Fractional Calculus and Applied Analysis (2023) 26:1465–1503
https://doi.org/10.1007/s13540-023-00161-4

ORIGINAL PAPER

# Queuing models with Mittag-Leffler inter-event times

**Jacob Butt**[1] · **Nicos Georgiou**[1] · **Enrico Scalas**[1,2]

Received: 23 January 2023 / Revised: 20 April 2023 / Accepted: 20 April 2023 /
Published online: 18 May 2023
© The Author(s) 2023

**Abstract**
We study three non-equivalent queueing models in continuous time that each generalise the classical M/M/1 queue in a different way. Inter-event times in all models are Mittag-Leffler distributed, which is a heavy tail distribution with no moments. For each of the models we answer the question of the queue being at zero infinitely often (the 'recurrence' regime) or not (the transient regime). Aside from this question, the different analytical properties of each models allow us to answer a number of questions such as existence and description of equilibrium distributions, mixing times, asymptotic behaviour of return probabilities and moments and functional limit theorems.

**Keywords** GI/ GI/1 queue · Mittag-Leffler queues · Queue length · Recurrence · Transience · Fractional derivatives · Time-changed queue · Semi-Markov process · Mixing times · Scaling limits for the queue length

**Mathematics Subject Classification** 60K25 · 60F17 (primary) · 60K15

## 1 Introduction

The theory and application of models for queuing systems became an integral part of many disciplines, playing an important role in areas such as health-care [15], telecommunications [18], and operations research [23] among others. The original theoretical

✉ Enrico Scalas
E.Scalas@sussex.ac.uk; enrico.scalas@uniroma1.it

Jacob Butt
J.Butt@sussex.ac.uk

Nicos Georgiou
N.Georgiou@sussex.ac.uk

[1] University of Sussex, Brighton, UK

[2] 'La Sapienza' University of Rome, Rome, Italy

basis for much of classical queuing theory lies in the extensively studied M/M/1 queue model [1, 7, 21]. This model uses exponentially distributed inter-arrival and service times for the customers, rendering it amenable to rigorous mathematical treatment and making it a good starting point for many situations. Being a simple model, however, means it lacks a number of properties such as preservation of memory, which are often desirable when studying more complex systems. There are several examples in which memory effects are necessary; for example in studying hysteresis effects in biological and epidemiological models.

One natural extension is to consider more general distributions for the arrival and service times, leading to GI/GI/1 queue models (see [28] section 7). The analysis of these models becomes more robust when inter-arrival and service times have moments, but it becomes more challenging in the absence of all moments. Nonetheless, heavy-tailed distributions appear naturally in queuing models, see for example [16, 42]. One particular question of interest is whether the queue empties infinitely often and the length has an equilibrium distribution when the inter-arrival and service times do not have all moments. This stability of the first-in, first out (or first come -first served) queue is characterised by the value of the load (or the traffic intensity), which is the ratio of the first moments of inter-arrival and service times so a different version of this characterisation is needed when both of these moments do not exist.

Different methods for handling queuing models with heavy-tailed distributions have begun to emerge of late [9], using techniques from the field of fractional calculus. These techniques are particularly useful when the model has Mittag-Leffler distributions for waiting times [20, 25, 31] as the connection of the Mittag-Leffler function to fractional calculus is well understood [26]. Its properties make the models that use the Mittag-Leffler distribution amenable to rigorous mathematical analysis and exact results. Moreover, it has become increasingly common to introduce fractional operators into standard models in order to introduce memory effects. These models are also developed with applications in mind; for example see [39] and [40] which give a good overview of some applications in engineering and finance respectively.

Recently, in [8], the Caputo fractional derivative appeared in the system of differential equations for the M/M/1 queue. In this fractional queue model, the inter-arrival and service times are now characterised using the Mittag-Leffler distribution with parameter $\alpha \in (0, 1]$, where setting $\alpha = 1$ recovers the classical model. It is shown that this model can be described analytically as a time-changed version of the standard M/M/1 queue through the use of an inverse stable subordinator. These subordinators are often used in the study of fractional models and introduce a random timescale and memory effects into the model [29, 30]. In particular, when $\alpha \in (0, 1)$, the model now loses the memoryless property associated with the classical M/M/1 queue and becomes a semi-Markov model. This queuing model is closely related to birth-death models, of which non-local versions have been well studied in the recent past using similar subordinator methods. One can see [5, 12, 34] for results on such fractional birth-death models. Also of particular note are works done on various modifications of these fractional M/M/1 and birth-death models. These include, among others, the addition of catastrophes to the model [2], as well as a related fractional Erlang queue model [3], and a fractional immigration-death process [4].

In this article we study three different ways to introduce a single server queue with Mittag-Leffler inter-event times. The first one is precisely the time-changed M/M/1 model in which the evolution of the system depends on events of a background Poisson process. Whether the Poisson event is an arrival or a departure is decided by a coin flip. Initial studies of this queue were performed in [3, 8, 12, 34]. Here we offer an expression and evolution equation for the moment generating function of the queue length, as well as asymptotic behaviour for return probabilities, moments and mixing times to equilibrium. The second model is a generalisation of a Gillespie-type version of the queue which has a renewal structure. After every event, competing independent waiting times re-start for arrivals and departures, and the fastest one decides the type of event. Finally, the third model is the fully generalised GI/GI/1 queue with Mittag-Leffler distributed inter-arrival and service times. A crucial question for all models is that of recurrence and transience of the queue length, where the notions are suitably interpreted for these heavy-tailed models.

The original motivation for this article came from a need to model certain financial markets. Over the past few decades numerous studies have been conducted aiming to examine the behaviour of times between trades in financial markets. In a number of cases (see for example [35, 37]) waiting times have been found to match nicely with heavy tailed distributions such as the Mittag-Leffler distribution (i.e. they exhibit no moments at all). As models for such markets (referring in particular to models such as the Double Auction models in [11, 36]) can be related to GI/GI/1 queue models, we expect the development of queueing systems with Mittag-Leffler distributed waiting times to be useful in the near future.

### 1.1 Structure of the article:

In Sect. 2 we describe three basic equivalent descriptions of the Markovian M/M/1. The equivalence of these three models hinges on the Markov property. Each of these will lead to a distributionally different fractional queuing model. In Sect. 3 we record the results for each of the models. The proofs of these results is the rest of the article. In Sect. 4 we revisit the fractional model in [8] where we prove the fractional evolution of the moment generating function as well as mixing times to equilibrium in the recurrent regime. The proofs that are closely mimicking existing proofs in the literature are postponed to Appendices A, B, C, however we do include them for completeness and easy access. Section 5 is dedicated to the renewal type queue. Because of the renewal structure this one can be partially studied with classical tools. Finally in Sect. 6 we study the fully generalised GI/GI/1 queue with Mittag-Leffler distributed arrival and service times. We establish and prove a functional limit theorem for this generalised model, and classify transient and recurrent regimes in this context.

## 2 Preliminaries

### 2.1 Three equivalent M/M/1 models

In this article we consider three generalisations of the classical M/M/1 queue. Each of the models we are considering are natural extensions of different equivalent ways to define the continuous time M/M/1 queue.

All three classical models can be viewed as a time-change of the discrete queue, which we present first, in full generality. Each of these equivalent models have been used extensively in various ways as classical examples. Model 1 is a standard Markov chain on an infinite state space that can be studied completely. Model 2 is used in simulations of Markov chains, such as the Gillespie algorithm. Model 3 is the natural description of a queuing system. However, when we change the waiting time distribution, these are no longer equivalent and they give rise to different queues that can be of interest.

Let $\{Q_n\}_{n\geq 0}$ denote a Markov chain on the non-negative integers, which represents the length of the queue at discrete time $t = n$. Initially, at $n = 0$ the queue has $Q_0 = i_0 \in \mathbb{Z}_+$ and after that it evolves according to the homogeneous transition probabilities

$$p_{i,j} = P\{Q_{n+1} = j | Q_n = i\},$$

given by

$$p_{k,k} = \beta, \tag{2.1}$$

$$p_{k,k+1} = \begin{cases} 1-\beta, & \text{if } k = 0 \\ (1-\beta)p, & \text{otherwise,} \end{cases} \tag{2.2}$$

$$p_{k,k-1} = \begin{cases} 0, & \text{if } k = 0 \\ (1-\beta)(1-p) & \text{otherwise.} \end{cases} \tag{2.3}$$

All other transition probabilities equal zero. Parameter $\beta$ can equal 0. This just means that there is either an arrival or a departure in any time interval in which the queue length is not empty. It can also equal 1 but in that case the queue does not evolve.

Ergodicity properties (or lack of) for the Markov chain are well-known (see, for instance, [33]), as the transition probabilities $\{p_{i,j}\}_{i,j\in\mathbb{Z}_+^2}$ represent a discrete *birth and death chain.* In particular we have that

$$Q_n \text{ is positive recurrent } \iff p < 1/2, \quad Q_n \text{ is null recurrent } \iff p = 1/2.$$

In these cases the queue empties infinitely often (as $Q_n = 0$ for infinitely many $n$) and when $Q_n$ is positive recurrent, the queue has an equilibrium and unique invariant distribution that is computable and given by

$$\lim_{n\to\infty} \mathbb{P}\{Q_n = j\} = \pi_j = \frac{1-2p}{1-p}\Big(\frac{p}{1-p}\Big)^j. \tag{2.4}$$

The discrete Markovian queue can be made into a continuous time Markovian queue in several equivalent ways, three of which we present here. In the sequel, when we create a time-fractional queue, we will have that each of these ways lead to a different (non-Markovian) queue.

**Model 1: Time-changed discrete chain.** A way to define the continuous time $M/M/1$ queue is by performing a global time change on the discrete model. To this end, let $N_1(t)$ denote a Poisson process with rate 1, and fix the parameters of the discrete model to $\beta = 0$ and $p > 0$. Then, the length $L_t^{(1)}$ is defined as

$$L_t^{(1)} = Q_{N_1(t)}. \tag{2.5}$$

With this time change we have the Poisson point process, with inter-event times to be i.i.d. Exponential(1). At the time of each of these events, an independent coin is flipped with success probability equal to $p$. If it the flip is a success, we interpret the Poisson event as an arrival, while if it is a failure, with probability $1 - p$ the Poisson event is interpreted as a departure from the queue. If no customer is at the queue, the departure event is ignored. This description ensures that the equality in (2.5) is almost sure, and the discrete chain becomes the embedded chain of this continuous time Markov chain.

There is nothing special about the rate of the Poisson process being 1, but it motivates the construction of Model 3 below.

**Model 2: The renewal approach.** Consider two independent i.i.d. sequences of exponential random variables with rate parameters $\lambda$ and $\mu$ respectively, $\{X_i^{(\lambda)}\}_{i\geq 1}$ and $\{X_i^{(\mu)}\}_{i\geq 1}$. Define

$$T_i = X_i^{(\lambda)} \wedge X_i^{(\mu)} \sim \text{Exp}(\lambda + \mu).$$

Then for any $t > 0$, define

$$N_{\lambda+\mu}(t) = \max\left\{k : \sum_{i=1}^{k} T_i \leq t < \sum_{i=1}^{k+1} T_i\right\}.$$

As above, $N(t)$ is a Poisson Process with rate $\lambda+\mu$, but constructed slightly differently from Model 1. Moreover, when the Poisson process ticks for the $n$-th time, we interpret the events as

$$\text{arrival of a customer at the } n\text{-th event} \iff X_i^{(\lambda)} \wedge X_i^{(\mu)} = X_i^{(\lambda)},$$

otherwise we have a completion of service. Then the queue length $L_t^{(2)}$ at time $t$ is given by

$$L_t^{(2)} = Q_{N_{\lambda+\mu}(t)}, \tag{2.6}$$

where the parameters for the discrete (embedded) chain $Q_n$ are $p = \lambda(\lambda+\mu)^{-1} = \mathbb{P}\{X_i^{(\lambda)} \wedge X_i^{(\mu)} = X_i^{(\lambda)}\}$ and $\beta = 0$.

**Model 3: Independent arrivals and departures.** For a distributional equality in (2.5) of Model 1, without looking at the evolution of $\{Q_n\}_{n\geq 1}$, go through the events of $N_1(t)$ and flip an independent coin for each one (and independent of $N_1(t)$), marking the Poisson events as success or failure. Let $p$ denote the probability of success and $1-p$ the probability of failure. This creates an arrival and a departure process $N_p(t)$ and $N_{1-p}(t)$ respectively. These are *thinned out* Poisson processes and are independent of each other.

Every time $N_p(t)$ ticks, there is an arrival in the system while every time $N_{1-p}(t)$ ticks, there is a completion of a service time and, if there is a customer in the system, they exit reducing the queue length by 1. The queue length in this case is

$$L_t^{(3)} = Q_{N_p(t)+N_{1-p}(t)} \overset{\mathcal{D}}{=} Q_{N_1(t)} = L_t^{(1)}. \tag{2.7}$$

This can be generalised to having two independent processes $N_\lambda(t)$ and $N_\mu(t)$ representing the arrival process and the departure process respectively. Then using $p = \lambda(\lambda+\mu)^{-1}$ and $\beta = 0$ for the discrete queue $\{Q_n\}_{n\geq 1}$ we have

$$L_t^{(3)} = Q_{N_\lambda(t)+N_\mu(t)} \overset{\mathcal{D}}{=} Q_{N_{\lambda+\mu}(t)}. \tag{2.8}$$

Moreover, using the interpretation of the model that we have two independent arrival and departure processes there is an explicit and useful formula calculating the queue length at any given time

$$L_t = \big(N_\lambda(t) - N_\mu(t)\big) - \inf_{0\leq s\leq t}\big\{N_\lambda(s) - N_\mu(s)\big\}. \tag{2.9}$$

The proof of this a.s. equality goes by induction on the number of Poisson points in the interval $[0, t]$. As such, the proof works with any two counting processes to represent arrivals and departures, not just when these are Poisson processes, as long as the number of events in any time interval is almost surely finite. A proof of equation (2.9) can be found in Appendix A. For a schematic representation see Fig. 1.

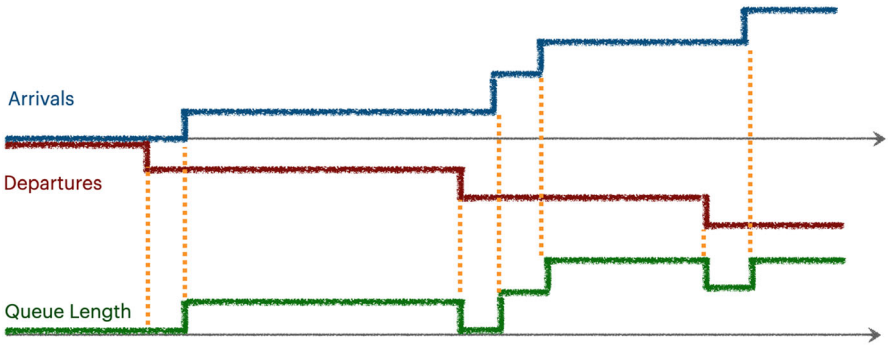


**Fig. 1** A schematic of the queue length formula as a reflected process of the difference between arrivals and departures

## 3 Results

### 3.1 Heavy tailed models

As stated earlier, we will be extending each of the three equivalent Markovian queuing systems into queuing systems that are generated from Mittag-Leffler inter-event times. The three generalisations however will not be equivalent, and this demonstrates the need to find the correct fractional model in applications that demonstrate heavy tailed inter-event times. We first record some properties of the Mittag-Leffler function and distribution that are used throughout this article, followed by our results in each of the three models. An extensive review of the Mittag-Leffler function can be found in [31] (see also the references therein).

