)$. We now consider $\underline{N}_{\underline{p}/\delta}$, $\mathcal{A}_n(t) \overset{\mathrm{d}}{=} A_n^m(t)$ and $\overline{N}_{\overline{p}/\delta}$ as arrival processes for three queues, all starting in 0 at $t-\delta$. Let us denote the three queues respectively as $\underline{Q}$, $Q_{n,\delta}$ and $\overline{Q}$). Then (3.4.9) implies

$$\underline{Q}(\delta n) \preceq Q_{n,\delta}(t) \preceq \overline{Q}(\delta n), \tag{3.4.10}$$

where $\underline{Q}(\cdot)$ and $\overline{Q}(\cdot)$ are $M/G/1$ queues. The arrival process of $\underline{Q}(\cdot)$ (resp. of $\overline{Q}(\cdot)$) has rate $\underline{p}/\delta$ (resp. $\overline{p}/\delta$). The service times in the three queues are given by the same random variables $(S_i)_{n=1}^{\infty}$. Equation (3.4.10) is equivalent to

$$\mathbb{P}(\underline{Q}(\delta n)>x)\le\mathbb{P}(Q_{n,\delta}(t)>x)\le\mathbb{P}(\overline{Q}(\delta n)>x). \tag{3.4.11}$$

Letting $n \to \infty$, $\underline{Q}(\delta n)$ (resp. $\overline{Q}(\delta n)$) converges to its stationary distribution $\underline{Q}$ (resp. $\overline{Q}$) [25, Section II.4.3]. Therefore,

$$\begin{aligned}\mathbb{P}(\underline{Q} > x) &\le \liminf_{n\to\infty} \mathbb{P}(Q_{n,\delta}(t) > x)\\ &\le \limsup_{n\to\infty} \mathbb{P}(Q_{n,\delta}(t) > x)\\ &\le \mathbb{P}(\overline{Q} > x). \end{aligned} \tag{3.4.12}$$

Since $Q_{n,\delta}(t)$ and $Q_n(t)$ have asymptotically the same distribution, the previous relation simplifies to

$$\begin{aligned}\mathbb{P}(\underline{Q} > x) &\le \liminf_{n\to\infty} \mathbb{P}(Q_n(t) > x)\\ &\le \limsup_{n\to\infty} \mathbb{P}(Q_n(t) > x)\\ &\le \mathbb{P}(\overline{Q} > x). \end{aligned} \tag{3.4.13}$$

Taking first $\varepsilon \searrow 0$ and then $\delta \searrow 0$, $|\mathbb{P}(\underline{Q} > x) - \mathbb{P}(\overline{Q} > x)| \to 0$ and both $\underline{Q}$ and $\overline{Q}$ converge to the stationary distribution of an $M/G/1$ queue $\mathcal{Q} = \overline{Q}(t)$ with arrival rate $f_T(t)$. Thus,

$$\lim_{n\to\infty} \mathbb{P}(Q_n(t) > x) = \mathbb{P}(Q(t) > x), \tag{3.4.14}$$

concluding the proof. □

3.5 Conclusions

This chapter concludes the analysis of the heavy-traffic standard $\Delta_{(i)}/G/1$ queue. We have shown that the limiting behavior of the embedded queue and the queue-length process is almost identical. In fact, the two processes differ by a linear time-change in the limit. We have crucially assumed that the variance of the service-time distribution is finite. Next, we drop this assumption and study the $\Delta_{(i)}/G/1$ queue with *heavy-tailed* services.

CHAPTER 4

Heavy-tailed services

In this chapter we investigate the heavy-tailed behavior of the $\Delta_{(i)}/G/1$ queue. Our starting point is the representation of the $\Delta_{(i)}/G/1$ queue-length process given in Chapter 3. We give a new representation of the arrival process as a thinned Poisson process with time-dependent thinning. Assuming that the tail of the service distribution decays as a power-law with exponent $\gamma \in (1,2)$, so that the mean of the service times is finite, but the variance is not, we show that the queue-length process converges to an γ-stable motion with negative quadratic drift.

4.1 Model description

For both the $M/G/1$ queue and the $\Delta_{(i)}/G/1$ queue it is clear that the queue-length process is strongly influenced by the service times and in particular depends on whether or not the service-time distribution is heavy tailed. For the $M/G/1$ queue, several heavy-traffic limit theorems have been established for heavy-tailed service-time distributions with infinite variance; see [6, 23, 96] and references therein. In this chapter we pursue similar limit theorems for the heavy-tailed $\Delta_{(i)}/G/1$ queue, although our thinned arrival process leads to vastly different results. A connection, however, with the classical work on the $M/G/1$ queue [6, 23, 96] is that also in the case of the $\Delta_{(i)}/G/1$ queue stable laws play a crucial role. For the $M/G/1$ queue, and in queueing theory in general, one typically distinguishes between light-tailed and heavy-tailed service-time distribution, and many models originally studied under light-tailed assumptions were later considered under heavy-tailed conditions. As such, this chapter should be regarded as the heavy-tailed extension of the light-tailed setting studied in Chapter 3.

For simplicity we will restrict ourselves to exponentially distributed arrival times. More precisely, we assume $(T_i)_{i=1}^n$ to be a sequence of i.i.d. exponential random variables with mean $1/\lambda$. The arrival times are then given by the order statistics of $(T_i)_{i=1}^n$. The job sizes are given by a sequence $(S_i)_{i=1}^n$ of i.i.d. random variables. The server works with speed c_n, so that the service time of customer i is given by $D_i := S_i/c_n$. We will take $c_n = n/(1+\beta n^{-\eta})$ for some yet unspecified $\eta > 0$ that is chosen appropriately. We denote the distribution function of S_1 by $F_S(\cdot)$. We say a function $\ell(\cdot)$ is *slowly varying* when $\lim_{t\to\infty} \ell(tc)/\ell(t) = 1$ for all $c > 0$. The service-time distribution is assumed to be in the domain of attraction of an γ-stable law, that is its tail decays as

$$\mathbb{P}(S > t) = 1 - F_S(t) = t^{-\gamma}\ell(t), \qquad \gamma \in (1,2), \tag{4.1.1}$$

for a slowly-varying function $\ell(\cdot)$. Assumption (4.1.1) implies, in particular, that $\mathbb{E}[S^k] = \infty$ for $k > \gamma$, and $\mathbb{E}[S^k] < \infty$ for $k < \gamma$. In this setting, our heavy-traffic condition simplifies to

$$n\lambda\mathbb{E}[D] = \lambda\mathbb{E}[S](1+\beta n^{-\eta}) = 1+\beta n^{-\eta}. \tag{4.1.2}$$

We study the queue after $X_n(0)$ customers have already joined, where $X_n(0)$ may depend on n and $X_n(0) \to \infty$. Since in our setting $X_n(0) \ll n$, without loss of generality we can assume that at time 0 there are (still) n customers in the pool. Before stating the main results of this chapter, let us introduce some notation. Recall that the queue-length process $Q_n(t)$ is given by

$$Q_n(t) = X_n(0) + \mathcal{A}_n(t) - \sigma_n(B_n(t)), \tag{4.1.3}$$

where $X_n(0)$ denotes the number of customers already in the queue at the beginning of the first service; See Section 3.3 for the detailed construction of $Q_n(\cdot)$.

Alternatively, $Q_n(\cdot)$ can be expressed as the reflection of a free process $X_n(\cdot)$ as follows [6, Proposition 2.2, p. 251]:

$$Q_n(t) = \phi(X_n)(t), \qquad t \geq 0, \tag{4.1.4}$$

where $X_n(\cdot)$ is given by

$$X_n(t) = X_n(0) + \mathcal{A}_n(t) - \sigma_n(B_n(t)) - I_n(t)/\mathbb{E}[S]. \tag{4.1.5}$$

Recall the definitions of $\mathcal{A}_n(\cdot)$ and $\sigma_n(\cdot)$ in (3.1.1) and (3.1.2). See Figure 1.4 for an example of a sample path of $X_n(\cdot)$.

We will consider the scaled processes given by

$$\widehat{X}_n(t) = n^{-\frac{1}{2\gamma-1}}\ell_2(n)X_n(\tau_n(t)), \tag{4.1.6}$$

$$\widehat{Q}_n(t) = \phi(\widehat{X}_n)(t), \tag{4.1.7}$$

$$\tau_n(t) = tn^{-\frac{\gamma-1}{2\gamma-1}}\ell_1(n), \tag{4.1.8}$$

where $\ell_1(\cdot)$ and $\ell_2(\cdot)$ are slowly-varying functions that depend on $\ell(\cdot)$ in (4.1.1). Using basic properties of slowly-varying functions [18, Proposition 1.3.6], the scaling constants can be rewritten as

$$n^{-\frac{\gamma-1}{2\gamma-1}}\ell_1(n) = n^{-\frac{(1+o(1))(\gamma-1)}{2\gamma-1}}, \qquad n^{-\frac{1}{2\gamma-1}}\ell_2(n) = n^{-\frac{1+o(1)}{2\gamma-1}}. \tag{4.1.9}$$

In particular, for $\gamma = 2$ the scaling exponents are asymptotically equal to the exponents for the finite variance case in Theorem 9. We can now state our main result:

Theorem 15 (The critically loaded $\Delta_{(i)}/G/1$ queue with heavy-tailed services). *Assume $X_n(0) = qn^{\frac{1}{2\gamma-1}}\ell_2^{-1}(n)$ for some $q \geq 0$. Assume further that $\eta = (\gamma-1)/(2\gamma-1)$ so that $c_n = n/(1+\beta n^{-(\gamma-1)/(2\gamma-1)})$. Then,*

$$\widehat{X}_n(\cdot) \xrightarrow{\text{d}} \widehat{X}(\cdot) \qquad \text{in } (\mathcal{D}, M_1), \tag{4.1.10}$$

where

$$\widehat{X}(t) = q + \beta\lambda t - \frac{\lambda^2}{2}t^2 + s_\gamma\mathcal{S}(t), \tag{4.1.11}$$

$s_\gamma = 1/\mathbb{E}[S]^{1+1/\gamma}$ and $\mathcal{S}(\cdot)$ is a spectrally positive γ-stable process. Moreover,

$$\widehat{Q}(\cdot) \xrightarrow{\text{d}} \phi(\widehat{X}\cdot) \qquad \text{in } (\mathcal{D}, M_1). \tag{4.1.12}$$

Convergence in $(\mathcal{D}, M_1)$ is a shorthand notation for convergence in distribution in the space of càdlàg functions $\mathcal{D}$ endowed with the M_1 topology. We elaborate on this later on. See Figure 1.5 for some sample paths of $\phi(\widehat{X})(\cdot)$ for different choices of γ for fixed q, β, λ, s_γ. See also Figure 1.6 for a graph of the first passage time as a function of the linear drift β, for fixed γ and q, and different values of the linear drift parameter β.

The following corollary of Theorem 15 characterizes the limiting distribution of $T_{\widehat{X}}(0)$:

Corollary 3 (Busy period convergence). *Under the assumptions of Theorem 15, as $n \to \infty$,*

$$T_{\widehat{X}_n}(0) \xrightarrow{\mathrm{d}} T_{\widehat{X}}(0). \tag{4.1.13}$$

Proof. Note that $t \mapsto \widehat{X}(t)$ only has positive jumps. Then, by [52, Chapter VI, Proposition 2.11], the functional $f \mapsto T_f(0)$ is continuous in $\widehat{X}$ with probability one when $\mathcal{D}$ is endowed with the M_1 topology. Indeed, it is continuous when $\mathcal{D}$ is endowed with the stronger J_1 topology. The conclusion follows by an application of the Continuous-Mapping Theorem. □

4.2 Preliminaries

In this section we introduce various results that will be useful for the proof of Theorem 15. In Section 4.2.1 we present an FCLT for the service-time process $\sigma(\cdot)$. In Section 4.2.2 we derive an alternative characterization of the arrival process of the $\Delta_{(i)}/G/1$ queue which reveals a connection with the Poisson process. Finally, in Section 4.2.3 we give a heuristic argument that motivates the scaling constants appearing in Theorem 15.

Since we deal with limit processes with *unmatched jumps*, we endow $\mathcal{D}$ with the M_1 topology. This topology is coarser than the usual J_1 topology, so that convergence with respect to the J_1 topology implies convergence with respect to the M_1 topology. When dealing with vector-valued functions (taking values, say, in $\mathbb{R}^k$) we make use of the *weak* M_1 topology M_1^{W}, which coincides with the product topology on $\mathcal{D} \times \mathcal{D} \times \cdots \times \mathcal{D} = \mathcal{D}^k$. For an in-depth discussion on the various Skorokhod topologies, see [96].

4.2.1 FCLT for a renewal process

We start by presenting an FCLT for the renewal process $\sigma_n(\cdot)$. To do so we exploit the well-known equivalence between the FCLT for partial sums and counting processes. Let $(S_i)_{i=1}^n$ be a sequence of non-negative random variables and let

$$\widehat{\Sigma}_n(t) := \frac{\sum_{i=1}^{\lfloor nt \rfloor} S_i - \mathbb{E}[S]nt}{d_n}, \tag{4.2.1}$$

where $(d_n)_{n=1}^{\infty}$ will be chosen appropriately later. Let $\widehat{\sigma}_n(\cdot)$ denote the rescaled renewal process associated with the service times, defined as

$$\widehat{\sigma}_n(t) := \frac{\sigma_n(t) - \mathbb{E}[S]^{-1}nt}{d_n}. \tag{4.2.2}$$

The relation between the scaling limits of $\widehat{\Sigma}_n(\cdot)$ and $\widehat{\sigma}_n(\cdot)$ is described in the following theorem:

Theorem 16 (FCLT equivalence [96, Theorem 7.3.2]). *Assume $(S_i)_{i=1}^{\infty}$ is a sequence of non-negative random variables, and $(d_n)_{n=1}^{\infty}$ is such that $d_n \to \infty$, $n/d_n \to \infty$. Then,*

$$\widehat{\Sigma}_n(\cdot) \xrightarrow{\mathrm{d}} \mathcal{S}(\cdot) \qquad \text{in } (\mathcal{D}, M_1) \tag{4.2.3}$$

for some process $\mathcal{S}(\cdot)$ if and only if

$$\widehat{\sigma}_n(\cdot) \xrightarrow{\mathrm{d}} -\mathbb{E}[S]^{-1}\mathcal{S} \circ \mathbb{E}[S]^{-1}\mathrm{id}(\cdot) \qquad \text{in } (\mathcal{D}, M_1), \tag{4.2.4}$$

where $\mathrm{id}(\cdot)$ *denotes the identity function.*

The topology M_1 plays a crucial role in Theorem 16. Indeed, it can be seen that while (4.2.3) holds in most cases in the J_1 topology, the convergence (4.2.4) can only take place in the M_1 topology when the limit process has positive jumps; See [96, Chapter 7.3.2] for a more detailed explanation. By assumption (4.1.1), the sequence $(S_i)_{i=1}^{\infty}$ is in the domain of attraction of an γ-stable motion, that is (4.2.3) holds, and $\mathcal{S}(\cdot)$ is a centered, spectrally positive γ-stable motion.

By Theorem 16, the process $\widehat{\sigma}_n(\cdot)$ is then also in the domain of attraction of an γ-stable motion. Note that the space scaling constants d_n in (4.2.1) and (4.2.2) are the same.

4.2.2 Poissonian representation of the arrival process

We now introduce an alternative characterization of the arrival process as a thinned, marked Poisson process. It is constructed as follows. Given $\Pi(\cdot)$, a rate λ homogeneous Poisson process, assign to each of its points a mark chosen uniformly in $[n] := \{1, \dots, n\}$. We then discard a point if it has a mark that has already been observed in the past. Therefore, conditioned on the marks $M_1, \dots, M_{k-1}$, the next point of $\Pi(\cdot)$ will be accepted with probability $(n - |\{M_1, \dots, M_{k-1}\}|)/n$. We denote this thinned process as $A_n^m(\cdot)$. Formally, $A_n^m(t)$ is given by

$$A_n^m(t) = \Pi(t) - R_n(t), \tag{4.2.5}$$

where $R_n(t)$ counts the number of *repeated* marks until time t. We emphasize that $\Pi(\cdot)$ and $R_n(\cdot)$ are *not* independent. The arrival process just defined is closely related with the i.i.d. sampling in the $\Delta_{(i)}/G/1$ queue. In fact, we will show that $\mathcal{A}_n(\cdot)$ and $A_n^m(\cdot)$ are equivalent. First, let us introduce some preliminary notation and results. Given a sequence of random variables $(X_i)_{i=1}^n$, recall that $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denote their order statistics. When $(X_i)_{i=1}^n$ are i.i.d. exponential random variables, the distribution of the order statistics is well known:

Lemma 28 (Order statistics of exponentials). *Let $E_1, \ldots, E_n$ be independent exponentially distributed random variables with mean one. Then,*

$$(E_{(j)})_{j=1}^n \overset{\text{d}}{=} \Big(\sum_{s=1}^{j} \frac{E_s}{n-s+1} \Big)_{j=1}^n. \tag{4.2.6}$$

See for example [31, Section 2.5] for a proof. Lemma 28 allows us to relate the process $A_n^m(\cdot)$ we just defined to the arrival process in the $\Delta_{(i)}/G/1$ queue.

Lemma 29. *For all $t \geq 0$,*

$$A_n^m(t) \overset{\text{d}}{=} \mathcal{A}_n(t/n). \tag{4.2.7}$$

Proof. The ordered arrival times in the $\Delta_{(i)}/G/1$ queue are precisely the order statistics of $(T_i)_{i=1}^n$ and the inter-arrival times are the differences between the order statistics. By Lemma 28, the distributions of the inter-arrival times are

$$\frac{1}{\lambda}(E_{(k)} - E_{(k-1)}) \overset{\text{d}}{=} \frac{E_k/\lambda}{n-k+1}, \qquad k \geq 1, \tag{4.2.8}$$

where we set $E_{(0)} = 0$ for convenience. Multiplying both sides by n, and noting that $E_i/\lambda = T_i$, gives

$$n(T_{(k)} - T_{(k-1)}) \overset{\text{d}}{=} \frac{E_k}{1 - \frac{k-1}{n}} \frac{1}{\lambda}. \tag{4.2.9}$$

Now consider the process $A_n^m(\cdot)$. Conditioned on the process up to the arrival $k-1$, the next point of $\Pi(\cdot)$ is accepted with probability $1 - \frac{k-1}{n}$. Then, since $\Pi(\cdot)$ is a rate λ Poisson process, the time at which the next point of $A_n^m(\cdot)$ occurs is distributed as an exponential random variable with rate $\lambda(1 - \frac{k-1}{n})$. Equation (4.2.9) then implies that the inter-arrival times in the process $t \mapsto A_n^m(t)$ have the same distribution as the inter-arrival times of $\sum_{i=1}^n \mathbb{1}_{\{nT_i \leq t\}} = \mathcal{A}_n(t/n)$. □

4.2.3 Determining the scaling constants

We now derive the space and time scalings in the limit process $\widehat{X}(\cdot)$ in (4.1.11), starting with the scaling of time $k = k(n)$. It is well known that, whenever the limit $\mathcal{S}(\cdot)$ in (4.2.3) is an γ-stable motion, the fluctuations of $\sum_{i=1}^{\lfloor nkt \rfloor} S_i$ around its mean are of the order $d_k = \ell_0(k)(nk)^{1/\gamma}$ (see e.g. [96, Theorem 4.5.1]), where $\ell_0(\cdot)$ is a slowly-varying function that is a priori different from $\ell(\cdot)$ in (4.1.1) (but can be determined from it). Moreover, in (4.3.25) below we show that the highest order contribution to the drift component $R_n(nkt)$ is $\Pi(nkt)^2/(2n) = O_{\mathbb{P}}(k^2 n)$, all the other terms being negligible. In the process $\widehat{X}(\cdot)$ both a drift and a random component appear, so that we must have

$$\ell_0(k)k^{1/\gamma}n^{1/\gamma} = k^2 n. \tag{4.2.10}$$

Equivalently,

$$\ell_0(k)^{-\frac{\gamma}{2\gamma-1}}k = n^{\frac{\gamma-1}{2\gamma-1}}, \tag{4.2.11}$$

where $\ell_0(\cdot)^{-\gamma/(2\gamma-1)}$ is, by basic properties of slowly-varying functions, again slowly varying. On the left-hand side of (4.2.11) we recognize a regularly-varying function with index 1. By [18, Theorem 1.5.12] each regularly-varying function with index γ admits an (asymptotic) inverse that is itself regularly varying, with index $1/\gamma$. Therefore, there exists a slowly-varying function $\rho(\cdot)$ such that

$$k = n^{\frac{\gamma-1}{2\gamma-1}}\rho(n^{\frac{\gamma-1}{2\gamma-1}}). \tag{4.2.12}$$

Any sequence $(k(n))_{n=1}^{\infty}$ that satisfies condition (4.2.12) is suitable for our purposes, so that we simply take $k(n) = n^{-\frac{\gamma-1}{2\gamma-1}}\ell_1(n)$, where $\ell_1(n) = \rho(n^{-\frac{\gamma-1}{2\gamma-1}})$. Note that $n \mapsto \ell_1(n)$ is again slowly varying. Therefore, the rescaled time parameter is defined as

$$\tau_n(t) := tn^{-\frac{\gamma-1}{2\gamma-1}}\ell_1(n). \tag{4.2.13}$$

We shall denote the time scaling factor by $\tau_n(1) = n^{-\frac{\gamma-1}{2\gamma-1}}\ell_1(n)$. In order to obtain the space-scaling sequence $(d_n)_{n=1}^{\infty}$, it is enough to insert $k = n^{-\frac{\gamma-1}{2\gamma-1}}\ell_1(n)$ into $f(k) := k^2 n$. Therefore, we define d_n as

$$d_n = ((n^{-\frac{\gamma-1}{2\gamma-1}}\ell_1(n))^2 n)^{-1} = \ell_1(n)^{-2}n^{-\frac{1}{2\gamma-1}} = \ell_2(n)n^{-\frac{1}{2\gamma-1}}, \tag{4.2.14}$$

where $\ell_2(n) := \ell_1(n)^{-2}$ is again slowly varying.

4.3 Proof of Theorem 15

In this section we carry out the proof of Theorem 15. We first prove Theorem 15 for $\beta = 0$, and later show how to extend it to the general case $\beta \neq 0$.

Rewriting equation (4.1.5) using (4.2.7) gives

$$\begin{aligned}X_n(t) &\overset{\mathrm{d}}{=} X_n(0) + (A_n^m(nt) - c_n t/\mathbb{E}[S]) + (c_n B_n(t)/\mathbb{E}[S] - \sigma_n(B_n(t)) \\ &= X_n(0) + (\Pi(nt) - nt/\mathbb{E}[S]) \\ &\quad + (nB_n(t)/\mathbb{E}[S] - \sigma_n(B_n(t))) - R_n(nt), \qquad (4.3.1)\end{aligned}$$

where we also used the equality $I_n(t) = c_n t - c_n B_n(t)$, with $c_n = n$, see (3.3.10). For simplicity, we introduce the scaled version of the arrival and service processes, and of the busy time, as

$$\begin{aligned}\widehat{\Pi}(t) &:= n^{-\frac{1}{2\gamma-1}} \ell_2(n)(\Pi(n\tau_n(t)) - c_n\tau_n(t)/\mathbb{E}[S]), \\ \widehat{R}_n(t) &:= n^{-\frac{1}{2\gamma-1}} \ell_2(n) R_n(n\tau_n(t)), \\ \widehat{\sigma}_n(t) &:= n^{-\frac{1}{2\gamma-1}} \ell_2(n)(c_n\tau_n(t)/\mathbb{E}[S] - \sigma_n(\tau_n(t))), \\ \bar{B}_n(t) &:= B_n(\tau_n(t))/\tau_n(1). \qquad (4.3.2)\end{aligned}$$

Assume that $X_n(0) = qn^{\frac{1}{2\gamma-1}}\ell_1(n)$, for some $q \geq 0$. After rescaling, (4.3.1) becomes

$$\widehat{X}_n(\tau_n(t)) = q + \widehat{\Pi}(t) + \widehat{\sigma}_n(\bar{B}_n(t)) - \widehat{R}_n(t). \qquad (4.3.3)$$

The proof of Theorem 15 proceeds as follows. First, the term $\widehat{\Pi}(\cdot)$ is shown to be negligible in the limit. Second, $\widehat{\sigma}_n(\cdot)$ converges to an γ-stable motion by (4.1.1) and Theorem 16. Third, $\widehat{R}_n(\cdot)$ is shown to converge to the parabolic drift $-\lambda^2/2t^2$. Finally, $\bar{B}_n(\cdot)$ is shown to converge to the identity function. All these results are then pieced together in Section 4.3.3. Convergence of the above processes is proven in $\mathcal{D}([0,T])$ for a fixed $T > 0$. Since T is arbitrary, this implies convergence in $\mathcal{D}([0,\infty])$ by [17, Lemma 3, p.174].

4.3.1 Stable limit

We start by showing that the process $\widehat{\Pi}(\cdot)$ does not contribute to the randomness of the limit process.

Lemma 30. *As $n \to \infty$,*

$$\sup_{t \le T} |\widehat{\Pi}(\tau_n(t))| \xrightarrow{\mathbb{P}} 0. \tag{4.3.4}$$

Proof. By the FCLT for the Poisson process,

$$\frac{\Pi(n\tau_n(\cdot)) - \lambda n\tau_n(\cdot)}{\sqrt{n\tau_n(1)}} \xrightarrow{\mathrm{d}} W(\cdot), \qquad \text{in } (\mathcal{D}, U), \tag{4.3.5}$$

where $W(\cdot)$ is a standard Brownian motion, since $1/\mathbb{E}[S] = \lambda$ by (4.1.2). By the Skorokhod Representation Theorem, this implies that we can couple $\Pi(\tau_n(\cdot))$ and $W(\cdot)$ in such a way that

$$\sup_{t \le T} \Big| \frac{\Pi(n\tau_n(t)) - \lambda n\tau_n(\cdot)}{\sqrt{n\tau_n(1)}} - W(t) \Big| \xrightarrow{\mathbb{P}} 0. \tag{4.3.6}$$

Moreover, for any $c > 0$ and n large enough,

$$c\sqrt{n\tau_n(1)} = cn^{\gamma/(4\gamma-2)} \ell_1(n)^{1/2} \le n^{1/(2\gamma-1)} \ell_2(n)^{-1}, \tag{4.3.7}$$

so that $k_n := n^{1/(2\gamma-1)} \ell_2(n)/\sqrt{\tau_n} \to \infty$ and

$$\begin{aligned}\sup_{t \le T} \Big| &\frac{\Pi(n\tau_n(t)) - \lambda n\tau_n(\cdot)}{n^{1/(2\gamma-1)} \ell_2^{-1}(n)} \Big| \\ &\le \frac{1}{k_n} \sup_{t \le T} \Big| \frac{\Pi(n\tau_n(t)) - \lambda n\tau_n(t)}{\sqrt{n\tau_n(1)}} - W(t) \Big| + \sup_{t \le T} \Big| \frac{W(t)}{k_n} \Big|. \end{aligned} \tag{4.3.8}$$

Since the right-hand side of (4.3.8) converges in probability to zero as $n \to \infty$, the claim follows. □

Next, we show convergence of the rescaled service process $\widehat{\sigma}_n(\cdot)$ to an γ-stable motion:

Lemma 31 (Stable limit)**.** *As $n \to \infty$,*

$$\widehat{\sigma}_n(\cdot) \xrightarrow{\mathrm{d}} s_\gamma \mathcal{S}(\cdot) \qquad \text{in } (\mathcal{D}, M_1), \tag{4.3.9}$$

where $s_\gamma = 1/\mathbb{E}[S]^{(\gamma+1)/\gamma}$ and $\mathcal{S}(\cdot)$ is a spectrally positive γ-stable motion.

Proof. By classical results, the rescaled partial sums of $(S_i)_{i=1}^{\infty}$ converge to a spectrally positive γ-stable motion, see e.g. [52] and [96, Theorem 4.5.3]. In particular (4.2.3) is satisfied. Theorem 16 implies (4.2.4), that is

$$\widehat{\sigma}_n(\cdot) \xrightarrow{\mathrm{d}} \frac{1}{\mathbb{E}[S]}\mathcal{S}\Big(\frac{\cdot}{\mathbb{E}[S]}\Big) \qquad \text{in } (\mathcal{D}, M_1). \tag{4.3.10}$$

By standard properties of stable motion $(\mathcal{S}(ct))_{t\geq 0} \overset{\mathrm{d}}{=} (c^{1/\gamma}\mathcal{S}(t))_{t\geq 0}$ for $c > 0$, so that the claim (4.3.9) follows. □

Remark 1. The stable law corresponding to $\gamma = 2$ is the standard normal distribution. In particular, its variance is finite. Although our results do not directly hold for $\gamma = 2$, it is still possible to enter $\gamma = 2$ in the formulas that we obtain, and what is obtained should be consistent with our results of Chapter 3. This is true, for example, for the coefficient of the stable motion in (4.3.9). Indeed, in Theorem 9 we proved that if $\mathbb{E}[S^2] = 1$, the standard deviation of the limiting Brownian motion is $\lambda^{3/2} = \mathbb{E}[S]^{-3/2}$.

4.3.2 Drift limit

The most difficult task in proving Theorem 15 is to deal with the complicated drift $\widehat{R}_n(\cdot)$ in (4.3.3). We will prove the following result:

Proposition 2 (Drift limit). *As $n \to \infty$ and for any $T > 0$,*

$$\sup_{t\leq T}\Big|\widehat{R}_n(t) - \frac{\lambda^2}{2}t^2\Big| \xrightarrow{\mathbb{P}} 0. \tag{4.3.11}$$

The proof will exploit upper and lower bounds for $\widehat{R}_n(\cdot)$, obtained by giving an equivalent representation of the drift process. First, note that the probability of extracting a mark that has already appeared at time $i > 0$ is $D_n(i-1)/n$, where $D_n(i)$ denotes the number of *different* marks seen up to the i-th arrival epoch in $\Pi(\cdot)$. Therefore, conditionally on $D_n(i-1)$, the thinning procedure is represented by a Bernoulli random variable with parameter $D_n(i-1)/n$. Since at time t a total of $\Pi(t)$ points have been accepted, we have

$$R_n(t) \overset{\mathrm{d}}{=} \sum_{i=1}^{\Pi(t)} \mathbb{1}_{\{U_i \leq \frac{D_n(i-1)}{n}\}}, \tag{4.3.12}$$

where $(U_i)_{i=1}^{\infty}$ are random variables that are uniformly distributed on $[0,1]$, and are independent of all other randomness. Then, $\mathbb{1}_{\{U_i\leq x\}}$ is distributed as a Bernoulli random variable with parameter x. Moreover, $D_n(i)$ is given explicitly as

$$D_n(i) = i - Z_n(i), \tag{4.3.13}$$

where $Z_n(i)$ is the number of *repeated* marks seen up to the time of the i-th arrival. In other words we have the crucial relation

$$D_n(i) \overset{\mathrm{d}}{=} i - R_n(\Pi^{-1}(i)), \tag{4.3.14}$$

where $\Pi^{-1}(i)$ is the arrival time of the i-th customer; see Figure 4.1.

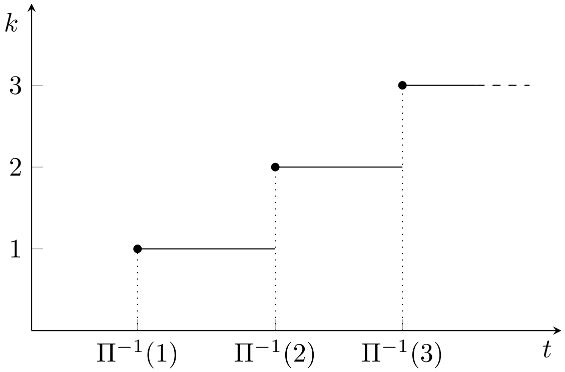

Figure 4.1: A sample path of the process $\Pi(\cdot)$.

Exploiting these ideas, we construct a process $(R_n(k))_{k=1}^{\infty}$ recursively, by setting $\tilde{R}_n(0) := 0$ and

$$\tilde{R}_n(k) := \tilde{R}_n(k-1) + \mathbb{1}_{\{U_k \leq \frac{k-1-\tilde{R}_n(k-1)}{n}\}}, \qquad k \geq 1. \tag{4.3.15}$$

Unraveling the recursion, we get

$$\tilde{R}_n(k) := \sum_{i=1}^{k} \mathbb{1}_{\{U_i \leq \frac{i-1-\tilde{R}_n(i-1)}{n}\}}, \qquad k \geq 1. \tag{4.3.16}$$

We see that

$$R_n(t) \overset{\mathrm{d}}{=} \tilde{R}_n(\Pi(t)). \tag{4.3.17}$$

As already mentioned, the processes $R_n(\cdot)$ and $\Pi(\cdot)$ are not independent. The distributional equality (4.3.17) reveals the dependency of $R_n(\cdot)$ on the process $\Pi(\cdot)$.

The next step is to construct an upper and a lower bound on $\tilde{R}_n(k)$. Since $\tilde{R}_n(k) \geq 0$, the upper bound is trivially

$$\mathbb{1}_{\{U_i \leq (i-1-\tilde{R}_n(i-1))/n\}} \leq \mathbb{1}_{\{U_i \leq \frac{i-1}{n}\}}, \tag{4.3.18}$$

so that, almost surely,

$$\tilde{R}_n(k) \leq \tilde{R}_n^{(\mathrm{up})}(k) := \sum_{i=1}^{k} \mathbb{1}_{\{U_i \leq \frac{i-1}{n}\}}. \tag{4.3.19}$$

The lower bound is more involved. By (4.3.19),

$$\mathbb{1}_{\{U_i \leq \frac{i-1-\tilde{R}_n(i-1)}{n}\}} \geq \mathbb{1}_{\{U_i \leq \frac{i-1-\tilde{R}_n^{(\mathrm{up})}(i-1)}{n}\}} \tag{4.3.20}$$

so that

$$\tilde{R}_n(k) \geq \tilde{R}_n^{(\mathrm{low})}(k) := \sum_{i=1}^{k} \mathbb{1}_{\{U_i \leq \frac{i-1-\tilde{R}_n^{(\mathrm{up})}(i-1)}{n}\}}. \tag{4.3.21}$$

Note that U_i is independent of $\tilde{R}_n^{(\mathrm{up})}(i-1)$. We have then constructed a coupling such that for all $t \geq 0$, almost surely,

$$R_n^{(\mathrm{low})}(t) \leq R_n(t) \leq R_n^{(\mathrm{up})}(t), \tag{4.3.22}$$

where $R_n^{(\mathrm{low})}(t) := \tilde{R}_n^{(\mathrm{low})}(\Pi(t))$ and $R_n^{(\mathrm{up})}(t) := \tilde{R}_n^{(\mathrm{up})}(\Pi(t))$. For the next and last step we prove uniform convergence of the upper and lower bounds to the same limit.

