Ann Oper Res (2016) 247:37–63
DOI 10.1007/s10479-015-1904-7

Retrial queueing models in discrete time: a short survey of some late arrival models

Rein Nobel[1]

Published online: 29 May 2015

Abstract This paper presents an overview of one-server queueing models with retrials in discrete-time. In all these models the number of primary customers arriving in a time slot follows a general probability distribution and the different numbers of primary arrivals in consecutive time slots are mutually independent. Each customer requires from the server a generally distributed number of slots for his service, and the service times of the different customers are independent. Only models with delayed access are considered, and the so-called late arrival setup is chosen. For all the models the steady-state behavior is studied through the generating function of the number of customers in the orbit. From the generating function several performance measures are deduced, like the average orbit size and the mean busy period.

Keywords Retrial queues · Discrete time · Late arrivals · Generating functions

1 Introduction

Queueing Theory (QT) is a vast well-established research area with more than a century of history, and *retrial queueing models* form only a tiny subset of the huge variety of models which have been studied during this century [see e.g. Jain et al. (2007) for a nice recent monograph]. Also most models discussed in QT are continuous-time models, and relatively few researchers have studied *discrete-time models*. We refer to Bruneel and Kim (1993), Miyazawa and Takagi (1994), and Takagi (1993) for overviews on discrete-time queueing models. So, it is safe to guess that papers dedicated to *discrete-time retrial models* are almost absent in the literature. Nevertheless, the last two decades people have started to study retrial models in discrete time, although sometimes the term ‘retrial’ in the title of these papers is used a little bit sloppy, i.e. strictly speaking many of these models should not be coined

✉ Rein Nobel
r.d.nobel@vu.nl

[1] Department of Econometrics, Vrije University, Amsterdam, The Netherlands

‘retrial models’. Therefore, let us start with a clear-cut description of what is meant with the term *retrial* in this paper (in the sequel we assume that the terms ‘orbit’, ‘primary arrival’, ‘secondary arrival’, ‘retrial time’, et cetera, are known to the reader): models are only called retrial models when customers who upon arrival find all the servers busy are sent into the orbit, *and subsequently they decide to approach the server(s) anew some random time later, independently of the other customers in the orbit and the state of the server(s)*. Essential is here that the *customer in the orbit takes the initiative to approach the server*. So, in this paper we exclude all models with so-called ‘generally distributed retrial times’ which turn out to be models in which the server starts searching for a customer in the orbit. It is very confusing to call these latter models also retrial models because the main difficulty in the study of *retrial* models, i.e. the complicated mixed arrival pattern of primary and secondary customers, is absent in models with servers who upon a service completion start to search for a customer. In these models the choice of the next customer to be served lies in the hands of the server and is not determined by the decision of a customer to approach the server from the orbit [see Atencia and Moreno (2004a), Wang and Zhao (2007) and Aboul-Hassan et al. (2008) for three examples in discrete time]. We think that the term retrial queue for these latter models is a misnomer, and we propose to rename these models as *queueing models with an actively searching server* (everybody who has a better name is welcome of course!) to avoid further misunderstanding. Of course, it is possible to consider retrial models with generally distributed retrial times, but these are usually intractable, because all the (dependent) residual retrial times of the customers in the orbit have to be incorporated in the state description of the system, and this mere fact hampers a feasible mathematical analysis. Hence, only approximative approaches can shed some light on these models [see e.g. Yang et al. (1994)].

An important modeling decision to make in discrete-time queues is to establish the precedence relations between arrivals, departures and start of service. In the literature we see two different systems. The first, called *early arrival systems* (EAS), give precedence to departures over arrivals, i.e. when in the same time slot both an arrival and a departure take place, the arriving customer(s) see the server idle. The other choice is to give precedence to arrivals over departures leading to so-called *late arrival systems* (LAS). In that case arriving customers see the server still busy when in the arrival slot a departure occurs. As a consequence, in a one-server retrial model customers who arrive in a slot in which also a departure occurs are all (re)sent into the orbit when the late arrival setup is chosen! For a nice detailed explanation of the difference between EAS and LAS we refer to Chaudhry (1993) or Chaudhry and Gupta (1997).

Also a decision must be made when to start a service. The possibilities are again twofold: (i) when upon arrival the customer(s) find a server idle, the service of (at least one of) the customer(s) starts immediately, i.e. the time slot of the arrival is also the first time slot of the service [we call this *immediate access* (IA)], or (ii) the service of (at least one of) the arrived customer(s) starts at the earliest only in the next time slot [we call this *delayed access* (DA)]. In the literature on discrete-time queues all variants (EAS-IA, EAS-DA, LAS-IA and LAS-DA) show up, but, curiously, almost all papers on discrete-time retrial queues choose the EAS setup [the first paper is Yang and Li (1995)]. In our opinion this is not the most natural choice, for two reasons. Firstly, although services and departures are best considered to occur only at time slot boundaries, arrivals can take place physically any time instant between the boundaries of the slot, so not necessarily at the slot boundaries. As a consequence, it is not clear whether the time slot of the arrival should be counted as the first slot of the service or not. In the DA setup this problem is absent. Secondly, but this is only relevant for retrial queues, in continuous-time retrial queues after a service completion there is always a time interval of positive length during which the server is idle, because there are no customers waiting in a queue and ready to be served: the server has to wait for the next arrival, coming either from the orbit or from the primary arrival stream. Using the EAS-IA setup in a discrete-time context (as most authors do) this characteristic gets lost, because the newly arrived customer will start his service immediately, i.e. in the time slot of his arrival, when a departure has taken place in the same time slot. Hence, to enforce more parallelism between the continuous-time retrial queues and their discrete-time counterparts we prefer the *late arrival setup with delayed access*, from now on LAS-DA.