The density of the Mittag-Leffler distribution of power index $\alpha \in (0, 1]$ is

$$f_\alpha(x) = x^{\alpha-1} E_{\alpha,\alpha}(-x^\alpha) = x^{\alpha-1} \sum_{k=0}^{\infty} \frac{(-1)^k x^{\alpha k}}{\Gamma(\alpha + \alpha k)},$$

with c.d.f. given by

$$F_\alpha(x) = \mathbb{P}\{X_\alpha \le x\} = 1 - E_{\alpha,1}(-x^\alpha) = 1 - \sum_{\ell=0}^{\infty} \frac{(-1)^\ell x^{\alpha\ell}}{\Gamma(1 + \alpha\ell)}.$$

Note that the power series above converge absolutely, since the generalised Mittag-Leffler function with parameters $\alpha$, $\beta$

$$E_{\alpha,\beta}(z) = \sum_{\ell=0}^{\infty} \frac{z^\ell}{\Gamma(\beta + \alpha\ell)}$$

also does.

When $\alpha = 1$ the density and c.d.f. above coincide with those of the exponential distribution of parameter 1. We can scale the Mittag-Leffler distribution the same way as the exponential distribution can be scaled, namely for $\lambda > 0$

$$X_{\alpha,\lambda} = \frac{1}{\lambda} X_\alpha. \tag{3.1}$$

The density of $X_\alpha$ is given by $f_\alpha$ and therefore the density and c.d.f. of $X_{\alpha,\lambda}$ are given by

$$f_{\alpha,\lambda}(x) = \lambda^\alpha x^{\alpha-1} E_{\alpha,\alpha}(-\lambda^\alpha x^\alpha), \quad F_\alpha(x) = 1 - E_{\alpha,1}(-\lambda^\alpha x^\alpha). \tag{3.2}$$

Note that $X_\alpha$ is the same as $X_{\alpha,1}$.

***Remark 3.1*** The usual way we see the scaling $\lambda$ in the literature is that the c.d.f. is given as

$$F_\alpha(x) = 1 - E_{\alpha,1}(-\lambda x^\alpha).$$

This would correspond to

$$X_{\alpha,\lambda} = \frac{1}{\lambda^{1/\alpha}} X_\alpha,$$

which is notationally slightly more cumbersome for our purposes. So in this article we are using (3.1) which will make the index $\alpha$ appear as in (3.2). □

We also utilise the following asymptotic property of $E_{\alpha,1}$,

$$1 - F_\alpha(x) = E_{\alpha,1}(-x^\alpha) \sim \frac{\sin(\alpha\pi)\Gamma(\alpha)}{\pi} x^{-\alpha}, \quad x \to \infty, \tag{3.3}$$

without any particular mention.

Finally, it is worth stating some basic background on subordinators and the fractional Poisson process for future reference. For $\alpha \in (0,1)$ define the $\alpha-$stable subordinator $(L_\alpha(t))_{t\geq 0}$ to be a positive valued Lévy process with Laplace transform

$$\mathbb{E}\left[e^{-\omega L_\alpha(t)}\right] = e^{-t\omega^\alpha}. \tag{3.4}$$

From this, we can naturally define the inverse $\alpha-$stable subordinator $(Y_\alpha(t))_{t\geq 0}$ to be given by

$$Y_\alpha(t) = \inf\{u \geq 0 : L_\alpha(u) > t\}. \tag{3.5}$$

This inverse $\alpha-$stable subordinator can be used to define the fractional Poisson process as a time change of a Poisson process $N_{\lambda^\alpha}(t)$, as was proved in [29], by setting

$$N_\lambda^{(\alpha)}(t) := N_{\lambda^\alpha}(Y_\alpha(t)). \tag{3.6}$$

***Remark 3.2*** The fractional Poisson process can also be viewed as a renewal counting process for which the inter-event times are i.i.d. Mittag-Leffler distributed (see for example [27, 29]). We will use this fact without any specific mention in the sequel.

One important property to note is that both $L_\alpha$ and $Y_\alpha$ are self-similar; i.e,

$$L_\alpha(t) \stackrel{d}{=} t^{1/\alpha} L_\alpha(1), \qquad Y_\alpha \stackrel{d}{=} t^\alpha Y_\alpha(1). \tag{3.7}$$

This self-similarity property naturally carries over to the fractional Poisson process, where it manifests as the following chain of equalities in distribution

$$N_\lambda^{(\alpha)}(t) \stackrel{d}{=} N_{\lambda^\alpha}(Y_\alpha(t)) \stackrel{d}{=} N_1(\lambda^\alpha Y_\alpha(t)) \stackrel{d}{=} N_1(Y_\alpha(\lambda t)) \stackrel{d}{=} N_1^{(\alpha)}(\lambda t). \tag{3.8}$$

While the distributional equalities above are stated for a fixed $t$, it is also convenient to emphasise that they hold at the process level as well.

### 3.2 The (single) time fractional queue

For this model, we use the semi-Markov approach taken in [17] for a global time change on the discrete queue process; the interested reader can see [19] for a more general overview of such methods. It is equivalent to the counting process being a Fractional Poisson Process $N_\lambda^{(\alpha)}(t)$, where inter-event times $T_i \sim X_{\alpha,\lambda}$ are Mittag-Leffler distributed with power index $\alpha$ and scale $\lambda$, with density given by (3.2).

Let the queue length for this model be defined by

$$L_{\alpha,\lambda}^{(1)}(t) = Q_{N_\lambda^{(\alpha)}(t)}. \tag{3.9}$$

For the discrete model in this case we assume that $\beta = 0$ in (C1) (just for simplicity) and that $L_{\alpha,\lambda}^{(1)}(0) = Q_{N_\lambda^{(\alpha)}(0)} = Q_0 = 0$. As such, we have the forward Kolmogorov equations for the queue length

$$\begin{aligned} p_{0,i}(t) &= \mathbb{P}\{L_{\alpha,\lambda}^{(1)}(t) = i\} \\ &= \bar{F}_{\alpha,\lambda}(t)\delta_0(i) + \sum_{k\in\{(i-1)\vee 0, i+1\}} q_{k,i} \int_0^t p_{0,k}(u) f_{\alpha,\lambda}(t-u)du. \end{aligned} \tag{3.10}$$

Above we shorthanded $\bar{F}_{\alpha,\lambda}(t) = 1 - F_{\alpha,\lambda}(t)$ for the survival function (c.c.d.f.) of $X_{\alpha,\lambda}$. Equation (3.10) holds because we assume that the first waiting time begins at time $t = 0$ where the queue is empty. The range of index $k$ comes from the the fact that $q_{k,i} = 0$ if $k$ is not in that range. Using Laplace transforms, one can show (see Appendix C for a direct derivation) that the temporal evolution of $p_i(t)$ satisfies the time-fractional differential equation

$$\frac{d^\alpha p_i(t)}{dt^\alpha} = -\lambda^\alpha p_i(t) + \lambda^\alpha p p_{i-1}(t) + \lambda^\alpha (1-p) p_{i+1}(t), \quad i \geq 1. \tag{3.11}$$

Similarly, we can find the boundary conditions

$$\frac{d^\alpha p_0(t)}{dt^\alpha} = -\lambda^\alpha p p_0(t) + \lambda^\alpha (1-p) p_1(t). \tag{3.12}$$

Operator $\dfrac{d^\alpha p_i(t)}{dt^\alpha}$ is the Caputo derivative [13] of $p_i(t)$. For any function $f$ it is defined by

$$\frac{d^\alpha}{dt^\alpha} f(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-u)^{-\alpha} \frac{d}{du} f(u)\, du.$$

As mentioned earlier, this queue is the same as the fractional queue studied by [8]. We would like to extend some of their results in this article to recover a full expression for the temporal evolution of the moments of the queue length, as well as provide a proof of convergence to invariant distribution and some asymptotic results.

We offer the formula for the Laplace transform of the moment generating function

$$M_{\alpha,\lambda}(z,t) = \mathbb{E}(e^{zL^{(1)}_{\alpha,\lambda}(t)})$$

where $L^{(1)}_{\alpha,\lambda}(t) = Q_{N^{(\alpha)}_{\lambda}(t)}$ when we assume that the queue is empty at time $t = 0$. At the same time, the first Mittag-Leffler inter-event time begins elapsing. Define

$$r_i(s) = \begin{cases} \dfrac{1+s^\alpha + \sqrt{(1+s^\alpha)^2 - 4p(1-p)}}{2p}, & i = 1 \\ \dfrac{1+s^\alpha - \sqrt{(1+s^\alpha)^2 - 4p(1-p)}}{2p}, & i = 2. \end{cases} \tag{3.13}$$

**Theorem 3.1** *Let $p_{0,0}(t) = \mathbb{P}_0\{L^{(1)}_{\alpha,\lambda}(t) = 0\}$. Then the moment generating function satisfies the fractional differential equation*

$$\frac{d^\alpha}{dt^\alpha} M_{\alpha,1}(z,t) = (1-p)(1-e^{-z})p_{0,0}(t) + \left(pe^z - 1 + (1-p)e^{-z}\right) M_{\alpha,1}(z,t)$$

*and has the Laplace transform*

$$\widetilde{M}_{\alpha,1}(z,s) = \mathcal{L}\{M_{\alpha,1}(z,t)\}(s) = \frac{-s^{\alpha-1}}{p(e^z - r_1(s))(1-r_2(s))}. \tag{3.14}$$

Particular uses of this are asymptotics for moments, the variance of the queue length and the probability of seeing an empty queue as $t \to \infty$ via Tauberian theorems. These can be found in Examples 4.1, 4.2 is Sect. 4.2.

***Remark 3.3*** In general, the queuing representation (2.9) also applies in this setting. Time-changing the two independent Poisson processes with a common inverse stable subordinator will lead to the same queuing system in distribution. The difference of two Fractional Poisson Processes with this common time change can be found in the literature under the name of a fractional Skellam process of type 2, see [22]. In this respect, the queue is a fractional Skellam process, partially reflected at 0. □

Since this queue is obtained by a direct time change of the Markovian one, it is straight forward to argue that when $p < 1/2$, as in Eq. (2.4),

$$\lim_{t\to\infty} p_i(t) = \frac{1-2p}{1-p}\left(\frac{p}{1-p}\right)^i.$$

Since the state space for the queue length is infinite and time is delayed due to the time change, it is natural to search for quantitative estimates for the time it takes for the $p_i(t)$ to be near its equilibrium value. For this we will be using the total variation distance. The total variation norm between two measures, $\mu$ and $\nu$, on a common probability space $\Omega$ with $\sigma-$algebra $\mathcal{F}$ is defined as

$$\|\mu - \nu\|_{TV} = \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)|. \tag{3.15}$$

The support space for the case of the queue length is $\mathbb{N}_0$ which is a countable space, so overall the total variation distance can be written more simply as

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in \mathbb{N}_0} |\mu(x) - \nu(x)|. \tag{3.16}$$

Irreducible discrete Markov chains that are positive recurrent converge to a unique equilibrium and this is the case with $Q_n$ in (2.4). The *$\varepsilon$-mixing time* $T^{\varepsilon}_{\text{mix}}$ for any chain $X_t$ with initial distribution $\mu_0$ and with a unique equilibrium distribution $\pi$ is defined by

$$T^{\varepsilon}_{\text{mix}} = \inf\{t > 0 : \|\mathbb{P}_{\mu_0}\{X_t \in \cdot\} - \pi\|_{TV} \leq \varepsilon\} \tag{3.17}$$

Then we have the following theorem.

**Theorem 3.2** *Let $L^{(1)}_{\alpha,\lambda}(t)$ be given by (3.9) and let $\pi$ be the equilibrium distribution given by (2.4). Assume that the initial queue length distribution $\mu_0$ is finitely supported. Then there exist positive constants $C_{\alpha,\lambda,\mu_0}, U_{\alpha,\lambda,\mu_0}$ such that for all $0 < \varepsilon < C_{\alpha,\lambda,\mu_0}$*

$$T^{\varepsilon}_{\text{mix}} = \inf\{t > 0 : \|\mathbb{P}_{\mu_0}\{L^{(1)}_{\alpha,\lambda}(t) \in \cdot\} - \pi\|_{TV} \leq \varepsilon\} \leq U_{\alpha,\lambda,\mu_0} \left( \frac{C_{\alpha,\lambda,\mu_0} - \varepsilon}{\varepsilon} \right)^{1/\alpha}.$$

### 3.3 The renewal queue

For this section we define

$$p^{\lambda,\mu}_{\alpha,\beta} = \mathbb{P}\{X_{\alpha,\lambda} < X_{\beta,\mu}\}. \tag{3.18}$$

The random variables $X_{\alpha,\lambda}$, $X_{\beta,\mu}$ are independent, Mittag-Leffler distributed. Consider two independent i.i.d. sequences $\{X^{(i)}_{\alpha,\lambda}\}_{i\in\mathbb{N}}$, $\{X^{(i)}_{\beta,\mu}\}_{i\in\mathbb{N}}$ and define the i.i.d. sequence

$$Y_i = X^{(i)}_{\alpha,\lambda} \wedge X^{(i)}_{\beta,\mu}.$$

The counting process for this model is now a renewal process $R_t$ with inter-event times $Y_i$. A renewal event will be classified as 'arrival' with probability $p^{\lambda,\mu}_{\alpha,\beta}$ and 'departure' otherwise. To describe $R_t$ in words, at time $t = 0$ we have two competing Mittag-Leffler distributions and see which one rings first. If $X_{\alpha,\lambda}$ rings first (i.e. $X_{\alpha,\lambda} < X_{\beta,\mu}$) then we interpret the event at time $t = Y_1$ as an arrival, otherwise it is a departure attempt. At time $t = Y_1$ the waiting time restarts.

An interesting aspect for this model is the fact that the competing Mittag-Leffler variables that define $Y_i$ they can allow the inter-event times to have moments. To see this consider

$$\begin{aligned}\mathbb{P}\{Y_i > t\} &= \mathbb{P}\{X^{(i)}_{\alpha,\lambda} \wedge X^{(i)}_{\beta,\mu} > t\} \\ &= \mathbb{P}\{X^{(i)}_{\alpha,\lambda} > t\}\mathbb{P}\{X^{(i)}_{\beta,\mu} > t\} \sim \frac{C_{\alpha,\beta,\lambda,\mu}}{t^{\alpha+\beta}}, \quad t \to \infty.\end{aligned}$$

Therefore, when $\alpha + \beta > 1$ the inter-event times have moments and the time-changed queue can be treated as with classical renewal queues with a first moment (but not second).

As in Model (1), the queue length is the time-changed queue length

$$L_t^{(2)} = Q_{R_t}$$

with an arrival event appearing with probability $p_{\alpha,\beta}^{\lambda,\mu}$.