Upper bound

In this section we will estimate the quantity

$$U_n(T) := \sup_{t \leq T} \left| n^{-1/(2\gamma-1)} \ell_2(n) R_n^{(\mathrm{up})}(n\tau_n(t)) - \frac{\lambda^2}{2} t^2 \right|, \tag{4.3.23}$$

We will prove the following:

Lemma 32 (Upper bound converges to zero). *As $n \to \infty$,*

$$U_n(T) \xrightarrow{\mathbb{P}} 0, \tag{4.3.24}$$

for every fixed $T > 0$.

Proof. We split the absolute value in (4.3.23) as

$$\begin{aligned}
U_n(T) &\le \Big| n^{-\frac{1}{2\gamma-1}} \ell_2(n) \sum_{i=1}^{\Pi(n\tau_n(t))} \Big(\mathbb{1}_{\{U_i \le \frac{i-1}{n}\}} - \frac{i-1}{n} \Big) \Big| \\
&\quad + \Big| n^{-\frac{1}{2\gamma-1}} \ell_2(n) \sum_{i=1}^{\Pi(n\tau_n(t))} \Big(\frac{i-1}{n} \Big) - \frac{\lambda^2}{2} t^2 \Big| \\
&\le \Big| n^{-\frac{1}{2\gamma-1}} \ell_2(n) \sum_{i=1}^{\Pi(n\tau_n(t))} \Big(\mathbb{1}_{\{U_i \le \frac{i-1}{n}\}} - \frac{i-1}{n} \Big) \Big| \\
&\quad + \Big| \frac{\Pi(n\tau_n(t))^2}{2n^{2\gamma/(2\gamma-1)} \ell_2^{-1}(n)} - \frac{\lambda^2}{2} t^2 \Big| + \varepsilon_n, \qquad (4.3.25)
\end{aligned}$$

where $\varepsilon_n = |\Pi(\tau_n(t))/2n|$ is an error term. By the strong FLLN for the Poisson process

$$\frac{\Pi(tn^{\gamma/(2\gamma-1)} \ell_1(n))}{n^{\gamma/(2\gamma-1)} \ell_2(n)^{-1/2}} \overset{\text{a.s.}}{\to} \lambda t, \qquad \text{in } (\mathcal{D}, U). \qquad (4.3.26)$$

We note that we have made explicit use of the specific form of the scaling functions $\ell_1(\cdot)$ and $\ell_2(\cdot)$ as determined above in (4.2.10). More specifically, by definition we have that $\ell_1(n)^{-2} = \ell_2(n)$. Moreover, the functional $x \mapsto x^2$ from $\mathcal{D}([0,T])$ to itself is almost surely continuous in $f(t) = \lambda t$ in the uniform topology. This implies that the second and third terms in (4.3.25) converge to zero uniformly for $t \le T$ as $n \to \infty$.

By the LLN for the Poisson process we have that $\Pi(s) \le (\lambda + \varepsilon)s$ with high probability for $s = O(n^{\gamma/(2\gamma-1)})$. The sum in the first term in (4.3.25) is then bounded on the event $\{\Pi(s) \le (\lambda+\varepsilon)s\}$ as

$$\begin{aligned}
&\sup_{s \le n\tau_n(T)} \Big| \sum_{i=1}^{\Pi(s)} \Big(\mathbb{1}_{\{U_i \le \frac{i-1}{n}\}} - \frac{i-1}{n} \Big) \Big| \\
&\qquad \le \sup_{s \le (\lambda+\varepsilon) n\tau_n(T)} \Big| \sum_{i=1}^{\lfloor s \rfloor} \Big(\mathbb{1}_{\{U_i \le \frac{i-1}{n}\}} - \frac{i-1}{n} \Big) \Big|. \qquad (4.3.27)
\end{aligned}$$

The right-hand side is the supremum of a martingale. In the following and future computations we shall denote $\bar{T} := T(\lambda + \varepsilon)$. Then, an application of Doob's L^2 martingale inequality [63, Theorem 11.2] gives

$$
\begin{aligned}
\mathbb{P}\Big(\sup_{s\leq \bar{T}n^{\gamma/(2\gamma-1)}\ell_1(n)}\Big|\sum_{i=1}^{\lfloor s\rfloor}\Big(\mathbb{1}_{\{U_i\leq\frac{i-1}{n}\}}-\frac{i-1}{n}\Big)\Big|\geq \varepsilon n^{\frac{1}{2\gamma-1}}\ell_2^{-1}(n)\Big)& \\
\leq \sum_{i=1}^{\bar{T}n^{\frac{\gamma}{2\gamma-1}}\ell_1(n)}\frac{\mathbb{E}[(\mathbb{1}_{\{U_i\leq\frac{i-1}{n}\}}-\frac{i-1}{n})^2]}{\varepsilon^2 n^{\frac{2}{2\gamma-1}}\ell_2^{-2}(n)}& \\
=\frac{1}{\varepsilon^2 n^{\frac{2}{2\gamma-1}}\ell_2^{-2}(n)}\sum_{i=1}^{\bar{T}n^{\frac{\gamma}{2\gamma-1}}\ell_1(n)-1}\Big(\frac{i}{n}-\frac{i^2}{n^2}\Big)& \\
\leq\frac{\bar{T}^2 n^{\frac{2\gamma}{2\gamma-1}}\ell_1^2(n)}{\varepsilon^2 n^{\frac{2\gamma+1}{2\gamma-1}}\ell_2^{-2}(n)}=O(n^{-\frac{1}{2\gamma-1}}\ell_2(n)),& \qquad (4.3.28)
\end{aligned}
$$

and this implies that the right-hand side of (4.3.27) is $o_{\mathbb{P}}(n^{1/(2\gamma-1)}\ell_2^{-1}(n))$. □

Lower bound

By (4.3.22) we also have

$$
R_n(t)\succeq R_n^{(\mathrm{low})}=\sum_{i=1}^{\Pi(t)}\mathbb{1}_{\{U_i\leq(i-1-\tilde{R_n}^{(\mathrm{up})}(i-1))/n\}}. \qquad (4.3.29)
$$

Consequently, we now estimate

$$
L_n(T):=\sup_{t\leq T}\Big|n^{-1/(2\gamma-1)}\ell_2(n)R_n^{(\mathrm{low})}(n\tau_n(t))-\frac{\lambda^2}{2}t^2\Big|. \qquad (4.3.30)
$$

Lemma 33 (Lower bound converges to zero). *As $n\to\infty$,*

$$
L_n(T)\xrightarrow{\mathbb{P}}0, \qquad (4.3.31)
$$

for every fixed $T>0$.

Proof. Similarly as before, conditioned on the event $\{\Pi(s) \leq (\lambda+\varepsilon)s\}$,

$$\begin{aligned}
&L_n(T)\\
&\leq \sup_{s\leq n\tau_n(\bar{T})}\Big|n^{-\frac{1}{2\gamma-1}}\ell_2(n)\sum_{i=1}^{\lfloor s\rfloor}\Big(\mathbb{1}_{\{U_i\leq\frac{i-1-\tilde{R}_n^{(\mathrm{up})}(i-1)}{n}\}}-\frac{i-1-\tilde{R}_n^{(\mathrm{up})}(i-1)}{n}\Big)\Big|\\
&\quad+\sup_{t\leq T}\Big|n^{-\frac{1}{2\gamma-1}}\ell_2(n)\sum_{i=1}^{\Pi(n\tau_n(t))}\frac{i-1}{n}-\frac{\lambda^2}{2}t^2\Big|\\
&\quad+\sup_{s\leq n\tau_n(\bar{T})}\Big|n^{-\frac{1}{2\gamma-1}}\ell_2(n)\sum_{i=1}^{\lfloor s\rfloor}\frac{\tilde{R}_n^{(\mathrm{up})}(i-1)}{n}\Big|. \qquad (4.3.32)
\end{aligned}$$

The first term in (4.3.32) is bounded as before, since it is the supremum of a martingale. Denote $Y_n(i) := (i-1-\tilde{R}_n^{(\mathrm{up})}(i-1))/n$ for convenience. By Doob's L^2 martingale inequality,

$$\begin{aligned}
\varepsilon^2 n^{\frac{2}{2\gamma-1}}\ell_2^{-2}(n)\mathbb{P}\Big(\sup_{s\leq n\tau_n(\bar{T})}\Big|\sum_{i=1}^{\lfloor s\rfloor}(\mathbb{1}_{\{U_i\leq Y_n(i)\}}-Y_n(i))\Big|\geq\varepsilon n^{\frac{1}{2\gamma-1}}\ell_2^{-1}(n)\Big)\\
\leq\mathbb{E}\Big[\Big(\sum_{i=1}^{n\tau_n(\bar{T})}\mathbb{1}_{\{U_i\leq Y_n(i)\}}-Y_n(i)\Big)^2\Big]\\
=\sum_{i=1}^{n\tau_n(\bar{T})}\mathbb{E}[(\mathbb{1}_{\{U_i\leq Y_n(i)\}}-Y_n(i))^2]. \qquad (4.3.33)
\end{aligned}$$

Since the variance of a Bernoulli random variable with parameter p is $p(1-p)$, we get

$$\mathbb{E}[(\mathbb{1}_{\{U_i\leq Y_n(i)\}}-Y_n(i))^2]=\mathbb{E}[Y_n(i)-Y_n(i)^2]\leq\mathbb{E}[Y_n(i)]\leq\frac{i}{n}. \qquad (4.3.34)$$

This implies that

$$\sup_{i\leq\bar{T}n^{\frac{\gamma}{2\gamma-1}}\ell_1(n)}\mathbb{E}[(\mathbb{1}_{\{U_i\leq Y_n(i)\}}-Y_n(i))^2]\leq\bar{T}n^{\frac{1-\gamma}{2\gamma-1}}\ell_1(n). \qquad (4.3.35)$$

In particular,

$$\begin{aligned}
&\sum_{i=1}^{n\tau_n(\bar{T})}\mathbb{E}[(\mathbb{1}_{\{U_i\leq Y_n(i)\}}-Y_n(i))^2]\\
&\qquad\leq\tau_n(\bar{T})\bar{T}n^{\frac{1-\gamma}{2\gamma-1}}\ell_1^2(n)=\bar{T}^2n^{\frac{1}{2\gamma-1}}\ell_1^2(n)=o(n^{\frac{2}{2\gamma-1}}). \qquad (4.3.36)
\end{aligned}$$

The second term in (4.3.32) has been shown to converge in (4.3.25) and (4.3.26). Since $t \mapsto \tilde{R}_n^{(\mathrm{up})}(t)$ is non-decreasing, we bound the third term as

$$\sup_{s \le n\tau_n(\bar{T})} \Big| \sum_{i=1}^{\lfloor s \rfloor} \frac{\tilde{R}_n^{(\mathrm{up})}(i-1)}{n} \Big| \le \bar{T} n^{\frac{1-\gamma}{2\gamma-1}} \ell_1(n) \tilde{R}_n^{(\mathrm{up})}(n\tau_n(\bar{T})). \tag{4.3.37}$$

Note that $\bar{T} n^{(1-\gamma)/(2\gamma-1)} \ell_1(n) \to 0$ as $n \to \infty$. Since

$$n^{-1/(2\gamma-1)} \ell_2(n) \tilde{R}_n^{(\mathrm{up})}(n\tau_n(\bar{T})) \xrightarrow{\mathbb{P}} 0 \tag{4.3.38}$$

by Lemma 32,

$$\begin{aligned} n^{-\frac{1}{2\gamma-1}} \ell_2(n) \sup_{s \le n\tau_n(\bar{T})} X\Big| \sum_{i=1}^{\lfloor s \rfloor} \frac{\tilde{R}^{(\mathrm{up})}(i-1)}{n} \Big| \\ \le (\bar{T} n^{\frac{1-\gamma}{2\gamma-1}} \ell_1(n)) n^{-\frac{1}{2\gamma-1}} \ell_2(n) \tilde{R}_n^{(\mathrm{up})}(\tau_n(\bar{T})) \xrightarrow{\mathbb{P}} 0 \end{aligned} \tag{4.3.39}$$

as $n \to \infty$. This concludes the proof of Lemma 33. □

Proof of Proposition 2. Since

$$\sup_{t \le T} |n^{-\frac{1}{2\gamma-1}} \ell_2(n) R_n(t n^{\frac{\gamma}{2\gamma-1}} \ell_1(n)) - \frac{1}{2} t^2| \tag{4.3.40}$$

$$\begin{aligned} &= \sup_{t \le T} (n^{-\frac{1}{2\gamma-1}} \ell_2(n) R_n(t n^{\frac{\gamma}{2\gamma-1}} \ell_1(n)) - \frac{1}{2} t^2)^+ \\ &\quad + \sup_{t \le T} (n^{-\frac{1}{2\gamma-1}} \ell_2(n) R_n(t n^{\frac{\gamma}{2\gamma-1}} \ell_1(n)) - \frac{1}{2} t^2)^- \end{aligned} \tag{4.3.41}$$

we get

$$\sup_{t \le T} |n^{-\frac{1}{2\gamma-1}} \ell_2(n) R_n(t n^{\frac{\gamma}{2\gamma-1}} \ell_1(n)) - \frac{1}{2} t^2| \le U_n(T) \vee L_n(T) \tag{4.3.42}$$

and both $U_n(T)$ and $L_n(T)$ converge in probability to zero by Lemma 32 and Lemma 33. This completes the proof of Proposition 2. □

4.3.3 Busy-time process limit

For the final step, we prove that the cumulative busy-time process converges to the identity function.

Lemma 34 (Cumulative idle time is negligible)**.** *As $n \to \infty$,*

$$\bar{B}_n(\cdot) \xrightarrow{\mathrm{d}} \mathrm{id}(\cdot), \qquad \text{in } (\mathcal{D}, U), \tag{4.3.43}$$

where $\mathrm{id}(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}^+$ *is the identity function.*

Proof. Since $B_n(t) = t - I_n(t)$, we will prove that $I_n(t) = \inf_{0\leq s\leq t}(P_n(s)^-)$ converges uniformly to zero, where $P_n(t)$ is the net-put process defined in (3.3.7). By continuity of the map $\psi(\cdot)$ given by $\psi : f(\cdot) \to \inf_{0\leq s\leq t}(f(s)^-)$, it is sufficient to prove that $P_n(\cdot)$ converges uniformly to zero, when appropriately rescaled. By manipulating (3.3.7) we immediately get

$$\begin{aligned}\frac{1}{\tau_n(1)} \sup_{t\leq T} |P_n(\tau_n(t))| &= \sup_{t\leq T} \Big| \frac{\mathcal{A}_n(\tau_n(t))}{\tau_n(1)} \frac{1}{A_n(\tau_n(t))} \sum_{i=1}^{\mathcal{A}_n(\tau_n(t))} S_i - 1 \Big| \\ &\leq \sup_{t\leq T} \Big| \frac{\mathcal{A}_n(\tau_n(t))}{\tau_n(1)} - \frac{1}{\mathbb{E}[S]} \Big| \frac{1}{\mathcal{A}_n(\tau_n(t))} \sum_{i=1}^{\mathcal{A}_n(\tau_n(t))} S_i \\ &\quad + \sup_{t\leq T} \Big| \frac{1}{\mathbb{E}[S]} \frac{1}{\mathcal{A}_n(\tau_n(t))} \sum_{i=1}^{\mathcal{A}_n(\tau_n(t))} S_i - 1 \Big|. \end{aligned} \tag{4.3.44}$$

Note that $\tau_n(t) \to \infty$ and $\mathcal{A}_n(\tau_n(t)) \xrightarrow{\mathbb{P}} \infty$ as $n \to \infty$. Then the second term converges to zero in probability by the LLN and the first one converges to zero by the LLN for the Poisson process. Indeed, since $\mathcal{A}_n(\tau_n(t)) = \Pi(\tau_n(t)) - R_n(\tau_n(t))$, we have that

$$\begin{aligned} &\sup_{t\leq T} \Big| \frac{\mathcal{A}_n(\tau_n(t))}{\tau_n(1)} - \frac{1}{\mathbb{E}[S]} \Big| \\ &\leq \sup_{t\leq T} \Big| \frac{\Pi(\tau_n(t))}{\tau_n(1)} - \frac{1}{\mathbb{E}[S]} \Big| + \frac{1}{\tau_n(1)} \sup_{t\leq T} |R_n(\tau_n(t))| \\ &= \sup_{t\leq T} \Big| \frac{\Pi(\tau_n(t))}{\tau_n(1)} - \frac{1}{\mathbb{E}[S]} \Big| + \frac{n^{\frac{1}{2\gamma-1}} \ell_n(n)^{-1}}{\tau_n(1)} \sup_{t\leq T} \frac{|R_n(\tau_n(t))|}{n^{\frac{1}{2\gamma-1}} \ell_n(n)^{-1}}. \end{aligned} \tag{4.3.45}$$

As shown above in Proposition 2, $n^{-1/(2\gamma-1)}\ell_2(n)R_n(\tau_n(t))$ converges uniformly to $-\lambda^2/2t^2$, and since $n^{1/(2\gamma-1)}\ell_n^{-1}/\tau_n(1) \to 0$, the second term in (4.3.45) is negligible. By the heavy-traffic assumption (4.1.2) and the LLN for the Poisson process the first term also converges to zero. □

We conclude the proof of Theorem 15 by collecting the results from the previous sections. First, we split the process $\widehat{X}_n(\cdot)$ in its martingale and drift components as in (4.3.3) to get

$$\widehat{X}_n(t) = q + \widehat{\Pi}(t) + \widehat{\sigma}_n(\bar{B}_n(t)) - \widehat{R}_n(t). \tag{4.3.46}$$

Since $\widehat{\Pi}(\cdot)$ and $\widehat{\sigma}_n(\cdot)$ are independent, and $\bar{B}_n(\cdot)$ and $\widehat{R}_n(\cdot)$ converge to deterministic limits in $\mathcal{D}$, we have

$$(\widehat{\Pi}(\cdot), \widehat{\sigma}_n(\cdot), \bar{B}_n(\cdot), \widehat{R}_n(\cdot)) \overset{\mathrm{d}}{\to} (0, s_\gamma \mathcal{S}(\cdot), \mathrm{id}(\cdot), \lambda^2/2\mathrm{id}(\cdot)^2). \tag{4.3.47}$$

in $(\mathcal{D}^4, M_1^{\mathrm{W}})$. This, together with the time-change theorem for processes with discontinuous sample paths [96, Theorem 13.2.3] implies

$$(\widehat{\Pi}(\cdot), \widehat{\sigma}_n(\bar{B}_n(\cdot)), \widehat{R}_n(\cdot)) \overset{\mathrm{d}}{\to} (0, s_\gamma \mathcal{S}(\cdot), \mathrm{id}(\cdot)^2 \lambda^2/2). \tag{4.3.48}$$

in $(\mathcal{D}^3, M_1^{\mathrm{W}})$. Note that [96, Theorem 13.2.3] does not hold in general in the finer J_1 topology. Since the three limit processes in (4.3.48) do not have common discontinuity points, we have that addition is continuous in $(0, 1/\mathbb{E}[S]^{(\gamma+1)/\gamma}\mathcal{S}(\cdot), \mathrm{id}(\cdot)^2\lambda^2/2)$ in the M_1 topology, so that

$$\widehat{X}_n(t) \overset{\mathrm{d}}{\to} q + s_\gamma \mathcal{S}(t) - \frac{\lambda^2}{2}t^2, \qquad \text{in } (\mathcal{D}, M_1). \tag{4.3.49}$$

The second claim (4.1.12) follows immediately from the Continuous-Mapping Theorem, since the reflection map is Lipschitz continuous in the M_1 topology by [96, Theorem 13.5.1]. □

Extension to general initial drift Now we assume that

$$c_n = n/(1 + \beta n^{-\frac{\gamma-1}{2\gamma-1}} \ell_2(n)^{-1}), \tag{4.3.50}$$

with $\beta \neq 0$. We rewrite (4.3.1) for a general service speed c_n as

$$\begin{aligned} X_n(t) &\overset{\mathrm{d}}{=} X_n(0) + (\Pi(nt) - \lambda c_n t) \\ &\quad + (c_n B_n(t)/\mathbb{E}[S] - \sigma_n(B_n(t))) - R_n(nt) \\ &= X_n(0) - \frac{\lambda}{1 + \beta n^{-\frac{\gamma-1}{2\gamma-1}}} t + (\Pi(nt) - \lambda nt) \\ &\quad + (c_n B_n(t)/\mathbb{E}[S] - \sigma_n(B_n(t))) - R_n(nt), \end{aligned} \tag{4.3.51}$$

where we have used assumption (4.1.2). By rescaling the process as in (4.3.3), we obtain

$$\widehat{X}_n(t) = q - \frac{\lambda t n^{\frac{\gamma-1}{2\gamma-1}}}{1 + \beta n^{-\frac{\gamma-1}{2\gamma-1}}} + \widehat{\Pi}(t) + \widehat{\sigma}_n(\bar{B}_n(t)) - \widehat{R}_n(t). \tag{4.3.52}$$

Since $1+\beta n^{-(\gamma-1)/(2\gamma-1)} \to 1$ as $n \to \infty$, the rescaled partial sums of the *double sequence* $(S_i/(1+\beta n^{-(\gamma-1)/(2\gamma-1)}))_{i=1}^{\infty}$ converge to the γ-stable motion $\mathcal{S}(\cdot)$, hence Theorem 16 holds and $\widehat{\sigma}_n(\cdot) \to s_\gamma \mathcal{S}(\cdot)$. Moreover, $\widehat{\Pi}(\cdot)$, $\bar{B}_n(\cdot)$, and $\widehat{R}_n(\cdot)$ converge as before, and as $n \to \infty$,

$$-\frac{\lambda t n^{\frac{\gamma-1}{2\gamma-1}}}{1+\beta n^{-\frac{\gamma-1}{2\gamma-1}}} \to \lambda\beta t. \tag{4.3.53}$$

Summarizing, we have shown that

$$\widehat{X}_n(t) \xrightarrow{\mathrm{d}} q + \lambda\beta t + s_\gamma \mathcal{S}(\cdot) - \frac{\lambda^2}{2}t^2, \qquad \text{in } (\mathcal{D}, M_1), \tag{4.3.54}$$

concluding the proof.

4.4 Conclusions

In this chapter we have extended the analysis of the $\Delta_{(i)}/G/1$ queue to the case of power-law service distributions. We have shown that the limiting behavior of the queue-length process is vastly different from the light-tailed case. In the heavy-tailed setting, the queue-length process is driven by large upward jumps and, consequently, is discontinuous almost everywhere. However, the depletion-of-points effect is again present in the form of a negative quadratic drift. Next, we analyze the $\Delta_{(i)}/G/1$ queue in the setting of *size-biased* arrival times.

CHAPTER 5

Biased arrivals and random graphs

In this chapter we consider a generalization of the $\Delta_{(i)}/G/1$ queue, which we call the $\Delta_{(i)}^{\alpha}/G/1$ queue. In this model, n customers arrive at the queue at times depending on their service requirement. A customer with stochastic service requirement S arrives to the queue after an exponentially distributed time with mean $S^{-\alpha}$ for some $\alpha \in [0,1]$; so larger service requirements trigger customers to join earlier. As α varies in $[0,1]$, this model interpolates between the $\Delta_{(i)}/G/1$ queue and the exploration process for inhomogeneous random graphs. We consider the asymptotic regime in which the pool size n grows to infinity and establish that the scaled embedded queue process converges to a diffusion process with a negative quadratic drift. While the form of the limit process is identical to the $\alpha = 0$ case, the coefficients of the drift and the Brownian component depend crucially on α. We also describe how this first busy period of the queue gives rise to a critically connected random forest.

5.1 Model description

The $\Delta_{(i)}^{\alpha}/G/1$ queue is defined as follows. Again n customers are triggered to join the queue after independent exponential times, but the rates of their exponential clocks depend on their service requirements.

More specifically, denoting again the service requirement of customer i by S_i, conditioned on S_i the arrival time T_i of i is distributed as a rate $S^{-\alpha}$ exponential random variable. We will initially take $\alpha \in [0,1]$. Then, when $\alpha = 0$, the arrival times are i.i.d. and when $\alpha \in (0,1]$ the arrival times decrease with the service requirement. The queue is attended by a single server that starts working at time zero, works at unit speed, and serves the customers in a first-come first-served order. At time zero, we allow for the possibility that i of the n customers have already joined the queue and are waiting for service. We will take $i \ll n$, so that without loss of generality we assume that there are still n customers waiting for service. These initial customers are numbered $1, \ldots, i$ and the customers that arrive later are numbered $i+1, i+2, \ldots$ in order of their arrival. Let $A_n(k)$ denote the number of customers arriving during the service time of the k-th customer. Note that the random variables $(A_n(k))_{k\geq 1}$ are not i.i.d. due to the finite-pool effect and the service-dependent arrival rates. Because of the complicated dependence structure of the $(A_n(k))_{k\geq 1}$, we will model and analyze this queue using the queue-length process embedded at service completions, generalizing the results of Chapter 2.

While the queueing process consists of alternating busy and idle periods, in the $\Delta_{(i)}^{\alpha}/G/1$ queue we naturally focus on the first busy period. The negative quadratic drift in the scaling captures the effect of a pool of potential customers that diminishes with time: after some time, the activity in the queue inevitably becomes negligible. The early phases of the process are therefore of primary interest, when the head start provided by the initial customers still matters and when the rate of newly arriving customers is still relatively high. The head start and strong influx together lead to a substantial first busy period, and essentially determine the relevant time of operation of the system.

We also consider the structural properties of the first busy period in terms of a (directed) random graph associated to the queueing process as follows. Say that the number of customers served in the first busy period, starting with i initial customers, is N and consider a graph with vertex set $\{1, 2, \ldots, N\}$ and in which two vertices r and s are joined by an edge if and only if the r-th customer arrives during the service time of the s-th customer. If $i = 1$, then the graph is a rooted tree with N labeled vertices, the root being labeled 1. If $i > 1$, then the graph is a forest consisting of i distinct rooted trees whose roots are labeled $1, \ldots, i$, respectively. The total number of vertices in the forest is N.

5.1.1 The $\Delta^{\alpha}_{(i)}/G/1$ queue

Let us now describe our assumptions in detail. We consider a sequence of queueing systems, each with a finite number n of potential customers labelled with indices $i \in [n] := \{1, \dots, n\}$. Customers have i.i.d. service requirements $S_1, S_2, \dots$ with cumulative distribution function $F_S(\cdot)$. We denote with S a generic random value with distribution $F_S(\cdot)$. In order to obtain meaningful limits as the system grows large, we assume that the service speed c_n scales as $c_n = n/(1+\beta n^{-1/3})$ with $\beta \in \mathbb{R}$ so that the service time of customer i is given by

$$D_i = \frac{S_i}{c_n} = \frac{S_i}{n}(1+\beta n^{-1/3}). \tag{5.1.1}$$

The assumptions above follow the assumptions on the $\Delta_{(i)}/G/1$ model in Chapter 2 closely. For the $\Delta^{\alpha}_{(i)}/G/1$ queue, we fix a parameter $\alpha \in [0,1]$ and we assume, crucially, that $\mathbb{E}[S^{2+\alpha}] < \infty$.

Conditioned on S_i, the arrival time T_i of customer i is assumed to be exponentially distributed with mean $1/(\lambda S_i^{\alpha})$, with $\lambda > 0$. Hence

$$T_i \overset{\mathrm{d}}{=} \frac{E_i}{\lambda S_i^{\alpha}}, \tag{5.1.2}$$

where $(E_i)_{i=1}^n$ denotes a family of independent mean one exponential random variables. Note that conditionally on the service times, the arrival times are independent. However they are not identically distributed. We introduce $c(1), c(2), \dots, c(n)$ as the indices of the customers in order of arrival, so that $T_{c(1)} \leq T_{c(2)} \leq T_{c(3)} \leq \dots$ almost surely.

We crucially assume that the queueing system is critically loaded. In this setting, the heavy-traffic condition for the load ρ_n turns out to be

$$\rho_n := \lambda_n \mathbb{E}[S^{1+\alpha}](1+\beta n^{-1/3}) = 1 + \beta n^{-1/3} + o_{\mathbb{P}}(n^{-1/3}), \tag{5.1.3}$$

where $\lambda = \lambda_n$ may depend on n and $f_n = o_{\mathbb{P}}(n^{-1/3})$ is such that $f_n n^{1/3} \xrightarrow{\mathbb{P}} 0$. The parameter β then determines the position of the system inside the critical window: the traffic intensity is greater than one for $\beta > 0$, so that the system is initially overloaded, while the system is initially underloaded for $\beta < 0$.

Our main object of study is the queue-length process embedded at service completions, given by $Q_n^e(0) = i$ and

$$Q_n^e(k) = (Q_n^e(k-1) + A_n(k) - 1)^+. \tag{5.1.4}$$

The number $A_n(k)$ of arrivals during the k-th service is given by

$$A_n(k) = \sum_{i \notin \nu_k} \mathbb{1}_{\{T_i \leq D_{c(k)}\}} \tag{5.1.5}$$

where $\nu_k \subseteq [n]$ denotes the set of customers who have been served or are in the queue at the start of the k-th service. Note that

$$|\nu_k| = (k-1) + Q^e_n(k-1) + 1 = k + Q^e_n(k-1). \tag{5.1.6}$$

We recall that $Q^e_n(\cdot)$ is also represented as the reflected version of a process $N_n(\cdot)$, as

$$Q^e_n(k) = \phi(N_n)(k), \tag{5.1.7}$$

with $N_n(\cdot)$ given by $N_n(0) = i$ and satisfying the recursion

$$N_n(k) = N_n(k-1) + A_n(k) - 1. \tag{5.1.8}$$

By construction, in this queueing system there are no idle periods. We assume that whenever the server finishes processing one customer, and the queue is empty, the customer to be placed into service is chosen according to the size-biased distribution

$$\mathbb{P}(\text{customer } j \text{ is placed in service} \mid \nu_{i-1}) = \frac{S^\alpha_j}{\sum_{l \notin \nu_{i-1}} S^\alpha_l}, \qquad j \notin \nu_{i-1}, \tag{5.1.9}$$

where we tacitly assumed that customer j is the i-th customer to be served. In fact, with definitions (5.1.5) and (5.1.9), the process (5.1.4) describes the $\Delta^\alpha_{(i)}/G/1$ queue with exponential arrivals (5.1.2) embedded at service completions.

Remark 2 (A directed random tree). The embedded queueing process in (5.1.4) and (5.1.7) gives rise to a certain directed rooted tree. To see this, associate a vertex i to customer i and let $c(1)$ be the root. Then, draw a directed edge to $c(1)$ from $c(2), \ldots, c(A_n(1)+1)$ so to all customers who have joined during the service time of $c(1)$. Then, draw an edge from all customers who have joined during the service time of $c(2)$ to $c(2)$, and so on. This procedure draws a directed edge from $c(i)$ to $c(i + \sum_{j=1}^{i-1} A_n(j)), \ldots, c(i + \sum_{j=1}^{i} A_n(j))$ if $A_n(i) \geq 1$. The procedure stops when the queue is empty and there are no more customers to serve. When $Q^e_n(0) = 1$, this gives a random directed rooted tree (resp. forest when $Q^e_n(0) = i$ with $i \geq 2$). The degree of vertex $c(i)$ is $1 + |A_n(i)|$ and the total number of vertices in the tree (forest) is given by

$$T_{Q^e_n}(0) = \inf\{k \geq 0 : Q^e_n(k) = 0\}, \tag{5.1.10}$$

the hitting time of zero of the process $Q_n^e(\cdot)$.

Remark 3 (An inhomogeneous random graph). If $\alpha = 1$, the random tree constructed as above is distributionally equivalent to the tree spanned by the exploration process of an inhomogeneous random graph. Let us elaborate on this. An inhomogeneous random graph is a set of vertices $V = [n]$ with (possibly random) weights $(\mathcal{W}_i)_{i=1}^n$ and edges between them. In a *Norros-Reittu random graph*, given $(\mathcal{W}_i)_{j=1}^n$, i and j share an edge with probability

$$p_{i\leftrightarrow j} := 1 - \exp\Big(- \frac{\mathcal{W}_i \mathcal{W}_j}{\sum_{j=1}^n \mathcal{W}_i}\Big). \tag{5.1.11}$$

The tree constructed from the $\Delta_{(i)}^1/G/1$ queue then corresponds to the exploration process of a rank-1 inhomogeneous random graph, defined as follows. Start with a first arbitrary vertex and reveal all its neighbors. Then the first vertex is discarded and we move to one (suitably chosen) neighbor, and reveal its neighbors. This process continues by exploring the neighbors of each revealed vertex, in order of appearance. By interpreting each vertex as a different customer, this exploration process can be coupled to a $\Delta_{(i)}^1/G/1$ queue, for a specific choice of $(\mathcal{W}_i)_{i=1}^n$ and λ_n. Indeed, if $\mathcal{W}_i = (1+\beta n^{-1/3})S_i$ for $i = 1, \dots, n$, we get

$$\begin{aligned} p_{j\leftrightarrow i} &= 1 - \exp\Big(- (1+\beta n^{-1/3})\frac{S_i}{n}\frac{S_j}{\sum_{l=1}^n S_l/n}\Big) \\ &= 1 - \exp\Big(- D_i S_j \frac{n}{\sum_{i=1}^n S_i}\Big) \\ &= \mathbb{P}(T_j \leq D_i \mid (S_i)_{j=1}^n), \end{aligned} \tag{5.1.12}$$

where

$$T_j \sim \frac{E_j}{S_j \lambda_n}, \tag{5.1.13}$$

and $\lambda_n = n/(\sum_{i=1}^n S_i)$. The rank-1 inhomogeneous random graph with weights $(S_i)_{i=1}^n$ is said to be *critical* (see [15, (1.13)]) if

$$\frac{\sum_{i=1}^n S_i^2}{\sum_{i=1}^n S_i} = \frac{\mathbb{E}[S^2]}{\mathbb{E}[S]} + o_{\mathbb{P}}(n^{-1/3}) = 1 + o_{\mathbb{P}}(n^{-1/3}). \tag{5.1.14}$$

Consequently, if $\beta = 0$ and $\lambda_n = n/\sum_{i=1}^n S_i$, the heavy-traffic condition (5.1.3) for the $\Delta_{(i)}^1/G/1$ queue implies the criticality condition (5.1.14) for the associated random graph, and vice versa.

Remark 4 (The embedded queue and the queue-length process). By definition, the embedded queue (5.1.4) neglects the idle time of the server. Via the time-change argument of Chapter 3 it is possible to prove that, in the limit, the cumulative idle time is negligible and the embedded queue process is arbitrarily close to the queue-length process uniformly over compact intervals. Indeed, the techniques developed earlier can be extended to the $\Delta_{(i)}^{\alpha}/G/1$ queue without additional difficulties. We refrain from doing it here.