A further distinction which can be made in the research on discrete-time queues concerns the methodology. The last decades many papers have been published using *Matrix Analytic Methods* (MAM) [see e.g. Alfa (2002) and Alfa (2006)]. This algorithmic approach has become quite popular because numerical results can be obtained for a great variety of models. In this paper we will only discuss results obtained by the *Generating Function Method* (GFM), following the tradition of e.g. Bruneel and Kim (1993) and Chaudhry (1993) or Chaudhry and Gupta (1997), and many others. The main reason for us to make this choice is a matter of taste: we prefer a formula, albeit a generating function, to an algorithm. And nowadays it is not a problem to find numerical results from generating functions using the discrete Fast Fourier Transform (FFT). Also, having available the generating function, for instance for the number of customers in the orbit, the mean and variance can be calculated explicitly, although as we will see not always in closed form. On the other hand we admit that in many cases the GFM will fail where MAM can still succeed. Nevertheless we believe that it is worthwhile to find the boundaries of tractability in using the GFM and to see the pros and cons of both methods.

As pointed out before, retrial queueing models received much less attention in the literature than the more traditional delay and loss models. It is probably fair to say that the unpopularity of the research on retrial models is partly due to their intractability, because from a practical point of view retrial models often describe a more realistic picture of many queueing situations than many of the other type of models, as loss and delay models. Notwithstanding the mathematical difficulties encountered in the study of retrial systems, the last three decades many retrial models have been studied and we refer to the monographs of Falin and Templeton (1997) and Artalejo and Gómez-Corral (2008) for a more or less complete overview of the main results so far. Only in the latter monograph some attention has been paid to discrete-time retrial queues (the Geo/G/1 and the Geo/Geo/c retrial queue). The authors choose the EAS setup for both models. The Geo/G/1 model is discussed using GFM [based on Yang and Li (1995)] and the Geo/Geo/c model is studied using MAM [based on Artalajo et al. (2008)]. From a didactic point of view it would be illustrative to compare an analysis of the Geo/PH/1 retrial queue based on the GFM with an alternative analysis using MAM, to illustrate the pros and cons of the two methodologies. The last twenty years several papers have been published on discrete-time retrial queueing models. We mention Yang and Li (1995), Choi and Kim (1997), Li and Yang (1998, 1999), Takahashi et al. (1999), Artalejo et al. (2005), Atencia and Moreno (2004a, b, 2006a, b, c). See further the references in the monograph Artalejo and Gómez-Corral (2008) for all the papers published until 2008. More recent papers are e.g. Aboul-Hassan et al. (2010), Atencia et al. (2010), Kim and Kim (2010), Artalejo and Li (2010), Amador and Moreno (2011), Wang and Zhang (2009) and Wang and Zhang (2009a), Wu et al. (2011) and (2013), to name just a few. Not surprisingly in all these papers the EAS setup is chosen and we will not discuss these papers any further in this survey. For a comparison between the EAS-setup and the LAS-setup we refer to Nobel and Moreno (2008) where the differences are analyzed in detail. As said before we prefer the LAS-setup, and therefore we will take the standard discrete-time one-server *delay model* of Bruneel and

Kim (1993) as our starting point and we will analyze several retrial variants of that model. So, to resume the setup very briefly [see Nobel and Moreno (2008) for more details], customers arrive in batches during a slot and the batch size follows a general probability distribution. Customers who upon arrival find the server busy are sent into the orbit. When upon arrival a customer finds the server idle, it can start its service only at the beginning of the next slot at the earliest. Of all the arriving customers during a slot with the server idle only one customer is selected (randomly) for service, while the others are (re)sent into the orbit. The batch sizes of primary arrivals in different slots are independent, and each customer requires a generally distributed number of slots for its service. During each slot every customer in the orbit tries to re-enter the system with a fixed retrial probability. This is our standard model, which will be discussed in more detail in Sect. 2. Two variants of this model will be discussed. In Sect. 3 we enrich the model with Bernoulli feedback and in Sect. 4 we add a second type of customers who have priority over the customers in the orbit. In Sect. 5 we allow the customers to leave the orbit before being served, and it turns out that for that model a general solution cannot be found. Only for special cases the generating function of the number of customers in the orbit can be calculated. In Sect. 6 we discuss a model with a so-called *tolerant* server, i.e. the server accepts all the arriving customers, who are served uninterruptedly as a mixed batch of primary and secondary arrivals. Finally, in Sect. 7 we draw some conclusions.