**Proposition 3.1** *For any choice of parameters $\alpha, \beta \in (0, 1)$, there exists a critical $\rho_{\alpha,\beta}^* \in (0, \infty)$ such that the embedded queue length is transient if and only if $\lambda/\mu > \rho_{\alpha,\beta}^*$.*

*Moreover, if $\lambda/\mu < \rho_{\alpha,\beta}^*$ then the queue length reaches an equilibrium with distribution (2.4) and parameter $p_{\alpha,\beta}^{\lambda,\mu}$.*

*In the case where $\lambda/\mu = \rho_{\alpha,\beta}^*$, the embedded Markov chain is null recurrent.*

### 3.4 The Mittag-Leffler GI/GI/1 queue

In this model, we consider two independent arrival and departure Fractional Poisson processes. The arrival process $N_\lambda^{(\alpha_1)}(t)$ has inter-event times which are Mittag-Leffler distributed with tail parameter $\alpha_1$ and scaling $\lambda > 0$. The departure process $N_\mu^{(\alpha_2)}(t)$ has potentially different tail exponent $\alpha_2$ and different scaling $\mu$.

A renewal point of the arrival process signifies that a customer joined the queue while a renewal point on the departure process suggests the completion of a service and if the queue length is strictly positive at that time it is reduced by 1, otherwise it remains at zero.

We denote the queue length at time $t$ by $L_{\lambda,\mu}^{\alpha_1,\alpha_2}(t)$. If the scalings are $\lambda = \mu = 1$ we omit them from the notation and simply write $L^{\alpha_1,\alpha_2}(t)$ for the length.

In Fig. 2 we see how the different tail indices can affect the queue length. The three cases show the qualitative differences for the queue length, and all these behaviours are explored in the theorems below in various ways.

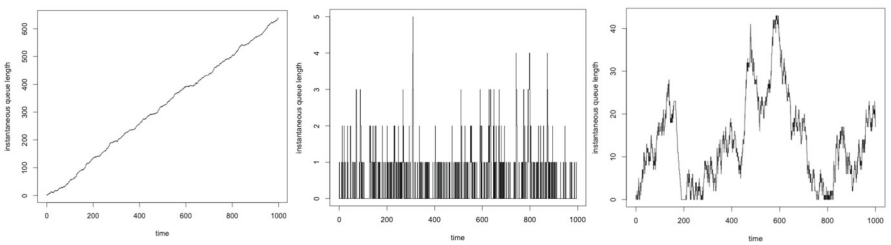


**Fig. 2** Three simulations for the Mittag-Leffler $GI/GI/1$ queue for large times. From left to right are the cases $\alpha_1 > \alpha_2, \alpha_1 < \alpha_2, \alpha_1 = \alpha_2$ respectively

We re-write equation (2.9) with this notation.

$$\left\{L_{\lambda,\mu}^{\alpha_1,\alpha_2}(t)\right\}_{t\geq 0} = \left\{\left(N_{\lambda}^{(\alpha_1)}(t) - N_{\mu}^{(\alpha_2)}(t)\right) - \inf_{0\leq s\leq t}\left(N_{\lambda}^{(\alpha_1)}(s) - N_{\mu}^{(\alpha_2)}(s)\right)\right\}_{t\geq 0}, \tag{3.19}$$

with $\alpha_1, \alpha_2 \in (0,1)$ and $\lambda, \mu > 0$. The first theorem in this section concerns the scaling limit of the queue length.

**Theorem 3.3** (Scaling limits) *Let $Y_\alpha(t)$, $\tilde{Y}_\alpha(t)$ be two independent copies of the inverse $\alpha-$stable subordinator defined in (3.5). Let $\gamma = \max\{\alpha_1, \alpha_2\} \in (0,1)$, then we have the following weak convergence as $t \to \infty$ with respect to the Skorohod $J_1$ topology:*

$$\left\{\frac{L_{\lambda,\mu}^{\alpha_1,\alpha_2}(t\tau)}{t^{\gamma}}\right\}_{\tau\geq 0} \xrightarrow{(w)} \begin{cases} \left\{\lambda^{\alpha_1} Y_\gamma(\tau)\right\}_{\tau\geq 0}, & \alpha_1 > \alpha_2, \\ \{0\}_{\tau\geq 0}, & \alpha_1 < \alpha_2, \\ \left\{\lambda^{\gamma} Y_\gamma(\tau) - \mu^{\gamma}\tilde{Y}_\gamma(\tau) - \inf_{0\leq s\leq \tau}\left(\lambda^{\gamma} Y_\gamma(\tau) - \mu^{\gamma}\tilde{Y}_\gamma(\tau)\right)\right\}_{\tau\geq 0}, & \alpha_1 = \alpha_2. \end{cases} \tag{3.20}$$

Theorem 3.3 already suggests the potential behaviour of the queue length in terms of recurrence and transience. Recurrence in this case would mean that the queue length would be zero infinitely often while transience means that the queue length will diverge to infinity as time goes by. Since the subordinator is non-negative and increasing, we expect that when $\alpha_1 > \alpha_2$ the queue length should divert to infinity. Similarly, when $\alpha_1 < \alpha_2$ and the limit is 0, we expect the queue length to not grow by much (though at this point the theorem only guarantees that it grows less than $t^{\gamma}$). See Fig. 2 for the corresponding behaviours when the queue length is unscaled.

Finally, when the tail indices are the same the behaviour should be guided by the scalings $\lambda$ and $\mu$ so the situation is slightly more delicate. This can also be seen in the simulations of Fig. 3.

The two following theorems explore these various transience and recurrence behaviour. See also Figs. 4 and 5.

**Theorem 3.4** *Let $L_{\lambda,\mu}^{\alpha_1,\alpha_2}$ denote the Mittag-Leffler* GI/GI/1 *queue length, as defined in equation (3.19). Then for any $\lambda, \mu > 0$ we have that*

1. *If $\alpha_1 < \alpha_2$ then*

$$\mathbb{P}\{L_{\lambda,\mu}^{\alpha_1,\alpha_2}(t) = 0 \ i.o.\} = 1.$$

2. *If $\alpha_1 > \alpha_2$ then*

$$\mathbb{P}\{\overline{\lim_{t\to\infty}} L_{\lambda,\mu}^{\alpha_1,\alpha_2}(t) = \infty\} = 1.$$

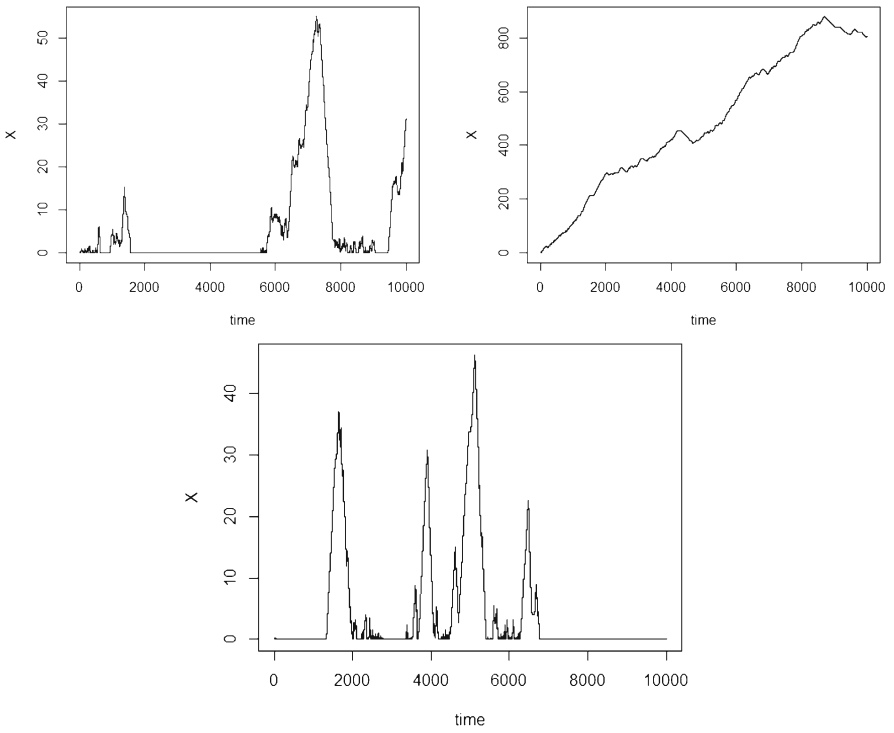


**Fig. 3** Three simulated realisations for the limiting process given in the last line of equation (3.20) for the case where $\alpha_1 = \alpha_2 = 0.9$. These figures come directly from the simulations of the inverse alpha-stable subordinator using the CMS formula [10]. From left to right we have the behaviours for (left) $\lambda = \mu = 1$, (right) $\lambda = 2$, $\mu = 1$ and (bottom) $\lambda = 1$, $\mu = 2$

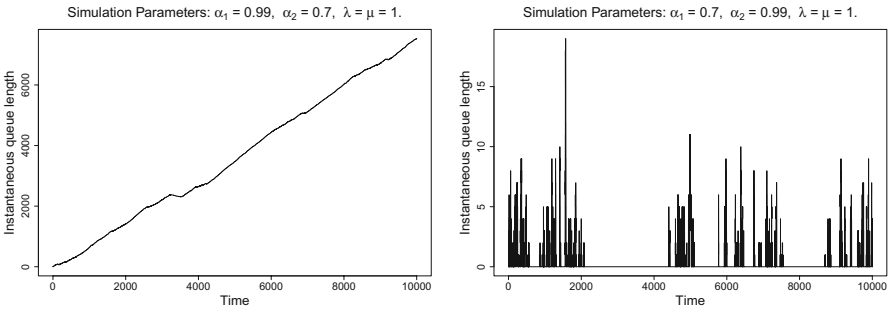


**Fig. 4** Two simulations for the Mittag-Leffler GI/GI/1 queue which highlight their recurrence and transience properties. The arrival and departure counting process have different tail index $\alpha$. These are described in Theorem 3.4

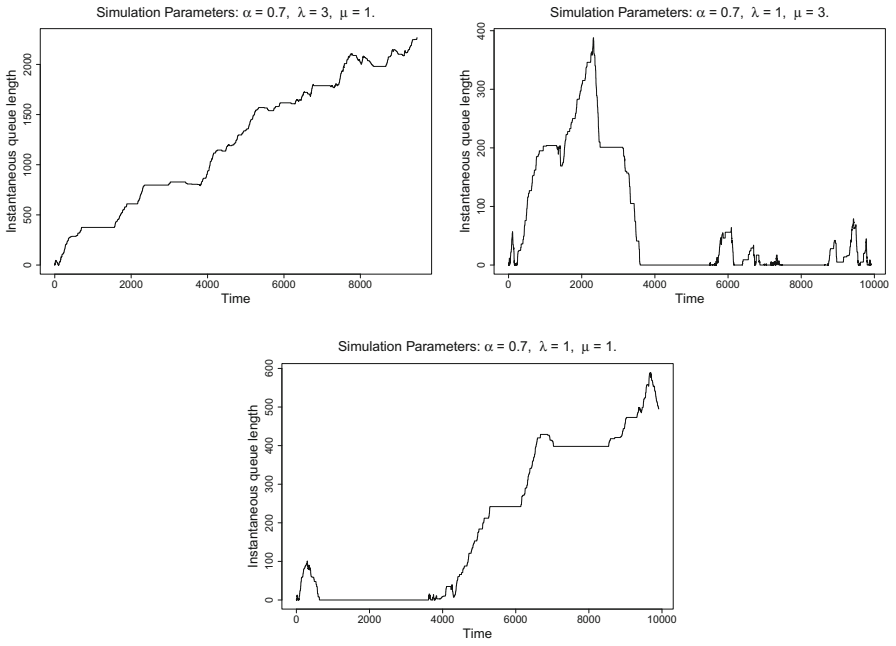


**Fig. 5** Three simulations for the Mittag-Leffler GI/GI/1 queue. The arrival and departure counting process have the same tail index $\alpha = 0.7$ but the arrival rates differ. This is the situation of Theorem 3.5

**Theorem 3.5** *Let* $L^{\alpha_1,\alpha_2}_{\lambda,\mu}$ *denote the Mittag-Leffler* GI/GI/1 *queue length, as defined in equation* (3.19)*. Then,*

1. *If* $\alpha_1 = \alpha_2 = \alpha$ *and* $\mu \geq \lambda$*, then*

$$\mathbb{P}\{L^{\alpha,\alpha}_{\lambda,\mu}(t) = 0 \; i.o. \; \} = 1.$$

2. *If* $\alpha_1 = \alpha_2 = \alpha$ *and* $\mu \leq \lambda$*, then we can find a positive constant* $c_{\mu,\lambda}$ *so that*

$$\mathbb{P}\Big\{ \overline{\lim_{t\to\infty}} \, L^{\alpha,\alpha}_{\lambda,\mu}(t) = \infty \Big\} > c_{\lambda,\mu}.$$

***Remark 3.4*** Note that in Fig. 5 it seems that when $\lambda > \mu$ we should be able to substitute $\overline{\lim}$ with a $\underline{\lim}$ in the statement of Theorem 3.5, or upgrade $c_{\lambda,\mu}$ with 1. Similarly, when $\lambda = \mu$ we should be able to say $c_{\lambda,\mu} = 1$. It seems that both of these arguments require a stronger coupling and comparisons with random walks than the one we actually develop in the sequel. The main difficulty is that these random walks have heavy-tailed continuous increments and are subject to unintuitive behaviours.

## 4 The (single) time fractional queue

### 4.1 Moment generating function

The first thing we would like to establish is that the queue length $L^{(1)}_{\alpha,\lambda}(t)$ with mass function $\{p_{0,i}(t)\}_{i\geq 0}$ has exponential moments in a neighbourhood of 0 for any fixed $t$.

To this end, let $s > 0$ and estimate

$$\mathbb{E}(e^{sL^{(1)}_{\alpha,\lambda}(t)}) = \mathbb{E}\Big(e^{sQ_{N^{(\alpha)}_{\lambda}(t)}}\Big) \leq \mathbb{E}(e^{sN^{(\alpha)}_{\lambda}(t)}) = E_{\alpha,1}(\lambda^{\alpha}t^{\alpha}(e^{s}-1)) < \infty,$$

since $Q_n \leq n$. This is enough to guarantee that the moment generating function is well-defined for $s \in \mathbb{R}$.

The interested reader can see a direct proof of Theorem 3.1 in Appendix B. The steps of the proof follow the methodology of [8] which is written for the probability generating function for the general fractional queue. These steps also imitate classical methodology for the M/M/1 queue [6]. The Laplace transform of the p.g.f. for the queue with starting length $i \in \mathbb{N}$ is explicitly computed in [8] and is given by the formula

$$\widetilde{G}_{\alpha,1}(z,s) = s^{\alpha-1}\frac{z^{i+1} - (1-z)r_2(s)^{i+1}(1-r_2(s))^{-1}}{-p(z-r_2(s))(z-r_2(s))}. \tag{4.1}$$

Here, $r_1(s), r_2(s)$ are those from equation (3.13). This is enough to give us the Laplace transform of all probabilities

$$\widetilde{p}_{0,n}(s) = \frac{s^{\alpha-1}}{p(1-r_2(s))r_1(s)}\frac{1}{r_1^n(s)} = \frac{\widetilde{p}_{0,0}(s)}{r_1^n(s)}, \quad n \geq 0, \tag{4.2}$$

in terms of $\widetilde{p}_{0,0}(s)$ which is given by (see also (B6))

$$\widetilde{p}_{0,0}(s) = \frac{s^{\alpha-1}}{p(1-r_2(s))r_1(s)} = \frac{r_2(s)s^{\alpha-1}}{(1-p)(1-r_2(s))} = \frac{1}{s} - \frac{p}{1-p}\frac{r_2(s)}{s}. \tag{4.3}$$

The proof of Theorem 3.1 is postponed to Appendix B as the steps are similar to those of the p.g.f. and we continue the section with applications and examples by considering equation (3.14) as given.