5.1.2 The scaling limit of the $\Delta_{(i)}^{\alpha}/G/1$ queue

In what follows, all the processes we consider are elements of the space $\mathcal{D} := \mathcal{D}([0,\infty))$. Recall that, for a discrete-time process $X(\cdot) : \mathbb{N} \to \mathbb{R}$, we write $X(t)$, with $t \in [0,\infty)$, instead of $X(\lfloor t \rfloor)$. In particular, a process defined in this way has càdlàg paths. In this setting, we endow the space $\mathcal{D}$ with the Skorokhod J_1 topology.

We are now able to state our main result:

Theorem 17 (Scaling limit for the $\Delta_{(i)}^{\alpha}/G/1$ queue)**.** *Assume that $\alpha \in [0,1]$, $\mathbb{E}[S^{2+\alpha}] < \infty$ and that the heavy-traffic condition* (5.1.3) *holds. Assume also that the arrival times $(T_i)_{i=1}^{n}$ satisfy* (5.1.2)*. If $Q_n^e(0) = qn^{1/3}$, then as $n \to \infty$,*

$$n^{-1/3}Q_n^e(\cdot n^{2/3}) \xrightarrow{\mathrm{d}} \phi(\widehat{N})(\cdot) \qquad \text{in } (\mathcal{D}, J_1), \tag{5.1.15}$$

where $\widehat{N}(\cdot)$ is the diffusion process

$$\widehat{N}(t) = q + \beta t - \lambda \frac{\mathbb{E}[S^{1+2\alpha}]}{2\mathbb{E}[S^{\alpha}]} t^2 + \sigma W(t), \tag{5.1.16}$$

with $\sigma^2 := \lambda^2 \mathbb{E}[S^{\alpha}]\mathbb{E}[S^{2+\alpha}]$ and $W(\cdot)$ is a standard Brownian motion.

As a straightforward consequence of Theorem 18 we have the following:

Theorem 18 (Number of customers in the first busy period)**.** *Assume that $\alpha \in [0,1]$, $\mathbb{E}[S^{2+\alpha}] < \infty$ and that the heavy-traffic condition* (5.1.3) *holds. If $Q_n^e(0) = qn^{1/3}$, then as $n \to \infty$,*

$$n^{-2/3}T_{Q_n^e}(0) \xrightarrow{\mathrm{d}} T_{\phi(\widehat{N})}(0), \tag{5.1.17}$$

where $\widehat{N}(\cdot)$ is given in (5.1.16).

In particular, denoting by $|F_n|$ the number of vertices in the forest constructed from the $\Delta^{\alpha}_{(i)}/G/1$ queue in Remark 2, we have that, as $n \to \infty$,

$$|F_n| \xrightarrow{\mathrm{d}} T_{\phi(\widehat{N})}(0). \tag{5.1.18}$$

Remark 5 (The parameter α). We restrict the parameter α in the interval $[0,1]$ because the extremes correspond to two well-known models, but it is also of interest to investigate the scaling limit for the $\Delta^{\alpha}_{(i)}/G/1$ model when α is any real number. It is clear from (5.1.16) that a *necessary* condition for Theorem 17 to hold is that $\mathbb{E}[\max\{S^{2+\alpha}, S^{1+2\alpha}, S^{\alpha}\}] < \infty$. If $\alpha \in [0,1]$, this is equivalent to $\mathbb{E}[S^{2+\alpha}] < \infty$, as can be seen from Figure 5.1. The same figure also clarifies the necessary conditions for convergence for the remaining values of α. In particular, if $\alpha \in [-1,0)$, $\mathbb{E}[S^{2+\alpha}] < \infty$ and $\mathbb{E}[S^{\alpha}] < \infty$ are both necessary conditions, since one does not imply the other. Analogously, if $\alpha \in (-2,-1)$, both $\mathbb{E}[S^{2+\alpha}] < \infty$ and $\mathbb{E}[S^{1+2\alpha}] < \infty$ are necessary conditions. In all cases, the moment condition in Figure 5.1 and the heavy-traffic assumption turn out to be *sufficient* for Theorem 17 to hold.

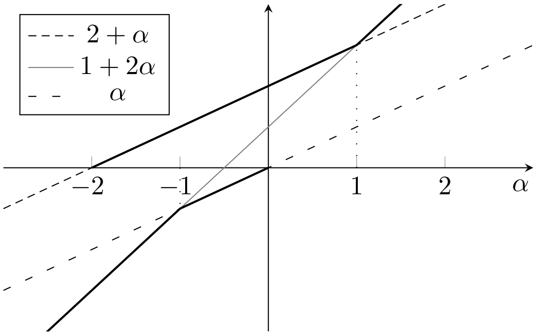

Figure 5.1: Exponents of the service requirement S in the limiting diffusion. The thick black line represents the moments that are required to be finite for the main theorem to hold.

Remark 6 (The $\Delta_{(i)}/G/1$ queue and related models). Let us now compare Theorem 17 with two known results. For $\alpha = 0$, the limit diffusion simplifies to

$$\widehat{N}(t) = \beta t - \frac{1}{2}t^2 + \sigma W(t), \tag{5.1.19}$$

with $\sigma^2 := \lambda^2 \mathbb{E}[S^2]$, in agreement with our results in Chapter 2.

In [15] it is shown that, for a general class of weights $(\mathcal{W}_i)_{i=1}^n$ that include i.i.d. weights and further assuming (5.1.14), the exploration process of the corresponding inhomogeneous random graph converges to

$$\widehat{N}(t) = \beta t - \frac{\mathbb{E}[\mathcal{W}^3]}{2\mathbb{E}[\mathcal{W}^2]^2} t^2 + \frac{\sqrt{\mathbb{E}[\mathcal{W}]\mathbb{E}[\mathcal{W}^3]}}{\mathbb{E}[\mathcal{W}^2]} W(t). \tag{5.1.20}$$

For $\alpha = 1$, (5.1.16) can be rewritten using (5.1.3) as

$$\widehat{N}(t) = \beta t - \frac{\mathbb{E}[S^3]}{2\mathbb{E}[S^2]^2} t^2 + \frac{\sqrt{\mathbb{E}[S]\mathbb{E}[S^3]}}{\mathbb{E}[S^2]} W(t). \tag{5.1.21}$$

Therefore the two processes coincide if $\mathcal{W}_i = S_i$, as expected.

Let us now draw a subtler connection, this time between the $\Delta_{(i)}^{\alpha}/G/1$ and the $\Delta_{(i)}/G/1$ queues. We will show that, in small time intervals and for large n, the $\Delta_{(i)}^{\alpha}/G/1$ queue is approximated by a $\Delta_{(i)}/G/1$ queue with different service and arrival time distributions. Despite being valid only at very small time scales, this approximation gives the correct leading order behavior of the queue. In particular, it will motivate our choice of the heavy-traffic parameter (5.1.3). The interarrival times in the $\Delta_{(i)}^{\alpha}/G/1$ queue are distributed as

$$T_{(k)} - T_{(k-1)} \overset{\mathrm{d}}{=} \frac{E_k}{\lambda \sum_{i \notin \mathfrak{S}_{k-1}} S_i^{\alpha}}, \tag{5.1.22}$$

where $(E_k)_{k=1}^n$ denote mean one exponential random variables and the sum is over the set $[n] \setminus \mathfrak{S}_{k-1}$, where $\mathfrak{S}_{k-1} = \{c(1), c(2), \ldots, c(k-1)\}$. Note that $|[n] \setminus \mathfrak{S}_{k-1}| = n - (k-1)$. When $k = k(n) \to \infty$ as $n \to \infty$, but $k = o(n)$ (say, $k = n^{2/3}$) we have

$$n(T_{(k)} - T_{(k-1)}) \xrightarrow{\mathrm{d}} \frac{E_k}{\lambda \mathbb{E}[S^{\alpha}]}. \tag{5.1.23}$$

Note that by (5.1.5), scaling the service requirement as S_i/c_n is equivalent to scaling the arrival times as $c_n T_i$, where in (5.1.23) we have taken $c_n = n$ for simplicity. It is also equivalent to studying the (unscaled) $\Delta_{(i)}^{\alpha}/G/1$ queue on a time interval $[0, t/c_n] = [0, t/n]$. Equation (5.1.23) implies that the arrival process of the $\Delta_{(i)}^{\alpha}/G/1$ queue is approximated for large n by the arrival process of a $\Delta_{(i)}/G/1$ queue with a different arrival rate parameter $\lambda^* = \lambda \mathbb{E}[S^{\alpha}]$, the equality holding in the limit as $n \to \infty$.

Let us now focus on the service times and compute the law $\mathcal{L}_t(S_i)$ of S_i conditioned on $T_i = t$. We will show that for any fixed i, $\mathcal{L}_t(S_i)$

is *different* from the law of S_i, and that asymptotically is independent from t. This result can be directly extended to a finite family $(\mathcal{L}_t(S_i))_{i=1}^k$ and thus holds for all arrival-service pairs occurring in a small time interval $[0, t/n]$. The probability density function of $\mathcal{L}_t(S_i)$ is given by $f_{S_i,T_i}(s,t)/f_{T_i}(t)$, where $f_{S_i,T_i}(s,t)$ is the joint density function of S_i and T_i. First we compute

$$\mathbb{P}(S_i \le s, T_i \le t) = \mathbb{P}(S_i \le s, \exp_i(\lambda S_i^\alpha) \le t) = \int_0^s (1 - \mathrm{e}^{-tx^\alpha}) f_S(x) \mathrm{d}x. \tag{5.1.24}$$

Taking $s \to \infty$ gives $\mathbb{P}(T_i \le t)$ as

$$\mathbb{P}(T_i \le t) = \int_0^\infty (1 - \mathrm{e}^{-tx^\alpha}) f_S(x) \mathrm{d}x. \tag{5.1.25}$$

Therefore,

$$\frac{f_{S_i,T_i}(s,t)}{f_{T_i}(t)} = \frac{s^\alpha f_S(s) \mathrm{e}^{-ts^\alpha}}{\int_0^\infty x^\alpha f_S(x) \mathrm{e}^{-tx^\alpha} \mathrm{d}x}. \tag{5.1.26}$$

By rescaling T_i as nT_i as in (5.1.23) we get that

$$\frac{f_{S_i,nT_i}(s,t)}{f_{nT_i}(t)} = \frac{f_{S_i,T_i}(s,t/n)}{f_{T_i}(t/n)} \to \frac{s^\alpha f_S(s)}{\int_0^\infty x^\alpha f_S(x) \mathrm{d}x}, \tag{5.1.27}$$

as $n \to \infty$. In this scaling regime, the (conditioned) service times of customers arriving in the time interval $[0, t/n]$ converge to i.i.d. random variables S_i^* with density function given by

$$f_{S_i^*}(s) = \frac{s^\alpha f_S(s)}{\int_0^\infty x^\alpha f_S(x) \mathrm{d}x}. \tag{5.1.28}$$

Note that $\mathbb{E}[S^*] = \mathbb{E}[S^{1+\alpha}]/\mathbb{E}[S^\alpha]$. Equation (5.1.27) implies that, under the scaling regime nT_i (i.e. on every time interval $[0, t/n]$) and for large n, the $\Delta_{(i)}^\alpha/G/1$ queue is approximated by a $\Delta_{(i)}/G/1$ queue with arrival rate parameter $\lambda^* = \lambda\mathbb{E}[S^\alpha]$ and a different service-time distribution S^*.

We are now able to show how this approximation motivates assumption (5.1.3). Recall from (2.1.2) that the heavy-traffic condition for the standard $\Delta_{(i)}/G/1$ queue is

$$\lambda^*\mathbb{E}[S^*](1 + \beta n^{-1/3}) = 1 + \beta n^{-1/3}. \tag{5.1.29}$$

Rewriting the left-hand side of (5.1.29) in terms of λ and S gives

$$\begin{aligned} \lambda^*\mathbb{E}[S^*](1 + \beta n^{-1/3}) &= \lambda\mathbb{E}[S^\alpha]\frac{\mathbb{E}[S^{1+\alpha}]}{\mathbb{E}[S^\alpha]}(1 + \beta n^{-1/3}) \\ &= \lambda\mathbb{E}[S^{1+\alpha}](1 + \beta n^{-1/3}), \end{aligned} \tag{5.1.30}$$

as in (5.1.3). The above computations explain the unusual heavy-traffic assumption (5.1.3), but also suggest that, if the service distribution and arrival times are suitably chosen, Theorem 17 should follow from Theorem 3 of Chapter 2. However, (5.1.23) is a *marginal* convergence result. For Theorem 17 we consider the first $tn^{2/3}$ arrival-service pairs, and for these we cannot prove a joint convergence like (5.1.23). In fact, it is this subtle dependence structure that gives rise to the quadratic drift in (5.1.16). We illustrate this by comparing the two limiting processes for the $\Delta_{(i)}/G/1$ queue and the $\Delta_{(i)}^{\alpha}/G/1$ queue. Equation (5.1.19) gives the general expression of the limit process for a $\Delta_{(i)}/G/1$ queue. We apply this to a $\Delta_{(i)}/G/1$ queue with rate λ^* exponential arrival clocks and service requests given by S^*. We see that the limit process is

$$\begin{aligned}\widehat{N}^*(t) &= \beta t - \frac{1}{2}t^2 + \lambda^* \sqrt{\mathbb{E}[S^{*2}]}W(t)\\ &= \beta t - \frac{1}{2}t^2 + \lambda\mathbb{E}[S^\alpha]\sqrt{\frac{\mathbb{E}[S^{2+\alpha}]}{\mathbb{E}[S^\alpha]}}W(t)\\ &= \beta t - \frac{1}{2}t^2 + \frac{\sqrt{\mathbb{E}[S^\alpha]\mathbb{E}[S^{2+\alpha}]}}{\mathbb{E}[S^{1+\alpha}]}W(t). \qquad (5.1.31)\end{aligned}$$

The variance of the Brownian motion is predicted correctly (compare with (5.1.16) and (5.1.21)), but the coefficient of the quadratic drift, accounting for the depletion-of-points effect, is *not*. The approximation only captures the leading order behavior of the queue, yielding for example the correct heavy-traffic assumption.

5.1.3 Numerical results

We now use Theorem 18 to obtain numerical results for the first busy period. We also use the explicit expression of the probability density function of the first passage time of zero of $\phi(\widehat{N})$ obtained by Martin-Löf [72], see also [45]. Let $\mathrm{Ai}(x)$ and $\mathrm{Bi}(x)$ denote the classical Airy functions [1]. We recall that the first passage time of zero of $\widehat{N}(t) = q + \beta t - 1/2t^2 + \sigma W(t)$ has probability density [72]

$$\begin{aligned}f(t;\beta,\sigma) = {}& \mathrm{e}^{-((t-\beta)^3+\beta^3)/6\sigma^2-\beta a} \qquad (5.1.32)\\ &\times \int_{-\infty}^{+\infty} \mathrm{e}^{tu}\frac{\mathrm{Bi}(cu)\mathrm{Ai}(c(u-a)) - \mathrm{Ai}(cu)\mathrm{Bi}(c(u-a))}{\pi(\mathrm{Ai}(cu)^2+\mathrm{Bi}(cu)^2)}\mathrm{d}u,\end{aligned}$$

where $c = (2\sigma^2)^{1/3}$ and $a = q/\sigma^2 > 0$. The result (5.1.32) can be extended to a diffusion with a general quadratic drift through the scaling relation

$\widehat{N}(\tau^2 t) = \tau(q/\tau + \beta\tau t - \tau^3 t^2/2 + \sigma W(t))$. Figure 5.2 shows the empirical density of $n^{-2/3} T_{Q_n^e}$, for increasing values of n and various values of α, together with the exact limiting value (5.1.32). Table 5.1 shows the mean busy period for different choices of α and different service time distributions. We computed the exact value for $n = \infty$ by numerically integrating (5.1.32). Observe that $\mathbb{E}[T_{Q_n^e}]$ decreases with α. This might seem counterintuitive, because the larger α, the more likely customers with larger service join the queue early, who in turn might initiate a large busy period. Let us explain this apparent contradiction. When the arrival rate λ is fixed, assumption (5.1.3) does not necessarily hold and $\mathbb{E}[T_{Q_n^e}]$ increases with α, as can be seen in Table 5.2. However, our heavy-traffic condition (5.1.3) implies that λ depends on α since $\lambda = 1/\mathbb{E}[S^{1+\alpha}]$. The interpretation of condition (5.1.3) is that, on average, one customer joins the queue during one service time. Notice that, due to the size-biasing,

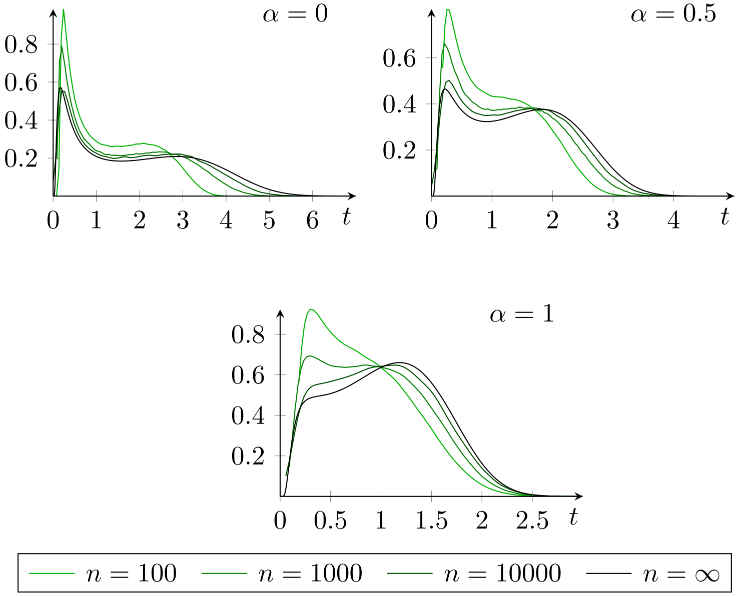

Figure 5.2: Density plot (black) and Gaussian kernel density estimates (colored) of the first busy period length. The plots for finite n were obtained by running 10^6 simulations of the $\Delta_{(i)}^{\alpha}/G/1$ queue. In all cases, the service times are exponentially distributed and $q = \beta = \mathbb{E}[S] = 1$.

	Deterministic	Exponential			Hyperexponential		
α	$0, 1/2, 1$	0	$1/2$	1	0	$1/2$	1
n							
10	1.1318	1.0359	0.8980	0.7429	0.8920	0.6356	0.5332
100	1.5842	1.3584	1.0924	0.8333	1.0959	0.7454	0.5525
1000	1.9188	1.6387	1.2506	0.9284	1.2936	0.8352	0.6134
10000	2.1474	1.8419	1.3925	1.0014	1.4960	0.9210	0.6554
∞	2.3374	2.0038	1.4719	1.0440	1.6242	0.9717	0.6881

Table 5.1: Numerical values of $n^{-2/3}\mathbb{E}[T_{Q^e_n}]$ for different population sizes and the exact expression for $n = \infty$ computed using (5.1.32). The service requirements are displayed in order of increasing coefficient of variation. In all cases $q = \beta = \mathbb{E}[S] = 1$. The hyperexponential service times follow a rate $\lambda_1 = 0.501$ exponential distribution with probability $p_1 = 1/2$ and a rate $\lambda_2 = 250.5$ exponential distribution with probability $p_2 = 1 - p_1 = 1/2$. Each value for finite n is the average of 10^4 simulations.

	Exponential				
α	0	$1/4$	$1/2$	$3/4$	1
n					
10	1.0854	1.0922	1.1053	1.1118	1.1306
100	5.9515	8.1928	11.4478	16.3598	22.0381

Table 5.2: Expected number of customers served in the first busy period of the *nonscaled* $\Delta^{\alpha}_{(i)}/G/1$ queue with mean one exponential service times and arrival rate $\lambda = 0.01$. In all cases $q = 1$. Each value is the average of 10^4 simulations.

the average service time is not $\mathbb{E}[S]$. Therefore, the number of customers that join during a (long) service is roughly equal to one as $\alpha \uparrow 1$. However, when customers with large services leave the system, they are not able to join any more. As $\alpha \uparrow 1$, customers with large services leave the system earlier. Therefore, as $\alpha \uparrow 1$, the resulting second order *depletion-of-points effect* causes shorter excursions as time progresses, see also Figure 1.10.

In the limit process, this phenomenon is represented by the fact that the coefficient of the negative quadratic drift increases as $\alpha \uparrow 1$, as shown in the following lemma:

Lemma 35. *Let*

$$\alpha \mapsto f(\alpha) := \frac{\mathbb{E}[S^{1+2\alpha}]}{\mathbb{E}[S^{\alpha}]\mathbb{E}[S^{1+\alpha}]}. \tag{5.1.33}$$

Then $f'(\alpha) \geq 0$.

Proof. Since

$$\begin{aligned} f'(\alpha) = {} & \frac{2\mathbb{E}[\log(S)S^{1+2\alpha}]}{\mathbb{E}[S^{\alpha}]\mathbb{E}[S^{1+\alpha}]} - \frac{\mathbb{E}[S^{1+2\alpha}]\mathbb{E}[\log(S)S^{\alpha}]}{\mathbb{E}[S^{\alpha}]^2\mathbb{E}[S^{1+\alpha}]} \\ & - \frac{\mathbb{E}[S^{1+2\alpha}]\mathbb{E}[\log(S)S^{1+\alpha}]}{\mathbb{E}[S^{\alpha}]\mathbb{E}[S^{1+\alpha}]^2}, \end{aligned} \tag{5.1.34}$$

$f'(\alpha) \geq 0$ if and only if

$$\begin{aligned} 2\mathbb{E}[\log(S)S^{1+2\alpha}]\mathbb{E}[S^{\alpha}]\mathbb{E}[S^{1+\alpha}] \geq {} & \mathbb{E}[S^{1+\alpha}]\mathbb{E}[S^{1+2\alpha}]\mathbb{E}[\log(S)S^{\alpha}] \\ & + \mathbb{E}[S^{\alpha}]\mathbb{E}[S^{1+2\alpha}]\mathbb{E}[\log(S)S^{1+\alpha}]. \end{aligned} \tag{5.1.35}$$

We split the left-hand side in two identical terms and show that each of them dominates one term on the right-hand side. That is

$$\mathbb{E}[\log(S)S^{1+2\alpha}]\mathbb{E}[S^{\alpha}]\mathbb{E}[S^{1+\alpha}] \geq \mathbb{E}[S^{1+\alpha}]\mathbb{E}[S^{1+2\alpha}]\mathbb{E}[\log(S)S^{\alpha}], \tag{5.1.36}$$

the proof of the second bound being analogous. The inequality (5.1.36) is equivalent to

$$\frac{\mathbb{E}[(\log(S)S^{1+\alpha})S^{\alpha}]}{\mathbb{E}[S^{\alpha}]} \geq \frac{\mathbb{E}[S^{1+\alpha}S^{\alpha}]}{\mathbb{E}[S^{\alpha}]}\frac{\mathbb{E}[\log(S)S^{\alpha}]}{\mathbb{E}[S^{\alpha}]}. \tag{5.1.37}$$

The term on the left and the two terms on the right can be rewritten as the expectation of a size-biased random variable W, so that (5.1.37) is equivalent to

$$\mathbb{E}[\log(W)W^{1+\alpha}] \geq \mathbb{E}[\log(W)]\mathbb{E}[W^{1+\alpha}]. \tag{5.1.38}$$

Finally, the inequality (5.1.38) holds because W is positive with probability one and $x \mapsto \log(x)$ and $x \mapsto x^{1+\alpha}$ are increasing functions [44, Lemma 2.14]. □

5.2 Overview of the proof

The proof of Theorem 17 uses the techniques developed in Chapter 2. However, the dependency structure of the arrival times complicate the analysis considerably. Customers with larger job sizes have a higher probability of joining the queue quickly, and this gives rise to a size-biased reordering of the service times.

5.2.1 Preliminaries

All the random variables that we consider are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For all our results, we condition on the entire sequence $(S_i)_{i=1}^n$. More precisely, we define a new probability space $(\Omega, \mathcal{F}_S, \mathbb{P}_S)$, with $\mathbb{P}_S(A) := \mathbb{P}(A|(S_i)_{i=1}^\infty)$ and $\mathcal{F}_S := \sigma(\{\mathcal{F}, (S_i)_{i=1}^\infty\})$, the σ-algebra generated by $\mathcal{F}$ and $(S_i)_{i=1}^\infty$. Correspondingly, for any random variable X on Ω we define $\mathbb{E}_S[X]$ as the expectation with respect to $\mathbb{P}_S$, and $\mathbb{E}[X]$ for the expectation with respect to $\mathbb{P}$. We say that a sequence of events $(\mathcal{E}_n)_{n=1}^\infty$ holds with high probability if $\mathbb{P}(\mathcal{E}_n) \to 1$ as $n \to \infty$.

The following well-known result will be useful on several occasions:

Lemma 36. *Assume $(X_i)_{i=1}^n$ is a sequence of positive i.i.d. random variables such that $\mathbb{E}[X] < \infty$. Then $\max_{i\in[n]} X_i = o_{\mathbb{P}}(n)$.*

Proof. We have the inclusion of events

$$\{\max_{i\in[n]} X_i \geq \varepsilon n\} \subseteq \bigcup_{i=1}^n \{X_i \geq \varepsilon n\}. \tag{5.2.1}$$

Therefore,

$$\mathbb{P}(\max_{i\in[n]} X_i \geq \varepsilon n) \leq \sum_{i=1}^n \mathbb{P}(X_i \geq \varepsilon n). \tag{5.2.2}$$

Since for any positive random variable Y, $\varepsilon \mathbb{1}_{\{Y\geq\varepsilon\}} \leq Y\mathbb{1}_{\{Y\geq\varepsilon\}}$ almost surely, it follows that

$$\mathbb{P}(\max_{i\in[n]} X_i \geq \varepsilon n) \leq \frac{\sum_{i=1}^n \mathbb{E}[X_i \mathbb{1}_{\{X_i\geq\varepsilon n\}}]}{\varepsilon n} = \frac{\mathbb{E}[X\mathbb{1}_{\{X\geq\varepsilon n\}}]}{\varepsilon}. \tag{5.2.3}$$

The right-most term tends to zero as $n \to \infty$ since $\mathbb{E}[X] < \infty$, and this concludes the proof. □

The intuitive notion that customers with larger service times are more likely to join earlier is formalized by the concept of *size-biased reordering*. Given a vector $\bar{x} = (x_1, x_2, \dots, x_n)$ with deterministic, real-valued entries, the size-biased ordering of $\bar{x}$ is a *random* vector $\bar{X} = (X_1, X_2, \dots, X_n)$ such that

$$\mathbb{P}(X_1 = x_j) = \frac{x_j}{\sum_{l=1}^n x_l}, \ \mathbb{P}(X_2 = x_j \mid X_1 = x_i) = \frac{x_j}{\sum_{l=1}^n x_l - x_1}, \ \dots \tag{5.2.4}$$

For any $\alpha \in \mathbb{R}$, the α-size-biased ordering of $\bar{x}$ is given by a vector $\bar{X}^{(\alpha)} = (X_1^{(\alpha)}, X_2^{(\alpha)}, \dots, X_n^{(\alpha)})$ such that

$$\mathbb{P}(X_1^{(\alpha)} = x_j) = \frac{x_j^\alpha}{\sum_{l=1}^n x_l^\alpha}, \ \mathbb{P}(X_2^{(\alpha)} = x_j \mid X_1^{(\alpha)} = x_i) = \frac{x_j^\alpha}{\sum_{l=1}^n x_l^\alpha - x_i^\alpha}, \ \dots \tag{5.2.5}$$

Finally, we denote by

$$\mathfrak{S}_k = \{c(1), \dots, c(k)\} \tag{5.2.6}$$

the set of the first k customers served. The following lemma is the first step in understanding the structure of the arrival process:

Lemma 37 (Size-biased reordering of the arrivals). *The order of appearance of customers is the α-size-biased ordering of their service times. In other words,*

$$\mathbb{P}_S(c(j) = i \mid \mathfrak{S}_{j-1}) = \frac{S_i^\alpha}{\sum_{l \notin \mathfrak{S}_{j-1}} S_l^\alpha}. \tag{5.2.7}$$

Proof. Conditioned on $(S_i)_{i=1}^n$, the arrival times are independent exponential random variables. By basic properties of exponentials, we have, for $i \notin \mathfrak{S}_{j-1}$,

$$\begin{aligned}\mathbb{P}_S&(c(j) = i \mid \mathfrak{S}_{j-1}) \\ &= \mathbb{P}_S(\min\{T_l : l \notin \mathfrak{S}_{j-1}\} = T_i \mid \mathfrak{S}_{j-1}) = \frac{S_i^\alpha}{\sum_{l \notin \mathfrak{S}_{j-1}} S_l^\alpha},\end{aligned} \tag{5.2.8}$$

concluding the proof. □

We remark that (5.2.7) differs from the classical size-biased reordering in that the weights are a *non-linear* function of the $(S_i)_{n=1}^n$.

The next lemma is crucial, establishing stochastic domination between the service requirements of the customers in order of appearance. In our definition of the queueing process (5.1.4)–(5.1.5), we do not keep track of the service requirements of the customers that join the queue, but only of their arrival times (5.1.2). Therefore, at the start of service, a customer's service requirement is a random variable that depends on the arrival time relative to the remaining customers. Lemma 37 then gives the precise distribution of the service requirement of the j-th customer entering service. Recall that X stochastically dominates Y (briefly $Y \preceq X$) if and only if there exists a probability space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$ and two random variables $\bar{X}, \bar{Y}$ defined on $\bar{\Omega}$ such that $\bar{X} \overset{\mathrm{d}}{=} X$, $\bar{Y} \overset{\mathrm{d}}{=} Y$ and $\bar{\mathbb{P}}(\bar{Y} \leq \bar{X}) = 1$.

Lemma 38. *Let $f : \mathbb{R}^+ \to \mathbb{R}$ be a function such that $\mathbb{E}[f(S)S^\alpha] < \infty$. Then there exists a constant $C_{f,S}$ such that almost surely, for n large enough,*

$$\mathbb{E}_S[f(S_{c(k)})] \leq C_{f,S} < \infty, \tag{5.2.9}$$

uniformly in $k \leq cn$, for a fixed $c \in (0,1)$.

Proof. We compute explicitly

$$\begin{aligned}
\mathbb{E}_S[f(S_{c(k)})] &= \mathbb{E}_S\Big[\frac{\sum_{j\notin\mathfrak{S}_{k-1}} f(S_j)S_j^\alpha}{\sum_{j\notin\mathfrak{S}_{k-1}} S_j^\alpha}\Big] \\
&= \mathbb{E}_S\Big[\frac{\sum_{j=1}^n f(S_j)S_j^\alpha - \sum_{j\in\mathfrak{S}_k} f(S_j)S_j^\alpha}{\sum_{j\notin\mathfrak{S}_{k-1}} S_j^\alpha}\Big] \\
&\leq \mathbb{E}_S\Big[\frac{1}{\sum_{j\notin\mathfrak{S}_{k-1}} S_j^\alpha}\Big] \sum_{j=1}^n f(S_j)S_j^\alpha. \qquad (5.2.10)
\end{aligned}$$

We have the almost sure bound

$$\begin{aligned}
\frac{1}{\sum_{j\notin\mathfrak{S}_{k-1}} S_j^\alpha} &= \frac{1}{\sum_{j=1}^n S_j^\alpha - \sum_{j\in\mathfrak{S}_{k-1}} S_j^\alpha} \leq \frac{1}{\sum_{j=1}^n S_j^\alpha - \sum_{j\in\mathfrak{S}_{k-1}} S_j^\alpha} \\
&\leq \frac{1}{\sum_{j=1}^n S_j^\alpha - \sum_{j=1}^{k-1} S_{(n-j+1)}^\alpha} = \frac{1}{\sum_{j=1}^{n-k+1} S_{(j)}^\alpha}, \qquad (5.2.11)
\end{aligned}$$

where $S_{(1)} \leq S_{(2)} \leq \ldots \leq S_{(n)}$ denote the order statistics of the sequence $(S_i)_{i=1}^n$. There exists $p \in (0,1)$ such that $n - k + 1 \geq pn$, for large enough n. Consequently,

$$\frac{1}{\sum_{j\notin\mathfrak{S}_{k-1}} S_j^\alpha} \leq \frac{1}{\sum_{j=1}^{\lfloor pn \rfloor} S_{(j)}^\alpha}, \tag{5.2.12}$$

so that

$$\mathbb{E}_S[f(S_{c(k)})] \leq \frac{\sum_{j=1}^{n} f(S_j)S_j^{\alpha}}{\sum_{j=1}^{\lfloor pn \rfloor} S_{(j)}^{\alpha}}. \tag{5.2.13}$$

Note that $S_{(\lfloor np \rfloor)} = F_{n,S}^{-1}(\lfloor np \rfloor / n)$, where $F_{n,S}(t) = \sum_{i=1}^{n} \mathbb{1}_{\{S_i \leq t\}}/n$ is the empirical distribution function of the $(S_i)_{i=1}^{n}$. Indeed, the assumption $f_S(\xi_p) > 0$ implies that $F_S(\cdot)$ is invertible in a neighborhood of ξ_p. Then, since $S_{(\lfloor pn \rfloor)} \overset{\text{a.s.}}{\to} \xi_p$ as $n \to \infty$,

$$\frac{1}{n}\Big|\sum_{j=1}^{n} S_j \mathbb{1}_{\{S_j \leq \xi_p\}} - \sum_{j=1}^{n} S_j \mathbb{1}_{\{S_j \leq S_{(\lfloor pn \rfloor)}\}}\Big| \overset{\text{a.s.}}{\to} 0, \tag{5.2.14}$$

as $n \to \infty$. Therefore, by the strong Law of Large Numbers, as $n \to \infty$,

$$\frac{\sum_{j=1}^{\lfloor pn \rfloor} S_{(j)}}{n} \overset{\text{a.s.}}{\to} \mathbb{E}[S\mathbb{1}_{\{S \leq \xi_p\}}]. \tag{5.2.15}$$

Then, choosing $C_{n,f,S} = \mathbb{E}[f(S)S^{\alpha}]/\mathbb{E}[S\mathbb{1}_{\{S \leq \xi_p\}}] + \varepsilon$, for an arbitrary $\varepsilon > 0$, gives the conclusion. □

When $\alpha > 0$ the proof of Lemma 38 shows that, uniformly in $k = O(n^{2/3})$,

$$\begin{aligned} \mathbb{E}_S[f(S_{c(k)})] &\leq \frac{\sum_{j=1}^{n} f(S_j)S_j^{\alpha}}{\sum_{j=1}^{\lfloor pn \rfloor} S_{(j)}^{\alpha}} \\ &= \frac{\sum_{j=1}^{n} f(S_j)S_j^{\alpha}}{\sum_{j=1}^{n} S_{(j)}^{\alpha}}\Big(1 + \frac{\sum_{j=\lfloor pn \rfloor}^{n} S_{(j)}^{\alpha}}{\sum_{j=1}^{\lfloor pn \rfloor} S_{(j)}^{\alpha}}\Big). \end{aligned} \tag{5.2.16}$$

Since $k = O(n^{2/3})$, we may take $p = p_n \to 1$ sufficiently slowly and therefore

$$\mathbb{E}_S[f(S_{c(k)})] \leq \mathbb{E}_S[f(S_{c(1)})](1 + o_{\mathbb{P}_S}(1)). \tag{5.2.17}$$

If $f(\cdot)$ is an increasing function, then (5.2.17) makes precise the intuition that, when $\alpha > 0$, customers with larger job sizes join the queue earlier. We will often make use of the expression (5.2.17).