2 The standard LARS-DA retrial model

In this section we present the precise description of the discrete-time one-server retrial queueing model with the late arrival setup and delayed access [the LARS-DA model, see Nobel and Moreno (2008) for more details]. In each time slot customers arrive in batches. The batch sizes are mutually independent and follow a general probability distribution $\{a_i\}_{i=0}^{\infty}$ with probability generating function (p.g.f.)

$$\mathcal{A}(z) = \sum_{i=0}^{\infty} a_i z^i.$$

We call these arrivals primary arrivals. Each individual customer requires a generally distributed service time, measured as a number of time slots. The service times of the different jobs are also mutually independent and all follow the discrete probability distribution $\{b_j\}_{j=1}^{\infty}$ with p.g.f.

$$\mathcal{B}(w) = \sum_{j=1}^{\infty} b_j w^j. \qquad b_0 = 0.$$

The batch sizes of the primary customers and the service times of the customers are all mutually independent. In every time slot each customer in the orbit approaches the server with the so-called retrial probability r, independently of the other customers in the orbit, the residual service time, and the primary arrivals. Last but not least, we give precedence to arrivals over departures (LAS), so an arriving customer sees the server busy when at his arrival epoch also a departure takes place. As a consequence, it has been pointed out already in the introduction, after a service completion the server stays idle for at least one slot! Also we choose for delayed access. So even when a customer finds the server idle upon arrival his service will start only in the next slot at the earliest.

In this paper we only discuss the steady-state behavior of the number of customers in the orbit [for other performance measures we refer again to Nobel and Moreno (2008)].

To analyze the LARS-DA model, we introduce a discrete-time Markov chain (DTMC) by observing the system at epochs $k-$, that is at the start of the time slots k just after, possibly, a service has started, but before the arrivals during time slot k have occurred. We define the following random variables,

$$\begin{aligned}H_k &= \text{the residual service time of the job in service at time } k-,\\ Q_k &= \text{the number of jobs in orbit at time } k-.\end{aligned}$$

We define $H_k = 0$ when at epoch $k-$ the server is idle. Then, due to the independencies stated in the description of the model, the stochastic process $\{(H_k, Q_k) : k = 0, 1, 2, \ldots\}$ is an irreducible aperiodic DTMC and under the stability (ergodicity) condition that $\mathcal{A}'(1)(\mathcal{B}'(1)+1) < 1$ it is positive recurrent, as can be easily seen from the fact that $\mathcal{A}'(1)$ is the mean number of primary arrivals per time slot, and $\mathcal{B}'(1)$ is the mean service time. Notice that the 'plus 1' is required because in our setup the server is always idle for at least one time slot immediately after the departure of a customer. Nobel and Moreno (2008) contains the formal proof of this ergodicity condition.

We will derive the joint generating function of the steady-state distribution of the DTMC introduced above. Under the stability condition $\mathcal{A}'(1)(\mathcal{B}'(1) + 1) < 1$ we can define the following limiting joint distribution

$$\pi(j, n) = \lim_{k\to\infty} \mathbb{P}(H_k = j; Q_k = n), \qquad j = 0, 1, 2, \ldots; \quad n = 0, 1, 2, \ldots,$$

with its associated two-dimensional generating function

$$\Pi(w, z) = \sum_{j=0}^{\infty}\sum_{n=0}^{\infty} \pi(j, n) w^j z^n.$$

In the following it is convenient to introduce also the partial generating function,

$$\Pi_0(z) = \sum_{n=0}^{\infty} \pi(0, n) z^n.$$

To find $\Pi(w, z)$ we write down the system of balance equations,

$$\begin{aligned}\pi(0, n) &= a_0(1 - r)^n \pi(0, n) + \sum_{k=0}^{n} a_k \pi(1, n - k), \qquad n = 0, 1, 2, \ldots,\\ \pi(j, n) &= \sum_{k=0}^{n} a_k \pi(j + 1, n - k) + b_j \sum_{k=1}^{n+1} a_k \pi(0, n + 1 - k)\\ &\quad + b_j a_0(1 - (1 - r)^{n+1})\pi(0, n + 1),\\ &\quad j = 1, 2, \ldots; \quad n = 0, 1, 2, \ldots\end{aligned}$$

From the above equations, following a parallel methodology as used in Yang and Li (1995), we get

$$\begin{aligned}\Pi_0(z) &= a_0 \frac{\mathcal{B}(\mathcal{A}(z)) - z}{\mathcal{A}(z)\mathcal{B}(\mathcal{A}(z)) - z} \Pi_0((1 - r)z),\\ \Pi(w, z) &= \left(1 - w \frac{1 - \mathcal{A}(z)}{w - \mathcal{A}(z)}\, \frac{\mathcal{B}(w) - \mathcal{B}(\mathcal{A}(z))}{z - \mathcal{B}(\mathcal{A}(z))}\right) \Pi_0(z). \qquad (1)\end{aligned}$$

Now we define the so-called *retrial function*

$$\mathcal{R}(z) = a_0 \frac{\mathcal{B}(\mathcal{A}(z)) - z}{\mathcal{A}(z)\mathcal{B}(\mathcal{A}(z)) - z}.$$

Then $\mathcal{R}(z)$ is analytic on $|z| < 1$ [see Nobel and Moreno (2008) for details]. Further, $\mathcal{R}(0) = 1$ and

$$\mathcal{R}(1) = a_0 \frac{1 - \mathcal{A}'(1)\mathcal{B}'(1)}{1 - \mathcal{A}'(1)(\mathcal{B}'(1) + 1)}.$$