### 4.2 Applications for asymptotics

At this particular point we can expand $\widetilde{M}_{\alpha,1}(z,s)$ into two different series, for two different applications. First,

$$\widetilde{M}_{\alpha,1}(z,s) = \frac{s^{\alpha-1}}{p(1-r_2(s))r_1(s)}\sum_{n=0}^{\infty}\frac{1}{r_1^n(s)}e^{nz}. \tag{4.4}$$

This is obtained by taking the Laplace transform of the series for $M_{\alpha,1}(z,t)$ and using (4.2), (4.3).

Second, we can expand (4.4) in powers of $z$. We have

$$\begin{aligned}\widetilde{M}_{\alpha,1}(z,s) &= \frac{s^{\alpha-1}}{p(1-r_2(s))r_1(s)}\sum_{n=0}^{\infty}\frac{1}{r_1^n(s)}e^{nz}\\ &= \frac{s^{\alpha-1}}{p(1-r_2(s))r_1(s)}\sum_{n=0}^{\infty}\frac{1}{r_1^n(s)}\sum_{k=0}^{\infty}\frac{n^k z^k}{k!}\\ &= \frac{s^{\alpha-1}}{p(1-r_2(s))r_1(s)}\sum_{k=0}^{\infty}\Big(\sum_{n=0}^{\infty}\frac{1}{r_1^n(s)}n^k\Big)\frac{z^k}{k!}. \end{aligned} \tag{4.5}$$

Then a coefficient comparison gives the Laplace transform for all

$$\begin{aligned}\mathcal{L}\{\mathbb{E}((L_{\alpha,1}^{(1)}(t))^k)\}(s) &= \frac{s^{\alpha-1}}{p(1-r_2(s))r_1(s)}\Big(\sum_{n=0}^{\infty}\frac{1}{r_1^n(s)}n^k\Big)\\ &= \frac{s^{\alpha-1}}{p(1-r_2(s))r_1(s)}\Phi\Big(\frac{1}{r_1(s)},-k,0\Big).\end{aligned}$$

The function $\Phi$ is the Hurwitz-Lerch transcendent. In our case, since $k$ is an integer, it can be computed explicitly with repeated derivatives of the geometric series. It can be directly verified that for any $a>1$

$$\Phi\Big(\frac{1}{a},-k,0\Big) = -a\frac{d}{da}\Phi\Big(\frac{1}{a},-k+1,0\Big),$$

which gives a fast way to compute the Laplace transform of the moments. We can then directly compute the Laplace transform of the mean and, by inverting it, its formula. We have

$$\mathcal{L}\{\mu_{L_{\alpha,1}^{(1)}(t)}\}(s) = \frac{2p-1}{s^{1+\alpha}} + \frac{1-p}{s^{\alpha}}\widetilde{p}_{0,0}(s) \tag{4.6}$$

$$\Longleftrightarrow \mu_{L_{\alpha,1}^{(1)}(t)} = \frac{2p-1}{\Gamma(1+a)}t^{\alpha} + \frac{1-p}{\Gamma(a)}\int_0^t (t-u)^{\alpha-1}p_{0,0}(u)\,du$$

$$= \frac{2p-1}{\Gamma(1+a)}t^{\alpha} + (1-p)J^{\alpha}p_{0,0}(t). \tag{4.7}$$

Above we used $J^{\alpha}p_{0,0}(t)$ to denote the Riemann-Liouville fractional integral of $p_{0,0}(t)$, in general defined by

$$J^{\alpha}f(t) = \frac{1}{\Gamma(\alpha)}\int_0^t (y-t)^{\alpha-1}f(y)\,dy.$$

Similarly, we can compute by hand the Laplace transform for the second moment of the queue length, given by

$$\begin{aligned}\mathcal{L}\{\mathbb{E}((L^{(1)}_{\alpha,1}(t))^2)\}(s) &= \widetilde{p}_{0,0}(s)\frac{r_1(s)(1+r_1(s))}{(r_1(s)-1)^3}\\ &= \frac{p^3}{s^{3a}}\widetilde{p}_{0,0}(s)r_1(s)(1+r_1(s))(1-r_2(s))^3. \qquad (4.8)\end{aligned}$$

Repeated applications of (B5) and (B6) can then 'simplify' the expression into

$$\begin{aligned}\mathcal{L}\{\mathbb{E}((L^{(1)}_{\alpha,1}(t))^2)\}(s) &= \frac{1}{s^{3a+1}}\Big(\theta_1(p)+\theta_2(p)s^{\alpha}+\theta_3(p)s^{2\alpha}\Big)\\ &\quad+\frac{1}{s^{3a}}\widetilde{p}_{0,0}(s)\Big(\theta_4(p)+\theta_5(p)s^{\alpha}+\theta_6(p)s^{2\alpha}\Big),\end{aligned}$$

which can the be inverted should a close formula be desirable.

***Example 4.1*** (Asymptotics for the probability of an empty queue) First assume $p \neq 1/2$. As $s \to 0$ we Taylor expand $\widetilde{p}_{0,0}(s)$

$$\begin{aligned}\widetilde{p}_{0,0}(s) &= \frac{1-2p}{2(1-p)}\frac{1}{s} - \frac{1}{2(1-p)s^{1-\alpha}} + \frac{|1-2p|}{2(1-p)s}\sqrt{1+\frac{2}{(1-2p)^2}s^{\alpha}+o(s^{\alpha})}\\ &= \Big(\frac{1-2p}{2(1-p)}+\frac{|1-2p|}{2(1-p)}\Big)\frac{1}{s}+\frac{1}{2(1-p)}\Big(\frac{1}{|1-2p|}-1\Big)\frac{1}{s^{1-\alpha}}+o(s^{\alpha-1}).\end{aligned} \qquad (4.9)$$

If $p < 1/2$, which means $Q_n$ is positive recurrent, the leading term above is $\frac{1-2p}{1-p}s^{-1}$ as $s \to 0$. Then from the Tauberian theorem of Laplace transforms we have

$$\frac{1-2p}{1-p} = \lim_{t\to\infty}\frac{\int_0^t p_{0,0}(u)\,du}{t} = \lim_{t\to\infty} p_{0,0}(t), \qquad (4.10)$$

which coincides with (2.4). A similar approach can prove the limiting distribution for $p_{0,n}$, via (4.2). For more on the equilibrium distribution, see the next section.

If $p > 1/2$, the $s^{-1}$ disappears from (4.9) and the leading term is $\frac{1}{2p-1}s^{\alpha-1}$. Therefore,

$$\frac{1}{(2p-1)\Gamma(2-\alpha)} = \lim_{t\to\infty}\frac{\int_0^t p_{0,0}(u)\,du}{t^{1-\alpha}} = \lim_{t\to\infty}\frac{p_{0,0}(t)}{(1-\alpha)t^{-\alpha}},$$

which gives

$$p_{0,0}(t) \sim \frac{1}{(2p-1)\Gamma(1-\alpha)}t^{-\alpha}, \quad t\to\infty.$$

Finally, focus on the particular case of $p = 1/2$. We have for $s \to 0$

$$\begin{aligned}\widetilde{p}_{0,0}(s) = \frac{1}{s} - \frac{r_2(s)}{s} &= -\frac{1}{s^{1-\alpha}} + \frac{\sqrt{2}s^{\alpha/2}\sqrt{1+s^\alpha/2}}{s} \\ &= -\frac{1}{s^{1-\alpha}} + \frac{\sqrt{2}s^{\alpha/2}(1+s^\alpha/4+o(s^\alpha))}{s} \\ &= \frac{\sqrt{2}}{s^{1-\alpha/2}} + o(s^{-1+\alpha/2}).\end{aligned}$$

Then as before,

$$p_{0,0}(t) \sim \frac{\sqrt{2}}{\Gamma(1-\alpha/2)} t^{-\alpha/2}, \quad t \to \infty. \tag{4.11}$$

***Example 4.2*** (Asymptotics for the mean queue length and variance) From Example 4.1 and equation (4.7), we obtain the Laplace transform asymptotics as $s \to 0$ for the queue length

$$\mathcal{L}\{\mu_{L^{(1)}_{\alpha,1}(t)}\}(s) = \begin{cases} \dfrac{p}{1-2p} s^{-1} + o(s^{-1}), & p < 1/2, \\ \dfrac{\sqrt{2}}{2s^{1+\alpha/2}} + o(s^{-1-\alpha/2}), & p = 1/2, \\ \dfrac{2p-1}{s^{1+\alpha}} + o(s^{-1-\alpha}), & p > 1/2. \end{cases} \tag{4.12}$$

This, via the Tauberian theorem for Laplace transforms, gives the asymptotics

$$\mu_{L^{(1)}_{\alpha,1}(t)} \sim \begin{cases} \dfrac{p}{1-2p}, & p < 1/2, \\ \dfrac{\sqrt{2}}{2\Gamma(1+\alpha/2)} t^{\alpha/2}, & p = 1/2, \\ \dfrac{2p-1}{\Gamma(1+\alpha)} t^{\alpha}, & p > 1/2. \end{cases} \tag{4.13}$$

By Taylor expanding (4.8) up to 3 terms in order to find the asymptotic of the second moment, we also obtain that

$$\mathrm{Var}(L^{(1)}_{\alpha,1}(t)) \sim \begin{cases} \mathcal{C}_1, & p < 1/2, \\ \mathcal{C}_2 t^{\alpha}, & p = 1/2, \\ \mathcal{C}_3 t^{2\alpha}, & p > 1/2. \end{cases} \tag{4.14}$$

Note that when $p \leq 1/2$ the Taylor approximation can only be with the first two terms of the expansion but the case $p > 1/2$ requires the third term.

***Remark 4.1*** Note that in the case $p > 1/2$ the orders of both the mean and the variance coincide with the corresponding orders of the fractional Poisson process. This is yet another indication of the transience of the queue length in this case.

### 4.3 Equilibrium and mixing times in the ergodic regime

There are several ways to compute the equilibrium probabilities as $t \to \infty$ when $p < 1/2$. For example, as in Example 4.1, we can perform the same calculations that led to (4.10) and obtain in general that the limiting mass function of the queue length coincides with that of $Q_n$ from (2.4), namely

$$\lim_{t\to\infty} p_{0,k}(t) = \pi_k = \frac{1-2p}{1-p}\Big(\frac{p}{1-p}\Big)^k, \quad k \in \mathbb{N}_0. \tag{4.15}$$

This is just a limiting statement for the probabilities for fixed $k$ and we would like to upgrade the statement in terms of mixing time for the chain, in terms of the total variation distance.

***Proof of Theorem 3.2*** Our starting point is Corollary 3.1 from [32] which states that for a spatial birth and death chain, potentially on an infinite space (like $Q_n$) we have the following estimate for the total variation distance between the invariant distribution $\{\pi_k\}_{k\in\mathbb{N}_0}$ and $\{\mathbb{P}_\gamma\{Q_n = k\}\}_{k\in\mathbb{N}_0}$. The initial distribution $\gamma$ for the chain must satisfy a non-degeneracy condition (see conditions in [32]) which is satisfied if for example $\gamma$ has a finite support. Using the result, we can find positive constants $c$ and $\theta < 1$ such that

$$\sup_{A\in\mathcal{F}} \Big|\pi(A) - \mathbb{P}_\gamma\{Q_n \in A\}\Big| \le c\,\theta^n. \tag{4.16}$$

Note that Corollary 3.1 is stated for continuous times birth and death chains, but indeed the same follows for discrete times by the concentration of the Poisson process around its mean.

We use this estimate to bound the total variation distance between $p_\gamma(t)$ and $\pi$. Condition on the events of the counting process to obtain

$$p_{\gamma,k}(t) = \sum_{n=0}^{\infty} \mathbb{P}_\gamma\{Q_n = k\}\mathbb{P}\{N^{(\alpha)}(t) = n\} \tag{4.17}$$

and then estimate

$$\begin{aligned}
\|p_\gamma(t) - \pi\|_{TV} &= \frac{1}{2}\sum_{k\in\mathbb{N}_0} |p_{\gamma,k}(t) - \pi_k| \\
&= \frac{1}{2}\sum_{k\in\mathbb{N}_0} |\sum_{n=0}^{\infty} \mathbb{P}_\gamma\{Q_n = k\}\mathbb{P}\{N^{(\alpha)}(t) = n\} - \pi_k(t)|, \quad \text{by (4.17)} \\
&\le \frac{1}{2}\sum_{k\in\mathbb{N}_0}\sum_{n=0}^{\infty} \big|\mathbb{P}_\gamma\{Q_n = k\} - \pi_k\big|\mathbb{P}\{N^{(\alpha)}(t) = n\},
\end{aligned}$$

$$
\begin{aligned}
&\text{by triangle inequality}\\
&= \sum_{n=0}^{\infty} \left\| \mathbb{P}_\gamma\{Q_n \in \cdot\} - \pi \right\|_{TV} \mathbb{P}\{N^{(\alpha)}(t) = n\}\\
&\leq \sum_{n=0}^{\infty} c\theta^n \mathbb{P}\{N^{(\alpha)}(t) = n\}, \quad \text{by (4.16)}\\
&= c\mathbb{E}(\theta^{N^{(\alpha)}(t)}) = cE_{\alpha,1}(-(1-\theta)t^\alpha).
\end{aligned}
$$

***Remark 4.2*** The above estimate shows that the total variation distance from the equilibrium decays like $t^{-\alpha}$, the decay rate of the Mittag-Leffler function. Note that the upper bound above is strictly monotone in $t$.

For any $\varepsilon > 0$ define $T_\varepsilon^{c,\theta}$ to be

$$T_\varepsilon^{c,\theta} = \inf\{t > 0 : cE_{\alpha,1}(-(1-\theta)t^\alpha) \leq \varepsilon\}.$$

Then we conclude that

$$T_\varepsilon^{\text{mix}} \leq T_\varepsilon^{c,\theta}. \tag{4.18}$$

Applying Theorem 4 from [38], we can say that for all time $t > 0$ we have

$$\frac{c}{1 + \Gamma(1-\alpha)(1-\theta)t^\alpha} \leq cE_{\alpha,1}(-(1-\theta)t^\alpha) \leq \frac{c}{1 + \Gamma(1-\alpha)^{-1}(1-\theta)t^\alpha}. \tag{4.19}$$

When $t = T_\varepsilon^{c,\theta}$, we have by continuity that $cE_{\alpha,1}(-(1-\theta)(T_\varepsilon^{c,\theta})^\alpha) = \varepsilon$. This then implies that

$$\varepsilon \leq \frac{c}{1 + \Gamma(1-\alpha)^{-1}(1-\theta)(T_\varepsilon^{c,\theta})^\alpha} \iff (T_\varepsilon^{c,\theta})^\alpha \leq \left(\frac{c}{\varepsilon} - 1\right) \Gamma(1-\alpha)^{-1}(1-\theta).$$

The theorem follows with constants $C_{\alpha,\lambda,\mu_0}$ given by $c$ in (4.16), and $U_{\alpha,\lambda,\mu_0}$ given by

$$U_{\alpha,\lambda,\mu_0} = \left(\frac{\Gamma(1+\alpha)}{1-\theta}\right)^{\frac{1}{\alpha}}. \tag{4.20}$$

□

## 5 The renewal queue

Recall $p_{\alpha,\beta}^{\lambda,\mu}$ from (3.18) is the probability of an arrival when a renewal event occurs. We can easily find some special values of $p_{\alpha,\beta}^{\lambda,\mu}$.