The following lemma will often prove useful in dealing with sums over a random index set:

Lemma 39 (Uniform convergence of random sums). *Let $(S_j)_{j=1}^{\infty}$ be a sequence of positive random variables such that $\mathbb{E}[S^{2+\alpha}] < +\infty$, for $\alpha \in (0,1)$. Then,*

$$\sup_{\substack{\mathcal{X}\subseteq[n]\\ |\mathcal{X}|=O_{\mathbb{P}}(n^{2/3})}} \frac{1}{n}\sum_{j\in\mathcal{X}} S_j^{\alpha} = o_{\mathbb{P}}(1). \tag{5.2.18}$$

Proof. By Lemma 36, $\max_{j\in[n]} S_j^{\alpha} = o_{\mathbb{P}}(n^{\alpha/(2+\alpha)})$. Then,

$$\begin{aligned}&\sup_{\substack{\mathcal{X}\subseteq[n]\\ |\mathcal{X}|=O_{\mathbb{P}}(n^{2/3})}} \frac{1}{n}\sum_{j\in\mathcal{X}} S_j^{\alpha}\\ &\quad\leq \frac{\max_{j\in[n]} S_j^{\alpha}}{n^{1/3}} O_{\mathbb{P}}(1) = o_{\mathbb{P}}(n^{\frac{\alpha-2/3-\alpha/3}{2+\alpha}}) = o_{\mathbb{P}}(n^{\frac{2}{3}\frac{\alpha-1}{2+\alpha}}).\end{aligned} \tag{5.2.19}$$

Since $\alpha - 1 \leq 0$ by assumption, the claim is proven. □

We now focus on the i-th customer joining the queue (for i large) and characterize the distribution of its service time. In particular, for $\alpha > 0$ this is different from S_i.

Lemma 40 (Size-biased distribution of the service times). *Let $f(\cdot)$ be a bounded, continuous function. Then, as $n \to \infty$,*

$$\mathbb{E}_S[f(S_{c(i)}) \mid \mathcal{F}_{i-1}] \xrightarrow{\mathbb{P}} \frac{\mathbb{E}[f(S)S^{\alpha}]}{\mathbb{E}[S^{\alpha}]}, \tag{5.2.20}$$

uniformly for $i = O_{\mathbb{P}_S}(n^{2/3})$. Moreover, as $n \to \infty$,

$$\mathbb{E}_S[f(S_{c(i)})] \to \frac{\mathbb{E}[f(S)S^{\alpha}]}{\mathbb{E}[S^{\alpha}]}, \qquad \text{for } i = O_{\mathbb{P}_S}(n^{2/3}). \tag{5.2.21}$$

Proof. First note that

$$\begin{aligned}\mathbb{E}_S[f(S_{c(i)}) \mid \mathcal{F}_{i-1}] &= \sum_{j\notin\mathfrak{S}_{i-1}} f(S_j)\mathbb{P}_S(c(i) = j \mid \mathcal{F}_{i-1})\\ &= \sum_{j\notin\mathfrak{S}_{i-1}} \frac{f(S_j)S_j^{\alpha}}{\sum_{l\notin\mathfrak{S}_{i-1}} S_l^{\alpha}}.\end{aligned} \tag{5.2.22}$$

This can be further decomposed as

$$\mathbb{E}_S[f(S_{c(i)}) \mid \mathcal{F}_{i-1}] = \frac{\sum_{j=1}^{n} f(S_j)S_j^{\alpha} - \sum_{j\in\mathfrak{S}_{i-1}} f(S_j)S_j^{\alpha}}{\sum_{l=1}^{n} S_l^{\alpha} - \sum_{l\in\mathfrak{S}_{i-1}} S_l^{\alpha}}. \tag{5.2.23}$$

Since $|\mathfrak{S}_{i-1}| = i - 1$ and $i = O_{\mathbb{P}}(n^{2/3})$, by the Law of Large Numbers and Lemma 39,

$$\frac{\sum_{j \notin \mathfrak{S}_{i-1}} f(S_j) S_j^{\alpha}}{n} \xrightarrow{\mathbb{P}} \mathbb{E}[f(S)S^{\alpha}], \qquad \frac{\sum_{l \notin \mathfrak{S}_{i-1}} S_l^{\alpha}}{n} \xrightarrow{\mathbb{P}} \mathbb{E}[S^{\alpha}]. \tag{5.2.24}$$

uniformly in $i = O_{\mathbb{P}}(n^{2/3})$. This gives the first claim.

Next, we bound $\mathbb{E}_S[f(S_{c(i)}) \mid \mathcal{F}_{i-1}]$ as

$$\mathbb{E}_S[f(S_{c(i)}) \mid \mathcal{F}_{i-1}] = \sum_{j \notin \mathfrak{S}_{i-1}} \frac{f(S_j) S_j^{\alpha}}{\sum_{l \notin \mathfrak{S}_{i-1}} S_l^{\alpha}} \leq \sup_{x \geq 0} f(x) < \infty. \tag{5.2.25}$$

Since $\mathbb{E}_S[f(S_{c(i)}] = \mathbb{E}_S[\mathbb{E}_S[f(S_{c(i)}) \mid \mathcal{F}_{i-1}]]$, using (5.2.20) and the Dominated Convergence Theorem the second claim follows. □

In Lemma 40 we have studied the distribution of the service time of the i-th customer, and we now focus on its (conditional) moments. The following lemma should be interpreted as follows: Because of the size-biased re-ordering of the customer arrivals, the service time of the i-th customer being served (for i large) is highly concentrated:

Lemma 41. *For any fixed $\gamma \in [-1, 1]$,*

$$\mathbb{E}_S[S_{c(i)}^{1+\gamma} \mid \mathcal{F}_{i-1}] = \frac{\mathbb{E}[S^{1+\gamma+\alpha}]}{\mathbb{E}[S^{\alpha}]} + o_{\mathbb{P}}(1) \quad \text{for } i = O_{\mathbb{P}_S}(n^{2/3}), \tag{5.2.26}$$

where the error term is uniform in $i = O_{\mathbb{P}_S}(n^{2/3})$. Moreover, the convergence holds in L^1, i.e.

$$\mathbb{E}_S\left[\left|\mathbb{E}_S[S_{c(i)}^{1+\gamma} \mid \mathcal{F}_{i-1}] - \frac{\mathbb{E}[S^{1+\gamma+\alpha}]}{\mathbb{E}[S^{\alpha}]}\right|\right] = o_{\mathbb{P}}(1), \tag{5.2.27}$$

uniformly in $i = O_{\mathbb{P}_S}(n^{2/3})$.

Proof. In order to apply Lemma 40, we first split

$$S_{c(i)}^{1+\gamma} = (S_{c(i)} \wedge K)^{1+\gamma} + ((S_{c(i)} - K)^{+})^{1+\gamma}, \tag{5.2.28}$$

where $K > 0$ is arbitrary, so that

$$\begin{aligned} &\mathbb{E}_S[S_{c(i)}^{1+\gamma} \mid \mathcal{F}_{i-1}] \\ &\quad = \mathbb{E}_S[(S_{c(i)} \wedge K)^{1+\gamma} \mid \mathcal{F}_{i-1}] + \mathbb{E}_S[((S_{c(i)} - K)^{+})^{1+\gamma} \mid \mathcal{F}_{i-1}]. \end{aligned} \tag{5.2.29}$$

The first term is bounded, and thus converges to $\mathbb{E}[(S \wedge K)^{1+\gamma} S^{\alpha}]/\mathbb{E}[S^{\alpha}]$ by Lemma 40. The second term can be bounded through Markov's inequality, as

$$\mathbb{P}_S(\mathbb{E}_S[((S_{c(i)} - K)^+)^{1+\gamma} \mid \mathcal{F}_{i-1}] \geq \varepsilon) \leq \frac{\mathbb{E}_S[((S_{c(i)} - K)^+)^{1+\gamma}]}{\varepsilon}. \tag{5.2.30}$$

Applying Lemma 38 to the function $f(x) = f_K(x) = ((x-K)^+)^{1+\gamma}$ we get

$$\mathbb{E}_S[((S_{c(i)} - K)^+)^{1+\gamma}] \leq C_{f_K,S}. \tag{5.2.31}$$

Therefore,

$$\begin{aligned}&\Big|\mathbb{E}_S[S_{c(i)}^{1+\gamma} | \mathcal{F}_{i-1}] - \frac{\mathbb{E}[S^{1+\gamma+\alpha}]}{\mathbb{E}[S^{\alpha}]}\Big| \\ &\qquad \leq \Big|\mathbb{E}_S[(S_{c(i)} \wedge K)^{1+\gamma} \mid \mathcal{F}_{i-1}] - \frac{\mathbb{E}[S^{1+\gamma+\alpha}]}{\mathbb{E}[S^{\alpha}]}\Big| + C_{f_K,S}.\end{aligned} \tag{5.2.32}$$

The proof of Lemma 38 shows that, for any $\varepsilon > 0$, $\lim_{K\to\infty} C_{f_K,S} \leq \varepsilon$, thus $\lim_{K\to\infty} C_{f_K,S} = 0$. Therefore, by letting $K \to \infty$ in (5.2.32), the claim (5.2.26) follows. Next split

$$\begin{aligned}&\mathbb{E}_S\Big[\Big|\mathbb{E}_S[S_{c(i)}^{1+\gamma} \mid \mathcal{F}_{i-1}] - \frac{\mathbb{E}[S^{1+\gamma+\alpha}]}{\mathbb{E}[S^{\alpha}]}\Big|\Big] \\ &\leq \mathbb{E}_S\Big[\Big|(S_{c(i)} \wedge K)^{1+\gamma} - \frac{\mathbb{E}[S^{1+\gamma+\alpha}]}{\mathbb{E}[S^{\alpha}]}\Big|\Big] + \mathbb{E}_S[((S_{c(i)} - K)^+)^{1+\gamma}].\end{aligned} \tag{5.2.33}$$

The second term is bounded as in (5.2.31). For the first term,

$$\begin{aligned}\mathbb{E}_S\Big[\Big|(S_{c(i)} \wedge K)^{1+\gamma} - \frac{\mathbb{E}[S^{1+\gamma+\alpha}]}{\mathbb{E}[S^{\alpha}]}\Big|\Big] &\leq \Big|\frac{\sum_{j=1}^{n}(S_j \wedge K)^{1+\gamma} S_j^{\alpha}}{\sum_{j=1}^{n} S_j^{\alpha}} - \frac{\mathbb{E}[S^{1+\gamma+\alpha}]}{\mathbb{E}[S^{\alpha}]}\Big| \\ &+ \mathbb{E}_S\Big[\Big|\frac{\sum_{j=1}^{n}(S_j \wedge K)^{1+\gamma} S_j^{\alpha} \sum_{l\in\mathfrak{S}_{i-1}} S_l^{\alpha}}{(\sum_{j=1}^{n} S_j^{\alpha})^2}\Big|\Big] \\ &+ \mathbb{E}_S\Big[\Big|\frac{\sum_{l=1}^{n} S_l^{\alpha} \sum_{j\in\mathfrak{S}_{i-1}}(S_j \wedge K)^{1+\gamma} S_j^{\alpha}}{(\sum_{j=1}^{n} S_j^{\alpha})^2}\Big|\Big],\end{aligned} \tag{5.2.34}$$

where we have used that $|(a-b)/(c-d) - a/c| \leq ad/c^2 + bc/c^2$, for positive a, b, c, d. The second and third terms converge uniformly over

$i = O_{\mathbb{P}_S}(n^{2/3})$ by Lemma 39. Summarizing,

$$\begin{aligned}\mathbb{E}_S\Big[\Big|\mathbb{E}_S[S_{c(i)}^{1+\gamma} \mid \mathcal{F}_{i-1}] - \frac{\mathbb{E}[S^{1+\gamma+\alpha}]}{\mathbb{E}[S^\alpha]}\Big|\Big]& \\ \leq \Big|\frac{\sum_{j=1}^n (S_j \wedge K)^{1+\gamma} S_j^\alpha}{\sum_{j=1}^n S_j^\alpha} &- \frac{\mathbb{E}[S^{1+\gamma+\alpha}]}{\mathbb{E}[S^\alpha]}\Big| \\ + \frac{\sum_{l=1}^n ((S_l - K)^+)^{1+\gamma}}{\sum_{j=1}^n S_j^\alpha} &+ o_{\mathbb{P}}(1). \qquad (5.2.35)\end{aligned}$$

Letting first $n \to \infty$ and then $K \to \infty$, the claim (5.2.27) follows. □

We will make use of Lemma 41 several times throughout the proof, with the specific choices $\gamma \in \{0, \alpha, 1\}$. The following lemma is of central importance in the proof of the uniform convergence of the quadratic drift:

Lemma 42. *As $n \to \infty$,*

$$n^{-2/3} \sup_{j \leq tn^{2/3}} \Big|\sum_{i=1}^j \Big(S_{c(i)}^{1+\alpha} - \frac{\mathbb{E}[S^{1+2\alpha}]}{\mathbb{E}[S]}\Big)\Big| \xrightarrow{\mathbb{P}} 0. \qquad (5.2.36)$$

Proof. By Lemma 41, (5.2.36) is equivalent to

$$n^{-2/3} \sup_{j \leq tn^{2/3}} \Big|\sum_{i=1}^j \Big(S_{c(i)}^{1+\alpha} - \mathbb{E}[S_{c(i)}^{1+\alpha} \mid \mathcal{F}_{i-1}]\Big)\Big| \xrightarrow{\mathbb{P}} 0. \qquad (5.2.37)$$

We split the event space to bound separately

$$n^{-2/3} \sup_{j \leq tn^{2/3}} \Big|\sum_{i=1}^j \Big(S_{c(i)}^{1+\alpha} \mathbb{1}_{\{S_{c(i)}^{1+\alpha} \leq K_n\}} - \mathbb{E}[S_{c(i)}^{1+\alpha} \mathbb{1}_{\{S_{c(i)}^{1+\alpha} \leq K_n\}} \mid \mathcal{F}_{i-1}]\Big)\Big| \qquad (5.2.38)$$

and

$$n^{-2/3} \sup_{j \leq tn^{2/3}} \Big|\sum_{i=1}^j \Big(S_{c(i)}^{1+\alpha} \mathbb{1}_{\{S_{c(i)}^{1+\alpha} > K_n\}} - \mathbb{E}[S_{c(i)}^{1+\alpha} \mathbb{1}_{\{S_{c(i)}^{1+\alpha} > K_n\}} \mid \mathcal{F}_{i-1}]\Big)\Big|, \qquad (5.2.39)$$

for a yet unspecified sequence $(K_n)_{n=1}^\infty$ such that $K_n \to \infty$. We start with (5.2.38). Since the sum inside the absolute value is a martingale as a function of j, (5.2.38) can be bounded through Doob's L^2 martingale inequality [63, Theorem 11.2] as

$$\begin{aligned}
&\mathbb{P}_S\Big(\sup_{j\le tn^{2/3}}\Big|\sum_{i=1}^{j}\Big(S_{c(i)}^{1+\alpha}\mathbb{1}_{\{S_{c(i)}^{1+\alpha}\le K_n\}}-\mathbb{E}_S[S_{c(i)}^{1+\alpha}\mathbb{1}_{\{S_{c(i)}^{1+\alpha}\le K_n\}}\mid\mathcal{F}_{i-1}]\Big)\Big|\ge\varepsilon n^{2/3}\Big)\\
&\le\frac{1}{\varepsilon n^{4/3}}\mathbb{E}_S\Big[\sum_{i=1}^{tn^{2/3}}(S_{c(i)}^{1+\alpha}\mathbb{1}_{\{S_{c(i)}^{1+\alpha}\le K_n\}}-\mathbb{E}_S[S_{c(i)}^{1+\alpha}\mathbb{1}_{\{S_{c(i)}^{1+\alpha}\le K_n\}}\mid\mathcal{F}_{i-1}])^2\Big]\\
&\le\frac{2}{\varepsilon n^{4/3}}\sum_{i=1}^{tn^{2/3}}\mathbb{E}_S[S_{c(i)}^{2+2\alpha}\mathbb{1}_{\{S_{c(i)}^{1+\alpha}\le K_n\}}]\le\frac{2}{\varepsilon n^{4/3}}\sum_{i=1}^{tn^{2/3}}K_n^{2\alpha}\mathbb{E}_S[S_{c(i)}^2]. \qquad (5.2.40)
\end{aligned}$$

Using Lemma 41 we approximate $\mathbb{E}_S[S_{c(i)}^2]$ uniformly by $\mathbb{E}[S^{2+\alpha}]/\mathbb{E}[S^\alpha]$, obtaining

$$\frac{2}{\varepsilon n^{4/3}}\sum_{i=1}^{tn^{2/3}}\Big(K_n^{2\alpha}\frac{\mathbb{E}[S^{2+\alpha}]}{\mathbb{E}[S^\alpha]}+o_{\mathbb{P}}(1)\Big)=\frac{tK_n^{2\alpha}}{\varepsilon n^{2/3}}O_{\mathbb{P}}(1), \qquad (5.2.41)$$

which converges to zero as $n\to\infty$ if and only if $K_n^\alpha/n^{1/3}$ converges to zero. We now turn to (5.2.39) and apply Doob's L^1 martingale inequality [63, Theorem 11.2] to obtain

$$\begin{aligned}
&\mathbb{P}_S\Big(\sup_{j\le tn^{2/3}}\Big|\sum_{i=1}^{j}\Big(S_{c(i)}^{1+\alpha}\mathbb{1}_{\{S_{c(i)}^{1+\alpha}>K_n\}}-\mathbb{E}_S[S_{c(i)}^{1+\alpha}\mathbb{1}_{\{S_{c(i)}^{1+\alpha}>K_n\}}\mid\mathcal{F}_{i-1}]\Big)\Big|\ge\varepsilon n^{2/3}\Big)\\
&\le\frac{1}{\varepsilon n^{2/3}}\mathbb{E}_S\Big[\Big|\sum_{i=1}^{tn^{2/3}}(S_{c(i)}^{1+\alpha}\mathbb{1}_{\{S_{c(i)}^{1+\alpha}>K_n\}}-\mathbb{E}_S[S_{c(i)}^{1+\alpha}\mathbb{1}_{\{S_{c(i)}^{1+\alpha}>K_n\}}\mid\mathcal{F}_{i-1}])\Big|\Big]\\
&\le\frac{2}{\varepsilon n^{2/3}}\sum_{i=1}^{tn^{2/3}}\mathbb{E}_S[S_{c(i)}^{1+\alpha}\mathbb{1}_{\{S_{c(i)}^{1+\alpha}>K_n\}}]\\
&\le\frac{2}{\varepsilon n^{2/3}}\sum_{i=1}^{tn^{2/3}}\mathbb{E}_S[S_{c(1)}^{1+\alpha}\mathbb{1}_{\{S_{c(1)}^{1+\alpha}>K_n\}}](1+O_{\mathbb{P}_S}(1))\\
&=\frac{2t}{\varepsilon}\mathbb{E}_S[S_{c(1)}^{1+\alpha}\mathbb{1}_{\{S_{c(1)}^{1+\alpha}>K_n\}}](1+O_{\mathbb{P}_S}(1))=o_{\mathbb{P}}(1). \qquad (5.2.42)
\end{aligned}$$

We have used Lemma 41 in the second inequality, and Lemma 38 with $f(x)=x^{1+\alpha}\mathbb{1}_{\{x^{1+\alpha}>K_n\}}$ in the third. The right-most term in (5.2.42) is $o_{\mathbb{P}}(1)$ as $n\to\infty$ by the strong Law of Large Numbers. Note that this side of the bound does not impose additional conditions on K_n, so that, if we take $K_n = n^c$, it is sufficient that $c < 1/3\alpha$, with the convention that $1/0 = \infty$. □

We conclude this section with a technical lemma concerning error terms in the computations of quadratic variations. Denote the density (resp. distribution function) of a rate λ exponential random variable by $f_E(\cdot)$ (resp. $F_E(\cdot)$):

Lemma 43. *We have that*

$$\mathbb{E}_S\Big[\sum_{h,q=1}^{n}\Big|F_E\Big(\frac{S_{c(i)}S_h^{\alpha}}{n}\Big)-\frac{\lambda S_{c(i)}S_h^{\alpha}}{n}\Big| \\ \times\Big|F_E\Big(\frac{S_{c(i)}S_q^{\alpha}}{n}\Big)-\frac{\lambda S_{c(i)}S_q^{\alpha}}{n}\Big|\mid\mathcal{F}_{i-1}\Big]=o_{\mathbb{P}}(1) \tag{5.2.43}$$

uniformly in $i = O(n^{2/3})$.

Proof. Since $|F_E(x) - x| = O(x^2)$, the bound

$$|\lambda S_{c(i)}S_h^{\alpha}/n - F_E(S_{c(i)}S_h^{\alpha}/n)| \leq C(S_{c(i)}S_h^{\alpha}/n)^{1+\varepsilon} \tag{5.2.44}$$

holds almost surely for $0 < \varepsilon < 1$ and $C > 0$, giving

$$\lambda^2\sum_{h,q=1}^{n}\mathbb{E}_S\Big[\Big(\frac{S_{c(i)}S_h^{\alpha}}{n}\Big)^{1+\varepsilon}\Big(\frac{S_q^{\alpha}S_{c(i)}}{n}\Big)^{1+\varepsilon}\mid\mathcal{F}_{i-1}\Big] \\ =\frac{\lambda^2}{n^{2+2\varepsilon}}\sum_{h,q=1}^{n}\mathbb{E}_S[S_{c(i)}^{2+2\varepsilon}\mid\mathcal{F}_{i-1}]S_h^{\alpha(1+\varepsilon)}S_q^{\alpha(1+\varepsilon)}. \tag{5.2.45}$$

Therefore,

$$\lambda^2\sum_{h,q=1}^{n}\mathbb{E}_S\Big[\Big(\frac{S_{c(i)}S_h^{\alpha}}{n}\Big)^{1+\varepsilon}\Big(\frac{S_q^{\alpha}S_{c(i)}}{n}\Big)^{1+\varepsilon}\mid\mathcal{F}_{i-1}\Big] \tag{5.2.46}$$

$$\leq\frac{\lambda^2}{n^{2+2\varepsilon}}\max_{j\in[n]}S_j^{2\varepsilon}\mathbb{E}_S[S_{c(i)}^2\mid\mathcal{F}_{i-1}]\sum_{h,q=1}^{n}S_h^{\alpha(1+\varepsilon)}S_q^{\alpha(1+\varepsilon)} \\ \leq\frac{\lambda^2\mathbb{E}[S^{2+\alpha}]}{\mathbb{E}[S^{\alpha}]}\frac{\max_{j\in[n]}S_j^{2\varepsilon}}{n^{2\varepsilon}}\frac{1}{n^2}\sum_{h,q=1}^{n}S_h^{\alpha(1+\varepsilon)}S_q^{\alpha(1+\varepsilon)}+o_{\mathbb{P}}(1),$$

where in the last step we have used Lemma 41. Note that, since $\mathbb{E}[S^{2+\alpha}] < \infty$, Lemma 36 gives $\max_{j\in[n]} S_j^{2\varepsilon} = o_{\mathbb{P}}(n^{2\varepsilon/(2+\alpha)})$. The right-most term in (5.2.46) then tends to zero as n tends to infinity as long as $0 < \varepsilon < \min\{1, 2/\alpha\}$. □

5.3 Proof of the scaling limit

We first establish various preliminary estimates on $N_n(\cdot)$ that will be crucial for the proof of convergence. We will upper bound the process $N_n(\cdot)$ by a simpler process $N_n^u(\cdot)$ in such a way that the increments of $N_n^u(\cdot)$ almost surely dominate the increments of $N_n(\cdot)$. We also show that, after rescaling, $N_n^u(\cdot)$ converges in distribution to $\widehat{N}(\cdot)$. The process $N_n^u(\cdot)$ is defined as $N_n^u(0) = N_n(0)$, and

$$N_n^u(k) = N_n^u(k-1) + A_n^u(k) - 1, \tag{5.3.1}$$

where

$$A_n^u(k) = \sum_{i \notin \mathfrak{S}_k} \mathbb{1}_{\{T_i \leq c_{n,\beta} S_{c(k)}/n\}}, \tag{5.3.2}$$

with

$$c_{n,\beta} = 1 + \beta n^{-1/3}, \tag{5.3.3}$$

and, as before,

$$T_i \stackrel{\mathrm{d}}{=} \frac{E_i}{\lambda S_i^\alpha}. \tag{5.3.4}$$

An interpretation of the process $N_n^u(\cdot)$ is that customers are not removed from the pool of potential customers until they have been served. Therefore, a customer could potentially join the queue more than once. The processes $N_n(\cdot)$ and $N_n^u(\cdot)$ are coupled as follows. Consider a sequence of arrival times $(T_i)_{i=1}^\infty$ and of service times $(S_i)_{i=1}^\infty$, then define $A_n(\cdot)$ as (5.1.5) and $A_n^u(\cdot)$ as (5.3.2). With this coupling we have that, almost surely,

$$A_n(k) \leq A_n^u(k) \qquad \forall\, k \geq 0. \tag{5.3.5}$$

Consequently,

$$N_n(k) \leq N_n^u(k) \qquad \forall k \geq 0, \tag{5.3.6}$$

and

$$Q_n^e(k) = \phi(N_n)(k) \leq \phi(N_n^u)(k) =: Q_n^{e,u}(k) \qquad \forall k \geq 0, \tag{5.3.7}$$

almost surely. Note that the reflection map $\phi(\cdot)$ is not monotone, thus (5.3.6) does not imply (5.3.7). On the other hand, almost sure step-size domination (5.3.5) guarantees that (5.3.7) holds.

While in general only the upper bounds (5.3.6) and (5.3.7) hold, the processes $N_n(\cdot)$ and $N_n^u(\cdot)$ (resp. $Q_n^e(\cdot)$ and $Q_n^{e,u}(\cdot)$) turn out to be close to each other for large n. We first prove a convergence result for $N_n^u(\cdot)$

and $Q_n^{e,u}(\cdot)$ because they are easier to handle. This allows us to prove that identical results hold for $N_n(\cdot)$ and $Q_n^e(\cdot)$.

In fact, we introduce the upper bound $N_n^u(\cdot)$ to deal with the complicated index set for the summation in (5.1.5). The difficulty arises as follows: in order to estimate $N_n(\cdot)$ one has to estimate $A_n(\cdot)$. To do this, one has to separately (uniformly) bound each element in the sum, and also estimate the number of elements in the sum. The first goal is accomplished, for example, through Lemma 41, while for the second the crude upper bound n is not sharp enough. However, estimating $|\nu_k|$ requires an estimate on $N_n(\cdot)$ itself, as (5.1.6) shows. To solve this circularity, we introduce a bootstrap argument: first, we upper bound $N_n(\cdot)$ and we obtain estimates on the upper bound, from this follows an estimate on $|\nu_k|$, and this in turn allows us to estimate $N_n(\cdot)$.

This technique can be applied to solve a recently found technical issue in the proof of the main result of [15]. In fact, the authors in [15] prove convergence of a process which upper bounds the exploration process of the graph. Therefore, their main result is analogous to Theorem 19. However, a further step is required to complete the proof of convergence of the exploration process, and this is provided by our approach.

The convergence of the process $N_n^u(\cdot)$ is given in the following theorem:

Theorem 19 (Convergence of the upper bound)**.** *As $n \to \infty$,*

$$n^{-1/3}N_n^u(\cdot n^{2/3}) \xrightarrow{\mathrm{d}} \widehat{N}(\cdot) \qquad \text{in } (\mathcal{D}, J_1), \tag{5.3.8}$$

where $\widehat{N}(\cdot)$ is the diffusion process in (5.1.16)*. In particular, as $n \to \infty$,*

$$n^{-1/3}Q_n^{e,u}(\cdot n^{2/3}) \xrightarrow{\mathrm{d}} \phi(\widehat{N})(\cdot) \qquad \text{in } (\mathcal{D}, J_1). \tag{5.3.9}$$

The next section is dedicated to the proof of Theorem 19.

5.3.1 Convergence of the upper bound

We use a classical martingale decomposition followed by a Martingale Functional Central Limit Theorem (MFCLT). The process $N_n^u(\cdot)$ in (5.3.1) is decomposed as $N_n^u(k) = M_n^u(k) + C_n^u(k)$, where $M_n^u(\cdot)$ is a martingale and $C_n^u(\cdot)$ is a drift term, as follows:

$$M_n^u(k) = \sum_{i=1}^{k}(A_n^u(i) - \mathbb{E}_S[A_n^u(i) \mid \mathcal{F}_{i-1}]),$$
$$C_n^u(k) = \sum_{i=1}^{k}(\mathbb{E}_S[A_n^u(i) \mid \mathcal{F}_{i-1}] - 1). \tag{5.3.10}$$

Moreover, $M_n^u(k)^2$ is rewritten as $M_n^u(k)^2 = Z_n^u(k) + V_n^u(k)$ with $Z_n^u(k)$ a martingale and $V_n^u(k)$ the compensator, or quadratic variation, of $M_n^u(k)$ given by

$$V_n^u(k) = \sum_{i=1}^{k}(\mathbb{E}_S[(A_n^u(i))^2 \mid \mathcal{F}_{i-1}] - \mathbb{E}_S[A_n^u(i) \mid \mathcal{F}_{i-1}]^2). \tag{5.3.11}$$

In order to prove convergence of $N_n^u(\cdot)$ we separately prove convergence of $C_n^u(\cdot)$ and of $M_n^u(\cdot)$. We prove the former directly, and the latter by applying the MFCLT [38, Theorem 7.1.4]. For this, we need to verify the following conditions, for every $\bar{t} \in \mathbb{R}^+$,

(i) $\sup_{t\leq\bar{t}} |n^{-1/3}C_n^u(tn^{2/3}) - \beta t + \lambda\frac{\mathbb{E}[S^{1+2\alpha}]}{2\mathbb{E}[S^\alpha]}t^2| \xrightarrow{\mathbb{P}} 0;$

(ii) $n^{-2/3}V_n^u(\bar{t}n^{2/3}) \xrightarrow{\mathbb{P}} \sigma^2\bar{t};$

(iii) $\lim_{n\to\infty} n^{-2/3}\mathbb{E}_S[\sup_{t\leq\bar{t}} |V_n^u(tn^{2/3}) - V_n^u(tn^{2/3}-)|] = 0;$

(iv) $\lim_{n\to\infty} n^{-2/3}\mathbb{E}_S[\sup_{t\leq\bar{t}} |M_n^u(tn^{2/3}) - M_n^u(tn^{2/3}-)|^2] = 0.$

Proof of (i) for the upper bound.