Notice that here the stability condition shows up! Using the retrial function we get by iteration from Eq. (1)

$$\begin{aligned}\Pi_0(z) &= \mathcal{R}(z)\Pi_0((1-r)z) = \mathcal{R}(z)\mathcal{R}((1-r)z)\Pi_0((1-r)^2 z)\\ &= \cdots = \prod_{i=0}^{n-1} \mathcal{R}((1-r)^i z)\Pi_0((1-r)^n z).\end{aligned}$$

and then for all $|z| \leq 1$

$$\lim_{n\to\infty} \prod_{i=0}^{n-1} \mathcal{R}((1-r)^i z)$$

exists because $0 \leq 1 - r < 1$ [see again Nobel and Moreno (2008) or Choi and Kim (1997) for details]. So, we get the following result for the partial generating function

$$\Pi_0(z) = \prod_{i=0}^{\infty} \mathcal{R}((1-r)^i z)\Pi_0(0). \tag{2}$$

Because we know that $\Pi_0(1)$ is the long-run fraction of time slots that the server is idle and by Little's law that $\mathcal{A}'(1)\mathcal{B}'(1)$ is the long-run fraction of time slots that the server is busy, we have $\Pi_0(1) = 1 - \mathcal{A}'(1)\mathcal{B}'(1)$. Using this we can get rid of $\Pi_0(0)$ in Eq. (2) and this gives

$$\Pi_0(z) = (1 - \mathcal{A}'(1)\mathcal{B}'(1)) \prod_{i=0}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)}. \tag{3}$$

Next, we look at the marginal distribution of the orbit size, say $\{q_n\}_{n=0}^{\infty}$. So, $q_n = \sum_{j=0}^{\infty} \pi(j, n)$.

Introduce the p.g.f. of this distribution,

$$\mathcal{Q}(z) := \Pi(1, z) = \sum_{n=0}^{\infty} q_n z^n.$$

Then we get

$$\mathcal{Q}(z) = \frac{1 - z}{\mathcal{B}(\mathcal{A}(z)) - z}\Pi_0(z).$$

From this p.g.f. the steady-state probabilities q_n can be calculated easily using the discrete FFT [see e.g. Press et al. (1986) and Tijms (2003)]. Notice that the formula for $\mathcal{Q}(z)$ is rigorous. Hence from $\mathcal{Q}(1) = 1$ we also would have found $\Pi_0(1) = 1 - \mathcal{A}'(1)\mathcal{B}'(1)$.

Now we can also calculate the 'long-run average number of jobs in orbit', say $\overline{Q}$ (seen at time epochs $k-$)

$$\overline{Q} = \mathcal{Q}'(1) = \lim_{z\to 1}\left\{\Pi_0'(z)\frac{z-1}{z-\mathcal{B}(\mathcal{A}(z))} + \Pi_0(z)\frac{1-\mathcal{B}(\mathcal{A}(z))-(1-z)\mathcal{B}'(\mathcal{A}(z))\mathcal{A}'(z)}{[\mathcal{B}(\mathcal{A}(z))-z]^2}\right\}.$$

Using L'Hôpital we get

$$\overline{Q} = \Pi_0'(1)\frac{1}{1-\mathcal{A}'(1)\mathcal{B}'(1)} + \frac{\mathcal{B}''(1)[\mathcal{A}'(1)]^2+\mathcal{B}'(1)\mathcal{A}''(1)}{2[1-\mathcal{A}'(1)\mathcal{B}'(1)]}.$$

It remains to calculate $\Pi_0'(1)$. From Eq. (3) we see that

$$\Pi_0'(z) = (1-\mathcal{A}'(1)\mathcal{B}'(1))\sum_{i=0}^{\infty}\frac{(1-r)^i\mathcal{R}'((1-r)^i z)}{\mathcal{R}((1-r)^i)}\prod_{j\neq i}\frac{\mathcal{R}((1-r)^j z)}{\mathcal{R}((1-r)^j)},$$

which gives

$$\Pi_0'(1) = (1-\mathcal{A}'(1)\mathcal{B}'(1))\sum_{i=0}^{\infty}\frac{(1-r)^i\mathcal{R}'((1-r)^i)}{\mathcal{R}((1-r)^i)}.$$

Of course, this series converges because for all $i \geq 1$ the factors $\mathcal{R}'((1-r)^i)/\mathcal{R}((1-r)^i)$ are bounded by the constant $\max\{\mathcal{R}'(x) : 0 \leq x \leq 1\}$ and $0 < 1-r < 1$. So, we find the following expression for the average orbit size,

$$\overline{Q} = \frac{\mathcal{B}''(1)[\mathcal{A}'(1)]^2+\mathcal{B}'(1)\mathcal{A}''(1)}{2[1-\mathcal{A}'(1)\mathcal{B}'(1)]} + \frac{\mathcal{R}'(1)}{\mathcal{R}(1)} + \sum_{i=1}^{\infty}\frac{(1-r)^i\mathcal{R}'((1-r)^i)}{\mathcal{R}((1-r)^i)}. \tag{4}$$

Notice that the first term is exactly the result for the average queue size in the corresponding delay model which is discussed in Bruneel and Kim (1993). The second term (independent of r) is accountable for the fact that after a service completion the server is always idle for at least one time slot, and the third term describes the difference between the delay model and the retrial model with respect to the number of jobs 'waiting' for service. For numerical results we refer to Nobel and Moreno (2008).