First, consider the case where either $\alpha$ or $\beta$ is equal to 1 and $\mu = \lambda$. Then we have

$$1 - p_{1,\beta}^{\lambda,\lambda} = 1 - p_{1,\beta}^{1,1} = P\{X_{1,1} > X_{\beta,1}\} = \int_0^\infty e^{-t} f_{X_\beta}(t)\, dt = \mathbb{E}(e^{-X_\beta}) = \frac{1}{2}.$$

So for this choice of parameters the embedded chain will correspond to a null-recurrent queue.

Then, if we factor in the scalings, we have

$$p_{1,\beta}^{\lambda,\mu} = \mathbb{P}\{\mu\lambda^{-1}X_1 < X_\beta\} = \int_0^\infty (1 - e^{-\lambda t/\mu}) f_{X_\beta}(t)\, dt = \frac{\lambda^\beta}{\mu^\beta + \lambda^\beta}.$$

Note that in this case $p_{1,\beta}^{\lambda,\mu} < 1/2 \iff \mu > \lambda$, i.e. the embedded chain is positive recurrent precisely when the departure process is faster.

Similarly, for any $\alpha \in (0, 1)$, by symmetry

$$p_{\alpha,\alpha}^{\lambda,\lambda} = \frac{1}{2},$$

and by a straightforward coupling argument,

$$p_{\alpha,\alpha}^{\lambda,\mu} < \frac{1}{2} \iff \mu > \lambda.$$

***Conjecture 5.1*** Let $\alpha, \beta$ in $(0, 1]$ and let $X_\alpha$ and $X_\beta$ be two independent Mittag-Leffler distributions with power indices $\alpha, \beta$ respectively and scalings $\lambda = \mu$. Then

$$\mathbb{P}\{X_\alpha < X_\beta\} = p_{\alpha,\beta}^{\lambda,\lambda} = \frac{1}{2}.$$

Then, the embedded discrete queue will be null recurrent if and only if $\lambda = \mu$ and positive recurrent if and only $\mu > \lambda$. I.e. the behaviour of the queue length depends solely on the time scales, not the power tails.

Unfortunately we have not been able to rigorously prove Conjecture 5.1, so we have Proposition 3.1. Simultaneously, the conjecture is strongly supported by high precision numerical calculations and you can see the results in Fig. 6.

***Proof of Proposition 3.1*** Since

$$p_{\alpha,\beta}^{\lambda,\mu} = p_{\alpha,\beta}^{\lambda/\mu,1}$$

define $\rho = \frac{\lambda}{\mu}$ to be a parameter in $(0, \infty)$, and consider the probabilities $p_{\alpha,\beta}^{\rho,1} = \mathbb{P}\{\rho^{-1}X_\alpha < X_\beta\}$. We have

$$p_{\alpha,\beta}^{\rho,1} = \int_0^\infty \mathbb{P}\{X_\alpha \le \rho t\} f_{X_\beta}(t)\, dt = \int_0^\infty F_{X_\alpha}(\rho t) f_{X_\beta}(t)\, dt.$$

With this representation we see that $p_{\alpha,\beta}^{\rho,1}$ is strictly monotonically increasing (from 0 to 1) as $\rho \to \infty$ by the complete monotonicity of $F_{X_\alpha}$ and the dominated convergence theorem. Then, there will be a critical $\rho_{\alpha,\beta}^*$ such that $p_{\alpha,\beta}^{\rho_{\alpha,\beta}^*,1} = 1/2$.

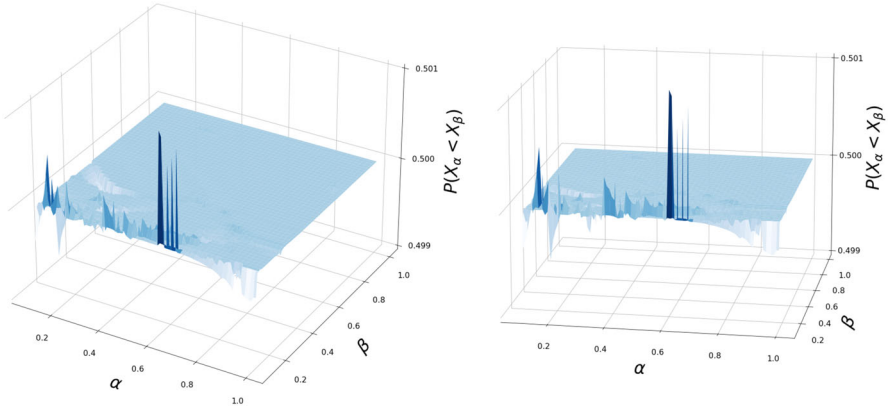


**Fig. 6** A numerical simulation for the value of $p^{\lambda,\mu}_{\alpha,\beta}$ when $\mu = \lambda = 1$ viewed from two angles. Note that even the minor discrepancies when $\alpha, \beta$ are small are still within 0.001 distance from 1/2 which is what we expect the true value to be. These discrepancies arise from numerical instabilities that occur when simulating values of $\alpha$ and $\beta$ near 0. As such the figure starts from values of $\alpha, \beta > 0.1$

The proposition then follows, since $\rho^*_{\alpha,\beta}$ will be the critical $\rho$ that separates positive recurrence and transience in the embedded chain. The embedded chain will be null recurrent when $\rho = \rho^*_{\alpha,\beta}$. □

***Remark 5.1*** Conjecture 5.1 says that $\rho^*_{\alpha,\beta} = 1$ for all values of $\alpha, \beta$. This is verified for the cases $\alpha \vee \beta = 1$ or $\alpha = \beta$. Note that in this case the characterisation of recurrence and transience of the embedded queue length would be equivalent to that of the M/M/1 queue. □

## 6 The Mittag-Leffler GI/GI/1 queue

### 6.1 Limit theorems

To prove these, we will first establish the convergence of the base fractional Poisson process, $N^{(\alpha)}_\nu(t)$, for $\alpha \in (0, 1)$ and $\nu > 0$.

**Lemma 6.1** *Let $\gamma \in (0, 1)$. Then we have the following convergence for the fractional Poisson process $N^{(\alpha)}_\nu$ in the Skorohod $J_1$ topology dependent on the value of $\gamma$:*

1. *$\gamma = \alpha$ implies that*

$$\left(\frac{N^{(\alpha)}_\nu(t\tau)}{t^\gamma}\right)_{\tau\geq 0} \xrightarrow{(w)} \left(\nu^\alpha Y_\alpha(\tau)\right)_{\tau\geq 0}, \tag{6.1}$$

2. *$\gamma > \alpha$ implies that*

$$\left(\frac{N^{(\alpha)}_\nu(t\tau)}{t^\gamma}\right)_{\tau\geq 0} \xrightarrow{(w)} \{0\}_{\tau\geq 0}. \tag{6.2}$$

***Proof*** The first statement follows from Theorem 6 of [24] and equations (3.6) - (3.8) from earlier.

Let us look at the second case, where $\gamma > \alpha$. In this case we can write $\gamma = \alpha + \delta$, with $\delta > 0$. Now, for fixed $t \geq 0$, we can write

$$\left(\frac{N_\nu^{(\alpha)}(t\tau)}{t^\gamma}\right)_{\tau\geq 0} \stackrel{d}{=} \left(\frac{N_\nu^{(\alpha)}(t\tau)}{t^\alpha} \cdot \frac{t^\alpha}{t^{\alpha+\delta}}\right)_{\tau\geq 0} \stackrel{d}{=} \left(\frac{N_\nu^{(\alpha)}(t\tau)}{t^\alpha}\right)_{\tau\geq 0} \cdot \frac{t^\alpha}{t^{\alpha+\delta}}. \tag{6.3}$$

It is known that multiplication in the Skorohod $J_1$ topology is continuous as long as the two functions share no common discontinuities. This is true in our case, since for every $t > 0$, $\frac{t^\alpha}{t^{\alpha+\delta}}$ can be viewed as a constant function with respect to $\tau$ and contains no discontinuities. It must also follow that the joint distribution of $\left(\frac{N_\nu^{(\alpha)}(t\tau)}{t^\alpha}\right)_{\tau\geq 0}$ and $\frac{t^\alpha}{t^{\alpha+\delta}}$ converges as $t \to \infty$, and we are free to invoke the Continuous Mapping theorem. Since $\frac{t^\alpha}{t^{\alpha+\delta}} \to 0$ as $t \to \infty$, it is clear that the whole expression above must converge to the constant function 0 with respect to the $J_1$ topology. □

***Proof of Theorem 3.3*** Now that we have established the convergence of the base fractional Poisson process, we can use results from [41] to push through to the whole queue length expression. First we draw attention to Theorem 4.1 of that paper, which states that the addition of two functions $x$, $y$ on a Skorohod space $\mathcal{D}$ is a continuous operation provided that $x$ and $y$ share no common discontinuities. Let $x_t$ and $y_t$ correspond to the fractional Poisson processes, $x_t(\tau) = \left\{\frac{N_\lambda^{(\alpha_1)}(t\tau)}{t^\gamma}\right\}_{\tau\geq 0}$ and $y_t(\tau) = \left\{\frac{N_\mu^{(\alpha_2)}(t\tau)}{t^\gamma}\right\}_{\tau\geq 0}$. These are two sequences of independent random variables and as such the joint distribution converges as $t \to \infty$. Since these processes share no common discontinuities we can then apply the continuous mapping theorem and Lemma 6.1 to show the convergence

$$\left\{\frac{N_\lambda^{(\alpha_1)}(t\tau) - N_\mu^{(\alpha_2)}(t\tau)}{t^\gamma}\right\}_{\tau\geq 0} \xrightarrow{(w)} \begin{cases} \left\{\lambda^{\alpha_1} Y_\gamma(\tau)\right\}_{\tau\geq 0} & \text{if } \alpha_1 > \alpha_2, \\ \left\{-\mu^{\alpha_2} Y_\gamma(\tau)\right\}_{\tau\geq 0} & \text{if } \alpha_1 < \alpha_2, \\ \left\{\lambda^\gamma Y_\gamma(\tau) - \mu^\gamma \tilde{Y}_\gamma(\tau)\right\}_{\tau\geq 0} & \text{if } \alpha_1 = \alpha_2. \end{cases} \tag{6.4}$$

Now we turn to Theorem 6.4 in [41]. Setting $c_n, c \equiv 0$ this states that if we have the convergence of $x_t(\tau) \to x(\tau)$ as $t \to \infty$ in the $J_1$ topology, then this implies the convergence $x_t^\downarrow(\tau) \to x^\downarrow(\tau)$, where

$$x^\downarrow(\tau) = x(\tau) - \inf_{0\leq s\leq \tau} x(s). \tag{6.5}$$

Combining this with the above result finally leads us to the expression

$$\left\{\frac{L_{\lambda,\mu}^{(\alpha_1,\alpha_2)}(t\tau)}{t^\gamma}\right\}_{\tau\geq 0} \xrightarrow{(w)}$$

$$\begin{cases} \left\{\lambda^{\alpha_1} Y_\gamma(\tau) - \inf_{0\le s\le\tau}\left(\lambda^{\alpha_1} Y_\gamma(s)\right)\right\}_{\tau\ge 0} & \text{if } \alpha_1 > \alpha_2, \\ \left\{-\mu^{\alpha_2} Y_\gamma(\tau) - \inf_{0\le s\le\tau}\left(-\mu^{\alpha_2} Y_\gamma(s)\right)\right\}_{\tau\ge 0} & \text{if } \alpha_1 < \alpha_2, \\ \left\{\lambda^{\gamma} Y_\gamma(\tau) - \mu^{\gamma}\tilde{Y}_\gamma(\tau) - \inf_{0\le s\le\tau}\left(\lambda^{\gamma} Y_\gamma(s) - \mu^{\gamma}\tilde{Y}_\gamma(s)\right)\right\}_{\tau\ge 0} & \text{if } \alpha_1 = \alpha_2. \end{cases}$$

To return to the original expression from the theorem, it suffices to observe that for all $\tau > 0$

$$\inf_{0\le s\le\tau}\left(\lambda Y_\gamma(s)\right) = 0,$$

and

$$\inf_{0\le s\le\tau}\left(-\mu^{\alpha_2} Y_\gamma(s)\right) = -\mu^{\alpha_2} Y_\gamma(\tau),$$

by the weak monotonicity of the subordinator. □

### 6.2 Recurrence results

In order to show the recurrence results of Theorems 3.4, 3.5, we need a few preliminary lemmas, as well as a coupling argument that will allow us to use a regeneration argument. Our first lemma works for any pair of counting process $N_\lambda(t)$, $N_\mu(t)$ and queue length given by

$$L_t = \left(N_\lambda(t) - N_\mu(t)\right) - \inf_{0\le s\le t}\left\{N_\lambda(s) - N_\mu(s)\right\}.$$

It offers a way to decide when we have an empty service.

**Lemma 6.2** *The following are equivalent.*

1. *Time $T$ is a discontinuity point of the infimum process, i.e.*

$$\inf_{0\le s\le T}\left\{N_\lambda(s) - N_\mu(s)\right\} = \inf_{0\le s\le T-}\left\{N_\lambda(s) - N_\mu(s)\right\} - 1. \tag{6.6}$$

2. *The departure process $N_\mu(t)$ has a renewal point at time $T$ and queue length satisfies*

$$L_{T-} = L_T = 0.$$

*In other words, at time $T$ there was an unused service time if and only if (6.6) is satisfied.*

***Proof*** We prove the direct implication. Since the infimum decreases by 1 at time $T$ by the assumption, we must have that $N_\mu(T)$ increased by 1 at time $T$, i.e. $T$ is a renewal point of $N_\mu$. Then we have that

$$0 \le L_{T-} = \left(N_\lambda(T-) - N_\mu(T-)\right) - \inf_{0\le s\le T-}\left\{N_\lambda(s) - N_\mu(s)\right\}$$

$$= N_\lambda(T) - (N_\mu(T) - 1) - \left( \inf_{0 \le s \le T} \{N_\lambda(s) - N_\mu(s)\} + 1 \right)$$
$$= N_\lambda(T) - N_\mu(T) - \inf_{0 \le s \le T} \{N_\lambda(s) - N_\mu(s)\} = L_T \le (L_{T-} - 1) \vee 0.$$

The last inequality holds because we know that the departure process has an event. But then, for the whole string of inequalities to hold, it must be that $L_T - = 0$.

The converse implication is left to the reader, but the approach can be found in the proof of (2.9) in Appendix A. □

***Proof of Theorem 3.4*** First consider the case $\alpha_1 < \alpha_2$. The proof is independent of the time scalings $\lambda$ and $\mu$; what is important is the tail exponent.