First we obtain an explicit expression for $\mathbb{E}[A_n^u(i) \mid \mathcal{F}_{i-1}]$, as

$$\begin{aligned}
\mathbb{E}_S[A_n^u(i) \mid \mathcal{F}_{i-1}] &= \sum_{j \notin \mathfrak{S}_{i-1}} \mathbb{P}_S(c(i) = j \mid \mathcal{F}_{i-1}) \sum_{l \notin \mathfrak{S}_{i-1} \cup \{j\}} F_E\Big(c_{n,\beta} \frac{S_j S_l^\alpha}{n}\Big) \\
&= \sum_{j \notin \mathfrak{S}_{i-1}} \mathbb{P}_S(c(i) = j \mid \mathcal{F}_{i-1}) \sum_{l=1}^{n} c_{n,\beta} \lambda \frac{S_j S_l^\alpha}{n} \\
&\quad - \sum_{j \notin \mathfrak{S}_{i-1}} \mathbb{P}_S(c(i) = j \mid \mathcal{F}_{i-1}) \sum_{l \in \mathfrak{S}_{i-1} \cup \{j\}} c_{n,\beta} \lambda \frac{S_j S_l^\alpha}{n} \\
&\quad + \sum_{j \notin \mathfrak{S}_{i-1}} \mathbb{P}_S(c(i) = j \mid \mathcal{F}_{i-1}) \\
&\quad \times \sum_{l \notin \mathfrak{S}_{i-1} \cup \{j\}} \Big(F_E\Big(c_{n,\beta} \frac{S_j S_l^\alpha}{n}\Big) - c_{n,\beta} \lambda \frac{S_j S_l^\alpha}{n}\Big). \qquad (5.3.12)
\end{aligned}$$

The third term is an error term. Indeed, for some $\zeta_n \in [0, S_{c(i)} S_l / n]$,

$$\begin{aligned}
&\mathbb{E}_S\Big[\Big| \sum_{l \notin \mathfrak{S}_{i-1} \cup \{j\}} F_E\Big(\frac{S_{c(i)} S_l^\alpha}{n}\Big) - \lambda \frac{S_{c(i)} S_l^\alpha}{n}\Big| \mid \mathcal{F}_{i-1}\Big] \qquad (5.3.13) \\
&\leq \sum_{l=1}^{n} \mathbb{E}_S\Big[\Big| F_E\Big(\frac{S_{c(i)} S_l^\alpha}{n}\Big) - \lambda \frac{S_{c(i)} S_l^\alpha}{n}\Big| \mid \mathcal{F}_{i-1}\Big] \\
&= \frac{1}{2n^2} \mathbb{E}_S[|F_E''(\zeta_n) S_{c(i)}^2| \mid \mathcal{F}_{i-1}] \sum_{l=1}^{n} S_l^{2\alpha} \leq \frac{\lambda^2}{2n^2} \mathbb{E}_S[S_{c(i)}^2 \mid \mathcal{F}_{i-1}] \sum_{l=1}^{n} S_l^{2\alpha},
\end{aligned}$$

since $|F_E''(x)| \leq \lambda^2$ for all $x \geq 0$. By Lemma 41 this is bounded by

$$\frac{\lambda^2}{2n^2} (c_n + o_{\mathbb{P}}(1)) \sum_{l=1}^{n} S_l^{2\alpha}, \qquad (5.3.14)$$

where c_n is bounded with high probability and the $o_{\mathbb{P}}(1)$ term is uniform in $i = O(n^{2/3})$. Therefore, the third term in (5.3.12) is $o_{\mathbb{P}}(n^{-1/3})$. The remaining terms in (5.3.12) are simplified as

$$\begin{aligned}
\mathbb{E}_S[A_n^u(i)|\mathcal{F}_{i-1}] - 1 &= c_{n,\beta}\lambda \sum_{j\notin\mathfrak{S}_{i-1}} \mathbb{P}_S(c(i)=j\mid\mathcal{F}_{i-1}) S_j \frac{\sum_{l=1}^n S_l^\alpha}{n} \\
&\quad - c_{n,\beta}\lambda \sum_{j\notin\mathfrak{S}_{i-1}} \mathbb{P}_S(c(i)=j\mid\mathcal{F}_{i-1}) \sum_{l\in\mathfrak{S}_{i-1}} S_j\frac{S_l^\alpha}{n} \\
&\quad - c_{n,\beta}\lambda \sum_{j\notin\mathfrak{S}_{i-1}} \mathbb{P}_S(c(i)=j\mid\mathcal{F}_{i-1}) \frac{S_j^{1+\alpha}}{n} - 1 + o_{\mathbb{P}}(n^{-1/3}) \\
&= \Big(c_{n,\beta}\lambda\frac{\sum_{l=1}^n S_l^\alpha}{n}\mathbb{E}[S_{c(i)}|\mathcal{F}_{i-1}] - 1\Big) - c_{n,\beta}\mathbb{E}_S[S_{c(i)}\mid\mathcal{F}_{i-1}]\sum_{l\in\mathfrak{S}_{i-1}}\lambda\frac{S_l^\alpha}{n} \\
&\quad - c_{n,\beta}\frac{\lambda}{n}\mathbb{E}_S[S_{c(i)}^{1+\alpha}\mid\mathcal{F}_{i-1}] + o_{\mathbb{P}}(n^{-1/3}). \qquad (5.3.15)
\end{aligned}$$

For the first term of (5.3.15), using $\frac{c}{a-b} = \frac{c}{a} + \frac{c}{a-b}\frac{b}{a}$, with $a = \sum_{l=1}^n S_l^\alpha$ and $b = \sum_{l\in\mathfrak{S}_{i-1}} S_l^\alpha$, we get

$$\begin{aligned}
&c_{n,\beta}\lambda\frac{\sum_{l=1}^n S_l^\alpha}{n}\mathbb{E}_S[S_{c(i)}\mid\mathcal{F}_{i-1}] - 1 \qquad (5.3.16)\\
&\quad= c_{n,\beta}\lambda\frac{\sum_{l=1}^n S_l^\alpha}{n}\sum_{j\notin\mathfrak{S}_{i-1}}\frac{S_j^{1+\alpha}}{\sum_{l=1}^n S_l^\alpha} - 1 \\
&\qquad+ c_{n,\beta}\lambda\frac{\sum_{l=1}^n S_l^\alpha}{n}\sum_{j\notin\mathfrak{S}_{i-1}}\frac{S_j^{1+\alpha}}{\sum_{l\notin\mathfrak{S}_{i-1}} S_l^\alpha}\frac{\sum_{s\in\mathfrak{S}_{i-1}} S_s^\alpha}{\sum_{l=1}^n S_l^\alpha} \\
&\quad= \Big(c_{n,\beta}\frac{\lambda}{n}\sum_{j\notin\mathfrak{S}_{i-1}} S_j^{1+\alpha} - 1\Big) + c_{n,\beta}\mathbb{E}_S[S_{c(i)}\mid\mathcal{F}_{i-1}]\sum_{s\in\mathfrak{S}_{i-1}}\lambda\frac{S_s^\alpha}{n}.
\end{aligned}$$

Note that the right-most term in (5.3.16) and the second term in (5.3.15) cancel out. This cancellation is what makes the analysis of $N_n^u(\cdot)$ considerably easier than the analysis of $N_n(\cdot)$.

Moreover, Lemma 41 implies that the third term in (5.3.15) is also

$o_{\mathbb{P}}(n^{-1/3})$. The expression in (5.3.12) is then simplified to

$$\begin{aligned}
&\mathbb{E}_S[A_n^u(i) \mid \mathcal{F}_{i-1}] - 1 \\
&\quad = c_{n,\beta}\frac{\lambda}{n}\sum_{j \notin \mathfrak{S}_{i-1}} S_j^{1+\alpha} - 1 + o_{\mathbb{P}}(n^{-1/3}) \\
&\quad = \Big(c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{n} S_j^{1+\alpha} - 1\Big) - c_{n,\beta}\frac{\lambda}{n}\sum_{j \in \mathfrak{S}_{i-1}} S_j^{1+\alpha} + o_{\mathbb{P}}(n^{-1/3}) \\
&\quad = \Big(c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{n} S_j^{1+\alpha} - 1\Big) - c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{i-1} S_{c(j)}^{1+\alpha} + o_{\mathbb{P}}(n^{-1/3}), \qquad (5.3.17)
\end{aligned}$$

and the $o_{\mathbb{P}}(n^{-1/3})$ term is uniform in $i = O(n^{2/3})$. We are now able to compute

$$\begin{aligned}
n^{-1/3}C_n^u(tn^{2/3}) &= n^{-1/3}\sum_{i=1}^{tn^{2/3}} (\mathbb{E}_S[A_n^u(i) \mid \mathcal{F}_{i-1}] - 1) \\
&= tn^{1/3}\Big(c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{n} S_j^{1+\alpha} - 1\Big) \\
&\quad - c_{n,\beta}\frac{\lambda}{n^{4/3}}\sum_{i=1}^{tn^{2/3}}\sum_{j=1}^{i-1} S_{c(j)}^{1+\alpha} + o_{\mathbb{P}}(1). \qquad (5.3.18)
\end{aligned}$$

Note that, since $\mathbb{E}[(S^{1+\alpha})^{(2+\alpha)/(1+\alpha)}] < \infty$, by the Marcinkiewicz and Zygmund Theorem [35, Theorem 2.5.8], if $\alpha \in (0,1]$,

$$\begin{aligned}
c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{n} S_j^{1+\alpha} &= c_{n,\beta}\lambda\mathbb{E}[S^{1+\alpha}] + o_{\mathbb{P}}(n^{-\frac{1}{2+\alpha}}) \\
&= 1 + \beta n^{-1/3} + o_{\mathbb{P}}(n^{-\frac{1}{2+\alpha}}). \qquad (5.3.19)
\end{aligned}$$

For $\alpha = 0$, by a similar result [35, Theorem 2.5.7], for all $\varepsilon > 0$,

$$\frac{1}{n}\sum_{j=1}^{n} S_j = \mathbb{E}[S] + o_{\mathbb{P}}(n^{-1/2}\log(n)^{1/2+\varepsilon}). \qquad (5.3.20)$$

In particular,

$$tn^{1/3}\Big(c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{n} S_j^{1+\alpha} - 1\Big) = t(\beta + o_{\mathbb{P}}(1)). \qquad (5.3.21)$$

By monotonicity, it follows that

$$\sup_{t\leq T}\Big|tn^{1/3}\Big(c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{n}S_j^{1+\alpha}-1\Big)-\beta t\Big|\xrightarrow{\mathbb{P}}0. \tag{5.3.22}$$

Therefore, for $\alpha\in[0,1]$,

$$n^{-1/3}C_n^{u}(tn^{2/3})=\beta t-c_{n,\beta}\frac{\lambda}{n^{4/3}}\sum_{i=1}^{tn^{2/3}}\sum_{j=1}^{i-1}S_{c(j)}^{1+\alpha}+o_{\mathbb{P}}(1). \tag{5.3.23}$$

Since $c_{n,\beta}=1+O(n^{-1/3})$, the second term in (5.3.23) converges uniformly to $-\lambda\frac{\mathbb{E}[S^{1+2\alpha}]}{2\mathbb{E}[S^{\alpha}]}t^2$ by Lemma 42.

Proof of (ii) for the upper bound.

Rewrite $V_n^u(k)$, for $k=O(n^{2/3})$, as

$$\begin{aligned}V_n^u(k)&=\sum_{i=1}^{k}(\mathbb{E}_S[A_n^u(i)^2\mid\mathcal{F}_{i-1}]-\mathbb{E}_S[A_n^u(i)\mid\mathcal{F}_{i-1}]^2)\\&=\sum_{i=1}^{k}(\mathbb{E}_S[A_n^u(i)^2\mid\mathcal{F}_{i-1}]-1)+O_{\mathbb{P}}(kn^{-1/3}),\end{aligned} \tag{5.3.24}$$

where we have used the asymptotics for $\mathbb{E}_S[A_n^u(i)|\mathcal{F}_{i-1}]$ obtained in (5.3.17)–(5.3.23). Moreover, we compute $\mathbb{E}_S[A_n^u(i)^2|\mathcal{F}_{i-1}]$ as

$$\mathbb{E}_S[A_n^u(i)^2\mid\mathcal{F}_{i-1}]=\mathbb{E}_S[(\sum_{h\notin\mathfrak{S}_i}\mathbb{1}_{\{T_h\leq c_{n,\beta}S_{c(i)}S_h/n\}})^2\mid\mathcal{F}_{i-1}] \tag{5.3.25}$$

$$=\mathbb{E}_S[A_n^u(i)\mid\mathcal{F}_{i-1}]+\mathbb{E}_S[\sum_{\substack{h,q\notin\mathfrak{S}_i\\h\neq q}}\mathbb{1}_{\{T_h\leq c_{n,\beta}S_{c(i)}S_h/n\}}\mathbb{1}_{\{T_q\leq c_{n,\beta}S_{c(i)}S_q/n\}}\mid\mathcal{F}_{i-1}].$$

Again by (5.3.17), $\mathbb{E}_S[A_n(i)|\mathcal{F}_{i-1}]=1+o_{\mathbb{P}}(1)$, uniformly in $i=O(n^{2/3})$, so that (5.3.24) simplifies to

$$\begin{aligned}V_n^u(k)=\sum_{i=1}^{k}\mathbb{E}_S[&\sum_{\substack{h,q\notin\mathfrak{S}_i\\h\neq q}}\mathbb{1}_{\{T_h\leq c_{n,\beta}S_{c(i)}S_h^{\alpha}/n\}}\mathbb{1}_{\{T_q\leq c_{n,\beta}S_{c(i)}S_q^{\alpha}/n\}}\mid\mathcal{F}_{i-1}]\\&+O_{\mathbb{P}}(kn^{-1/3}).\end{aligned} \tag{5.3.26}$$

We then focus on the second term in (5.3.25), which we compute as

$$\begin{aligned}
&\sum_{\substack{h,q\notin\mathfrak{S}_i\\ h\neq q}} \mathbb{E}_S[\mathbb{1}_{\{T_h\leq c_{n,\beta}S_{c(i)}S_h^\alpha/n\}}\mathbb{1}_{\{T_q\leq c_{n,\beta}S_{c(i)}S_q^\alpha/n\}} \mid \mathcal{F}_{i-1}]\\
&= \sum_{j\notin\mathfrak{S}_{i-1}} \mathbb{P}_S(c(i)=j \mid \mathcal{F}_{i-1})\\
&\quad\times \sum_{\substack{h,q\notin\mathfrak{S}_{i-1}\cup\{j\}\\ h\neq q}} \mathbb{E}_S[\mathbb{1}_{\{T_h\leq c_{n,\beta}S_jS_h^\alpha/n\}}\mathbb{1}_{\{T_q\leq c_{n,\beta}S_jS_q^\alpha/n\}} \mid \mathcal{F}_{i-1}].
\end{aligned} \tag{5.3.27}$$

By Lemma 43,

$$\begin{aligned}
(5.3.27) &= \sum_{j\notin\mathfrak{S}_{i-1}} \frac{S_j^\alpha}{\sum_{l\notin\mathfrak{S}_{i-1}} S_l^\alpha}\frac{1}{n^2}\sum_{\substack{h,q\notin\mathfrak{S}_{i-1}\cup\{j\}\\ h\neq q}} (c_{n,\beta}^2\lambda^2 S_j^2 S_h^\alpha S_q^\alpha + o_{\mathbb{P}}(n^{-2}))\\
&= (c_{n,\beta}\lambda)^2\mathbb{E}_S[S_{c(i)}^2 \mid \mathcal{F}_{i-1}]\frac{1}{n^2}\sum_{\substack{h,q\notin\mathfrak{S}_{i-1}\cup\{c(i)\}\\ h\neq q}} S_h^\alpha S_q^\alpha + o_{\mathbb{P}}(1)\\
&= \frac{(c_{n,\beta}\lambda)^2}{n^2}\mathbb{E}_S[S_{c(i)}^2 \mid \mathcal{F}_{i-1}]\sum_{h,q=1}^n S_h^\alpha S_q^\alpha\\
&\qquad - \frac{(c_{n,\beta}\lambda)^2}{n^2}\mathbb{E}_S[S_{c(i)}^2 \sum_{\substack{h,q\in\mathfrak{S}_{i-1}\cup\{c(i)\}\\ \cup\{h=q\}}} S_h^\alpha S_q^\alpha \mid \mathcal{F}_{i-1}] + o_{\mathbb{P}}(1).
\end{aligned}$$

The leading contribution to $V_n^u(k)$ is given by the first term, while the second term is an error term by Lemma 39. We have shown that $V_n^u(\cdot)$ can be rewritten as

$$V_n^u(k) = \Big(\frac{\lambda}{n}\sum_{h=1}^n S_h^\alpha\Big)^2\sum_{i=1}^k \mathbb{E}_S[S_{c(i)}^2 \mid \mathcal{F}_{i-1}] + o_{\mathbb{P}}(k). \tag{5.3.28}$$

Thus,

$$n^{-2/3}V_n^u(n^{2/3}u) \xrightarrow{\mathbb{P}} \lambda^2\mathbb{E}[S^\alpha]\mathbb{E}[S^{2+\alpha}]u, \tag{5.3.29}$$

concluding the proof of (ii). □

Proof of (iii) for the upper bound.

The jumps of $V_n^u(k)$ are given by

$$\begin{aligned} V_n^u(i) - V_n^u(i-1) &= \mathbb{E}_S[A_n^u(i)^2 \mid \mathcal{F}_{i-1}] - \mathbb{E}_S[A_n^u(i) \mid \mathcal{F}_{i-1}]^2 \\ &= \mathbb{E}_S\Big[\sum_{\substack{h,q \notin \mathfrak{S}_i \\ h \neq q}} \mathbb{1}_{\{T_h \leq c_{n,\beta} S_{c(i)} S_h^\alpha / n\}} \mathbb{1}_{\{T_q \leq c_{n,\beta} S_{c(i)} S_q^\alpha / n\}} \mid \mathcal{F}_{i-1}\Big] \\ &\quad + (\mathbb{E}_S[A_n^u(i) \mid \mathcal{F}_{i-1}] - \mathbb{E}_S[A_n^u(i) \mid \mathcal{F}_{i-1}]^2) \end{aligned} \tag{5.3.30}$$

In particular, $V_n^u(i) - V_n^u(i-1) \geq 0$. Since $\mathbb{E}_S[A_n^u(i) \mid \mathcal{F}_{i-1}] = 1 + O_{\mathbb{P}}(n^{-1/3})$ for $i = O_{\mathbb{P}}(n^{2/3})$ by (5.3.17), the second term is of order $O_{\mathbb{P}}(n^{-1/3})$, uniformly in $i = O_{\mathbb{P}}(n^{2/3})$. The first term was computed in (5.3.27). Therefore,

$$\begin{aligned} &V_n^u(i) - V_n^u(i-1) \\ &= \frac{(c_{n,\beta}\lambda)^2}{n^2} \mathbb{E}_S[S_{c(i)}^2 \mid \mathcal{F}_{i-1}] \sum_{h,q=1}^{n} S_h^\alpha S_q^\alpha \\ &\quad - \frac{(c_{n,\beta}\lambda)^2}{n^2} \mathbb{E}_S\Big[S_{c(i)}^2 \sum_{\substack{h,q \in \mathfrak{S}_{i-1} \cup \{c(i)\} \\ \cup \{h=q\}}} S_h^\alpha S_q^\alpha \mid \mathcal{F}_{i-1}\Big] + o_{\mathbb{P}}(1) \\ &\leq \frac{(c_{n,\beta}\lambda)^2}{n^2} \mathbb{E}_S[S_{c(i)}^2 \mid \mathcal{F}_{i-1}] \sum_{h,q=1}^{n} S_h^\alpha S_q^\alpha + o_{\mathbb{P}}(1). \end{aligned} \tag{5.3.31}$$

After rescaling and taking the expectation, we obtain the bound

$$\begin{aligned} &n^{-2/3} \mathbb{E}_S\Big[\sup_{i \leq \bar{t} n^{2/3}} |V_n^u(i) - V_n^u(i-1)|\Big] \\ &\qquad \leq \frac{(c_{n,\beta}\lambda)^2}{n^{2/3}} \mathbb{E}_S\Big[\sup_{i \leq \bar{t} n^{2/3}} S_{c(i)}^2\Big] \Big(\sum_{h,q=1}^{n} \frac{S_h^\alpha}{n}\Big)^2. \end{aligned} \tag{5.3.32}$$

Lemma 44. *Assume that* $\mathbb{E}[S^{2+\alpha}] < \infty$. *Then,*

$$\mathbb{E}_S\Big[\sup_{k \leq t n^{2/3}} S_{c(k)}^2\Big] = o_{\mathbb{P}}(n^{2/3}). \tag{5.3.33}$$

Proof. For $\varepsilon > 0$ split the expectation as

$$\mathbb{E}_S\Big[\Big(\sup_{k \leq t n^{2/3}} S_{c(k)}\Big)^2\Big] \leq \mathbb{E}_S\Big[\sup_{k \leq t n^{2/3}} S_{c(k)}^2 \mathbb{1}_{\{S_{c(k)} > \varepsilon n^{1/3}\}}\Big] + \varepsilon^2 n^{2/3}. \tag{5.3.34}$$

We bound the expected value in the first term as

$$\begin{aligned}
&\mathbb{E}_S[\sup_{k\leq tn^{2/3}} S^2_{c(k)}\mathbb{1}_{\{S_{c(k)}>\varepsilon n^{1/3}\}}] \\
&\quad\leq \sum_{k\leq tn^{2/3}} \frac{1}{n^{2/3}}\mathbb{E}_S[S^2_{c(k)}\mathbb{1}_{\{S_{c(k)}>\varepsilon n^{1/3}\}}] \\
&\quad\leq n^{2/3}t\mathbb{E}_S[S^2_{c(1)}\mathbb{1}_{\{S_{c(1)}>\varepsilon n^{1/3}\}}](1+O_{\mathbb{P}_S}(1)),
\end{aligned} \tag{5.3.35}$$

where we have used Lemma 38 with the function $f(x) = x^2\mathbb{1}_{\{x>\varepsilon n^{1/3}\}}$. Computing the expectation explicitly we get

$$\begin{aligned}
t\mathbb{E}_S[S^2_{c(1)}\mathbb{1}_{\{S_{c(1)}>\varepsilon n^{1/3}\}}] &= t\sum_{i=1}^n S_i^2\mathbb{1}_{\{S_i>\varepsilon n^{1/3}\}}\mathbb{P}(c(1)=i) \\
&= t\sum_{i=1}^n S_i^2\mathbb{1}_{\{S_i>\varepsilon n^{1/3}\}}\frac{S_i^\alpha}{\sum_{j=1}^n S_j^\alpha},
\end{aligned} \tag{5.3.36}$$

so that the left-hand side of (5.3.34) is bounded by

$$\frac{t}{\sum_{j=1}^n S_j^\alpha}\sum_{i=1}^n S_i^{2+\alpha}\mathbb{1}_{\{S_i>\varepsilon n^{1/3}\}} + \Big(\sum_{i=1}^n \frac{S_i^\alpha}{n}\Big)^2\varepsilon^2, \tag{5.3.37}$$

which tends to zero as $n\to\infty$ since $\mathbb{E}[S^{2+\alpha}]<\infty$ and $\varepsilon>0$ is arbitrary. □

By Lemma 44 the right-hand side of (5.3.32) converges to zero, and this concludes the proof of (iii). □

Proof of (iv) for the upper bound.

First we split

$$\begin{aligned}
&\mathbb{E}_S[\sup_{k\leq tn^{2/3}} (M_n^u(k)-M_n^u(k-1))^2] \\
&= \mathbb{E}_S[\sup_{k\leq tn^{2/3}} (A_n^u(k)-\mathbb{E}_S[A_n^u(k)\mid \mathcal{F}_{k-1}])^2] \\
&\leq \mathbb{E}_S[\sup_{k\leq tn^{2/3}} |A_n^u(k)|^2] + \mathbb{E}_S[\sup_{k\leq tn^{2/3}} \mathbb{E}[A_n^u(k)\mid \mathcal{F}_{k-1}]^2] \\
&\leq 2\mathbb{E}_S[\sup_{k\leq tn^{2/3}} |A_n^u(k)|^2].
\end{aligned} \tag{5.3.38}$$

We then stochastically dominate $(A_n^u(k))_{k=1}^{tn^{2/3}}$ by a sequence of Poisson processes $(\Pi_k)_{k=1}^{tn^{2/3}}$, according to

$$A_n^u(k) \preceq \Pi_k\Big(c_{n,\beta} S_{c(k)} \sum_{i=1}^n \frac{S_i^\alpha}{n}\Big) =: A_n'(k). \tag{5.3.39}$$

Given n exponential random variables $E_1/\lambda_1, E_2/\lambda_2, \dots, E_n/\lambda_n$ with parameters respectively $\lambda_1, \lambda_2, \dots, \lambda_n$, there exists an explicit coupling with a Poisson process $\Pi(\cdot)$ such that $\sum_{i\le n} \mathbb{1}_{\{E_i/\lambda_i \le t\}} \le \Pi(\sum_{i\le n}\lambda_i t)$. The coupling is constructed as follows. Each random variable E_i/λ_i is coupled with a rate one Poisson process $\Pi^{(i)}$ in such a way that $\mathbb{1}_{\{E_i/\lambda_i\le t\}} \le \Pi^{(i)}(\lambda_i t)$. Moreover, by basic properties of the Poisson process $\sum_{i=1}^n \Pi^{(i)}(\lambda_i t) \overset{\mathrm{d}}{=} \Pi(\sum_{i=1}^n \lambda_i t)$.

We bound (5.3.39) through martingale techniques applied to the decomposition

$$\begin{aligned} n^{-2/3}\mathbb{E}_S[\sup_{k\le tn^{2/3}} |A_n^u(k)|^2] &\le 2n^{-2/3}\mathbb{E}_S\Big[\Big(\sup_{k\le tn^{2/3}} |A_n'(k) - c_{n,\beta}S_{c(k)}\sum_{i=1}^n \frac{S_i^\alpha}{n}|\Big)^2\Big] \\ &+ 2n^{-2/3}\mathbb{E}_S\Big[\Big(c_{n,\beta}\sup_{k\le tn^{2/3}} S_{c(k)}\sum_{i=1}^n \frac{S_i^\alpha}{n}\Big)^2\Big] \end{aligned} \tag{5.3.40}$$

Applying Doob's L^2 martingale inequality [63, Theorem 11.2] to the first term we see that it converges to zero, since

$$\begin{aligned} n^{-2/3}\mathbb{E}_S\Big[\Big(\sup_{k\le tn^{2/3}} \Big|A_n'(k) - S_{c(k)}\sum_{i=1}^n \frac{S_i^\alpha}{n}\Big|\Big)^2\Big] &\le 4n^{-2/3}\mathbb{E}_S\Big[\Big|A_n'(tn^{2/3}) - S_{c(tn^{2/3})}\sum_{i=1}^n \frac{S_i^\alpha}{n}\Big|^2\Big] \\ &= 4n^{-2/3}\mathbb{E}_S\Big[S_{c(tn^{2/3})}\sum_{i=1}^n \frac{S_i^\alpha}{n}\Big]. \end{aligned} \tag{5.3.41}$$

The last equality follows from the expression for the variance of a Poisson random variable. The right-most term converges to zero by Lemma 41.

We now bound the second term in (5.3.40), as

$$\begin{aligned} n^{-2/3}\mathbb{E}_S\Big[\Big(\sup_{k\le tn^{2/3}} S_{c(k)}\sum_{i=1}^n \frac{S_i^\alpha}{n}\Big)^2\Big] \\ = \Big(\sum_{i=1}^n \frac{S_i^\alpha}{n}\Big)^2 n^{-2/3}\mathbb{E}_S[(\sup_{k\le tn^{2/3}} S_{c(k)})^2] \end{aligned} \tag{5.3.42}$$

By Lemma 44 the right-hand side of (5.3.42) converges to zero, concluding the proof of (iv). □

5.3.2 Convergence of the embedded queue

As a consequence of (5.3.7) and Theorem 19 we see that $Q_n^e(k) = O_{\mathbb{P}}(n^{1/3})$ for $k = O(n^{2/3})$. Moreover, the following lemma shows that the process $n^{-1/3}Q_n^e(\cdot n^{2/3})$ is tight:

Lemma 45. *Fix $\bar{t} > 0$. The sequence $n^{-1/3}\sup_{t\le\bar{t}} Q_n^e(tn^{2/3})$ is tight.*

Proof. The supremum function $f(\cdot) \mapsto \sup_{t\le\bar{t}} f(t)$ is continuous in $(\mathcal{D}, J_1)$ by [96, Theorem 13.4.1]. In particular,

$$n^{-1/3}\sup_{t\le\bar{t}} Q_n^{e,u}(tn^{2/3}) \xrightarrow{\text{d}} \sup_{t\le\bar{t}} \widehat{N}(t). \tag{5.3.43}$$

Since $Q_n^e(k) \le Q_n^{e,u}(k)$, the conclusion follows. □

As an immediate consequence of (5.1.6) and Lemma 45, we have the following important corollary. Recall that ν_i is the set of customers who have left the system or are in the queue at the beginning of the i-th service, so that $|\nu_i| = i + Q_n^e(i)$. Recall also that $0 \le Q_n^e(t) \le Q_n^{e,u}(t)$.

Corollary 4. *As $n \to \infty$,*

$$|\nu_i| = i + o_{\mathbb{P}}(i), \qquad \text{uniformly in } i = O_{\mathbb{P}}(n^{2/3}). \tag{5.3.44}$$

Intuitively, this implies that the main contribution to the downwards drift in the queue-length process comes from the customers that have left the system, and not from the customers in the queue. Alternatively, the order of magnitude of the queue length, that is $n^{1/3}$, is negligible with respect to the order of magnitude of the customers who have left the system, which is $n^{2/3}$.

In order to prove Theorem 17 we proceed as in the proof of Theorem 19, but we now need to deal with the more complicated drift term. As before, we decompose $N_n(k) = M_n(k) + C_n(k)$ and $M_n(k)^2 = Z_n(k) + V_n(k)$, where

$$\begin{aligned} M_n(k) &= \sum_{i=1}^{k} (A_n(i) - \mathbb{E}_S[A_n(i) \mid \mathcal{F}_{i-1}]), \\ C_n(k) &= \sum_{i=1}^{k} (\mathbb{E}_S[A_n(i) \mid \mathcal{F}_{i-1}] - 1), \\ V_n(k) &= \sum_{i=1}^{k} (\mathbb{E}_S[A_n(i)^2 \mid \mathcal{F}_{i-1}] - \mathbb{E}_S[A_n(i) \mid \mathcal{F}_{i-1}]^2). \end{aligned} \tag{5.3.45}$$

As before, we separately prove the convergence of the drift $C_n(k)$ and of the martingale $M_n(k)$, by verifying the conditions (i)–(iv) in Section 5.3.1. Verifying (i) proves to be the most challenging task, while the estimates for (ii)–(iv) in Section 5.3.1 carry over without further complications.

Proof of (i) for the embedded queue.

By expanding $\mathbb{E}_S[A_n(i) \mid \mathcal{F}_{i-1}] - 1$ as in (5.3.15), we get

$$\begin{aligned} \mathbb{E}_S[A_n(i) \mid \mathcal{F}_{i-1}] - 1 = &\Big(c_{n,\beta} \lambda \frac{\sum_{l=1}^{n} S_l^{\alpha}}{n} \mathbb{E}_S[S_{c(i)} \mid \mathcal{F}_{i-1}] - 1 \Big) \\ &- c_{n,\beta} \mathbb{E}_S[S_{c(i)} \mid \mathcal{F}_{i-1}] \sum_{l \in \nu_i \setminus \{c(i)\}} \lambda \frac{S_l^{\alpha}}{n} \\ &- c_{n,\beta} \frac{\lambda}{n} \mathbb{E}_S[S_{c(i)}^{1+\alpha} \mid \mathcal{F}_{i-1}] + o_{\mathbb{P}}(n^{-1/3}). \end{aligned} \tag{5.3.46}$$

By further expanding the first term in (5.3.46) as in (5.3.16), we get

$$\begin{aligned} \mathbb{E}_S[A_n(i) \mid \mathcal{F}_{i-1}] - 1 = &\Big(c_{n,\beta} \frac{\lambda}{n} \sum_{j \notin \mathfrak{S}_{i-1}} S_j^{1+\alpha} - 1 \Big) \\ &- c_{n,\beta} \mathbb{E}_S[S_{c(i)} \mid \mathcal{F}_{i-1}] \sum_{l=i+1}^{i+1+Q_n^e(i-1)} \lambda \frac{S_{c(l)}^{\alpha}}{n} \\ &- c_{n,\beta} \frac{\lambda}{n} \mathbb{E}_S[S_{c(i)}^{1+\alpha} \mid \mathcal{F}_{i-1}] + o_{\mathbb{P}}(n^{-1/3}), \end{aligned} \tag{5.3.47}$$

where in the first equality we have used (5.1.6). Comparing equation (5.3.47) with equation (5.3.17), one sees that the drift can be rewritten as

$$C_n(k) = C_n^u(k) - c_{n,\beta}\lambda \sum_{i=1}^{k} \mathbb{E}_S[S_{c(i)} \mid \mathcal{F}_{i-1}] \sum_{l=i+1}^{i+1+Q_n^e(i-1)} \frac{S_{c(l)}^{\alpha}}{n}. \tag{5.3.48}$$

Therefore, to conclude the proof of (i) it is enough to show that the second term vanishes, after rescaling. This is proven in the following lemma:

Lemma 46. *As $n \to \infty$,*

$$n^{-1/3}c_{n,\beta}\lambda \sum_{i=1}^{tn^{2/3}} \mathbb{E}_S[S_{c(i)} \mid \mathcal{F}_{i-1}] \sum_{l=i+1}^{i+1+Q_n^e(i-1)} \frac{S_{c(l)}^{\alpha}}{n} \xrightarrow{\mathbb{P}} 0. \tag{5.3.49}$$

Proof. By Lemma 45, $\sup_{i\leq tn^{2/3}} Q_n^e(i) \leq c_1 n^{1/3}$ with high probability for a large constant c_1, and by Lemma 41, $\sup_{i\leq tn^{2/3}} \mathbb{E}_S[S_{c(i)} \mid \mathcal{F}_{i-1}] \leq c_2$ with high probability for another large constant c_2. This implies that, with high probability,

$$\begin{aligned} n^{-1/3}c_{n,\beta}\lambda \sum_{i=1}^{tn^{2/3}} \mathbb{E}_S[S_{c(i)} \mid \mathcal{F}_{i-1}] \sum_{l=i+1}^{i+1+Q_n^e(i-1)} \frac{S_{c(l)}^{\alpha}}{n} \\ \leq c_{n,\beta}\lambda C_2 \sum_{i=1}^{tn^{2/3}} \sum_{l=i+1}^{i+1+C_1n^{1/3}} \frac{S_{c(l)}^{\alpha}}{n^{4/3}}. \end{aligned} \tag{5.3.50}$$

We rewrite the double sum as

$$\begin{aligned} c_{n,\beta}\lambda C_2 \sum_{i=1}^{tn^{2/3}} \sum_{l=i+1}^{i+1+c_1n^{1/3}} \frac{S_{c(l)}^{\alpha}}{n^{4/3}} &\leq c_{n,\beta}\lambda c_2 \sum_{j=1}^{tn^{2/3}+c_1n^{1/3}} \min\{j, c_1n^{1/3}\}\frac{S_{c(j)}^{\alpha}}{n^{4/3}} \\ &\leq c_{n,\beta}\lambda c_1 c_2 \sum_{j=1}^{(t+c_1)n^{2/3}} \frac{S_{c(j)}^{\alpha}}{n}. \end{aligned} \tag{5.3.51}$$

The right-hand side term converges to zero in probability as $n \to \infty$ by Lemma 42, concluding the proof. □

Since

$$\begin{aligned} n^{-1/3}C_n(tn^{2/3}) &= n^{-1/3}C_n^u(tn^{2/3}) \\ &\quad - n^{-1/3}c_{n,\beta}\lambda \sum_{i=1}^{tn^{2/3}} \mathbb{E}_S[S_{c(i)} \mid \mathcal{F}_{i-1}] \sum_{l=i+1}^{i+1+Q_n^e(i-1)} \frac{S_{c(l)}^{\alpha}}{n}, \end{aligned} \tag{5.3.52}$$

Lemma 46 and the analogous convergence result for $n^{-1/3}C_n^u(\cdot n^{2/3})$ in Section 5.3.1 conclude the proof of (i). □

Proof of (ii), (iii) and (iv) for the embedded queue

Proceeding as in Section 5.3.1, we find that

$$\begin{aligned} V_n(k) &= \sum_{i=1}^{k}(\mathbb{E}_S[A_n(i)^2 \mid \mathcal{F}_{i-1}] - \mathbb{E}_S[A_n(i) \mid \mathcal{F}_{i-1}]^2) \\ &= \sum_{i=1}^{k}(\mathbb{E}_S[A_n(i)^2 \mid \mathcal{F}_{i-1}] - 1) + O_{\mathbb{P}}(kn^{-1/3}), \end{aligned} \tag{5.3.53}$$

where

$$\begin{aligned} \mathbb{E}_S[A_n(i)^2 \mid \mathcal{F}_{i-1}] &= \mathbb{E}_S[A_n(i) \mid \mathcal{F}_{i-1}] \\ &\quad + \mathbb{E}_S[\sum_{\substack{h,q \notin \mathcal{V}_{i-1} \\ h \neq q}} \mathbb{1}_{\{T_h \leq S_{c(i)}S_h/n\}} \mathbb{1}_{\{T_q \leq S_{c(i)}S_q/n\}} \mid \mathcal{F}_{i-1}]. \end{aligned} \tag{5.3.54}$$

Similarly as in Section 5.3.1, we get

$$\begin{aligned} &\sum_{\substack{h,q \notin \mathcal{V}_{i-1} \\ h \neq q}} \mathbb{E}_S[\mathbb{1}_{\{T_h \leq S_{c(i)}S_h^\alpha/n\}} \mathbb{1}_{\{T_q \leq S_{c(i)}S_q^\alpha/n\}} \mid \mathcal{F}_{i-1}] \\ &\qquad = \mathbb{E}_S[S_{c(i)}^2 \mid \mathcal{F}_{i-1}]\lambda^2\Big(\sum_{h=1}^{n}\frac{S_h^\alpha}{n}\Big)^2 \\ &\qquad\quad - \mathbb{E}_S\Big[\lambda^2\frac{S_{c(i)}^2}{n^2}\sum_{\substack{h,q \in \mathcal{V}_{i-1}\cup\{c(i)\} \\ \cup\{h=q\}}} S_h^\alpha S_q^\alpha \mid \mathcal{F}_{i-1}\Big] + o_{\mathbb{P}}(1). \end{aligned} \tag{5.3.55}$$

The second term is an error term by Lemma 39 and Corollary 4. This implies that $V_n(\cdot)$ can be rewritten as

$$V_n(k) = \Big(\frac{\lambda}{n}\sum_{h=1}^{n}S_h^\alpha\Big)^2\sum_{i=1}^{k}\mathbb{E}_S[S_{c(i)}^2 \mid \mathcal{F}_{i-1}] + o_{\mathbb{P}}(k), \tag{5.3.56}$$

so that

$$n^{-2/3}V_n(n^{2/3}u) \xrightarrow{\mathbb{P}} \lambda^2\mathbb{E}[S^\alpha]\mathbb{E}[S^{2+\alpha}]u, \tag{5.3.57}$$

which concludes the proof of (ii). □

To conclude the proof of Theorem 17, we are left to verify (iii) and (iv). However, the estimates in Sections 5.3.1 and 5.3.1 also hold for $V_n(\cdot)$ and $M_n(\cdot)$, since they rely respectively on (5.3.32) and (5.3.39) to bound the lower-order contributions to the drift. This concludes the proof of Theorem 17. □

5.4 Conclusions

In this chapter we have analyzed the critical $\Delta_{(i)}^{\alpha}/G/1$ queue. We have also shown that a (directed) tree can be associated to the $\Delta_{(i)}^{\alpha}/G/1$ queue in a natural way. The heavy-traffic assumption for the queue then corresponds to assuming that the associated random tree is critical.