Finally, we notice that the mean busy period $\overline{L}$ can be easily found by the theory of regenerative processes. Recall from the previous section that $\pi(0,0)$ is the long-run fraction of time slots that the system is empty, i.e. the server is idle and no jobs in the orbit. Consider the regenerative cycle between two consecutive time slots in which a batch arrives while the orbit is empty and the server is idle. Then, we have immediately by the Renewal Reward Theorem,

$$\pi(0,0) = \frac{1/(1-a_0)}{1/(1-a_0)+\overline{L}}.$$

From Eq. (3) we have

$$\pi(0,0) = \Pi_0(0) = \frac{1-\mathcal{A}'(1)\mathcal{B}'(1)}{\prod_{i=0}^{\infty}\mathcal{R}((1-r)^i)}.$$

So we get

$$\overline{L} = \frac{1}{1-a_0}\left[\frac{\prod_{i=0}^{\infty}\mathcal{R}((1-r)^i)}{1-\mathcal{A}'(1)\mathcal{B}'(1)} - 1\right].$$

For other performance measures of the standard LARS-DA model we refer again to Nobel and Moreno (2008).

3 The LARS-DA model with Bernoulli feedback

This section is based on Nobel and Moreno (2005). We enrich the LARS-DA model with Bernoulli feedback: with probability θ a customer will be sent (back) to the orbit after service completion, whereas with probability $1-\theta$ the customer leaves the system forever. A customer sent back to the orbit after a service completion has to compete for another service with the other customers in the orbit and his new service time is considered independent of his previously completed service time. So, the total service time required for a customer is the sum of a geometrically distributed number of mutually independent identically distributed service times.

We are again interested in the steady-state behavior of this, say LARS-DA/BF model. We restrict ourselves to finding the p.g.f. of the steady-state distribution of the number of customers in the orbit and the mean busy period. The analysis of the LARS-DA/BF model is very similar to the analysis of the standard LARS-DA model: we can use the same DTMC $\{(H_k, Q_k) : k = 0, 1, 2, \ldots\}$ as in Sect. 2. Only the stability condition is slightly different: the DTMC $\{(H_k, Q_k) : k = 0, 1, 2, \ldots\}$ is positive recurrent if and only if $\mathcal{A}'(1)(\mathcal{B}'(1) + 1) < 1 - \theta$, as can be proved using Foster's criterion. So, we can give the new balance equations without further comments.

$$\begin{aligned}
\pi(0,n) &= a_0(1-r)^n\pi(0,n) + (1-\theta)\sum_{k=0}^{n} a_k\pi(1,n-k) + \theta\sum_{k=0}^{n-1} a_k\pi(1,n-1-k),\\
&\quad n = 0,1,2,\ldots,\\
\pi(j,n) &= \sum_{k=0}^{n} a_k\pi(j+1,n-k) + b_j\sum_{k=1}^{n+1} a_k\pi(0,n+1-k)\\
&\quad + b_j a_0(1-(1-r)^{n+1})\pi(0,n+1),\\
&\quad j = 1,2,\ldots;\quad n = 0,1,2,\ldots.
\end{aligned}$$

Defining again the *retrial function*

$$\mathcal{R}_\theta(z) = a_0\frac{(1-\theta+\theta z)\mathcal{B}(\mathcal{A}(z)) - z}{(1-\theta+\theta z)\mathcal{A}(z)\mathcal{B}(\mathcal{A}(z)) - z}.$$

we find by iteration, exactly as in Sect. 2

$$\Pi_0(z) = \prod_{i=0}^{\infty} \mathcal{R}_\theta((1-r)^i z)\Pi_0(0). \tag{5}$$

Because we know that $\Pi_0(1)$ is the long-run fraction of time slots that the server is idle, we also have by Little's law $\Pi_0(1) = 1 - \mathcal{A}'(1)\mathcal{B}'(1)/(1-\theta)$, we can get rid of $\Pi_0(0)$ in (5) and this gives

$$\Pi_0(z) = \left(1 - \frac{\mathcal{A}'(1)\mathcal{B}'(1)}{1-\theta}\right)\prod_{i=0}^{\infty}\frac{\mathcal{R}_\theta((1-r)^i z)}{\mathcal{R}_\theta((1-r)^i)}. \tag{6}$$

Then along the same lines as in Sect. 2 we find

$$\Pi(w,z) = \left(1 - \frac{\mathcal{A}'(1)\mathcal{B}'(1)}{1-\theta}\right)\left(1 - w\frac{1-\mathcal{A}(z)}{w-\mathcal{A}(z)}\,\frac{\mathcal{B}(w)-\mathcal{B}(\mathcal{A}(z))}{z-(1-\theta+\theta z)\mathcal{B}(\mathcal{A}(z))}\right)$$
$$\times \prod_{i=0}^{\infty} \frac{\mathcal{R}_\theta((1-r)^i z)}{\mathcal{R}_\theta((1-r)^i)}.$$