We begin with the following estimate which will be valid for any $s > 0$.

$$\begin{aligned}
&\mathbb{P}\Big\{ \inf_{0 \le s \le t} \{N^{(\alpha_1)}(s) - N^{(\alpha_2)}(s)\} > -t^{\alpha_1/2} \Big\} \\
&\qquad \le \mathbb{P}\big\{ N^{(\alpha_1)}(t) - N^{(\alpha_2)}(t) > -t^{\alpha_1/2} \big\} \\
&\qquad \le e^{st^{\alpha_1/2}} \mathbb{E}(e^{sN^{(\alpha_1)}(t)}) \mathbb{E}(e^{-sN^{(\alpha_2)}(t)}) \text{ via a Chernoff bound,} \\
&\qquad = e^{st^{\alpha_1/2}} E_{\alpha_1,1}((e^s - 1)t^{\alpha_1}) E_{\alpha_2,1}((e^{-s} - 1)t^{\alpha_2}).
\end{aligned}$$

At this point we make a particular choice for $s$ since we can choose any positive value for it. To have a bound that tends to 0 as $t \to \infty$, set $s = t^{-\alpha_1}$. Then we obtain

$$\begin{aligned}
&\mathbb{P}\Big\{ \inf_{0 \le s \le t} \{N^{(\alpha_1)}(t) - N^{(\alpha_2)}(t)\} > -t^{\alpha_1/2} \Big\} \\
&\qquad \le e^{t^{-\alpha_1/2}} E_{\alpha_1,1}((e^{t^{-\alpha_1}} - 1)t^{\alpha_1}) E_{\alpha_2,1}((e^{-t^{-\alpha_1}} - 1)t^{\alpha_2}) \\
&\qquad = e^{t^{-\alpha_1/2}} E_{\alpha_1,1}(1 + O(-t^{-\alpha_1})) E_{\alpha_2,1}(-t^{\alpha_2-\alpha_1} + o(t^{\alpha_2-\alpha_1})) \\
&\qquad \le C \frac{1}{t^{\alpha_2-\alpha_1}}, \quad \text{for } t \text{ large enough.}
\end{aligned}$$

To see the last inequality, observe that as $t$ grows, $e^{t^{-\alpha_1/2}}$ converges to 1 and $E_{\alpha_1,1}(1 + O(-t^{-\alpha_1}))$ converges by continuity to $E_{\alpha_1,1}(1)$). Therefore both of these terms are bounded by some constant $C_1$ for large $t$. The third term is the one who dictates the behaviour at infinity. For $t$ large enough, and by the monotonicity of $E_{\alpha_2,1}$ we have

$$\begin{aligned}
E_{\alpha_2,1}(-t^{\alpha_2-\alpha_1} + o(t^{\alpha_2-\alpha_1})) &\le E_{\alpha_2,1}\Big( -\frac{1}{2}\Big( t^{\frac{\alpha_2-\alpha_1}{\alpha_2}} \Big)^{\alpha_2} \Big) \\
&\le \frac{C_2}{\Big( t^{\frac{\alpha_2-\alpha_1}{\alpha_2}} \Big)^{\alpha_2}} = \frac{C_2}{t^{\alpha_2-\alpha_1}}.
\end{aligned}$$

Then, define the events

$$\mathcal{A}_n = \Big\{ \inf_{0 \le s \le n} \{N^{(\alpha_1)}(s) - N^{(\alpha_2)}(s)\} \le -n^{\alpha_1/2} \Big\}, \quad \mathbb{P}\{\mathcal{A}_n\} \ge 1 - \frac{C}{n^{\alpha_2-\alpha_1}}.$$

We compute

$$\mathbb{P}\{\mathcal{A}_n \text{ i.o.}\} = \mathbb{P}\Big\{\bigcap_{k=1}^{\infty}\bigcup_{n=k}^{\infty}\mathcal{A}_n\Big\} = \lim_{k\to\infty}\mathbb{P}\Big\{\bigcup_{n=k}^{\infty}\mathcal{A}_n\Big\} \geq \lim_{k\to\infty}\mathbb{P}\Big\{\mathcal{A}_k\Big\}$$
$$\geq \lim_{k\to\infty}\Big(1-\frac{C}{k^{\alpha_2-\alpha_1}}\Big) = 1.$$

Therefore, on a full probability event, we can find a sequence of integers $\{k_j\}_{j\geq 1}$ so that

$$f(j) = \inf_{0\leq s\leq k_j}\{N^{(\alpha_1)}(s) - N^{(\alpha_2)}(s)\}$$

is a strictly decreasing integer function. As such, by Lemma 6.2 we know that we can find an infinite sequence of times $t_j$ for which the departure process had an unused service time and that happens if and only if the queue length was already zero. Therefore the queue length becomes zero infinitely often with probability 1.

For the case $\alpha_1 > \alpha_2$ we reason similarly, but for $s < 0$.

$$\mathbb{P}\Big\{N^{(\alpha_1)}(t) - N^{(\alpha_2)}(t) < t^{\alpha_2/2}\Big\}$$
$$\leq e^{-st^{\alpha_2/2}}\mathbb{E}(e^{sN^{(\alpha_1)}(t)})\mathbb{E}(e^{-sN^{(\alpha_2)}(t)}) \text{ via a Chernoff bound,}$$
$$= e^{-st^{\alpha_2/2}}E_{\alpha_1,1}((e^s-1)t^{\alpha_1})E_{\alpha_2,1}((e^{-s}-1)t^{\alpha_2}).$$

At this point we make a particular choice for $s$ since we can choose any positive value for it. To have a bound that tends to 0 as $t\to\infty$, set $s = -t^{-\alpha_2}$. Then we obtain

$$\mathbb{P}\Big\{N^{(\alpha_1)}(t) - N^{(\alpha_2)}(t) < t^{\alpha_2/2}\Big\}$$
$$\leq e^{t^{-\alpha_2/2}}E_{\alpha_2,1}((e^{t^{-\alpha_2}}-1)t^{\alpha_2})E_{\alpha_1,1}((e^{-t^{-\alpha_2}}-1)t^{\alpha_1})$$
$$\leq e^{t^{-\alpha_2/2}}E_{\alpha_2,1}(1+O(-t^{-\alpha_2}))E_{\alpha_1,1}(-t^{\alpha_1-\alpha_2}+o(t^{\alpha_1-\alpha_2}))$$
$$\leq C\frac{1}{t^{\alpha_1-\alpha_2}}, \quad \text{for } t \text{ large enough.}$$

In order to conclude, use the fact that $L^{\alpha_1,\alpha_2}(t) \geq N^{\alpha_1}(t) - N^{\alpha_2}(t)$, which gives

$$\mathbb{P}\{L^{\alpha_1,\alpha_2}(t) \geq t^{\alpha/2}\} \geq 1 - \frac{C}{t^{\alpha_1-\alpha_2}}.$$

Finally, for any $n\in\mathbb{N}$, define

$$\mathcal{B}_n = \{L^{\alpha_1,\alpha_2}(n) \geq n^{\alpha/2}\}$$

and notice that

$$\mathbb{P}\{\overline{\lim_{t\to\infty}} L^{\alpha_1,\alpha_2}(t) = \infty\} \geq \mathbb{P}\{\mathcal{B}_n \text{ i.o. }\} \geq \lim_{n\to\infty}\Big(1-\frac{C}{n^{\alpha_1-\alpha_2}}\Big) = 1. \tag{6.7}$$

□

### 6.3 The proof of Theorem 3.5

The situation in Theorem 3.5 requires a slightly more delicate approach and two coupling arguments which we present first. We begin with a construction of a 'sped-up' queue which also provides a regeneration time for the dynamics.

Let $N_a(t)$ denote the arrival counting process and $N_d(t)$ denote the departure counting process. Let $L_t$ denote the corresponding queue length. Assume that at time $T$ we have a departure event, and assume that the next arrival event is at time $T + \eta$. Define new processes

$$\widetilde{N}_a(t) = \begin{cases} N_a(t), & t < T-, \\ N_a(\eta + t), & t \geq T, \end{cases}$$

which just speeds up the arrival process by $\eta$. Then we define a new queue length at time $T$

$$\widetilde{L}_t = \begin{cases} L_t, & t \leq T- \\ 1 + \widetilde{N}_a(t) - N_d(t) - \inf_{0 \leq s \leq t} \left\{ \widetilde{N}_a(s) - N_d(s) \right\}, & t \geq T. \end{cases} \tag{6.8}$$

This just says that at time $T$ the departure event happened before we brought forward the next departure event. Note that $\widetilde{L}_{t-T}, t \geq T$ has the same distribution as $L_t$ with an initial condition of $\widetilde{L}_T = L_T + 1$.

**Lemma 6.3** *Consider the queue $\widetilde{L}_t$ defined in (6.8). Let $T_0$ be the departure time event after which we use $\widetilde{N}_a(t)$ and let $T_1$ be the first unused service time after $T_0$ for the process $L_t$. (a priori $T_1$ could be infinity, and that is also fine). Then,*

$$L_t \leq \widetilde{L}_t, \quad \textit{for all } t < T_1. \tag{6.9}$$

Intuitively the lemma is clear, since we are just speeding up the arrival process by the random value $\eta$, however one should acknowledge the possibility that by doing so, a customer who appeared earlier may have benefited for a previously unused service event, thus making the second queue length shorter. Indeed by the construction this will not happen until the first unused departure time and this is shown in the next proof. See also Fig. 7 for a pictorial explanation.

***Proof of Lemma 6.3*** Since no modification happens up to time $T_0$, we have that $L_t = \widetilde{L}_t$ for all $t < T_0$. At time $T_0$ we have a departure event. If $L_{T_0-} > 0$, by construction we have that

$$L_{T_0} = L_{T_0-} - 1 = \widetilde{L}_{T_0} - 1 < \widetilde{L}_{T_0}.$$

Similarly, if $L_{T_0-} = 0$, the departure at $T_0$ would have been unused and $L_{T_0} = 0$ while in the modified queue $\widetilde{L}_{T_0} = 1$.

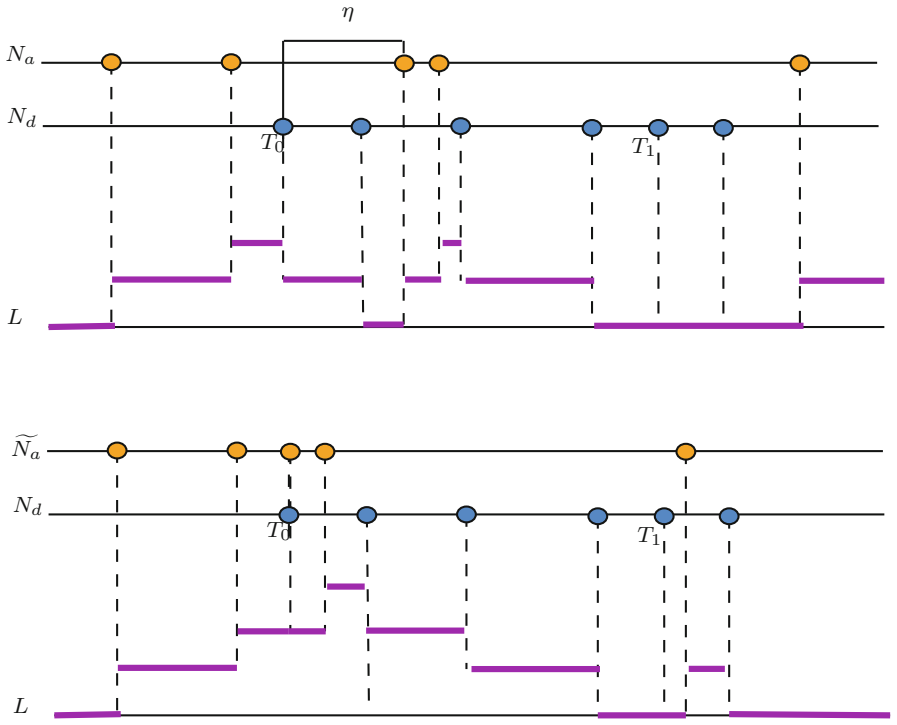


**Fig. 7** Schematic for the proof of Lemma 6.3. Time $T_0$ is a departure time (up until that point the two queue lengths are equal). $\eta$ is the time until next arrival in the original arrival process $N_a$ which becomes 0 in the sped up process $\tilde{N}_a$. $T_1$ is the first unused service after $T$. Then up to $T_1$ the queue length of the sped up process is not smaller than the original one. As the picture suggests, it is possible for this ordering to persist for longer, but that depends on the realisation of the two processes and it may break down (as is also shown in the picture)

We now have the claim at the initial time $T_0$ and we proceed by induction on the departure events and keep track of both queue lengths $L_t, \widetilde{L}_t$. Let $t_1 < t_2 < \ldots$ be departure events, $t_1 > T_0$. Assume that in $[T_0, t_1]$ the original arrival process had $k_1 \geq 0$ arrivals. That would make the queue length $L_{t_1} = (L_{T_0} + k_1 - 1) \vee 0$. The sped-up process $\widetilde{N}_a$ rang $k_1$ times (with the first ring now at $T_0$) and because of the shift by $\eta$, it may have rang even more times. Therefore, up to time $t_1$, $L_{t_1} \leq \widetilde{L}_{t_1}$ since only one departure occurred.

Now assume that up to time $t_n$ the claim remains true and $L_{t_n} \leq \widetilde{L}_{t_n}$ while no unused service time for $L_{t_n}$ occurred. We see what happens at time $t_{n+1}$.

We have the following inequalities

$$k_{n+1} = N_a(t_{n+1}) - N_a(T_0-) \leq \widetilde{N}_a(t_{n+1}) - \widetilde{N}_a(T_0-) = \tilde{k}_{n+1}.$$

Assume that $t_{n+1}$ is not an unused service time for $L_t$, otherwise we have nothing to show. This also implies that no unused service times occurred for $\widetilde{L}_t$ since it was

always no less than $L_t$. Then

$$\widetilde{L}_{t_{n+1}} = L_{T_0-} + 1 + \tilde{k}_{n+1} - n - 1 = L_{T_0-} + \tilde{k}_{n+1} - n \geq L_{T_0-} + k_{n+1} - n = L_{t_{n+1}},$$

as required. □

**Lemma 6.4** *Assume $\alpha_1 = \alpha_2 = \alpha$ and $\mu = \lambda$. Let $\{X^{(i)}_{\alpha,\lambda}\}_{i\geq 1}$, $\{Y^{(i)}_{\alpha,\lambda}\}_{i\geq 1}$ be the inter-event times of the arrival and departure process respectively that are i.i.d. Mittag-Leffler distributed. Consider the random walk*

$$S_n = \sum_{i=1}^{n} (X^{(i)}_{\alpha,\lambda} - Y^{(i)}_{\alpha,\lambda}), \quad S_0 = 0$$

*Then*

1. *$S_n$ will change sign infinitely often, $\overline{\lim_{n\to\infty}} S_n = +\infty$, $\underline{\lim_{n\to\infty}} S_n = -\infty$.*
2. *Assume that $S_n < 0$ and $S_{n+1} > 0$. Then there exists a time $t < t_{n+1}$ for which the two processes rang exactly n times and the next event in the process is a departure attempt.*

***Proof of Lemma 6.4*** Define

$$S^X_n = \sum_{i=1}^{n} X^{(i)}_{\alpha,\lambda}, \quad S^Y_n = \sum_{i=1}^{n} Y^{(i)}_{\alpha,\lambda}.$$

Each of these represent the time of the $n$-th arrival and the time of the $n$-th departure attempt respectively.