Lemma 40 implies that the distribution of the service time of the first $O(n^{2/3})$ customers to join the queue converges to the α-size-biased distribution of S, irrespectively of the precise time at which the customers arrive. This suggests that it is possible to prove Theorem 17 by approximating the $\Delta_{(i)}^{\alpha}/G/1$ queue via a $\Delta_{(i)}/G/1$ queue with service-time distribution S^* such that

$$\mathbb{P}(S^* \in \mathcal{A}) = \mathbb{E}[S^{\alpha}\mathbb{1}_{\{S\in\mathcal{A}\}}]/\mathbb{E}[S^{\alpha}], \tag{5.4.1}$$

and i.i.d. arrival times distributed as $T_i \sim \exp(\lambda\mathbb{E}[S^{\alpha}])$. This conjecture is supported by two observations. First, the heavy-traffic conditions for the two queues coincide. Second, the standard deviation of the Brownian motion is the same in the two limiting diffusions. However, this approximation fails to capture the higher-order contributions to the queue-length process. Because of this, the coefficients of the negative quadratic drift in the two queues are different. Therefore, Theorem 17 cannot be deduced by the analogous theorem for the $\Delta_{(i)}/G/1$ queue.

Surprisingly, the assumption that α lies in the interval $[0,1]$ plays no role in our proof. On the other hand, we see from (5.1.16) that

$$\max\{\mathbb{E}[S^{2+\alpha}], \mathbb{E}[S^{1+2\alpha}], \mathbb{E}[S^{\alpha}]\} < \infty \tag{5.4.2}$$

is a necessary condition for Theorem 17 to hold. From this we conclude that Theorem 17 remains true as long as $\alpha \in \mathbb{R}$ is such that (5.4.2) is satisfied. From a modelling point of view, the assumption $\alpha > 1$ represents a situation in which customers with larger job sizes have an even stronger incentive to join the queue. On the other hand, when $\alpha < 0$, the queue models a situation in which customers with large job sizes are lazy and favour joining the queue later. We remark that the *form* of the limiting diffusion is the same for all $\alpha \in \mathbb{R}$, but different values of α yield different fluctuations (standard deviation of the Brownian motion), and a different quadratic drift.

CHAPTER 6

Finite-population queues

In this chapter we study the $\Delta_{(i)}/G/1$ queue in the finite-population regime. We assume that the number n of customers in the pool at the start of the queue is fixed and finite. For tractability, we further assume that the arrival and service times are exponentially distributed, leading to a two-dimensional absorbing Markov process in the positive quadrant. In contrast to earlier chapters, we do not scale the service speed with n and are interested in exact rather than asymptotic results. The resulting Markov Process has inhomogeneous transition rates and is thus outside of the reach of classical methods for the analysis of time-dependent behavior. To overcome this, we develop novel ad-hoc combinatorial techniques to recursively express quantities of interest, such as the distribution of the number of customers served in the first busy period.[1]

6.1 Introduction

The goal of this chapter is to study the combinatorial structure of the $\Delta_{(i)}/G/1$ queue-length process for a fixed, small number of initial customers n. This analysis is meant to complement the asymptotic results we have presented so far. Indeed, on one hand the formulas we obtain are typically unwieldy for large values of n, further motivating the need

[1] This chapter contains the results of an ongoing collaboration with Jori Selen and Alessandro Zocca.

for asymptotic approximation schemes. On the other hand, when n is not too large, our results provide workable (exact) expressions for various performance measures of the $\Delta_{(i)}/G/1$ queue. For tractability, we assume that the arrival epochs and service times of the n customers are i.i.d. exponential random variables with rates respectively λ and μ. This assumptions lead to the definition of a two-dimensional Markov process $X(\cdot)$, describing the number of completed services and the total number of customers who have joined the queue. The process $X(\cdot)$ is both absorbing (the sink state is (n,n)) and time-inhomogeneous, since the transition rates crucially depend on the current state of the process.

Juneja and Shimkin [56] studied a finite-population queue (for a fixed population n) in the context of the concert queueing game. In their setting each customer independently chooses when to arrive (possibly at a random time) in order to minimize a certain linear cost functional. In the Nash equilibrium, all customers sample their arrival times from the same distribution, leading to the $\Delta_{(i)}/G/1$ queue. They show that the unique equilibrium distribution has a complicated form for finite n, but tends to a uniform distribution as $n \to \infty$.

A large number of probabilistic tools are unsuitable for the analysis of absorbing Markov processes, since no non-trivial stationary distribution exists. Nonetheless, in some cases it is possible to define a so-called *quasi-stationary* distribution and to compute it explicitly via matrix-theoretic techniques. Darroch and Seneta [29, 30] introduce various definitions of quasi-stationary distributions for absorbing Markov processes and discuss how they are related to each other and to the classical notion of stationary distribution. One approach consists in defining a new Markov process, with identical transition rates to the original one, and additional (small, say ε) transition rates from the sink states to the transient states. This new Markov process admits a proper stationary distribution and, conditioned on the process being in a transient state, it is independent of ε. In our setting, this would correspond to allow a transition from the absorbing state (n,n) to the starting state $(0,0)$. Barlett [8] applied this technique in the context of (diminishing) population models. When $\varepsilon > 0$, the modified process can also be studied using techniques from regenerative process theory. As $\varepsilon \to 0$, transitions out of the sink state(s) become less likely and one expects to obtain information on the original $\Delta_{(i)}/G/1$ queue-length process. In this context, see Keilson's Theorem for the asymptotic exponentiality of rare events in regenerative processes [41]. It is not obvious, however, how the two processes are related.

Another powerful approach to the study of random walks is via combinatorial methods, such as lattice path counting [65, 93] and generating functions [37, 81]. Takács [89, 90, 92] has pioneered the application of combinatorial methods to the study of queueing models. In [90], Takács studies various combinatorial models and their probabilistic conterparts. In particular, he derives the distribution of the number of customers served in the first busy period of a simplified $\Delta_{(i)}/G/1$ queue. However, these combinatorial techniques rely crucially on some assumed symmetries of the underlying probabilistic model. Perhaps closer to our setting is the OK Corral model studied by Kingman [61, 62]. In this model, two (diminishing) populations of gunmen shoot at each other until one of the two is eliminated. The transition rates of the resulting two-dimensional Markov process depend on the state of the process, similarly as in the $\Delta_{(i)}/G/1$ queue. However, the analysis in [61] crucially exploits the symmetric structure of the problem (the two populations are interchangeable). In fact, this symmetry is what ultimately allows for explicit expressions in the OK Corral model.

6.2 Model description

In this section we briefly recall the definition of the $\Delta_{(i)}/G/1$ queue and introduce various notations that simplify the treatment of the queue in the finite-population regime. Consider a single-server queue that serves customers in a first-come first-served manner. A finite pool of n customers will enter the system only once. Each customer independently joins the queue after an exponential time with rate λ and requires a service time that is exponentially distributed with rate μ. We define

$$\lambda_i := \lambda(n-i) \tag{6.2.1}$$

as the arrival rate of customers to the system if i customers have already arrived to the system. Denote by $X_1(t)$ the number of completed services at time t and let $X_2(t)$ be the number of customers that have joined the system up until time t. The state of the system at time t is $X(t) := (X_1(t), X_2(t))$. The process $X(\cdot)$ is a Markov process on the state space

$$\mathcal{B} := \{(i,j) \in \mathbb{N}_0^2 : 0 \leq i \leq n,\ 0 \leq j \leq i\}. \tag{6.2.2}$$

The transition rate diagram is depicted in Figure 6.1.

We denote by B be the number of customers served in the first busy period. Abbreviate $b_n := \mathbb{P}(B = n)$ as the probability that exactly n customers are served in the first busy period. We denote by $\mathcal{D}_i := \{(0,i),(1,i+1),\ldots,(n-i,n)\}$, $i \geq 1$ the set of states on the n-th superdiagonal of $\mathcal{S}$. For the diagonal we define $\mathcal{D}_0 := \{(1,1),(2,2),\ldots,(n,n)\}$.

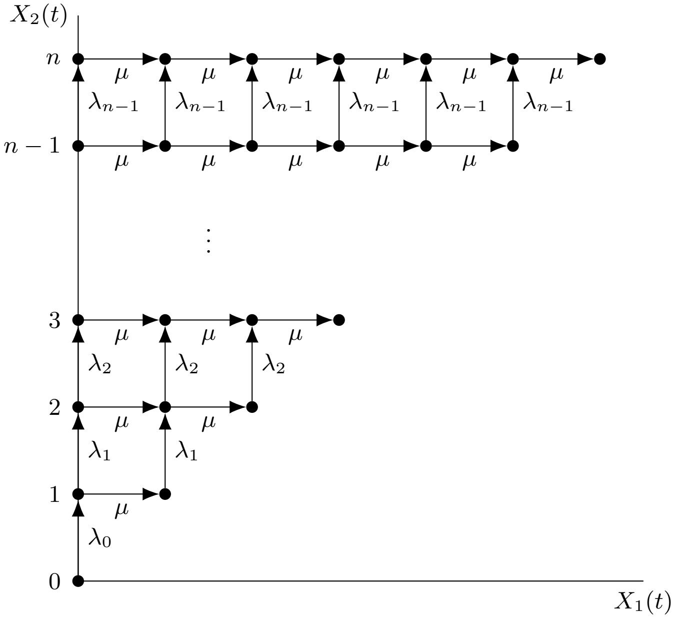

Figure 6.1: Transition rate diagram of the Markov process $X(\cdot)$.

By phase i we refer to the set of states $\mathcal{P}_i := \{(0,i),(1,i),\ldots,(i,i)\}$. Given any stochastic process Y, we let $\mathbb{E}_y[f(Y)]$ represent the expectation of a functional of Y, conditional on $Y(0)=y$ and similarly for $\mathbb{P}_y(\cdot)$. Denote by G_p a geometric random variable with support $\{0,1,\ldots\}$, failure probability p and probability generating function

$$\mathcal{G}_p(z) = \frac{1-p}{1-pz}, \quad |z| < \frac{1}{p}. \tag{6.2.3}$$

We define for each set $\mathcal{A} \subsetneq \mathcal{B}$ the hitting-time random variables

$$H_X(\mathcal{A}) := \inf\{t > 0 : X(t) \in \mathcal{A}\} \tag{6.2.4}$$

as the first time $X(\cdot)$ makes a transition into the set $\mathcal{A}$. For a singleton x, $H(x)$ should be understood to mean $H(\{x\})$. The probabilities b_n can be expressed in terms of the hitting-time random variables as

$$b_n = \mathbb{P}_{(0,1)}(X(H_{\mathcal{D}_0}) = (n,n)). \tag{6.2.5}$$

6.3 The number of customers in the first busy period

In this section we derive a recursion for the probabilities b_n by employing generating functions. To that end, we define, for $0 \leq i \leq j-1$, $2 \leq j \leq n$,

$$p_j(i) := \mathbb{P}_{(0,0)}(H(\mathcal{P}_j) < H(\mathcal{D}_0),\ X(H(\mathcal{P}_j)) = (i,j)). \tag{6.3.1}$$

as the probability that, conditional on $X(0) = (0,0)$, the Markov process $X(\cdot)$ reaches phase n in state (i,j) without residing in $\mathcal{D}_0$ before phase j is reached. Note that $p_j(j-1) = 0$. Define its generating function, for $z \in \mathbb{C}$,

$$P_j(z) := \sum_{i=0}^{j-2} p_j(i) z^i, \quad 2 \leq j \leq n. \tag{6.3.2}$$

Clearly, if $n = 1$, then $s_1 = 1$. For $n > 1$, we have by the strong Markov property that $s_1 = \rho_1$ and for $2 \leq j \leq n-1$,

$$\rho_j^j P_j(\rho_j^{-1}) = b_j, \quad P_n(1) = b_n, \tag{6.3.3}$$

where for convenience we have abbreviated

$$\rho_j := \frac{\mu}{\mu + \lambda_j}. \tag{6.3.4}$$

The following theorem identifies $P_j(z)$:

Theorem 20. *For $1 \leq j \leq n-1$, the generating functions are explicitly given by*

$$P_{j+1}(z) = \prod_{i=1}^{j} \mathcal{G}_{\rho_j}(z) - \sum_{i=1}^{j} b_i z^i \prod_{k=i}^{n} \mathcal{G}_{\rho_k}(z), \quad |z| < \frac{1}{\rho_j}. \tag{6.3.5}$$

Proof. We start by expressing $P_{j+1}(z)$ in terms of $P_j(z)$. From the strong Markov property at time $H(\mathcal{P}_j)$ we can write

$$p_{j+1}(i) = \sum_{k=0}^{i} p_j(k) \rho_j^{i-k} (1-\rho_j), \quad 0 \leq i \leq j-2, \tag{6.3.6}$$

$$p_{j+1}(j-1) = \sum_{k=0}^{j-2} p_j(k) \rho_j^{j-1-k} (1-\rho_j). \tag{6.3.7}$$

Multiply both sides of (6.3.6) by z^i and sum over all i with $0 \leq i \leq j-2$ and multiply both sides of (6.3.7) by z^{j-1}. Sum the two resulting expressions to get

$$P_{j+1}(z) \\ = \sum_{i=0}^{j-2}\sum_{k=0}^{i} p_j(k)\rho_j^{i-k}(1-\rho_j)z^i + \sum_{k=0}^{j-2} p_j(k)\rho_j^{j-1-k}(1-\rho_j)z^{j-1}. \quad (6.3.8)$$

Switch the order of the double summation to obtain

$$\begin{aligned} P_{j+1}(z) & \\ &= (1-\rho_j)\Big(\sum_{k=0}^{j-2} p_j(k)\sum_{i=k}^{j-2}\rho_j^{i-k}z^i + \sum_{k=0}^{j-2} p_j(k)\rho_j^{j-1-k}z^{j-1}\Big) \\ &= (1-\rho_j)\Big(\sum_{k=0}^{j-2} p_j(k)\sum_{l=0}^{j-2-k}\rho_j^{l}z^{k+l} + \sum_{k=0}^{j-2} p_j(k)\rho_j^{j-1-k}z^{j-1}\Big). \quad (6.3.9) \end{aligned}$$

Performing the inner summation over l and rewriting yields

$$\begin{aligned} P_{j+1}(z) &= (1-\rho_j)\Big(\sum_{k=0}^{j-2} p_j(k)\frac{z^k-\rho_j^{j-1-k}z^{j-1}}{1-\rho_j z} + \sum_{k=0}^{j-2} p_j(k)\rho_j^{j-1-k}z^{j-1}\Big) \\ &= \frac{1-\rho_j}{1-\rho_j z}\Big(\sum_{k=0}^{j-2} p_j(k)z^k - \rho_j^j z^j\sum_{k=0}^{j-2} p_j(k)\rho_j^{-k}\Big) \\ &= \mathcal{G}_{\rho_j}(z)(P_j(z)-b_j z^j). \quad (6.3.10) \end{aligned}$$

By iterating the relation (6.3.10) we obtain

$$P_{j+1}(z) = P_2(z)\prod_{i=2}^{n}\mathcal{G}_{\rho_i}(z) - \sum_{i=2}^{n} b_i z^i \prod_{k=i}^{j}\mathcal{G}_{\rho_k}(z). \quad (6.3.11)$$

We can further simplify (6.3.11) by noting that

$$P_2(z) = p_2(0) = 1-\rho_1 = (1-\rho_1 z)\mathcal{G}_{\rho_1}(z). \quad (6.3.12)$$

Since $b_1 = \rho_1$, we finally obtain (6.3.5). Notice that the radii of convergence of $\mathcal{G}_{\rho_j}(z)$ are decreasing in j, i.e., the radii of convergence are ordered: $\rho_1^{-1} > \rho_2^{-1} > \cdots > \rho_n^{-1} > 1$. □

Theorem 20 allows us to obtain an explicit recursion for the distribution of B, as detailed in the following corollary.

Corollary 5. *For $n > 1$ and $2 \leq j \leq n-1$, the probabilities b_j satisfy the recursion*

$$b_j = \rho_j^j \binom{n-1}{j-1} - \sum_{i=1}^{j-1} b_i \rho_j^{j-i} \binom{n-i}{j-i}, \quad b_n = 1 - \sum_{i=1}^{n-1} b_i \tag{6.3.13}$$

with initial term $b_1 = \rho_1$.

Proof. Combining the result of Theorem 20 with (6.3.3) yields the following recursion, for $2 \leq j \leq n-1$,

$$b_j = \rho_j^j \prod_{i=1}^{j-1} \mathcal{G}_{\rho_i}(\rho_j^{-1}) - \sum_{i=1}^{j-1} b_i \rho_j^{j-i} \prod_{k=i}^{j-1} \mathcal{G}_{\rho_k}(\rho_j^{-1}), \quad b_n = 1 - \sum_{i=1}^{n-1} b_i. \tag{6.3.14}$$

Since

$$\mathcal{G}_{\rho_k}(\rho_n^{-1}) = \frac{1-\rho_k}{1-\frac{\rho_k}{\rho_n}} = \frac{1-\frac{\mu}{\mu+\lambda_k}}{1-\frac{\mu+\lambda_n}{\mu+\lambda_k}} = \frac{\lambda_k}{\lambda_k-\lambda_n} = \frac{N-k}{n-k}, \tag{6.3.15}$$

we can simplify

$$\begin{aligned}\prod_{k=l}^{j-1} \mathcal{G}_{\rho_k}(\rho_j^{-1}) &= \frac{n-l}{j-l}\frac{n-l-1}{j-l-1}\frac{n-l-2}{j-l-2}\cdots\frac{n-j+1}{1} \\ &= \binom{n-l}{j-l},\end{aligned} \tag{6.3.16}$$

which proves the claim. □

Equation (6.3.5) has an appealing interpretation, which we now discuss. To this end, we introduce a Markov process $t \mapsto U(t)$ on the state space

$$\mathcal{B}_{\text{unb}} := \{(i,j) \in \mathbb{N}_0^2 : 1 \leq j \leq n\} \cup \{(0,0)\}, \tag{6.3.17}$$

which is unbounded in its first dimension. The transition rate diagram of $U(\cdot)$ is similar to the transition rate diagram of $X(\cdot)$ with the exception that for phases 1 until n there is no boundary that stops the process from transitioning further to the right, see Figure 6.2.

We interpret (6.3.5) in terms of the process $U(\cdot)$. The rough idea is as follows. The term $\mathcal{G}_{\rho_j}(z)$ corresponds to the number of transitions to the right performed by the process $U(\cdot)$ in phase j, before reaching phase $j+1$. So, the first term in (6.3.5) is a combinatorial object that

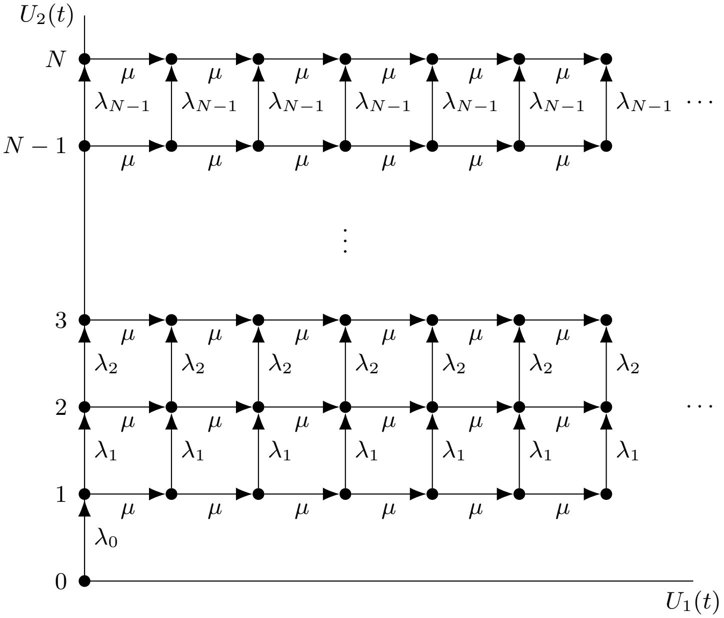
$U_2(t)$
N
μ μ μ μ μ μ
λ_{N-1} λ_{N-1} λ_{N-1} λ_{N-1} λ_{N-1} λ_{N-1} λ_{N-1} $\cdots$
$N-1$
μ μ μ μ μ μ
$\vdots$
3
μ μ μ μ μ μ
λ_2 λ_2 λ_2 λ_2 λ_2 λ_2 λ_2
2
$\cdots$
μ μ μ μ μ μ
λ_1 λ_1 λ_1 λ_1 λ_1 λ_1 λ_1
1
μ μ μ μ μ μ
λ_0
0
$U_1(t)$

Figure 6.2: Transition rate diagram of the Markov process $U(\cdot)$.

corresponds to all the ways in which the process $U(\cdot)$ can reach phase $j+1$, starting from $(0,0)$ or equivalently, $(0,1)$. On the other hand, the second term in (6.3.5) corresponds to only those paths that hit the diagonal $\mathcal{D}_0$ for the first time in state (i,i) for some $i = 1,2,\dots,j$. Intuitively, by subtracting these two terms, one is left with the trajectories of $U(\cdot)$ that do not intersect the diagonal and thus hit phase $j+1$ in one of the states $\{(0,j+1),(1,j+1),\dots,(j-1,j+1)\}$. Formally, we see that

$$P_{j+1}(z) = \mathbb{E}[z^{\sum_{i=1}^{n} G_{\rho_i}}] - \mathbb{E}[z^{B_{[j]}+\sum_{j=B_{[j]}}^{n} G_{\rho_j}}], \tag{6.3.18}$$

where the (defective) distribution of the random variable $B_{[j]}$ is given by $\mathbb{P}(B_{[j]} = k) = \mathbb{P}(B = k)$ for $k = 1,2,\dots,j$ and $\mathbb{P}(B_{[j]} = k) = 0$ otherwise. Another perspective on (6.3.18) is the following:

$$P_{j+1}(z) = \sum_{k=0}^{\infty} \Big(\mathbb{P}\Big(\sum_{i=1}^{j} G_{\rho_i} = k\Big) - \mathbb{P}\Big(B_{[j]} + \sum_{i=B_{[j]}}^{j} G_{\rho_i} = k\Big)\Big) z^k. \tag{6.3.19}$$

We have the inclusion of events

$$\Big\{B_{[j]} + \sum_{i=B_{[j]}}^{n} G_{\rho_i} = k\Big\} \subset \Big\{\sum_{i=1}^{j} G_{\rho_i} = k\Big\}, \tag{6.3.20}$$

so that

$$P_{n+1}(z) = \sum_{k=0}^{\infty} \mathbb{P}\Big(\Big\{\sum_{i=1}^{j} G_{\rho_i} = k\Big\} \setminus \Big\{B_{[j]} + \sum_{i=B_{[j]}}^{j} G_{\rho_i} = k\Big\}\Big) z^k. \tag{6.3.21}$$

Equation (6.3.21) formalizes the intuitive interpretation of (6.3.5) given above. In yet other words, relation (6.3.5) is a consequence of a decomposition of the sample space Ω associated with the process $U(\cdot)$. The space is decomposed as $\Omega = (\Omega \cap B) \cup (\Omega \cap B^c)$, where B denotes the event in which the process $U(\cdot)$ does not hit $\mathcal{D}_0$ before hitting phase $j+1$. The event B^c is then further decomposed in the disjoint events corresponding to the process $U(\cdot)$ hitting $\mathcal{D}_0$ at different phases $i = 1, \ldots, j$.

6.4 The maximum queue length in the first busy period

In this section, we investigate the distribution of the maximum number of customers during the first busy period, i.e., the quantity

$$M := \max_{0 \le t \le H(\mathcal{D}_0)} (X_2(t) - X_1(t)) \tag{6.4.1}$$

under the measure $\mathbb{P}(\,\cdot \mid X(0) = (0,0))$. We will compute

$$b_{m,j} := \mathbb{P}(M \le m,\ B = j), \quad 1 \le m, j \le n. \tag{6.4.2}$$

Summing over all relevant values of j, we get

$$\mathbb{P}(M = m) = \sum_{j=m}^{n} (b_{m,j} - b_{m-1,j}). \tag{6.4.3}$$

We adopt the generating function approach of Section 6.3. Define, for $2 \le m, j \le n$ and $\max(0, j-m) \le i \le j-2$,

$$p_{m,j}(i) := \mathbb{P}_{(0,0)}(H(\mathcal{P}_j) < H(\mathcal{D}_0 \cup \mathcal{D}_{m+1}),\ X(H(\mathcal{P}_j)) = (i,n)) \tag{6.4.4}$$

as the probability that, conditional on $X(0) = (0,0)$, the Markov process $X(\cdot)$ reaches phase j in state (i,j) without residing in $\mathcal{D}_0 \cup \mathcal{D}_{m+1}$ before phase j is reached. Define its generating function, for $z \in \mathbb{C}$, by

$$P_{m,j}(z) := \sum_{i=j-m}^{j-2} p_{m,j}(i) z^i, \quad 2 \leq m, j \leq n. \tag{6.4.5}$$

For $m \geq j$, notice that $P_{m,j}(z) = P_j(z)$ and $b_{m,j} = b_j$, which makes the case $m \geq j$ not interesting. If $n = 1$, then $b_{1,1} = 1$. For $n > 1$, we have by the strong Markov property that $b_{1,1} = \rho_1$ and for $2 \leq m \leq j \leq n-1$,

$$\rho_j^j P_{m,j}(\rho_j^{-1}) = b_{m,j}, \quad P_{k,j}(1) = b_{k,j},\ 2 \leq k \leq j. \tag{6.4.6}$$

The following theorem identifies $P_{m,j}(z)$:

Theorem 21. *For $2 \leq m < j+1 \leq n$, the generating functions are explicitly given by*

$$\begin{aligned} P_{m,j+1}(z) = P_m(z) \prod_{k=m}^{j} \mathcal{G}_{\rho_k}(z) - \sum_{i=m}^{j} b_{m,i} z^i \prod_{k=i}^{j} \mathcal{G}_{\rho_k}(z) \\ - \sum_{i=m}^{j} (1-\rho_i) p_{m,i}(i-m) z^{i-m} \prod_{k=i+1}^{j} \mathcal{G}_{\rho_k}(z) \end{aligned} \tag{6.4.7}$$

with the convention that the empty product $\prod_{k=j+1}^{j}(\cdot) = 1$.

Proof. We proceed similarly as in the proof of Theorem 20. Figure 6.3 can be used as a visual aid. We assume that $3 \leq m < j+1$ for m fixed; the case $m = 2$ is similar. From the strong Markov property at time $H(\mathcal{P}_j)$ we can write for i such that $j+1-m \leq i \leq j-2$

$$p_{m,j+1}(i) = \sum_{k=j-m}^{i} p_{m,j}(k) \rho_j^{i-k} (1-\rho_j), \tag{6.4.8}$$

$$p_{m,j+1}(j-1) = \sum_{k=j-m}^{j-2} p_{m,j}(k) \rho_j^{j-1-k} (1-\rho_j). \tag{6.4.9}$$

Multiply both sides of (6.4.8) by z^i and sum over all i with $j+1-m \leq i \leq j-2$. Further, multiply both sides of (6.4.9) by z^{j-1}. Sum the two

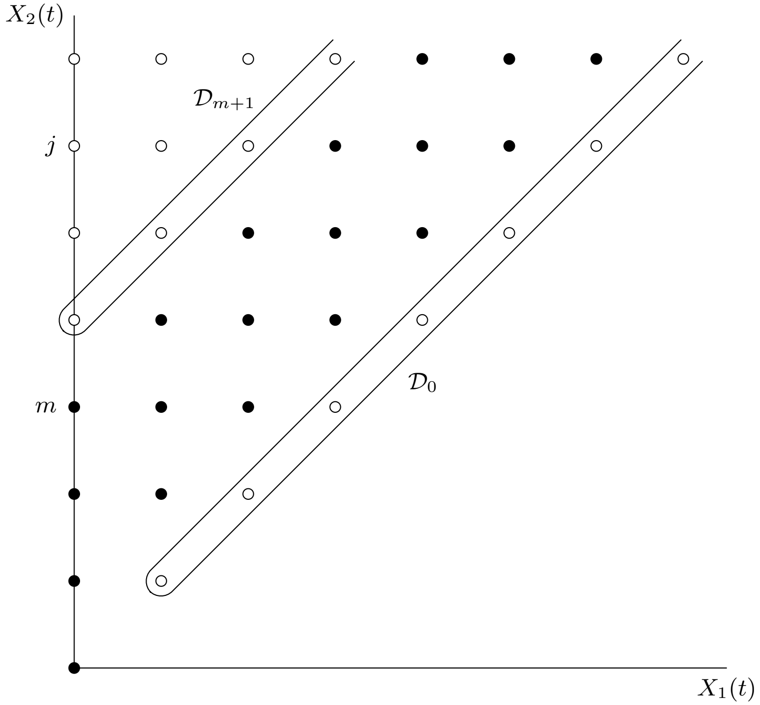

Figure 6.3: Visual aid for determining M. The process is only allowed to hit the black states.

resulting expressions to get

$$\begin{aligned} P_{m,j+1}(z) = &\sum_{i=j+1-m}^{j-2} \sum_{k=j-m}^{i} p_{m,j}(k)\rho_n^{i-k}(1-\rho_j)z^i \\ &+ \sum_{k=j-m}^{j-2} p_{m,j}(k)\rho_j^{j-1-k}(1-\rho_j)z^{j-1}. \end{aligned} \tag{6.4.10}$$

Isolate the summand corresponding to $k = j - m$ in the double summation and switch the order of the double summation to obtain

$$\begin{aligned}P_{m,j+1}(z) &= (1-\rho_j)\\ &\times \Big(\sum_{k=j+1-m}^{j-2} p_{m,j}(k) \sum_{i=k}^{j-2} \rho_j^{i-k} z^i + p_{m,j}(j-m) \sum_{i=j+1-m}^{j-2} \rho_j^{i-(j-m)} z^i \\ &\quad + \sum_{k=j-m}^{j-2} p_{m,j}(k) \rho_j^{j-1-k} z^{j-1} \Big). \end{aligned} \tag{6.4.11}$$

Simplifying the geometric series yields

$$\begin{aligned}P_{m,j+1}(z) &= (1-\rho_j)\\ &\times \Big(\sum_{k=j+1-m}^{j-2} p_{m,j}(k) \frac{z^k - \rho_j^{j-1-k} z^{j-1}}{1-\rho_j z} + p_{m,j}(j-m) \frac{\rho_j z^{j+1-m} - \rho_j^{m-1} z^{j-1}}{1-\rho_j z} \\ &\quad + \sum_{k=j-m}^{j-2} p_{m,j}(k) \rho_j^{j-1-k} z^{j-1} \Big). \end{aligned} \tag{6.4.12}$$

Rewriting the expression produces

$$\begin{aligned}P_{m,j+1}(z) = \frac{1-\rho_j}{1-\rho_j z} \Big(&\sum_{k=j-m}^{j-2} p_{m,j}(k) z^j - (1-\rho_j z) p_{m,j}(j-m) z^{j-m} \\ &- \rho_j^j z^j \sum_{k=j-m}^{j-2} p_{m,j}(k) \rho_j^{-k} \Big). \end{aligned} \tag{6.4.13}$$

By recognizing the generating function of a geometric random variable, the definition of the generating function $P_{m,j}(z)$, and the probability $b_{m,j}$, we finally obtain the relation

$$P_{m,j+1}(z) = \mathcal{G}_{\rho_j}(z)(P_{m,j}(z) - b_{m,j} z^j) - (1-\rho_j) p_{m,j}(j-m) z^{j-m}. \tag{6.4.14}$$

Iterating (6.4.14) and using the fact that $P_{m,m}(z) = P_m(z)$ proves the claim. □

Theorem 21 gives the following:

Corollary 6. *For $n > 1$, the probabilities $b_{m,j}$ satisfy the recursion on j, for $2 \le m < j \le n-1$,*

$$\begin{aligned} b_{m,j} = \rho_j^j \binom{n-1}{j-1} - \sum_{i=1}^{j-1} b_{m,i} \rho_j^{j-i} \binom{n-i}{j-i} \\ - \sum_{i=m}^{j-1} (1-\rho_i) p_{m,i}(i-m) \rho_j^{m+j-i} \binom{n-(i+1)}{j-(i+1)} \end{aligned} \tag{6.4.15}$$

and for $2 \le m < n$,

$$b_{m,n} = 1 - \sum_{i=1}^{n-1} b_{m,i} - \sum_{i=m}^{n-1} (1-\rho_i) p_{m,i}(i-m). \tag{6.4.16}$$

with initial terms $b_1, b_2, \ldots, b_m$ calculated from Corollary 5.