Next, we look at the marginal distribution of the orbit size. Introduce

$$\mathcal{Q}_\theta(z) = \Pi(1,z) = \sum_{n=0}^{\infty} q_n z^n.$$

Then we get

$$\mathcal{Q}_\theta(z) = \frac{(1-z)[1-\theta\mathcal{B}(\mathcal{A}(z))]}{(1-\theta+\theta z)\mathcal{B}(\mathcal{A}(z))-z}\Pi_0(z). \tag{7}$$

So we can find again the 'long-run average number of jobs in orbit', say $\overline{Q_\theta}$ (seen at time epochs $k-$). We only give the final result

$$\overline{Q_\theta} = \frac{(1-\theta)\left[\mathcal{B}''(1)[\mathcal{A}'(1)]^2 + \mathcal{B}'(1)\mathcal{A}''(1)\right] + 2\theta[\mathcal{B}'(1)\mathcal{A}'(1)]^2}{2(1-\theta)[1-\theta-\mathcal{A}'(1)\mathcal{B}'(1)]} + \frac{\mathcal{R}'_\theta(1)}{\mathcal{R}_\theta(1)}$$
$$+\sum_{i=1}^{\infty} \frac{(1-r)^i \mathcal{R}'_\theta((1-r)^i)}{\mathcal{R}_\theta((1-r)^i)}. \tag{8}$$

In this result we see the same type of decomposition as we have seen for the standard LARS-DA model.

We can also give a formula for the mean *busy period*, say $\overline{L}_\theta$ of the LARS-DA/BF model, using the same reasoning as in Sect. 2. Because we now know that

$$\pi(0,0) = \Pi_0(0) = \left(1 - \frac{\mathcal{A}'(1)\mathcal{B}'(1)}{1-\theta}\right)\prod_{i=0}^{\infty}\frac{1}{\mathcal{R}_\theta((1-r)^i)}.$$

we get

$$\overline{L}_\theta = \frac{1}{1-a_0}\left[\frac{1-\theta}{1-\theta-\mathcal{A}\prime(1)\mathcal{B}\prime(1)}\prod_{i=0}^{\infty}\mathcal{R}_\theta((1-r)^i) - 1\right].$$

For further details and numerical results for the LARS-DA/BF models we refer to Nobel and Moreno (2005).

4 The LARS-DA model with priority customers

In this section we discuss a priority loss/retrial model with one server and two types of arrivals. The section is based on Nobel and Moreno (2005a). Starting point is again the standard LARS-DA model, but we *add* a new arrival stream of customers, from now on type-1 customers. In this section the original arrival stream will be called the stream of type-2 customers, and they obey all the arrival and service time characteristics as described in Sect. 2. The type-1 customers, arrive *singly* at the server. The inter-arrival times follow a geometric distribution with mean $1/p$. When a type-1 customer finds the server idle upon arrival, this customer can start its service in the next slot and all type-2 customers arrived in the same slot, if any, are sent into the orbit. So we give priority to the type-1 customers over the type-2 customers. When upon arrival a type-1 customer finds the server busy, the customer is rejected forever and does not play any role in the future. So the priority of the type-1 customers is non-preemptive. We will call this model the PL/LARS-DA model. The service times of the different type-1 customers are mutually independent and follow a general probability distribution $\{c_j\}_{j=1}^{\infty}$. The corresponding p.g.f. is denoted by

$$\mathcal{C}(w) = \sum_{j=1}^{\infty} c_j w^j.$$

This completes the description of the PL/LARS-DA model.

As usual we are interested in the steady-state distribution of the number of type-2 customers in orbit, among others. As in the previous sections we can still use the DTMC $\{(H_k, Q_k) : k = 0, 1, 2, \ldots\}$, because for our analysis it is not important to know whether a type-1 or a type-2 customer is in service. So, H_k is simply the residual service time of the ongoing service as before at time $k-$ and Q_k is the number of *type-2* customers in orbit at time $k-$. Under the stability condition that

$$\mathcal{A}'(1)((1-p)\mathcal{B}'(1) + p\mathcal{C}'(1) + 1) < 1 - p$$

this DTMC is positive recurrent. See Nobel and Moreno (2005a) for an intuitive argument. Of course, using Foster's criterion a formal proof can be given. As for the previous models we will derive the joint generating function of the steady-state distribution of the DTMC introduced above. Although the model considered in this paper is much more complicated than the standard LARS-DA model, it turns out that as far as the steady-state distribution of the number of type-2 customers in the orbit is concerned we can follow a similar approach. This is due to the fact that it is not necessary to include the type of the customer in service in the state-description. Surprisingly, the residual service time is sufficient to describe the probabilistic behavior of the system. When the stability condition holds we can define the same limiting joint distribution $\pi(j, n)$ as in the Sects. 2 and 3. So, we write down immediately the system of balance equations,

$$\pi(0, n) = (1-p)a_0(1-r)^n \pi(0, n) + \sum_{k=0}^{n} a_k \pi(1, n-k), \quad n = 0, 1, 2, \ldots, \tag{9}$$

$$\begin{aligned} \pi(j, n) = {} & \sum_{k=0}^{n} a_k \pi(j+1, n-k) + p c_j \sum_{k=0}^{n} a_k \pi(0, n-k) \\ & + (1-p) b_j \left(\sum_{k=1}^{n+1} a_k \pi(0, n+1-k) + a_0(1-(1-r)^{n+1}) \pi(0, n+1) \right), \\ & j = 1, 2, \ldots; \quad n = 0, 1, 2, \ldots. \end{aligned} \tag{10}$$