1. Since $X^{(i)}_{\alpha,\lambda} - Y^{(i)}_{\alpha,\lambda}$ is a symmetric random variable, the random walk is an oscillating random walk. The claim then follows from Feller (See Section 12.2 in [14]).
2. Define
$$t_{n+1} = S^Y_{n+1} = \sum_{i=1}^{n+1} Y^{(i)}_{\alpha,\lambda} \tag{6.10}$$
At this time the departure process has its $n+1$-th event. Since $S_n < 0$ we have that the $n$-th arrival occurring at time $S^X_n$ happened before (or with) the $n$-th (and therefore before the $n+1$) departure attempt. The 'with' here is in general a measure 0 event, except at the beginning when we have renewal point for both processes. Since $S_{n+1} > 0$ we also have that $n+1$ arrival happens after $t_{n+1}$. So by time $t_{n+1}$, both processes jumped $n$ times, and at time $t_{n+1}$ we have the $n+1$ departure attempt.

□

***Proof of Theorem 3.5*** We first treat the case $\alpha_1 = \alpha_2 = \alpha$ and $\mu = \lambda$. To begin with we show that we can find a time $T_1$ for which $L_{T_1} = 0$ with probability 1.

Assume the queue begins at $L_0 = 1$, with both arrival and departure renewals beginning at 0 and let the position of the random walks

$$S_n^X = \sum_{i=1}^{n} X_{\alpha,\lambda}^{(i)}, \quad S_n^Y = \sum_{i=1}^{n} Y_{\alpha,\lambda}^{(i)}.$$

denote the time of the $n$-th arrival and departure respectively. First consider the first sign change of their difference from negative to positive which will occur infinitely often by Lemma 6.4. This occurs as we said above when both processes rang the same amount of times. At time $t_{n+1}$ (from (6.10)) we have the $n+1$ departure event. At that point, we either have a queue of length 1 and this will be a service time (since both processes rang $n$ times) making $L_{t_{n+1}} = 0$ or, if the queue has length $L_{t_{n+1}-} \geq 2$ we have a departure. But if the queue had length more than 1, then we must have had an unused departure event before, say at time $t_k$ for some $k \leq n$. This can only happen if the queue was already empty at time $t_k$. In either case, we have found a departure time $T_1 > 0$ at which the queue was empty. Departure time $T_1$ exists with probability 1.

Define

$$T_1 = \inf\left\{t > 0 : N_{\alpha,\lambda}^Y(t) \text{ rings }, L_{t-}^{\alpha,\lambda} = 0\right\}.$$

At time $T_1$ (an event that clears the queue) we apply Lemma 6.3. Let $\eta > 0$ be such that $T_1 + \eta$ is the time of the next arrival event, and speed up the arrival process by $\eta$. Per the construction in Lemma 6.3, the new queue length $\tilde{L}$ starts at 1 and we have renewal events at the beginning making the new length equal to 1. But then, either $\tilde{L}_t$ will hit 0 with probability 1 if the original process does not have an unused service and therefore the original process hits 0 by the inequality in Lemma 6.3, or the original process will have an unused service which also means it hit 0. Therefore, in either case, the original process will hit 0 at a time $T_2 > T_1$. Iterate this argument ad infinitum to see that the queue will be empty infinitely often. This proves the case $\mu = \lambda$.

When $\mu > \lambda$ we can write that the departure events occur faster since

$$S_n^{Y,\mu} = \sum_{i=1}^{n} Y_{\alpha,\mu}^{(i)} \overset{(d)}{=} \frac{\lambda}{\mu} \sum_{i=1}^{n} Y_{\alpha,\lambda}^{(i)} < S_n^{Y,\lambda}. \tag{6.11}$$

Therefore, the queue empties faster (in distribution). In terms of the difference random walk, it tends to be positive and tending to infinity suggesting a lot of unused services. As such the queue hits 0 infinitely often. The coupling argument utilises (6.11) and goes by arguing that by the time $S_n^{Y,\lambda}$ has an unused time, then already at a prior time $S_n^{Y,\mu}$ had an unused service time, and then Lemma 6.3 is utilised at the time the queue hit 0. The details are left to the interested reader.

Finally, to prove the second part of the theorem for $\lambda \geq \mu$. We can repeat the arguments in the proof of Theorem 3.4 but now the scalings $\lambda$ and $\mu$ play a role. We have $N_\lambda^{(\alpha)}(t)$ denote the arrival renewal process and $N_\mu^{(\alpha)}(t)$ denote the departure

process with Mittag-Leffler $\alpha$ parameter and scalings $\lambda$, $\mu$ respectively. For any $s < 0$

$$\begin{aligned}\mathbb{P}\left\{N_\lambda^{(\alpha)}(t) - N_\mu^{(\alpha)}(t) < t^{\alpha/2}\right\} &= \mathbb{P}\left\{s(N_\lambda^{(\alpha)}(t) - N_\mu^{(\alpha)}(t)) > st^{\alpha/2}\right\} \\ &\leq e^{-st^{\alpha/2}}\mathbb{E}(e^{sN_\mu^{(\alpha)}(t)})\mathbb{E}(e^{-sN_\lambda^{(\alpha)}(t)}) \text{ by a Chernoff bound,} \\ &= e^{st^{\alpha/2}}E_{\alpha,1}((e^s-1)\lambda^\alpha t^\alpha)E_{\alpha,1}((e^{-s}-1)\mu^\alpha t^\alpha).\end{aligned}$$

At this point we make a particular choice for $s$ since we can choose any negative value for it. Select $s = -c^\alpha t^{-\alpha}$. Then the bound above becomes

$$\begin{aligned}&\mathbb{P}\left\{N_\lambda^{(\alpha)}(s) - N_\mu^{(\alpha)}(s) < t^{\alpha/2}\right\} \\ &\qquad \leq e^{c^\alpha t^{-\alpha/2}}E_{\alpha,1}(-(c\lambda)^\alpha + o(t^{-\alpha}))E_{\alpha,1}((c\mu)^\alpha + o(t^{-\alpha})).\end{aligned}$$

Then we can find $T_0(\mu,\lambda), c_0(\mu,\lambda) > 0$ such that for $t > T_0$ and $0 < c < c_0$ so that the Taylor expansion at $c = 0$ of the upper bound above becomes

$$\begin{aligned}&\left(1 + \frac{c^\alpha}{t^{\alpha/2}} + \varepsilon_1\right)\left(1 - \frac{(c\lambda)^\alpha}{\Gamma(1+\alpha)} + \varepsilon_2\right)\left(1 + \frac{(c\mu)^\alpha}{\Gamma(1+\alpha)} + \varepsilon_3\right) \\ &\qquad = 1 + c^\alpha\frac{\mu^\alpha - \lambda^\alpha}{\Gamma(1+\alpha)} + o(c^\alpha) \vee o(t^{-\alpha/2}) < 1.\end{aligned}$$

The last inequality follows only when $\mu > \lambda$. Errors $\varepsilon_1, \varepsilon_2$ and $\varepsilon_3$ are all of order $o(c^\alpha) \vee o(t^{-\alpha/2})$. In order to conclude, use the fact that $L^{\alpha_1,\alpha_2}(t) \geq N^{\alpha_1}(t) - N^{\alpha_2}(t)$. The rest of the proof proceeds as in Theorem 3.4, where we obtain that

$$\mathbb{P}\{\overline{\lim_{t\to\infty}} L_{\lambda,\mu}^{\alpha,\alpha} = +\infty\} \geq 1 - c_0^\alpha\frac{\lambda^\alpha - \mu^\alpha}{\Gamma(1+\alpha)} > 0.$$

Similarly, if $\mu = \lambda$ we obtain that

$$\mathbb{P}\{\overline{\lim} L_{\lambda,\lambda}^{\alpha,\alpha} = +\infty\} \geq 1 - c^{2\alpha}\frac{\lambda^{2\alpha}}{\Gamma(1+\alpha)^2} > 0.$$

□

**Acknowledgements** Jacob Butt, Nicos Georgiou and Enrico Scalas would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme Fractional Differential Equations (FDE2) where work on this paper was undertaken. This work was supported by EPSRC grant no EP/R014604/1. Nicos Georgiou and Enrico Scalas were also partially funded by the Dr. Perry James (Jim) Browne Research Center at the Department of Mathematics, University of Sussex. Finally, we want to thank Dr. Vladislav Vysotsky for stimulating discussions and literature references.

**Data access statement** Code and data used in this paper are freely available from https://github.com/Jacob-Butt/Fractional-queues.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## Appendix A Inductive proof of equation (2.9)

*Proof of equation* (2.9) Let $N_a(t)$ and $N_d(t)$ denote the two counting processes representing the arrivals and departures attempts respectively and assume that $N_a(0) = N_d(0) = 0$.

We will prove the result by induction over all jump events in the time interval $[0, t]$, of which by assumption there are at most finitely many. Order all the jump times in the interval $[0, t]$ and label them using $0 < \tau_1 < \tau_2 < \ldots < \tau_M \leq t$.

The queue length remains at 0 until time $\tau_1-$, since no process jumped. At time $\tau_1$ we have a jump. If the jump is an arrival, then $N_a(\tau_1) = 1$ while still $N_d(\tau_1) = 0$. The difference $N_a(\tau_1) - N_d(\tau_1) = 1$ while the infimum of the difference is still attained at $t = 0$ and it equals 0. So the right-hand side equals 1, which equals the queue length since the queue experienced an arrival. In the case where the jump was for departures, we have that

$$N_a(\tau_1) - N_d(\tau_1) = -1 = \inf_{0 \leq s \leq \tau_1} \{N_a(s) - N_d(s)\},$$

where the infimum is attained precisely at $\tau_1$. If $\tau_1 = \tau_2$ and both processes jump, arrivals happen before departures, so $L_{\tau_1} = 0$ and $N_a(\tau_1) - N_d(\tau_1) = 0$.

This forms the base case of the induction. Now assume that the equation holds up to event $\tau_k$. Up until $t < \tau_{k+1}$ no process jumps, so the equality is maintained up to $\tau_{k+1}-$. At $\tau_{k+1}$ at least one process jumps.

If it is $N_a$, the infimum does not change since its argument increased at this time. So both $L_{\tau_{k+1}}$ and $N_a(\tau_{k+1}) - N_d(\tau_{k+1})$ increased by 1 and equality is preserved.

If the process that jumped is $N_d$, there are two cases to consider. If $L_{\tau_{k+1}-} = 0$ then this does not change with a jump of $N_d$. For the right-hand side

$$N_a(\tau_{k+1}-) - N_d(\tau_{k+1}-) = \inf_{0 \leq s < \tau_{k+1}} \{N_a(s) - N_d(s)\}.$$

In this situation, a jump of $N_d$ actually reduces the infimum by 1 since

$$\inf_{0 \leq s \leq \tau_{k+1}} \{N_a(s) - N_d(s)\} = \inf_{0 \leq s < \tau_{k+1}} \{N_a(s) - N_d(s)\} \wedge (N_a(\tau_{k+1}) - N_d(\tau_{k+1}))$$

$$= N_a(\tau_{k+1}-) - N_d(\tau_{k+1}-) - 1,$$

which also shows the right hand side is 0.

The other option is that $L_{\tau_{k+1}-} > 0$. Then

$$N_a(\tau_{k+1}-) - N_d(\tau_{k+1}-) - 1 \geq \inf_{0\leq s<\tau_{k+1}} \{N_a(s) - N_d(s)\}$$

which leads to

$$\begin{aligned}\inf_{0\leq s\leq\tau_{k+1}} \{N_a(s) - N_d(s)\} &= \inf_{0\leq s<\tau_{k+1}} \{N_a(s) - N_d(s)\} \wedge (N_a(\tau_{k+1}) - N_d(\tau_{k+1})) \\ &= \inf_{0\leq s<\tau_{k+1}} \{N_a(s) - N_d(s)\} \wedge (N_a(\tau_{k+1}-) - N_d(\tau_{k+1}-) - 1) \\ &= \inf_{0\leq s<\tau_{k+1}} \{N_a(s) - N_d(s)\}.\end{aligned}$$

In other words, in this case the infimum remains unchanged, while $N_a(\tau_{k+1}) - N_d(\tau_{k+1}) = N_a(\tau_{k+1}-) - N_d(\tau_{k+1}-) - 1$ and the right hand-side is reduced by 1, as is the queue length.

The case of the simultaneous jump is treated similarly. □

***Remark A.1*** While in the article we are dealing with continuous processes with the probability of both jumping at the same time equal to 0, it could be that other models might exhibit simultaneous jumps. In that case, the formula still remains true if we adopt the convention that arrivals happen before departures (instantaneously). It wouldn't necessarily hold if departures were happening before arrivals. The problem would come from events where the queue length process is already at 0:

$$N_a(\tau_{k+1}-) - N_d(\tau_{k+1}-) = \inf_{0\leq s<\tau_{k+1}} \{N_a(s) - N_d(s)\}.$$

Then if at $\tau_{k+1}$ we have a simultaneous jump and thus

$$N_a(\tau_{k+1}-) - N_d(\tau_{k+1}-) = N_a(\tau_{k+1}) - N_d(\tau_{k+1}),$$

while

$$\begin{aligned}\inf_{0\leq s\leq\tau_{k+1}} \{N_a(s) - N_d(s)\} &= \inf_{0\leq s<\tau_{k+1}} \{N_a(s) - N_d(s)\} \wedge (N_a(\tau_{k+1}) - N_d(\tau_{k+1})) \\ &= N_a(\tau_{k+1}-) - N_d(\tau_{k+1}-) = N_a(\tau_{k+1}) - N_d(\tau_{k+1}).\end{aligned}$$

Therefore if the formula remained true, the queue length would have been zero. However, if arrivals are happening after departures we have that queue length would equal 1. □

## Appendix B Proof of Theorem 3.1

***Proof of Theorem 3.1*** Let $M_{\alpha,\lambda}(z,t) = \mathbb{E}(e^{zL^{(1)}_{\alpha,\lambda}(t)})$ denote the moment generating function of the queue length at time $t$. For simplicity we assume $\lambda = 1$; the only effect of a different $\lambda$ below would be that the right hand side is multiplied with an extra $\lambda^\alpha$.

$$\begin{aligned}
\frac{d^\alpha}{dt^\alpha} M_{\alpha,1}(z,t) &= \frac{d^\alpha}{dt^\alpha}\mathbb{E}(e^{zL^{(1)}_{\alpha,1}(t)}) = \sum_{n=0}^{\infty} e^{zn}\frac{d^\alpha}{dt^\alpha} p_{0,n}(t) \\
&= \frac{d^\alpha}{dt^\alpha} p_{0,0}(t) + \sum_{n=1}^{\infty} e^{zn}(-p_{0,n}(t) + pp_{0,n-1}(t) + (1-p)p_{0,n+1}(t)) \\
&= \frac{d^\alpha}{dt^\alpha} p_{0,0}(t) - (M_{\alpha,1}(z,t) - p_{0,0}(t)) \\
&\quad + pe^z M_{\alpha,1}(z,t) + (1-p)e^{-z}(M_{\alpha,1}(z,t) - p_{0,0}(t) - e^z p_{0,1}(t))) \\
&= -pp_{0,0}(t) + (1-p)p_{0,1}(t) + p_{0,0}(t) - (1-p)e^{-z}p_{0,0}(t) \\
&\quad - (1-p)p_{0,1}(t) + \left(pe^z - 1 + (1-p)e^{-z}\right)M_{\alpha,1}(z,t) \\
&= (1-p)(1-e^{-z})p_{0,0}(t) + \left(pe^z - 1 + (1-p)e^{-z}\right)M_{\alpha,1}(z,t).
\end{aligned}$$

Multiply through by $e^z$ to obtain

$$e^z\frac{d^\alpha}{dt^\alpha} M_{\alpha,1}(z,t) = (1-p)(e^z-1)p_{0,0}(t) + \left(pe^{2z} - e^z + (1-p)\right)M_{\alpha,1}(z,t). \quad \text{(B1)}$$

We compute the Laplace transform (for the $t$ variable) and we see

$$e^z\left(s^\alpha \widetilde{M}_{\alpha,1}(z,s) - s^{\alpha-1}\right) = (1-p)(e^z-1)\widetilde{p}_{0,0}(s) + \left(pe^{2z} - e^z + (1-p)\right)\widetilde{M}_{\alpha,1}(z,s). \quad \text{(B2)}$$

Solving for $\widetilde{M}_{\alpha,1}(z,s)$ we obtain

$$\widetilde{M}_{\alpha,1}(z,s) = \frac{e^z s^{\alpha-1} + (1-p)(e^z-1)\widetilde{p}_{0,0}(s)}{e^z s^\alpha - \left(pe^{2z} - e^z + (1-p)\right)}. \quad \text{(B3)}$$

Set $e^z = u$. The roots of the denominator $r_1(s), r_2(s)$, with $r_1(s) > r_2(s)$ for the $u$ variable are given by

$$r_{1,2}(s) = \frac{1 + s^\alpha \pm \sqrt{(1-2p)^2 + 2s^\alpha + s^{2\alpha}}}{2p} = \frac{1 + s^\alpha \pm \sqrt{(1+s^\alpha)^2 - 4p(1-p)}}{2p}. \quad \text{(B4)}$$

The following relations hold, and we will be using them without particular mention

$$r_1 + r_2 = \frac{1+s^\alpha}{p}, \quad r_1 r_2 = \frac{1-p}{p}, \quad (1-r_1)(1-r_2) = \frac{-s^a}{p}. \quad \text{(B5)}$$

The first expression (B4) shows that the roots are real numbers, while the second shows that $r_2(s) > 0$. Moreover it is straightforward to check that $r_2(s) < 1$ while the third equality in (B5) gives $r_1(s) > 1$.