Proof. Combining the result of Theorem 21 with (6.4.6) yields the following recursion, for $2 \le m < j \le n-1$,

$$\begin{aligned} b_{m,j} = \rho_j^j P_m(\rho_j^{-1}) \prod_{i=m}^{j-1} \mathcal{G}_{\rho_i}(\rho_j^{-1}) - \sum_{i=m}^{j-1} b_{m,i} \rho_j^{j-i} \prod_{k=i}^{j-1} \mathcal{G}_{\rho_k}(\rho_j^{-1}) \\ - \sum_{i=m}^{j-1} (1-\rho_i) p_{m,i}(i-m) \rho_j^{m+j-i} \prod_{k=i+1}^{j-1} \mathcal{G}_{\rho_k}(\rho_j^{-1}) \end{aligned} \tag{6.4.17}$$

and for $2 \le m < n$,

$$b_{m,n} = P_m(1) - \sum_{i=m}^{n-1} b_{m,i} - \sum_{i=m}^{n-1} (1-\rho_i) p_{m,i}(i-m). \tag{6.4.18}$$

Notice that the term $P_m(1)$ in (6.4.18) is the probability that the process reaches phase m before it reaches the diagonal $\mathcal{D}_0$, so $P_m(1) = 1 - \sum_{i=1}^{m-1} b_i$.

The term $\rho_j^j P_m(\rho_j^{-1})$ in (6.4.17) can be simplified by using Theorem 20 and the fact that

$$\prod_{k=l}^{L} \mathcal{G}_{\rho_k}(\rho_j^{-1}) = \frac{\binom{n-l}{L+1-l}}{\binom{j-l}{L+1-l}}, \quad l \le L < j. \tag{6.4.19}$$

Employing this simplification and using that $b_n = b_{m,n}$ if $m \ge n$, we find the claimed result after some rewriting. □

Note that equation (6.4.16) can be written as

$$\sum_{i=1}^{n} b_{m,i} = 1 - \sum_{i=m}^{n-1} (1-\rho_i) p_{m,i}(i-m). \tag{6.4.20}$$

The interpretation of this relation is as follows. The left-hand side equals the probability $\mathbb{P}(M \leq m)$. The term $(1-\rho_i)p_{m,i}(i-m)$ at the right-hand side sum is equal to the probability that the process visits the superdiagonal $\mathcal{D}_{m+1}$ for the first time at phase $i \geq m$.

6.5 Coupling different finite-pool queues

In the previous sections we have developed various recursions for the $\Delta_{(i)}/G/1$ queue by relating the distribution of the process at different *phases*. In this section we expand on this idea by developing recursions that involve a different number of *initial customers in the pool*. We use the superscript $x^{(n)}$ whenever we want to emphasize the dependence of a certain quantity x on the initial number n of customers in the pool. For example, $X_2^{(n)}(t)$ denotes the number of customers who have joined the system by time t, when there were n customers in the pool at time zero. For our proofs, we construct an explicit coupling between $X^{(n+1)}(\cdot)$ and $X^{(j)}(\cdot)$, for $j \leq n$. Note that

$$\lambda_i^{(n+1)} = \lambda((n+1)-i) = \lambda(n-(i-1)) = \lambda_{i-1}^{(n)}. \tag{6.5.1}$$

Equation (6.5.1) expresses in precise terms the simple observation that, when we consider a $\Delta_{(i)}/G/1$ queue with $n+1$ initial customers and we disregard the first customer that arrives at the queue, we obtain a $\Delta_{(i)}/G/1$ queue with n initial customers. Therefore, we couple $X^{(n)}(\cdot)$ and $X^{(n+1)}(\cdot)$ by considering the state space $\mathcal{S}^{(n)}$ as a subset of $\mathcal{S}^{(n+1)}$ and by letting the transition probabilities be determined by the same rate $\lambda_i^{(n+1)}$, $i = 2, \dots, n+1$ (resp. rate μ) exponential clocks.

6.5.1 The number of customers in the first busy period

In this section we develop another recursive expression for the distribution of the number of customers in the first busy period. We will express $b_j^{(n)}$ as a function of $b_i^{(l)}$ for $l = 1, \dots, n-1$ and $i = 1, \dots, l$. We define the probability generating function of the number $B^{(n)}$ of customers served in the first busy period in a system with n total customers as

$$P_B^{(n)}(z) := \sum_{i=1}^{n} \mathbb{P}(B^{(n)} = i) z^i. \tag{6.5.2}$$

Next we show that $P_B^{(n)}(\cdot)$ solves a recursion formula similar to (6.3.10).

Theorem 22. *For $z \in \mathbb{C}$,*

$$P_B^{(n)}(z) = \Big(\rho_1^{(n)} + (1-\rho_1) \sum_{k=1}^{n-1} s_k^{(n-1)} z^k P_B^{(n-1-k)}(z) \Big) z. \tag{6.5.3}$$

Proof. We split $P_B^{(n)}(z)$ as

$$P_B^{(n)}(z) = \rho_1^{(n)} z + \sum_{i=1}^{n-1} \mathbb{P}(B^{(n)} = 1+i) z^{1+i}. \tag{6.5.4}$$

By the coupling given by (6.5.1), we have that, for $i \geq 1$,

$$\begin{aligned} &\mathbb{P}(S^{(n)} = 1+i) \\ &\quad = (1-\rho_1) \sum_{k=1}^{i} \mathbb{P}(S^{(n-1)} = k) \mathbb{P}(S^{(n-1-k)} = i-k), \end{aligned} \tag{6.5.5}$$

with the convention that $\mathbb{P}(S^{(0)} = 0) = 1$. See Figure 6.4 for an example. This allows us to rewrite (6.5.4) as

$$\begin{aligned} &P_B^{(n)}(z) \\ &= \rho_1^{(n)} z + (1-\rho_1) \sum_{i=1}^{n-1} \sum_{k=1}^{i} \mathbb{P}(S^{(n-1)} = k) \mathbb{P}(S^{(n-1-k)} = i-k) z^{1+i} \\ &= \rho_1^{(n)} z + (1-\rho_1) \sum_{k=1}^{n-1} \mathbb{P}(S^{(n-1)} = k) z \sum_{i=k}^{n-1} \mathbb{P}(S^{(n-1-k)} = i-k) z^i \\ &= z \Big(\rho_1^{(n)} + (1-\rho_1) \sum_{k=1}^{n-1} s_k^{(n-1)} z^k \sum_{i=k}^{n-1} \mathbb{P}(S^{(n-1-k)} = i-k) z^{i-k} \Big). \end{aligned} \tag{6.5.6}$$

The inner sum can be rewritten using the definition (6.5.2), leading to

$$P_B^{(n)}(z) = z \Big(\rho_1^{(n)} + (1-\rho_1) \sum_{k=1}^{n-1} s_k^{(n-1)} z^k P_B^{(n-1-k)}(z) \Big), \tag{6.5.7}$$

and concluding the proof. □

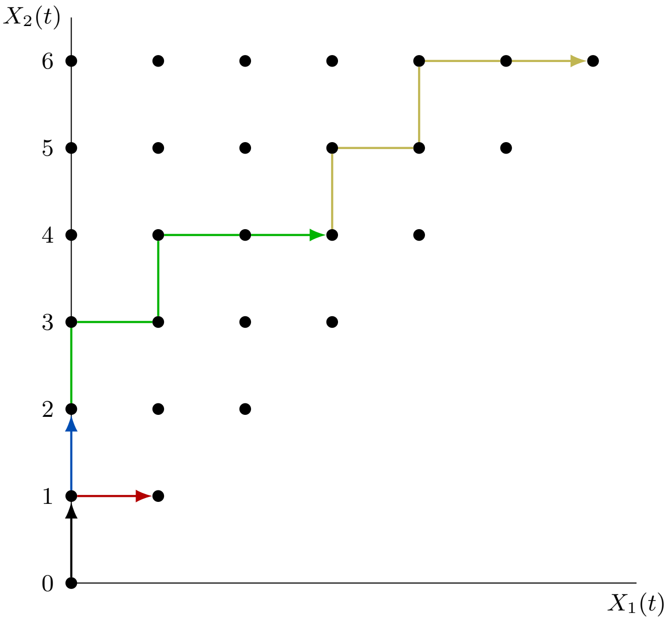

Figure 6.4: An example of the sample-path coupling obtained by (6.5.1) with $n = 6$. We have that $\mathbb{P}(\rightarrow) = \rho_1^{(n)}$ and $\mathbb{P}(\uparrow) = 1 - \rho_1^{(n)}$. In the first event, the first busy period consists of one service. In the second event, we split the busy period in the busy periods of two coupled $\Delta_{(i)}/G/1$ queues. The $\rightarrow$ path represents the first busy period of a queue with $n-1$ initial customers. The $\rightarrow$ path represents the first busy period of a queue with $n-3 = 3$ initial customers.

6.5.2 The total number of busy periods

The recursive techniques that we have introduced in the previous section can be used to obtain further insight on the performance of the $\Delta_{(i)}/G/1$ queue. As an example, in this section we compute the distribution of the total number of busy periods before the pool of customers depletes. We do so by recursively conditioning on the number of services in all of the n busy periods.

Denote by $K^{(n)}$ the total number of busy periods until all customers have been served when initially the number of customers in the pool is n. Trivially, $1 \leq K^{(n)} \leq n$ almost surely. Our aim is to determine $\mathbb{P}(K^{(n)} = j)$

for $j = 1, 2, \dots, n$. We see that $\mathbb{P}(K^{(n)} = 1) = b_n^{(n)}$, so we focus on the cases $j = 2, 3, \dots, n$. We use the recursive structure of the $\Delta_{(i)}/G/1$ queue to develop an expression for the probability generating function of $K^{(n)}$. To that end, define $P_K^{(0)}(z) := 1$ and

$$P_K^{(n)}(z) := \sum_{j=1}^{n} \mathbb{P}(K^{(n)} = j) z^j, \quad z \in \mathbb{C},\ n \geq 1. \tag{6.5.8}$$

The next theorem shows that $P_K^{(n)}(\cdot)$ satisfies a recursion in n:

Theorem 23. *For $z \in \mathbb{C}$*

$$P_K^{(n)}(z) = \Big(\sum_{k_1=1}^{n} b_{k_1}^{(n)} P_K^{(n-k_1)}(z) \Big) z. \tag{6.5.9}$$

Proof. Let $B^{(n)}$ denote the number of services during the first busy period if initially the number of customers in the pool is n. By conditioning on the number of services during the first busy period we obtain the expression

$$\mathbb{P}(K^{(n)} = j) = \sum_{k_1=1}^{n-(j-1)} \mathbb{P}(K^{(n)} = j \mid B^{(n)} = k_1) \mathbb{P}(B^{(n)} = k_1). \tag{6.5.10}$$

Invoking the strong Markov property at the time at which the queue empties for the first time, we see that

$$\mathbb{P}(K^{(n)} = j \mid B^{(n)} = k_1) = \mathbb{P}(K^{(n-k_1)} = j - 1) \tag{6.5.11}$$

and thus

$$\mathbb{P}(K^{(n)} = j) = \sum_{k_1=1}^{n-(j-1)} b_{k_1}^{(n)} \, \mathbb{P}(K^{(n-k_1)} = j - 1). \tag{6.5.12}$$

Multiply both sides of (6.5.12) by z^j, sum over all $j = 2, 3 \dots, n$, and add the equality $\mathbb{P}(K^{(n)} = 1) z = b_n^{(n)} z$ to obtain

$$P_K^{(n)}(z) = b_n^{(n)} z + \sum_{j=2}^{n} \sum_{k_1=1}^{n-(j-1)} b_{k_1}^{(n)} \, \mathbb{P}(K^{(n-k_1)} = j - 1) z^j. \tag{6.5.13}$$

Switch the order of the double summation and simplify using the definition of the probability generating function. This yields

$$\begin{aligned} P_K^{(n)}(z) &= \Big[b_n^{(n)} + \sum_{k_1=1}^{n-1} b_{k_1}^{(n)} P_K^{(n-k_1)}(z)\Big] z \\ &= \Big(\sum_{k_1=1}^{n} b_{k_1}^{(n)} P_K^{(n-k_1)}(z)\Big) z, \end{aligned} \tag{6.5.14}$$

concluding the proof. □

Theorem 23 can be used to obtain performance meaures. For example, taking derivatives with respect to z on both sides of (6.5.9) and setting $z = 1$ provides the mean number of busy periods until the system is empty:

$$\begin{aligned} \mathbb{E}[K^{(n)}] &= \sum_{k_1=1}^{n} b_{k_1}^{(n)} P_K^{(n-k_1)}(1) + \sum_{k_1=1}^{n} b_{k_1}^{(n)} \mathbb{E}[K^{(n-k_1)}] \\ &= 1 + \sum_{k_1=1}^{n-1} b_{k_1}^{(n)} \mathbb{E}[K^{(n-k_1)}], \end{aligned} \tag{6.5.15}$$

with the convention that $\mathbb{E}[K^{(0)}] = 0$. Iterating the recursion (6.5.15) gives the expression

$$\begin{aligned} \mathbb{E}[K^{(n)}] = 1 + (1 - b_n^{(n)}) + \Big(1 - \sum_{k_1=1}^{n-1} b_{k_1}^{(n)} b_{n-k_1}^{(n-k_1)}\Big) + \dots \qquad (6.5.16) \\ + \Big(1 - \sum_{k_1=1}^{2} \sum_{k_2=1}^{3-k_1} \cdots \sum_{k_{n-2}=1}^{(n-1)-\dots-k_{n-3}} b_{k_1}^{(n)} b_{k_2}^{(n-k_1)} \cdots b_{n-k_1-\dots-k_{n-2}}^{(n-k_1-\dots-k_{n-2})}\Big) \end{aligned}$$

The second factorial moment can be obtained by taking derivatives twice with respect to z on both sides of (6.5.9) and setting $z = 1$. This yields

$$\begin{aligned} \mathbb{E}[K^{(n)}(K^{(n)} - 1)] &= 2 \sum_{k_1=1}^{n} b_{k_1}^{(n)} \mathbb{E}[K^{(n-k_1)}] \\ &\quad + \sum_{k_1=1}^{n} b_{k_1}^{(n)} \mathbb{E}[K^{(n-k_1)}(K^{(n-k_1)} - 1)]. \end{aligned} \tag{6.5.17}$$

Combining (6.5.15) and (6.5.17) gives

$$\begin{aligned}
\mathrm{Var}(K^{(n)}) &= \mathbb{E}[K^{(n)}(K^{(n)}-1)] + \mathbb{E}[K^{(n)}] - (\mathbb{E}[K^{(n)}])^2 \\
&= \sum_{k_1=1}^{n} b_{k_1}^{(n)} \mathbb{E}[K^{(n-k_1)}(K^{(n-k_1)}-1)] \\
&\quad + \sum_{k_1=1}^{n} b_{k_1}^{(n)} \mathbb{E}[K^{(n-k_1)}] - \Big(\sum_{k_1=1}^{n} b_{k_1}^{(n)} \mathbb{E}[K^{(n-k_1)}] \Big)^2. \\
&= \sum_{k_1=1}^{n} b_{k_1}^{(n)} \mathbb{E}[(K^{(n-k_1)})^2] - \Big(\sum_{k_1=1}^{n} b_{k_1}^{(n)} \mathbb{E}[K^{(n-k_1)}] \Big)^2. \qquad (6.5.18)
\end{aligned}$$

Equations (6.5.18) has a simple interpretation. We see that the right-hand side is equal to $\mathbb{E}[(K^{(n-B^{(n)})})^2] - \mathbb{E}[K^{(n-B^{(n)})}]^2 = \mathrm{Var}(K^{(n-B^{(n)})})$. Noting then that $K^{(n)} \overset{\mathrm{d}}{=} 1 + K^{(n-B^{(n)})}$ gives a more straightforward proof of (6.5.18).

6.6 Conclusions

In this chapter we have studied the $\Delta_{(i)}/G/1$ queue for a finite and fixed number of customers n. Assuming exponentially distributed arrival and service times, we have analyzed the two-dimensional Markov process representing the number of completed services and the number of customers who have joined the queue. Exploiting the recursive structure of the Markov chain, we have derived an explicit expression for the join probability mass function of the number of customers served and the maximum queue length in the first busy period. We have also illustrated how the recursive structure can be exploited further to obtain explicit expressions for other quantities of interest.

CHAPTER 7

Open problems

In this thesis we have studied the $\Delta_{(i)}/G/1$ queue, a model for a queue where only a finite number n of customers can join. We have focused on the *heavy-traffic regime*, defined as follows: We let the customer pool n grow, while speeding up the service time so that, on average, approximately one customer arrives during one service. This assumption must hold at the peak of congestion of the queue, and we assume that this happens at time 0. Our heavy-traffic assumption gives rise to a negative quadratic (resp. polynomial) drift in the $n \to \infty$ limit of the queue-length process. This *depletion-of-points effect* exactly describes the influence of the finite pool of customers on the dynamics of the queue. Surprisingly, our results reveal that the depletion-of-points effect becomes relevant only after n^δ services with $\delta < 1$, that is, when there are still $n - n^\delta \approx n$ customers left in the pool. We have investigated this behavior for various $\Delta_{(i)}/G/1$ models and found that it holds with remarkable generality. Our results suggest that, when the variance of the service times is finite, the depletion-of-points effect invariably appears as a negative quadratic (resp. polynomial) drift. On the other hand, when the variance of the service times is infinite, the depletion-of-points effect is more subtle and requires a delicate analysis. In the remainder of the section we discuss this issue and other interesting open problems that arise from the study of the $\Delta_{(i)}/G/1$ queue.

Size-biased arrival times, infinite-variance service times. In Chapter 5 we studied in detail the $\Delta^{\alpha}_{(i)}/G/1$ queue with size-biased exponential arrival clocks. For this model,

$$\mathbb{P}(i \text{ joins during service of } j | D_i, D_j) \approx \frac{D_i^{\alpha} D_j}{n}, \tag{7.1}$$

for some $\alpha \in [0,1]$. If we interpret customers as vertices, and the service time of customer i as a *weight* associated to vertex i, we see that the $\Delta^{\alpha}_{(i)}/G/1$ queue is equivalent to the exploration process of an inhomogeneous random graph. For $\alpha = 1$ we retrieve the well-known rank-1 inhomogeneous random graph. In Chapter 5 we have shown that, if the variance of the *size-biased* service times is finite, i.e. $\mathbb{E}[S^{2+\alpha}] < \infty$, the $\Delta^{\alpha}_{(i)}/G/1$ queue-length process converges to the same limit as the standard $\Delta_{(i)}/G/1$ queue ($\alpha = 0$). When the third moment of the weights is infinite, Bhamidi, van der Hofstad and van Leeuwaarden [16] have shown that, for a certain choice of *deterministic weights*, the exploration process of the rank-1 inhomogeneous random graph converges to a so-called *thinned Lévy process* defined as

$$\mathcal{S}(t) = b + ct + \sum_{i=1}^{\infty} b i^{-1/\gamma} (\mathcal{I}_i(t) - a t i^{-1/\gamma}), \tag{7.2}$$

where $a, b > 0$ and $c \in \mathbb{R}$ are constants, γ is the power-law tail exponent as in (4.1.1), $\mathcal{I}_i(t) = \mathbb{1}_{\{E_i \leq a t i^{-1/\gamma}\}}$, and E_i are mean one i.i.d. exponential random variables. See also [32] for an analogous result for the configuration model. The authors of [32] also argue that, when the degrees are given by an i.i.d. sequence $(\mathcal{W}_i)_{i=1}^{n}$ following a power-law distribution, conditioned on $(\mathcal{W}_i)_{i=1}^{n}$ the limit (7.2) holds with the following modification. The term i in the sum is replaced by $\Gamma_i := \sum_{j=1}^{i} \bar{E}_j$, where $\bar{E}_j$ are i.i.d. rate one exponential random variables. This substitution gives

$$\mathcal{S}(t) = b + ct + \sum_{i=1}^{\infty} b \Gamma_i^{-1/\gamma} (\mathcal{I}_i(t) - a t \Gamma_i^{-1/\gamma}). \tag{7.3}$$

and

$$\mathcal{I}_i(t) := \mathbb{1}_{\{E_i \leq a t \Gamma_i^{-1/\gamma}\}}. \tag{7.4}$$

Note that $\mathbb{E}[\Gamma_i] = i$. Roughly speaking, the first term $b\Gamma_i^{-1/\gamma}$ in the summation in (7.2) represents the size of the jump of the exploration process when a high-weight vertex is found. On the other hand, the second term $(\mathcal{I}_i(t) - at\Gamma_i^{-1/\gamma})$ represents the arrival process of the high-weight vertices. Consider now our $\Delta_{(i)}^{\alpha}/G/1$ model with service times S_i such that

$$1 - F_S(t) = ct^{-\gamma} \tag{7.5}$$

with $t > c^{\gamma}$ and $\gamma \in (1+\alpha, 2+\alpha)$, $\alpha \in (0,1)$. As a generalization of (7.3), we conjecture that, *conditioned on the service times* $(S_i)_{i=1}^n$, the embedded queue process $Q_n^e(\cdot)$ converges to

$$\mathcal{S}_\alpha(t) = b + ct + \sum_{i=1}^{\infty} b\Gamma_i^{-1/\gamma}(\mathcal{I}_{i,\alpha}(t) - at\Gamma_i^{-\alpha/\gamma}) \tag{7.6}$$

with $\mathcal{I}_{i,\alpha}(t) := \mathbb{1}_{\{E_i \le at\Gamma_i^{-\alpha/\gamma}\}}$. It is reasonable to expect that setting $\alpha = 0$ in (7.6) yields a γ-stable motion with quadratic drift, consistently with our results in Chapter 4. To see this, we resort to the series representation for γ-stable random measures [85, Chapter 3.10]. The representation holds for processes defined on $[0,1]$ and is given by

$$X_\gamma(t) = \sum_{i=1}^{\infty}(\Gamma_i^{-1/\gamma}\mathbb{1}_{\{U_i \le t\}} - d_i t), \qquad \gamma \in (1,2), \tag{7.7}$$

where $(U_i)_{i=1}^n$ are i.i.d. uniform random variables and d_i is such that $d_i \sim i^{-1/\gamma}$. The process $X_\gamma(\cdot)$ in (7.7) is then a γ-stable motion. Let us define $F_E(t) = 1 - \mathrm{e}^{-t}$. Applying the time-change $t \mapsto F_E^{-1}(t)$ to (7.6) and approximating $\Gamma_i^{-1/\gamma} \approx i^{-1/\gamma} \approx d_i$ we get

$$\mathcal{S}_\alpha(F_E^{-1}(t)) \overset{\mathrm{d}}{=} b + cF_E^{-1}(t) + b\sum_{i=1}^{\infty}\Gamma_i^{-1/\gamma}(\mathbb{1}_{\{U_i \le t\}} - aF_E^{-1}(t)). \tag{7.8}$$

It turns out that $a = 1$ when $\alpha = 0$. By Taylor expanding $F_E^{-1}(t)$ as $F_E^{-1}(t) = t + t^2/2 + o(t^2)$, we obtain

$$\mathcal{S}_\alpha(t) \approx b + ct + bX_\gamma(t) - \frac{t^2}{2}\sum_{i=1}^{\infty} b\Gamma_i^{-1/\gamma}, \qquad t \ll 1. \tag{7.9}$$

Note that the coefficient of the quadratic drift in (7.9) is *random*. In fact, the limit process we obtained in Theorem 17 is an annealed version of (7.9), averaged over the laws of the Γ_i. The expression (7.9) is only formal, since the summation on the right-hand side diverges. This suggests that the heuristic argument above is too crude, and a more subtle analysis is needed. Indeed, it would be interesting to formalize this argument, and in particular to understand why it yields the expected result only for small t.

Joseph [55] investigates the configuration model with i.i.d. heavy-tailed weights. In particular, he shows that the exploration process converges to $Y(\cdot) + A(\cdot)$, where $A(\cdot) = -ct^{\gamma-1}$ is a negative deterministic drift and c is given in terms of the Gamma function. Moreover, $Y(\cdot)$ is uniquely characterized by having independent increments and Fourier transform

$$\mathbb{E}[\exp(iuY(t)] = \exp\Big(\int_0^t \int_0^\infty (\mathrm{e}^{iux} - 1 - iux) a \frac{1}{x^\gamma} \mathrm{e}^{-bxs} \mathrm{d}s\mathrm{d}x\Big). \quad (7.10)$$

Here $a, b \in \mathbb{R}$ are unimportant constants. It is rather straightforward to adapt the arguments in [55] to extend this result to our $\Delta^\alpha_{(i)}/G/1$ model. We find that, when the service times follow the power-law (7.5), the $\Delta^\alpha_{(i)}/G/1$ embedded queue process converges to the process $\widehat{Q}^e(\cdot) = \widehat{Y}(\cdot) + \widehat{A}(\cdot)$, where $\widehat{Y}(\cdot)$ has independent increments and Fourier transform

$$\mathbb{E}[\exp(iu\widehat{Y}(t))] = \exp\Big(\int_0^t \int_0^\infty (\mathrm{e}^{iux} - 1 - iux) a \frac{1}{x^{\gamma+1-\alpha}} \mathrm{e}^{-bx^\alpha s} \mathrm{d}s\mathrm{d}x\Big), \quad (7.11)$$

and $a, b \in \mathbb{R}$ are again unimportant constants. The term $\mathrm{e}^{-bx^\alpha s}$ in (7.11) accounts for the α-size-biased order of arrival of the customers. Moreover, the techniques of [55] imply that, when $\gamma \in (1 + 2\alpha, 2 + \alpha)$, the deterministic drift $A(\cdot)$ is given by

$$\widehat{A}(t) = -\frac{\mathbb{E}[X^{1+2\alpha}]}{2\mathbb{E}[X^\alpha]} t^2. \quad (7.12)$$

Interestingly, the drift (7.12) coincides with the drift for the $\mathbb{E}[S^{2+\alpha}] < \infty$ case, when $\lambda = 1$. This is consistent with the fact that, when $\alpha < 1$, it is possible to choose γ such that $2 + \alpha > \gamma > 1 + 2\alpha$ and thus, even if $\mathbb{E}[S^{2+\alpha}] = \infty$, $\mathbb{E}[S^{1+2\alpha}]$ and $\mathbb{E}[X^\alpha]$ are finite. However, when $\alpha = 1$ we have $2 + \alpha = 1 + 2\alpha$ and there is a continuous phase transition from a drift with degree 2 to a drift with degree $\gamma - 1$, with $\gamma \in (2, 3)$. It is not clear what the drift, or even the limit process, should be when $\gamma \in (1 + \alpha, 1 + 2\alpha)$. We remark that, in order to obtain (7.11) , we have not conditioned on the service times $(S_i)_{i=1}^n$. It should be possible to show that the law of $\widehat{Y}(t) + \widehat{A}(t)$ coincides with the law of (7.6) when the latter is averaged over the Γ_i.

Excursions of drifted stable processes. Let us go back to the setting of Chapter 4, where we dealt with the $\Delta_{(i)}/G/1$ queue with heavy-tailed service times. We have shown that the free process $X_n(\cdot)$ converges in distribution to $\widehat{X}(\cdot)$, a γ-stable motion with negative quadratic drift. Very little is known about this class of processes. In particular, it is not known whether the excursions above past minima of $\widehat{X}(\cdot)$ can be ordered, that is if there is a well-defined maximal excursion above zero of the process $\phi(\widehat{X})(\cdot)$. This result would be instrumental in proving that, for $K \in \mathbb{N}$ and for a sufficiently large head start q, the first busy period of $\widehat{Q}(\cdot)$ is one of the K largest ones with probability close to 1. A striking property of $\phi(\widehat{X})(\cdot)$ is that $\sup_{t\geq 0} \phi(\widehat{X})(\cdot) = \infty$ almost surely. Indeed, it is well known that for any Lévy process $X(\cdot)$ with unbounded Lévy measure

$$\mathbb{P}(\forall N \in \mathbb{N},\ \forall T > 0\ \exists t \geq T : \Delta X(t) \geq N) = 1, \tag{7.13}$$

where $\Delta X(t) := X(t) - \lim_{s\to t^-} X(s)$. However, due to the parabolic drift the excursions of $\phi(\widehat{X})(\cdot)$ containing a large jump become smaller as time passes. This can be justified heuristically as follows. Recall that $\phi(\widehat{X})(t) = \phi(q + \beta t + \mathcal{S}(t) - 1/2t^2)$, where $\mathcal{S}(\cdot)$ is a γ-stable motion. Let us set $q = \beta = 0$ for simplicity. Let $(t_n)_{n=1}^{\infty}$ a sequence of time instants such that $t_n \to \infty$ and $\mathcal{S}(\cdot)$ performs a 'typical large jump' in t_n. With 'typical large jump' we mean that $\mathcal{S}(t_n) \approx t_n^{1/\alpha}$. We look for $0 < t \ll 1$ such that $\phi(X_n)(t_n + t)$ is zero for the first time after t_n. Equivalently, we look for $0 < t \ll 1$ such that

$$X_n(t_n + t) - X_n(t_n) = -t_n^{1/\alpha}. \tag{7.14}$$

Rewriting the left-hand side of (7.14) using the definition of $X_n(\cdot)$ gives

$$\begin{aligned}\mathcal{S}(t_n + t) - \frac{(t_n+t)^2}{2} - \left(\mathcal{S}(t_n) - \frac{t_n^2}{2}\right)\\ = (\mathcal{S}(t_n + t) - \mathcal{S}(t_n)) - \frac{t^2}{2} - t_n t.\end{aligned} \tag{7.15}$$

The first term is a mean-zero stable random variable and thus we ignore it at a first approximation. We also ignore the second term, since t^2 is of lower order than t. It follows that $t > 0$ should be such that $-t_n t = -t_n^{1/\alpha}$, that is $t = t_n^{(1-\alpha)/\alpha}$. Since $\alpha \in (1, 2)$, this suggests that the average excursion length decreases over time (as $t_n \to \infty$). In particular, by formalizing this argument it should be possible to show that the excursions of $\phi(\widehat{X})(\cdot)$ can be ordered by their length and that the largest one is finite.

Large deviations. Recall that the net-put process $P_n(\cdot)$ is defined as

$$P_n(t) := \sum_{i=1}^{\mathcal{A}_n(t)} S_i - c_n t, \tag{7.16}$$

where $\mathcal{A}_n(t) = \sum_{i=1}^{n} \mathbb{1}_{\{T_i \leq t\}}$ and c_n is the rescaled service rate. Equation (7.16) can be cast in a simpler form as

$$P_n(t) \overset{\mathrm{d}}{=} \sum_{i=1}^{n} (S_i \mathbb{1}_{\{T_i \leq t\}} - t), \tag{7.17}$$

where we have taken $c_n = n$. In (7.17), $P_n(t)$ is represented as a partial sum of i.i.d. random variables. It seems that much information can be gained by exploiting the representation (7.17). For fixed $n \in \mathbb{N}$ and assuming that the S_i follow a subexponential distribution we can estimate the probability $\mathbb{P}(P_n(t) > x)$ for fixed t and large x. Indeed, we have

$$\mathbb{P}(P_n(t) > x) \leq \mathbb{P}\Big(\sum_{i=1}^{n} S_i > x + t\Big) \sim n\mathbb{P}(S > x + t), \tag{7.18}$$

where $f(x) \sim g(x)$ means $\lim_{x\to\infty} f(x)/g(x) = 1$. In the large n regime we can be more precise, and we expect the following to hold

$$\mathbb{P}(P_n(t) > x) \sim nF_T(t)\mathbb{P}(S > x + t). \tag{7.19}$$

To obtain the heuristic (7.19) we have replaced $\mathcal{A}_n(t)$ by $nF_T(t)$ in (7.16). On the other hand, if the cumulant generating function $s \mapsto \Lambda(s)$ of $X_i(t) := S_i \mathbb{1}_{\{T_i \leq t\}} - t$ is finite for some $s > 0$, then (7.17) allows us to estimate, as $n \to \infty$,

$$\mathbb{P}(P_n(t) > nx) \approx \mathrm{e}^{-nI(x)}, \tag{7.20}$$

for fixed t and x and a rate function $I(\cdot)$ given by the Legendre transform of $\Lambda(\cdot)$. It should be possible to refine these basic results to obtain asymptotic estimates for the workload process $L_n(t) := \phi(P_n)(t)$. Finally, another interesting venue of investigation are asymptotics for the length of the busy period. For the G/G/1 queue with regularly varying service distribution, it is known that the probability of a large busy period is related to the probability of a large cycle maximum [102]. The $\Delta_{(i)}/G/1$ queue is transitory, thus the focus lies on the first busy period. Nevertheless, it should still be possible to relate the busy period to the cycle maximum. Results on the maximum of a Brownian motion with parabolic drift are given in [39, 40, 53].

Heavy traffic in $t_c > 0$. Throughout this thesis we have assumed that the $\Delta_{(i)}/G/1$ queue satisfies the heavy-traffic condition

$$\max_{t\geq 0} f_T(t)\mathbb{E}[S] = f_T(0)\mathbb{E}[S] = 1, \tag{7.21}$$

where $f_T(t)$ is the distribution function of the arrival time and $1/\mathbb{E}[S]$ is the service rate. Assumption (7.21) implies that the queue is never overloaded (i.e. it is never the case that $f_T(t)\mathbb{E}[S] > 1$), and is critically loaded at the moment of peak congestion t_c. Additionally, (7.21) implies that t_c is attained in 0. We have showed that the first assumption reveals the depletion-of-points effect of the $\Delta_{(i)}/G/1$ queue. On the other hand, the assumption $t_c = 0$ is of a technical nature and does not lend itself to a satisfying justification. We note that $t_c = 0$ is satisfied for the important case of exponential clocks T_i. However, it is often the case that the instant of peak congestion is significantly later than the instant in which service starts. It is therefore also of great practical interest to investigate the heavy-traffic behavior of the $\Delta_{(i)}/G/1$ queue for $t_c > 0$.

Mandelbaum and Massey [71, Theorem 3.4] give one of the first results for the case $t_c > 0$ in the context of the $M_t/M_t/1$ queue. They introduce an additional time parameter T_0 such that $T_0 \leq t_c$. Conditioning on $Q_n(T_0) = 0$, and assuming that $T_0 \nearrow t_c$ after a suitable scaling, they show that the distribution of the queue length $Q_n(t)$ in $t \geq T_0$ (for t sufficiently close to t_c) is that of a Brownian motion with negative quadratic (resp. polynomial) drift starting in 0 in T_0. In fact, Mandelbaum and Massey scale T_0 in such a way that it is inside the *critical window*, that is the queue is in heavy-traffic in T_0. Conditioning on $Q_n(T_0) = 0$ implies that there is no backlog of work in T_0 and thus the queue behaves as if $T_0 = t_c = 0$. It would be interesting to understand what happens to the queue-length process close to t_c when T_0 is fixed or, more generally, when $T_0 \nearrow t_c$ but T_0 lies outside the critical window. Recall that the negative quadratic drift can be interpreted as the effect of the transition of the queue from being critical to being subcritical. Restricting ourselves to the finite-variance case, we wish to study $Q_n(t_c + n^{-1/3}t)$ for $t \in ((T_0 - t_c)n^{1/3}, \infty)$ conditioned on $Q_n(T_0) = 0$. Then, fixing $T_0 < t_c$ and letting $n \to \infty$ intuitively corresponds to conditioning the queue to be 0 at $-\infty$. In this intuitive picture, in the time interval $(-\infty, 0)$ the queue transitions from being subcritical to being critical. Moreover, conditioned on $Q_n(0) = q$, we expect the queue for $t > 0$ to follow a Brownian path around a negative parabolic (resp. polynomial) drift, representing again the transition away from criticality.