From these balance equations we can find the two-dimensional p.g.f. $\Pi(w, z)$ along the same lines as in the previous two sections. For the details we refer to Nobel and Moreno (2005a). Introduce the retrial function

$$\mathcal{R}_p(z) = (1-p)a_0 \frac{\mathcal{B}(\mathcal{A}(z)) - z}{\mathcal{A}(z)[(1-p)\mathcal{B}(\mathcal{A}(z)) + pz\mathcal{C}(\mathcal{A}(z))] - z}.$$

Notice again that $\mathcal{R}_p(z)$ is analytic on $|z| \leq 1$, and that $\mathcal{R}_p(0) = 1$ and

$$\mathcal{R}_p(1) = (1-p)a_0 \frac{1 - \mathcal{A}'(1)\mathcal{B}'(1)}{1 - p - \mathcal{A}'(1)[(1-p)\mathcal{B}'(1) + p\mathcal{C}'(1) + 1]}. \tag{11}$$

As in Sects. 2 and 3 it is illustrative to see that the stability condition shows up again in the denominator! As before, using the retrial function $\mathcal{R}_p$ we find

$$\Pi_0(z) = \prod_{i=0}^{\infty} \mathcal{R}_p((1-r)^i z)\Pi_0(0) \tag{12}$$

and because $1 - \Pi_0(1)$ is the long-run fraction of time slots that the server is busy, we have by the Renewal Reward Theorem,

$$1 - \Pi_0(1) = \frac{\mathcal{C}'(1) + \mathcal{A}'(1)\mathcal{B}'(1)\frac{1}{p}}{\mathcal{C}'(1) + \frac{1}{p}} = \frac{p\mathcal{C}'(1) + \mathcal{A}'(1)\mathcal{B}'(1)}{p\mathcal{C}'(1) + 1}.$$

Then, after simple but tedious algebra we get the joint probability generating function of the limiting distribution $\{\pi(j, n)\}$,

$$\begin{aligned}\Pi(w, z) \\ &= \frac{1-\mathcal{A}'(1)\mathcal{B}'(1)}{1+p\mathcal{C}'(1)} \left(1+w \frac{\begin{bmatrix} [(1-p)\mathcal{A}(z)-1] \\ \times[\mathcal{B}(\mathcal{A}(z))-\mathcal{B}(w)] \end{bmatrix} + p\mathcal{A}(z) \begin{bmatrix} [\mathcal{B}(\mathcal{A}(z))-z]\mathcal{C}(w) \\ -[\mathcal{B}(w)-z]\mathcal{C}(\mathcal{A}(z)) \end{bmatrix}}{[w-\mathcal{A}(z)][\mathcal{B}(\mathcal{A}(z))-z]} \right) \\ &\quad \times \prod_{i=0}^{\infty} \frac{\mathcal{R}_p((1-r)^i z)}{\mathcal{R}_p((1-r)^i)}.\end{aligned}$$

Next, we look again at the marginal distribution of the orbit size $\{q_n\}_{n=0}^{\infty}$. Introduce the p.g.f. of the orbit-size distributions,

$$\mathcal{Q}_p(z) = \Pi(1, z) = \sum_{n=0}^{\infty} q_n z^n.$$

Then we get for the PL/LARS-DA model

$$\mathcal{Q}_p(z) = \left(1 + \frac{\begin{bmatrix} [(1-p)\mathcal{A}(z)-1] \\ \times[\mathcal{B}(\mathcal{A}(z))-1] \end{bmatrix} + p\mathcal{A}(z) \begin{bmatrix} [\mathcal{B}(\mathcal{A}(z))-z] \\ -[1-z]\mathcal{C}(\mathcal{A}(z)) \end{bmatrix}}{[1-\mathcal{A}(z)][\mathcal{B}(\mathcal{A}(z))-z]} \right) \Pi_0(z). \tag{13}$$

Now we can also calculate the 'long-run average number of type-2 jobs in orbit', say $\overline{Q}_p$, (seen at time epochs $k-$). Using L'Hôpital repeatedly we get after long and tedious calculations [see Nobel and Moreno (2005a) for more details]

$$\begin{aligned}\overline{Q}_p &= \frac{\mathcal{B}''(1)[\mathcal{A}'(1)]^2 + \mathcal{B}'(1)\mathcal{A}''(1)}{2[1-\mathcal{A}'(1)\mathcal{B}'(1)]} + p\mathcal{A}'(1)\left(\frac{\mathcal{C}''(1) + 2\mathcal{C}'(1)}{2[1+p\mathcal{C}'(1)]}\right) + \frac{\mathcal{R}'_p(1)}{\mathcal{R}_p(1)} \\ &\quad + \sum_{i=1}^{\infty} \frac{(1-r)^i \mathcal{R}'_p((1-r)^i)}{\mathcal{R}_p((1-r)^i)}.\end{aligned} \tag{14}$$

It is interesting to notice that the first term is exactly the result for the average queue size in the corresponding delay model in which type-2 customers wait in queue instead of going into orbit, the second term describes the effect on the orbit size due to the presence of the type-1 priority jobs. The third term (independent of r) is accountable for the fact that after a service completion the server is always idle for at least one time slot, and the last term describes the difference between the delay model and the retrial model with respect to the number of type-2 jobs ‘waiting’ for service. Compare this decomposition result with the analogous results found in the previous sections. Finally, we remark that we get the formula for $\overline{Q}$ presented in Sect. 2 if we take $p = 0$, just as expected.