The expression for $\widetilde{p}_{0,0}$ can be calculated using Rouche's Theorem for the probability generating function [6]. Briefly, equation (B3) can be converted to an equivalent equation for the p.g.f again using the substitution $e^z = u$; then Rouche's Theorem can be applied to show that the denominator contains a root inside the unit disc, which must match with a root of the numerator. This was done in [8], giving

$$r_2(s)s^{a-1} + (1-p)(r_2(s)-1)\widetilde{p}_{0,0}(s) = 0 \Longleftrightarrow \widetilde{p}_{0,0}(s) = \frac{r_2(s)s^{\alpha-1}}{(1-p)(1-r_2(s))}.$$

In particular, we obtain the Laplace transform

$$\widetilde{p}_{0,0}(s) = \frac{s^{a-1}}{p(1-r_2(s))r_1(s)} = \frac{r_2(s)s^{a-1}}{(1-p)(1-r_2(s))} = \frac{1}{s} - \frac{p}{1-p}\frac{r_2(s)}{s}. \tag{B6}$$

***Remark B.1*** In the case $\alpha = 1$ the inverse Laplace transform can be given in terms of elementary functions and we obtain the formulas for the M/M/1 queue given in [6] The paper gives results for the p.g.f. but one can check for the m.g.f. by making the substitution $u = e^z$.

Substitute $\widetilde{p}_{0,0}(s)$ in (B3) to see

$$\widetilde{M}_{\alpha,1}(z,s) = \frac{-s^{\alpha-1}}{p(e^z - r_1(s))(1-r_2(s))},$$

which concludes the proof of the theorem. □

## Appendix C Derivation of the fractional evolution equations

Let $q_{i,j}$ denote the transition probabilities of moving from $i$ customers to $j$ customers at an event time. Then we can say

$$q_{k,k} = \beta, \tag{C1}$$

$$q_{k,k+1} = \begin{cases} 1-\beta & \text{if } k = 0 \\ (1-\beta)\cdot p & \text{o/w,} \end{cases} \tag{C2}$$

$$q_{k,k-1} = \begin{cases} 0 & \text{if } k = 0 \\ (1-\beta)\cdot(1-p) & \text{o/w,} \end{cases} \tag{C3}$$

with all other probabilities equal to zero.

Now define $p_i(t) = \mathbb{P}\{X(t) = i\}$ to be the probability of finding $i$ customers in the queue at time $t$. We then have that $p_i(t)$ satisfies the forward equations

$$p_i(t) = \bar{F}_T^{(\alpha)}(t)\delta_0 + \sum_{k\in\{i-1,i,i+1\}} q_{k,i} \int_0^t p_k(u) f_T^{(\alpha)}(t-u)du. \tag{C4}$$

We now want to Laplace transform equation to get

$$\begin{aligned}\tilde{p}_i(s) = \tilde{\bar{F}}_T^{(\alpha)}(s)\delta_0 + \beta \tilde{f}_T^{(\alpha)}(s)\tilde{p}_i(s) \\ + ((1-\beta)p)\ \tilde{f}_T^{(\alpha)}(s)\tilde{p}_{i-1}(s) + ((1-\beta)(1-p))\ \tilde{f}_T^{(\alpha)}(s)\tilde{p}_{i+1}(s)\end{aligned}$$

Now let us multiply both sides of this equation through by $s$ and then subtract $\delta_0$ from both sides, which yields

$$\mathcal{L}\left(\frac{dp_i(t)}{dt}; s\right) = s\tilde{\bar{F}}_T^{(\alpha)}(s)\delta_0 - \delta_0 + s\beta\, \tilde{f}_T^{(\alpha)}(s)\tilde{p}_i(s) + s\ ((1-\beta)p)\ \tilde{f}_T^{(\alpha)}(s)\tilde{p}_{i-1}(s) \tag{C5}$$

$$\begin{aligned}&+ s\ ((1-\beta)(1-p))\ \tilde{f}_T^{(\alpha)}(s)\tilde{p}_{i+1}(s) \\ &= \frac{s^\alpha}{\lambda^\alpha + s^\alpha}\delta_0 - \delta_0 + \frac{s\lambda^\alpha\beta}{\lambda^\alpha + s^\alpha}\tilde{p}_i(s) + \frac{s\lambda^\alpha}{\lambda^\alpha + s^\alpha}\ ((1-\beta)p)\ \tilde{p}_{i-1}(s) \\ &+ \frac{s\lambda^\alpha}{\lambda^\alpha + s^\alpha}\ ((1-\beta)(1-p))\ \tilde{p}_{i+1}(s).\end{aligned}$$

Rearranging this gives

$$\begin{aligned}\frac{\lambda^\alpha + s^\alpha}{\lambda^\alpha s}\mathcal{L}\left(\frac{dp_i(t)}{dt}; s\right) = -\frac{1}{s}\delta_0 + \beta\tilde{p}_i(s) + ((1-\beta)p)\ \tilde{p}_{i-1}(s) \\ + (1-\beta)(1-p)\tilde{p}_{i+1}(s).\end{aligned} \tag{C6}$$

The inverse Laplace transform of $\frac{y+s^\alpha}{s}$ is given by

$$\mathcal{L}^{-1}\left(\frac{y+s^\alpha}{s}; t\right) = \frac{t^{-\alpha}}{\Gamma(1-\alpha)} + y, \tag{C7}$$

and by substituting this in with $y = \lambda^\alpha$ we obtain

$$\frac{d^\alpha p_i(t)}{dt^\alpha} = \lambda^\alpha\ (\beta - 1)\ p_i(t) + \lambda^\alpha(1-\beta)pp_{i-1}(t) + \lambda^\alpha(1-\beta)(1-p)p_{i+1}(t). \tag{C8}$$

Similarly, we can find the boundary conditions

$$\frac{d^\alpha p_0(t)}{dt^\alpha} = -\lambda^\alpha(1-\beta)pp_0(t) + \lambda^\alpha(1-\beta)(1-p)p_1(t). \tag{C9}$$

## References


1. Abate, J., Whitt, W.: Transient behavior of the M/M/1 queue via Laplace transforms. Adv. Appl. Probability **20**, 145–178 (1988). https://doi.org/10.2307/1427274
2. Ascione, G., Leonenko, N., Pirozzi, E.: Fractional queues with catastrophes and their transient behaviour. Mathematics **6** (2018). https://doi.org/10.3390/math6090159
3. Ascione, G., Leonenko, N., Pirozzi, E.: Fractional Erlang queues. Stochastic Processes Appl. **130**, 3249–3276 (2020). https://doi.org/10.1016/j.spa.2019.09.012
4. Ascione, G., Leonenko, N., Pirozzi, E.: Fractional immigration-death processes. Journal of Mathematical Analysis and Applications **495**(2), 124768 (2021). https://doi.org/10.1016/j.jmaa.2020.124768. https://www.sciencedirect.com/science/article/pii/S0022247X20309318
5. Ascione, G., Leonenko, N., Pirozzi, E.: Non-local solvable birth-death processes. J. Theor. Probability **35**, 1–40 (2021)
6. Bailey, N.T.: A continuous time treatment of a simple queue using generating functions. J. Royal Stat. Soc.: Series B (Methodological) **16**(2), 288–291 (1954)
7. Bose, S.K.: An Introduction to Queueing Systems. Springer Science & Business Media, Berlin (2013)
8. Cahoy, D.O., Polito, F., Phoha, V.V.: Transient behavior of fractional queues and related processes. Methodology and Computing in Applied Probability **17**, 739–759 (2015). https://doi.org/10.1007/s11009-013-9391-2. arXiv:1303.6695
9. Carpinteri, A., Mainardi, F.: Fractals and Fractional Calculus in Continuum Mechanics. Springer, Berlin (1997)
10. Chambers, J.M., Mallows, C.L., Stuck, B.W.: A method for simulating stable random variables. J. Am. Stat. Assoc. **71**, 340–344 (1976). https://doi.org/10.1080/01621459.1976.10480344
11. Cont, R., Stoikov, S., Talreja, R.: A stochastic model for order book dynamics. Op. Res. **58**, 549–563 (2010). https://doi.org/10.1287/opre.1090.0780
12. Curinao, J.L.: Asymptotic behavior and quasi-limiting distributions on time-fractional birth and death processes. J. Appl. Probability **59**(4), 1199–1227 (2022)
13. Daftardar-Gejji, V.: Fractional Calculus. Alpha Science International Limited (2013)
14. Feller, W.: An Introduction to Probability Theory and its Applications, vol. 2. Wiley, Hoboken (2008)
15. Fomundam, S.F., Herrmann, J.W.: A survey of queuing theory applications in healthcare. https://drum.lib.umd.edu/handle/1903/7222 (2007)
16. Foss, S., Korshunov, D.: Heavy tails in multi-server queues. Queing Syst. **52**, 31–48 (2006)
17. Georgiou, N., Kiss, I.Z., Scalas, E.: Solvable non-Markovian dynamic network. Phys. Rev. E **92**(4), 042801 (2015)
18. Giambene, G.: Queuing Theory and Telecommunications, vol. 585. Springer, Berlin (2014)
19. Gikhman, I.I., Skorokhod, A.V.: The Theory of Stochastic Processes II. Springer Science & Business Media, Berlin (2004)
20. Gorenflo, R., Kilbas, A.A., Mainardi, F., Rogosin, S.V.: Mittag-Leffler Functions, Related Topics and Applications. Springer (2020). https://doi.org/10.1007/978-3-662-43930-2. http://link.springer.com/10.1007/978-3-662-43930-2
21. Kendall, D.G.: Stochastic processes occurring in the theory of queues and their analysis by the method of the embedded Markov chain. The Annals of Mathematical Statistics pp. 338–354 (1953)
22. Kerss, A., Leonenko, N., Sikorskii, A.: Fractional Skellam processes with applications to finance. Fract. Calc. Appl. Anal. **17**, 532–551 (2014). https://doi.org/10.2478/s13540-014-0184-2
23. Larson, R.C., Odoni, A.R.: Urban Operations Research. Prentice-Hall, Hoboken (1981)
24. Leonenko, N., Scalas, E., Trinh, M.: Limit theorems for the fractional nonhomogeneous Poisson process. J. Appl. Probability **56**, 246–264 (2019). https://doi.org/10.1017/jpr.2019.16
25. Mainardi, F.: On some properties of the Mittag–Leffler function $E^{\alpha}(-t^{\alpha})$, completely monotone for $t>0$ with $0<\alpha<1$. Discret. Contin. Dyn. Syst.- Series B **19**, 2267–2278 (2014). https://doi.org/10.3934/dcdsb.2014.19.2267
26. Mainardi, F.: Why the Mittag–Leffler function can be considered the queen function of the fractional calculus? Entropy **22**, 1–29 (2020). https://doi.org/10.3390/e22121359
27. Mainardi, F., Gorenflo, R., Scalas, E.: A fractional generalization of the Poisson processes. Vietnam J. Math. **32**, 53–64 (2004)
28. Medhi, J.: Stochastic Models in Queueing Theory. Elsevier, Amsterdam (2002)
29. Meerschaert, M., Nane, E., Vellaisamy, P.: The fractional Poisson process and the inverse stable subordinator. Electron. J. Probability **16**, 1600–1620 (2011). https://doi.org/10.1214/EJP.v16-920

30. Meerschaert, M., Straka, P.: Inverse stable subordinators. Math. Modelling Nat. Phenom. **8**, 1–16 (2013). https://doi.org/10.1051/mmnp/20138201
31. Van Mieghem, P.: The Mittag-Leffler function. arXiv:2005.13330 (2020)
32. Møller, J.: On the rate of convergence of spatial birth-and-death processes. Ann. Inst. Stat. Math. **41**(3), 565–581 (1989)
33. Norris, J.R.: Markov Chains. 2. Cambridge University Press, Cambridge (1998)
34. Orsingher, E., Polito, F.: On a fractional linear birth–death process. Bernoulli **17**, 114–137 (2011). https://doi.org/10.3150/10-BEJ263
35. Raberto, M., Scalas, E., Mainardi, F.: Waiting-times and returns in high-frequency financial data: an empirical study. Physica A **314**, 749–755 (2002). www.elsevier.com/locate/physa
36. Radivojević, T., Anselmi, J., Scalas, E.: Ergodic transition in a simple model of the continuous double auction. PLoS ONE **9** (2014). https://doi.org/10.1371/journal.pone.0088095
37. Sabatelli, L., Keating, S., Dudley, J., Richmond, P.: Waiting time distributions in financial markets. Eur. Phys. J. B **27**, 273–275 (2002). https://doi.org/10.1140/epjb/e20020151
38. Simon, T.: Comparing Fréchet and positive stable laws. Electron. J. Probability **19**, 1–25 (2014). https://doi.org/10.1214/EJP.v19-3058
39. Sun, H.G., Zhang, Y., Baleanu, D., Chen, W., Chen, Y.Q.: A new collection of real world applications of fractional calculus in science and engineering. Commun. Nonlinear Sci. Numer. Simul. **64**, 213–231 (2018). https://doi.org/10.1016/j.cnsns.2018.04.019
40. Tarasov, V.E.: On history of mathematical economics: Application of fractional calculus. Mathematics **7** (2019). https://doi.org/10.3390/math7060509
41. Whitt, W.: Some useful functions for functional limit theorems. Math. Op. Res. **5**, 67–85 (1980)
42. Whitt, W.: The impact of a heavy-tailed service-time distributions upon the m/gi/s waiting-distribution. Queing Syst. **36**, 71–87 (2000)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.