Keller [58] gives a heuristic solution to the problem above in the context of the heavy-traffic $M_t/M_t/1$ queue. Because of the deep relation between the $\Delta_{(i)}/G/1$ queue and the $M_t/M_t/1$ queue, we conjecture that an analogous result holds for the $\Delta_{(i)}/G/1$ queue. Let us then describe the result of Keller. Recall that in the $M_t/M_t/1$ queue, arrivals (resp. services) occur according to a time-dependent rate $\lambda(t)$ (resp. $\mu(t)$). Keller scales the arrival and service rates as $\lambda(t/n)$ and $\mu(t/n)$, and focuses on the behavior of the system when $n \to \infty$. He then expresses the transient probability mass function $P(q,t,n) := \mathbb{P}(Q_n(t) = q)$ as a power series, for which he computes the first few terms. He shows that, when the queue is in heavy-traffic in t_c, $P(q,t,n)$ satisfies

$$P(n^{1/3}q, t_c + n^{-1/3}t, n) = c_1 n^{-1/3} P_1(c_1 q, c_2 t, 0) + O(n^{-2/3}), \qquad (7.22)$$

where c_1, c_2 are constants and with an abuse of notation we have written $n^{1/3}q$ instead of $\lfloor n^{1/3}q \rfloor$. Note that the first $c_1 n^{-1/3}$ term on the right-hand side is a normalization constant. Moreover, the probability mass function $(x,t) \mapsto P_1(x,t,0)$ is given by the solution of the following partial differential equation

$$\frac{\partial P_1}{\partial t} = \frac{1}{2}\frac{\partial^2 P_1}{\partial x^2} - t\frac{\partial P_1}{\partial x}, \qquad x > 0, \qquad (7.23)$$

with nonstandard boundary conditions

$$\frac{1}{2}\frac{\partial P_1}{\partial x}(0,t,0) - tP_1(0,t,0) = 0, \qquad (7.24)$$

$$P_1(x,t,0) \sim -2t\mathrm{e}^{-(-2t)x}, \qquad \text{as } t \to -\infty. \qquad (7.25)$$

Note that, the solution of (7.23) with boundary condition (7.24) and $P_1(x,0,0) = \delta_a(x)$ is the probability density function of a reflected Brownian motion with negative quadratic drift, starting in $a \geq 0$. Here $x \mapsto \delta_a(x)$ is the Dirac measure centered in a. The nonstandard boundary term (7.25) can be interpreted as follows: when $t \to -\infty$, the queue gradually moves outside of the critical window; for $t \ll -1$ the queue is approximately subcritical. It can be shown that the queue-length process of the subcritical $M_t/M_t/1$ queue converges pointwise to a geometric random variable G_t with parameter $1 - \rho(t)$. This is the stationary distribution of an associated M/M/1 queue; see Chapter 3.4 where we prove this for the $\Delta_{(i)}/G/1$ queue. As $t \to t_c$, $\rho(t) \to 1$, and G_t is well approximated by an exponential random variable. Therefore, the boundary condition (7.25) forces a continuous transition between the subcritical ($t = -\infty$) and the critical ($t = O(1)$) regimes of the queue. Note also that, as $t \to -\infty$, $-2te^{2tx} \to \delta_0(x)$, which is consistent with our previous intuition of conditioning the queue to be zero at $-\infty$.

The function $P_1(x,t,0)$ describes the evolution of the probability mass distribution as the queue evolves from subcritical, through criticality, to subcritical again. In Chapter 5 we have established a connection between the $\Delta_{(i)}^{\alpha}/G/1$ queue and the Norros-Reittu random graph. It would be interesting to investigate whether $P_1(x,t,0)$ also describes the evolution of an appropriate random graph process.

Higher-order contact criticality for random graphs. We now take the connection between the finite-pool queues and random graphs further, and focus on the ℓ-th order contact introduced in Section 2.4. By exploiting this connection it should be possible to develop a theory of *ℓ-th order criticality* for random graphs. More precisely, it would be interesting to investigate for which random graphs the exploration process converges to a stochastic process with negative polynomial (not quadratic) drift. Let us focus on the Erdős-Rényi random graph with connection probability $p = 1/n$ for simplicity. The number of vertices visited by the exploration process $t \mapsto X_n(t)$ grows linearly in time; in $[t, t+\delta]$ the exploration process visits $O(\delta)$ vertices. Therefore at time t the number of visited vertices is approximately t and the number of potential neighbors of the vertex currently being explored is approximately $n - t = n(1 - t/n)$. Consequently, the expected number of neighbors of the vertex being explored at time t is approximately $1 - t/n$. Here we have ignored the contribution from the active vertices that have not been explored yet. The cumulative effect of the negative contribution $-t/n$ gives the negative parabolic drift $-t^2/2$ in the limit. This suggests that a random graph model will be *ℓ-th order critical* when the expected number of new active vertices discovered by the exploration process at time t is approximately $1 + c(t/n)^{\ell}$ for some constant $c \in \mathbb{R}$ and $\ell \in \mathbb{N}$. Note that when $c = -1$ and $\ell_1 < \ell_2$, we have $1 - (t/n)^{\ell_1} > 1 - (t/n)^{\ell_2}$; when ℓ is larger, the critical window shrinks. The computations in Section 2.4 suggest that, as $\ell \to \infty$, the critical window converges to $\beta n^{-1/2}$.

Critical digraphs. Recall that the graph constructed from the $\Delta_{(i)}^{\alpha}/G/1$ queue with $\alpha \in [0,1)$ is a directed graph (briefly, *digraph*). We now abstract the construction of the random digraph in Chapter 5. We are led to consider a random digraph defined as follows: To each vertex $i \in [n]$ we assign two weights $\mathcal{W}_{i,\text{out}}$ and $\mathcal{W}_{i,\text{in}}$. Conditionally on the weights, we draw a directed edge from vertex i to vertex j with probability

$$\mathbb{P}(i \to j) = 1 - \exp\Big(- \frac{\mathcal{W}_{i,\text{out}}\mathcal{W}_{j,\text{in}}}{n}\Big). \tag{7.26}$$

Note that for the random graph generated from the $\Delta_{(i)}^{\alpha}/G/1$ queue it holds $\mathcal{W}_{i,\text{out}} = \mathcal{W}_{i,\text{in}}^{\alpha}$. We wish to investigate the critical behavior of this random graph by determining the size of the largest *strongly connected* components $\mathcal{C}_1, \mathcal{C}_2, \ldots$. A strongly connected component (briefly, a component) $\mathcal{C}_i$ is a (maximal) set of vertices such that each vertex in $\mathcal{C}_i$ is reachable from each other vertex in $\mathcal{C}_i$ by following the directed arrows. Very little is currently known about critical digraphs. Most of the literature focuses on the supercritical and subcritical phases of various digraph models. This procedure identifies the critical window, but gives no additional information on the structure of the strongly connected components. Let us briefly summarize the previous literature.

Luczak [69] considers a graph sampled uniformly at random from the set of simple graphs on n vertices, with M edges present (directed Erdős-Rényi random graph). He proves that the critical threshold for the emergence of a size $O(n)$ component is $M/n = 1$ (in the undirected Erdős-Rényi random graph this is $1/2$). See also Karp [57]. Luczak [70] identifies the precise critical window for this model. He proves that the connectivity structure changes when $np = 1 + \varepsilon_n$, where $\varepsilon_n = o(1)$ and $\varepsilon_n = \Theta(n^{-1/3})$. They also show that when $np = 1 + \varepsilon n^{-1/3}$ the strongly connected components are essentially a $O_{\mathbb{P}}(1)$ number of cycles of length $O_{\mathbb{P}}(n^{1/3})$ 'glued together'. Bloznelis, Götze and Jaworski [19] derive similar results for a large class of inhomogeneous random graphs, by building on the celebrated paper by Bollobás, Janson and Riordan [22]. Their proof is based on the relation between the giant strongly connected component and *two* different branching processes, describing respectively the descendants and the ancestors of a uniformly chosen vertex. See [26] for analogous results in the context of the directed configuration model.

The classical approach to the study of the component sizes of a critical random graph is via an exploration process [5, 15, 16, 33, 32]. For a random *directed* graph this approach does not seem to be appropriate. Intuitively, this is because the exploration process only provides local information on a neighborhood of a uniformly chosen vertex. On the other hand, due to the directed edges, the strongly connected components depend on the *global structure* of the graph. One of the most successful approaches to the study of the global structure of connected components relies on Aldous's theory of continuum random trees [4] and was pioneered by Addario-Berry, Broutin and Goldschdmit for the Erdős-Rényi random graph [2]. In this approach the connected components are seen as metric measured spaces. The distance is given by the usual graph distance, rescaled so as the distance between two neighbouring vertices converges to zero and the measure is the usual counting measure. In [2], the spanning tree of a single connected component conditioned on its size is encoded by a *depth-first* exploration. The 'surplus edges' (the edges that are in the component but not in the spanning tree) are described by (random) marks on the area below the graph of the depth-first exploration. See also [14], where this approach is applied to study the critical inhomogeneous random graph.

This technique seems to be well suited for the study of critical inhomogeneous *digraphs*. Each connected component is now to be equipped with three measures $\mu_n^{\text{out}}(\cdot)$, $\mu_n^{\text{in}}(\cdot)$, $\mu_n^{e}(\cdot)$. The first two describe respectively the empirical out-degree distribution and the empirical in-degree distribution of the vertices in the component. The third measure $\mu_n^{e}(\cdot)$ describes the number of edges in a component. Conditioned on the number of descendants of a vertex v, the depth-first exploration then describes the structure obtained by considering the outgoing edges from v. The ingoing edges are added as marks on the depth-first exploration similarly as for the 'surplus edges' in the Erdős-Rényi random graph in [2]. This procedure depends crucially on the in-degrees and out-degrees of the vertices encountered in the depth-first exploration. This information is encoded in the measures $\mu_n^{\text{out}}(\cdot)$ and $\mu_n^{\text{in}}(\cdot)$. Note that, differently from the undirected case, this construction does not directly yield a (strongly) connected component. Instead, the connected component is contained in the resulting graph; see Figure 7.1, where the strongly connected component consists of two directed cycles glued together at the vertex v. The measure $\mu_n^{e}(\cdot)$ is then necessary in order to identify the limiting structure of the strongly connected component. Finally, the number of descendants of a vertex v is obtained through a convergence result for the breadth-first exploration; this would entail generalizing our main theorem in Chapter 5 to an arbitrary out-degree distribution $\mathcal{W}_{i,\text{out}}$.

The ideas outlined above originated from personal communications with Shankar Bhamidi and Souvik Dhara.

Finite-population queues and random trees. In Chapter 5 and in the previous paragraphs we have focused on the connection between the $\Delta_{(i)}/G/1$ queue and the exploration process of certain random graphs. However, the exploration process also generates a *spanning tree* of the

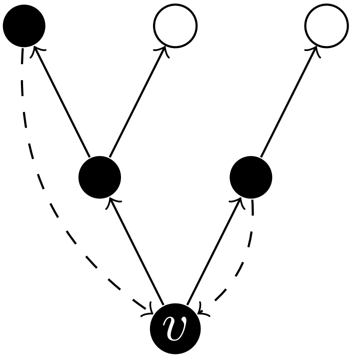

Figure 7.1: The depth-first exploration of the descendants of a vertex v and the associated strongly connected component (black vertices). The dashed arrows are ingoing edges.

random graph, which can be studied independently of the original graph. In fact, there is a natural connection between queues and random trees, given by the following general scheme. Consider a queue with one server which serves according to a First-Come-First-Served discipline. The first customer begins service at time $t = 0$. Let us denote the number of customers served in the first busy period by γ_1. We identify each customer $i \in \{1, \ldots, n\}$ with a (labelled) vertex, and label the customer in the queue at time $t = 0$ as the root. We define ν_i as the number of customers that join the queue during the i-th service, conditioned on $\gamma_1 = n$. Next we draw an edge from the root to all the ν_1 customers that joined during the first service. Then, we draw an edge between the second customer in the queue and all the ν_2 customers that join during the second service. By iterating this procedure we construct a *labelled, rooted* random tree with n vertices. We remark that in this context the queueing model is determined by the random variables $(\nu_i)_{i=1}^{\infty}$, which are in principle completely general. The difference with the similar procedure presented in Chapter 5 is twofold. First, here the emphasis lies on the family of random variables $(\nu_i)_{i=1}^{\infty}$ rather than on the precise arrival and service processes. Second, in this setting the total number of vertices in the tree is fixed. We use distinct notation from Chapter 5 in order to emphasize the difference between the two settings. We see that there is a correspondence between the construction of the random tree and the queue-length process embedded at service completions. The latter is defined as $\zeta(0) = 1$ and $\zeta(k) = (\zeta(k-1) + \nu_k - 1)^+$. The problem we present here can be broadly formulated as follows: can queueing techniques help to understand the geometry of the random tree associated with the random variables $(\nu_i)_{i=1}^{\infty}$? In turn, can a deeper understanding of the tree structure reveal new interesting features of the associated queueing model?

In fact, for the specific cases of i.i.d. and exchangeable $(\nu_i)_{i=1}^{\infty}$ the answer is positive, as Takács has shown in [90, 92]. In [90] Takács obtains exact (not asymptotic) explicit results by using combinatorial arguments on simple examples. His main tool is a simple formula for $\mathbb{P}(\gamma_1 = n)$ in the case of exchangeable $(\nu_i)_{i=1}^{\infty}$. His techniques allow him to give an explicit expression for the embedded queue length distribution for the $M/M/1$ queue since in this case $(\nu_i)_{i=1}^{\infty}$ are i.i.d. random variables. Interestingly, Takács is also able to give explicit formulas for the Erdős-Rényi random graph by a careful choice of the $(\nu_i)_{i=1}^{\infty}$. In [92], Takács studies the geometry of the random tree as $n \to \infty$. He assumes that $(\nu_i)_{i=1}^{\infty}$ are i.i.d. so that the process $(\zeta_k)_{k=1}^{\infty}$ is Markov. He proves that, when $\mathbb{E}[\nu] = 1$, the rescaled process $(\zeta_k)_{k=1}^{\infty}$ converges to a Brownian excursion. This fundamental result reveals a deep connection between the queue and the random tree. In fact, it turns out that various functionals of the embedded process such as the maximum queue length and the maximum number of arrivals during one service are asymptotically equivalent to certain quantities related to the random tree, such as the width of the tree and the height of the tree. Therefore, by exploiting the asymptotic results for the queueing process, Takács is able to characterize the limiting distribution of the width and the height of the associated random tree, in terms of functionals of a Brownian excursion. We remark that the Markov structure is crucial for the arguments of Takács. The arguments above can be easily carried over to the $M/G/1$ queue, since $(\nu_i)_{i=1}^{\infty}$ are again i.i.d. random variables.

It would be interesting to understand if the method outlined above is robust enough to be extended to a non-Markovian setting. We present two concrete problems that could guide future research efforts. First, given a queue in which the inter-arrival times and service times are generally distributed with unit mean and finite second moment, we ask what can be said on the associated random tree in the asymptotic regime $n \to \infty$. See also [92], where a similar issue is raised. Note that in this setting $(\nu_i)_{i=1}^{\infty}$ are not i.i.d. We conjecture that, conditioned on hitting zero at time n, the embedded queue-length process again converges to a Brownian excursion. In fact, the results of Takács [92] can be interpreted as a *conditioned Invariance Principle*, for which the

Markovian structure is not necessary. It is not yet clear, however, if it would be possible to relate the functionals of the queue and geometrical properties of the random tree as in the i.i.d. case. Second, it would be interesting to apply these techniques to gain further insights in the exploration process ζ_k of the Erdős-Rényi random graph. In this case, the ν_i have a strong dependency structure. We ask if new properties of the associated random spanning tree can be obtained by exploiting this connection. In fact, this issue is investigated in a different context in [2]. The authors characterize the distribution of the spanning tree of the Erdős-Rényi random graph conditioned on being connected. However, the spanning tree they consider is associated with the *depth-first* exploration process, and thus it has not a clear queueing interpretation.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Courier Corporation, 1964.

[2] L. Addario-Berry, N. Broutin, and C. Goldschmidt. The continuum limit of critical random graphs. *Probability Theory and Related Fields*, 152(3-4):367–406, 2012.

[3] E. Aidékon, R. van der Hofstad, S. Kliem, and J. S. H. van Leeuwaarden. Large deviations for power-law thinned Lévy processes. *Stochastic Processes and their Applications*, 126(5):1353–1384, 2016.

[4] D. Aldous. The Continuum Random Tree II: An overview. *Stochastic Analysis*, 167:23–70, 1991.

[5] D. Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Annals of Probability*, 25(2):812–854, 1997.

[6] S. Asmussen. *Applied Probability and Queues*. Springer Science & Business Media, 2003.

[7] S. Asmussen and H. Albrecher. *Ruin Probabilities*. World Scientific, 2010.

[8] M. S. Barlett. *Stochastic population models in ecology and epidemiology*. Springer, 1960.

[9] G. Bet. An alternative approach to heavy-traffc limits for finite-pool queues. 2017.

[10] G. Bet, R. van der Hofstad, and J. S. H. van Leeuwaarden. Heavy-traffic analysis through uniform acceleration of queues with diminishing populations. *arXiv:1412.5329*, 2015.

[11] G. Bet, R. van der Hofstad, and J. S. H. van Leeuwaarden. Big jobs arrive early: From critical queues to random graphs. *arXiv:1704.03406*, 2017.

[12] G. Bet, R. van der Hofstad, and J. S. H. van Leeuwaarden. Finite-pool queues with heavy-tailed services. *Journal of Applied Probability*, 54(3), 2017.

[13] S. Bhamidi, A. Budhiraja, and X. Wang. The augmented multiplicative coalescent, bounded size rules and critical dynamics of random graphs. *Probability Theory and Related Fields*, 160(3-4):733–796, 2014.

[14] S. Bhamidi, S. Sen, and X. Wang. Continuum limit of critical inhomogeneous random graphs. *Probability Theory and Related Fields*, 2016.

[15] S. Bhamidi, R. van der Hofstad, and J. S. H. van Leeuwaarden. Scaling limits for critical inhomogeneous random graphs with finite third moments. *Electronic Journal of Probability*, 15:1682–1702, 2010.

[16] S. Bhamidi, R. van der Hofstad, and J. S. H. van Leeuwaarden. Novel scaling limits for critical inhomogeneous random graphs. *Annals of Probability*, 40(6):2299–2361, 2012.

[17] P. Billingsley. *Convergence of Probability Measures*. Wiley, 1999.

[18] N. Bingham, C. Goldie, and J. Teugels. *Regular variation*. Cambridge University Press, 1987.

[19] M. Bloznelis, F. Götze, and J. Jaworski. Birth of a strongly connected giant in an inhomogeneous random digraph. *Journal of Applied Probability*, 49:601–611, 2012.

[20] B. Bollobás. On the evolution of random graphs. *Transactions of the American Mathematical Society*, 286(1), 1984.

[21] B. Bollobás. Random Graphs. In *Modern Graph Theory*. Springer, 1998.

[22] B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

[23] O. J. Boxma and J. W. Cohen. The M/G/1 queue with heavy-tailed service time distribution. *IEEE journal on selected areas in communications*, 16(5):749—-763, 1998.

[24] H. Chen and D. D. Yao. *Fundamentals of Queueing Networks. Performance, Asymptotics, and Optimization*. Springer Science & Business Media, 2001.

[25] J. W. Cohen. *The Single Server Queue*. Elsevier, 1982.

[26] C. Cooper and A. Frieze. The size of the largest strongly connected component of a random digraph with a given degree sequence. *Combinatorics, Probability and Computing*, 13(3):319–337, 2004.

[27] M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

[28] M. E. Crovella, M. S. Taqqu, and A. Bestavros. Heavy-tailed probability distributions in the World Wide Web. *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, 1998.

[29] J. N. Darroch and E. Seneta. On quasi-stationary distributions in absorbing discrete-time finite Markov chains. *Journal of Applied Probability*, 2(1):88–100, 1965.

[30] J. N. Darroch and E. Seneta. On quasi-stationary distributions in absorbing continuous-time finite Markov chains. *Journal of Applied Probability*, 4(1):3903–192–196, 1967.

[31] H. A. David and H. N. Nagaraja. *Order Statistics*. Wiley, New York, 2003.

[32] S. Dhara, R. van der Hofstad, J. S. H. van Leeuwaarden, and S. Sen. Heavy-tailed configuration models at criticality. *arXiv:1612.00650*, 2016.

[33] S. Dhara, R. van der Hofstad, J. S. H. van Leeuwaarden, and S. Sen. Critical window for the configuration model: finite third moment degrees. *Electronic Journal of Probability*, 22, 2017.

[34] T. Duquesne and J.-F. Le Gall. *Random Trees, Levy Processes and Spatial Branching Processes*. Societe mathematique de France, 2002.

[35] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2010.

[36] P. Erdős and A. Rényi. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(17-61), 1960.

[37] P. Eschenfeldt, B. Gross, and N. Pippenger. Stochastic service systems, random interval graphs and search algorithms. *Random Structures & Algorithms*, 45(3):421–442, 2014.

[38] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, New York, 1986.

[39] P. Groeneboom. Brownian motion with a parabolic drift and Airy functions. *Probability Theory and Related Fields*, 81:79–109, 1989.

[40] P. Groeneboom. The maximum of Brownian motion minus a parabola. *Electronic Journal of Probability*, 15:1930–1937, 2010.

[41] R. Grubel and M. Reich. Rarity and exponentiality: An extension of Keilson′ s theorem, with applications. *Journal of Applied Probability*, 42(2):393–406, 2005.

[42] G. Guadagni, S. Ndreca, and B. Scoppola. Queueing systems with pre-scheduled random arrivals. *Mathematical Methods of Operations Research*, 73(1):1–18, 2011.

[43] R. Hassin and S. Mendel. Scheduling arrivals to queues: A single-server model with no-shows. *Management Science*, 54(3):565–572, 2008.

[44] R. van der Hofstad. *Random Graphs and Complex Networks*. Cambridge University Press, 2016.

[45] R. van der Hofstad, A. Janssen, and J. S. H. van Leeuwaarden. Critical epidemics, random graphs, and Brownian motion with a parabolic drift. *Advances in Applied Probability*, 42(4):1187–1206, 2010.

[46] R. van der Hofstad, S. Kliem, and J. S. H. van Leeuwaarden. Cluster tails for critical power-law inhomogeneous random graphs. *arXiv:1404.1727*, 2014.

[47] H. Honnappa. Rare events of transitory queues. *arXiv:1705.08410*, 2017.

[48] H. Honnappa, R. Jain, and A. R. Ward. A queueing model with independent arrivals, and its fluid and diffusion limits. *Queueing Systems*, 80(1), 2015.

[49] H. Honnappa and A. R. Ward. On transitory queueing. *arXiv:1412.2321*, 2014.

[50] D. L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic. I. *Advances in Applied Probability*, 2(2):355–369, 1970.

[51] D. L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic. II: sequences, networks, and batches. *Advances in Applied Probability*, 2(2):355–369, 1970.

[52] J. Jacod and A. Shiryaev. *Limit Theorems for Stochastic Processes*. Springer, 2003.

[53] S. Janson, G. Louchard, and A. Martin-Lof. The maximum of Brownian motion with parabolic drift. *Electronic Journal of Probability*, 15:1893–1929, 2010.

[54] S. Janson, T. Luczak, and A. Rucinski. *Random Graphs*. John Wiley & Sons, 2000.

[55] A. Joseph. The component sizes of a critical random graph with a given degree sequence. *Annals of Applied Probability*, 24(6):2560–2594, 2014.

[56] S. Juneja and N. Shimkin. The concert queueing game: Strategic arrivals with waiting and tardiness costs. *Queueing Systems*, 74(4):369–402, 2013.

[57] R. M. Karp. The transitive closure of a random digraph. *Random Structures & Algorithms*, 1(1):73–93, 1990.

[58] J. B. Keller. Time-dependent queues. *SIAM Review*, 24(4):401–412, 1982.

[59] D. G. Kendall. Some problems in the theory of queues. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 151–185, 1951.

[60] P. Kihong and W. Willinger. *Self-Similar Network Traffic and Performance Evaluation*. Wiley, New York, 2000.

[61] J. F. C. Kingman. Martingales in the OK Corral. *Bulletin of the London Mathematical Society*, 31:601–606, 1999.

[62] J. F. C. Kingman and S. E. Volkov. Solution to the OK Corral Model via Decoupling of Friedman's Urn. *Journal of Theoretical Probability*, 16(1):267–276, 2003.

[63] A. Klenke. *Probability Theory: A Comprehensive Course*. Springer, London, 2008.

[64] J.-F. Le Gall. Random trees and applications. *Probability Surveys*, 2:245–311, 2005.

[65] J. S. H. van Leeuwaarden, M. S. Squillante, and E. M. M. Winands. Quasi-birth-and-death processes, lattice path counting, and hypergeometric functions. *Journal of Applied Probability*, 46(2):507–520, 2009.

[66] V. Limic. A LIFO queue in heavy traffic. *Annals of Applied Probability*, 11(2):301–331, 2001.

[67] G. Louchard. Large finite population queueing systems. The single-server model. *Stochastic Processes and their Applications*, 53:117–145, 1994.

[68] T. Luczak. Component behavior near the critical point of the random graph process. *Random Structures & Algorithms*, 1(3):287–310, 1990.

[69] T. Luczak. The phase transition in the evolution of random digraphs. *Journal of Graph Theory*, 14(2):217–223, 1990.

[70] T. Luczak and T. G. Seierstad. The critical behavior of random digraphs. *Random Structures & Algorithms*, 35(3):271–293, 2009.

[71] A. Mandelbaum and W. A. Massey. Strong approximations for time-dependent queues. *Mathematics of Operations Research*, 20(1):33–64, 1995.

[72] A. Martin-Löf. The final size of a nearly critical epidemic, and the first passage time of a Wiener process to a parabolic barrier. *Journal of Applied Probability*, 35(3):671–682, 1998.

[73] W. A. Massey. *Non-Stationary Queues*. PhD thesis, Stanford University, 1982.

[74] W. A. Massey. Asymptotic analysis of the time dependent M/M/1 queue. *Mathematics of Operations Research*, 10(2):305–327, 1985.

[75] G. F. Newell. Queues with time-dependent arrival rates I: The transition through saturation. *Journal of Applied Probability*, 5(2):436–451, 1968.

[76] G. F. Newell. Queues with time-dependent arrival rates II: The maximum queue and the return to equilibrium. *Journal of Applied Probability*, 5(3):579–590, 1968.

[77] G. F. Newell. Queues with time-dependent arrival rates III: A mild rush hour. *Journal of Applied Probability*, 5(3):591–606, 1968.

[78] G. F. Newell. *Applications of Queueing Theory*. Chapman & Hall, 1982.

[79] I. Norros and H. Reittu. On a conditionally Poissonian graph process. *Advances in Applied Probability*, 38(1):59–75, 2006.

[80] J. Pender. The truncated normal distribution: Applications to queues with impatient customers. *Operations Research Letters*, 43(1):40–45, 2015.

[81] N. Pippenger. Random interval graphs. *Random Structures & Algorithms*, 1998.

[82] B. Pittel. On the largest component of the random graph at a nearcritical stage. *Journal of Combinatorial Theory, Series B*, 82(2):237–269, 2001.

[83] M. I. Roberts. The probability of unusually large components in the near-critical Erdős-Rényi graph. *arXiv:1610.05485*, 2016.

[84] M. I. Roberts and B. Sengul. Exceptional times of the critical dynamical Erdős-Rényi graph. *arXiv:1610.06000*, 2016.

[85] G. Samorodnitsky and M. Taqqu. *Stable Non-Gaussian Processes*. Chapman & Hall, 1994.

[86] K.-I. Sato. *Levy Processes and Infinitely Divisible Distributions*. Cambridge University Press, 1999.

[87] T. Simon. Hitting densities for spectrally positive stable processes. *Stochastics: An International Journal of Probability and Stochastic Processes*, 83(2):203–214, 2011.

[88] A. V. Skorokhod. Limit theorems for stochastic processes. *Theory of Probability and its Applications*, I(3):261–290, 1956.

[89] L. Takacs. *Combinatorial Methods in the Theory of Stochastic Processes*. John Wiley & Sons, New York, 1967.

[90] L. Takács. Queues, random graphs and branching processes. *Journal of Applied Mathematics and Simulation*, 1(3):223–243, 1988.

[91] L. Takács. Limit distributions for queues and random rooted trees. *Journal of Applied Mathematics and Stochastic Analysis*, 6(3):189–216, 1993.

[92] L. Takács. Queueing methods in the theory of random graphs. In *Advances in Queueing Theory, Methods, and Open Problems*. CRC Press, 1995.

[93] L. A. V. Vianen, A. F. Gabor, and J.-K. van Ommeren. Waiting times in classical priority queues via elementary lattice path counting. *Queueing Systems*, 84(3-4):295–307, 2016.

[94] M. Virginia, A. Iovanella, C. Lancia, G. Lulli, and B. Scoppola. A model of inbound air traffic: The application to Heathrow airport. *Journal of Air Transport Management*, 34:116–122, 2014.

[95] W. Whitt. Some useful functions for functional limit theorems. *Mathematics of Operations Research*, 5(1):67–85, 1980.

[96] W. Whitt. *Stochastic-Process Limits. An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York, 2002.

[97] W. Whitt. Proofs of the martingale FCLT. *Probability Surveys*, 4:268–302, 2007.

[98] W. Whitt. Heavy-traffic limits for a single-server queue leading up to a critical point. *Operations Research Letters*, 44(6):796–800, 2016.

[99] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similar through high-variability: statistical analysis of Eternet LAN traffic at the source level. *IEEE/ACM Trans. on Networking.*, 5(1):71–86, 1997.

[100] Y. Yang and C. Knessl. Asymptotic analysis of the M/G/1 queue with a time-dependent arrival rate. *Queueing Systems*, 26:23–68, 1997.

[101] A. P. Zwart. *Queueing systems with heavy tails*. PhD thesis, Eindhoven University of Technology, 2001.

[102] A. P. Zwart. Tail asymptotics for the busy period in the GI/G/1 queue. *Mathematics of Operations Research*, 26(3):485–493, 2001.

Summary

This thesis studies the $\Delta_{(i)}/G/1$ queue, a model for a queueing system that serves only a finite number of customers. In the $\Delta_{(i)}/G/1$ queue, as time passes more customers have joined the system and thus fewer can potentially join, leading to a highly inhomogeneous arrival process. This modelling assumption of a diminishing population of customers gives rise to a class of reflected stochastic processes that lack a stationary distribution, and instead display relevant behavior only during a finite time window. The $\Delta_{(i)}/G/1$ queue is a model for numerous real-world situations, such as hospital out-patient wards and queues outside of concert halls, but is also useful in the study of the time-dependent behavior of classical ergodic models. Moreover, the $\Delta_{(i)}/G/1$ queue constitutes a good approximation for more complicated time-inhomogeneous queueing models. Chapter 1 introduces the $\Delta_{(i)}/G/1$ queue, presents the contents of this thesis and discusses the relevant results in the literature.

Chapters 2 and 3 deal with the standard $\Delta_{(i)}/G/1$ queue. In this model, n customers independently sample their arrival time from a common distribution and upon arrival join a common queue. The resulting queue-length process, describing the number of customers waiting to be served, is not Markovian since the evolution of the system crucially depends on the history of the process, that is on how many customers have been served. As a consequence, the exact study of the $\Delta_{(i)}/G/1$ queue is difficult, and we develop asymptotic approximations for the queue-length process. The proposed approximation rests upon the crucial assumption that the queueing system is critical, that is, we require the initial traffic intensity to be roughly one. Under the additional assumption that the service times are light-tailed, we show that, as the number of customers in the pool grows, the rescaled queue-length process converges to a Brownian motion with negative quadratic drift. The limiting negative drift encodes the *depletion-of-points effect* caused by the diminishing pool of customer.

Chapters 4 and 5 generalize the results of the previous chapters. Chapter 4 concerns the setting of heavy-tailed service times. More precisely, assuming that the service times follow a power-law distribution, we show that the rescaled queue-length process of the critical $\Delta_{(i)}/G/1$ queue converges to a stable motion with negative quadratic drift. The scaling exponents depend crucially on the precise distribution of the service times, in particular on the power-law exponent. The depletion-of-points effect contributes to the limiting process again in the form of a negative quadratic drift.

When the arrival times are exponential, the $\Delta_{(i)}/G/1$ queue can alternatively be seen as describing the exploration process of an appropriate random graph. In this analogy, customers are seen as vertices, and edges are traced between two vertices whenever one of the corresponding customers joins the queue during the service of the other. In Chapter 5 this connection is studied by introducing a new queueing model that we name the $\Delta_{(i)}^{\alpha}/G/1$ queue. As the parameter α varies in $(0,1)$, this model interpolates between the standard $\Delta_{(i)}/G/1$ queue and the exploration process of the *Norros-Reittu random graph*. The scaling limit of the $\Delta_{(i)}^{\alpha}/G/1$ queue then provides insight in the structure of the corresponding critical random graph.

When the number of customers n is held fixed, no scaling of the queue is needed. If, additionally, the arrival and service times are assumed to be exponentially distributed, then the vector representing the queue length and the number of customers in the pool at time t is a two-dimensional Markov process. Chapter 6 presents exact results for this process. In particular, we derive an explicit expression for the distribution of the number of customers served in the first busy period by exploiting the recursive structure of the embedded Markov chain.

Lastly, Chapter 7 collects various open problems that originated from the research conducted for this thesis. These problems represent fundamental questions in queueing theory as well as random graph theory. We outline the most promising approach for the solution of each problem and leave the details to future research.

About the author

Gianmarco Bet was born in Conegliano, Italy, on October 1, 1989. He completed his secondary education in 2008 at "Liceo Scientifico A. Cornaro" in Padova, Italy, and then started his studies in Mathematics at the University of Padova. After obtaining his Bachelor's degree in July 2011, he continued his studies by pursuing a Master's degree in Mathematics at the University of Padova. In July 2013, he obtained his Master's degree.

In September 2013, he started a PhD project at Eindhoven University of Technology in the Stochastic Operations Research group under the supervision of Johan van Leeuwaarden and Remco van der Hofstad. His PhD research focused on the critical scaling of time-inhomogeneous queueing systems.

Gianmarco will defend his PhD thesis at Eindhoven University of Technology on September 11, 2017.