Another performance measure of interest is the *loss probability* of type-1 jobs, say π_{rej}, i.e. the long-run fraction of type-1 jobs that is lost due to the fact that upon arrival the server is found busy. Because the type-1 jobs arrive according to a Bernoulii process we can use the discrete-time analogue of the PASTA property, say Bernoulli Arrivals See Time Averages (BASTA), and we get immediately

$$\pi_{rej} = 1 - \Pi_0(1) = \frac{p\mathcal{C}'(1) + \mathcal{A}'(1)\mathcal{B}'(1)}{p\mathcal{C}'(1) + 1}.$$

Finally, we notice that we can calculate the mean delay of the type-2 jobs in the orbit, say $\overline{D}_p$, by invoking Little’s law

$$\overline{D}_p = \frac{\overline{Q}_p}{\mathcal{A}'(1)}.$$

As in the previous sections we conclude with a formula for the *mean busy period*, say $\overline{L}_p$ of the discrete-time PL/LARS-DA model. Using the same argument as in Sect. 2 we get

$$\overline{L}_p = \frac{1}{1-(1-p)a_0}\left[\frac{1+p\mathcal{C}'(1)}{1-\mathcal{A}'(1)\mathcal{B}'(1)}\prod_{i=0}^{\infty}\mathcal{R}_p((1-r)^i) - 1\right].$$

For numerical results we refer to Nobel and Moreno (2005a).

A variant of the PL/LARS-DA model as discussed in this section is a mixed delay/retrial model in which the high-priority [type-1] customers arrive in batches and are put in a queue when they find the server busy upon arrival, *ceteris paribus*. The server only starts the service of a type-2 customer when the queue of type-1 customers is empty. It turns out that the analysis of this, say PD/LARS-DA model is much more complicated than the analysis of the PI/LARS-DA model. We refer to Nobel (2015) for further details.

5 Discrete-time models with abandonments

A very important characteristic of queueing situations is the phenomenon of ‘abandonments’: customers waiting in line (or staying in the orbit) decide to abandon the system forever before being served. When abandonments are incorporated in a queueing model (not necessarily a retrial model), the analysis of the model becomes much more difficult. This is in correspondence with our earlier observation that queueing models in which customers play an active role, for instance by having the option of reentering the system as in retrial models, are much more complicated than models in which the role of the customers is limited, as for instance in the well-studied loss and delay models. Also models with abandonments (or non-persistent customers as some people prefer to say) suffer from this ‘curse of intractability’. So, not surprisingly, as we have remarked already concerning the research on retrial models, models with abandonments have not received very much attention in the literature. Palm was the first author who studied the $M/M/c$ model enriched with the option of leaving the system from the queue without being served [see Palm (1953)]. More recently, mainly triggered by the application area of the call-center industry, more authors have studied queueing models with abandonments [see e.g. Garnett et al. (2002) and Mandelbaum et al. (2002); Mandelbaum and Zeltyn (2007) and references therein].

Needless to say that *retrial models with abandonments* have been studied even less. In this section we want to discuss the LARS-DA model enriched with the feature that in every time slot every customer in the orbit leaves the orbit forever with a probability θ independent of the other customers in the orbit or the state of the server. Because we choose the modeling assumption that abandonments will have precedence over arrivals we call this model a model with *early abandonments*. So we coin the model as the LARS-DA/EAb model. To be more precise, an arriving customer stays at least one time slot in the system and when abandonments and arrivals occur simultaneously, then upon arrival customers 'see the orbit' *without* the abandoned customers, they have left the system just before the arrivals.

To illustrate the difficulties in analyzing this LARS-DA/EAb model, we start this section with a discrete-time *delay* model with abandonments in Sect. 5.1. In Sect. 5.2 we will discuss the LARS-DA/EAb model, and we will see that we have reached the boundaries of tractability.

5.1 A discrete-time delay queueing model with abandonments

In this subsection we discuss a discrete-time analogue of the so-called Erlang-A model [see e.g. Mandelbaum and Zeltyn (2007)]. This continuous-time queueing model is a variant of the well-known $M/M/c$ model.

The discrete-time analogue discussed here is an extension of the standard 'late-arrival model with delayed access' as discussed in Bruneel and Kim (1993) in which the customers in the queue are allowed to abandon the queue. As usual we present a Markov chain analysis to study the steady-state behavior of the model using the GFM. The main problem in this approach turns out to be the calculation of the limiting probability that the system is empty. We will show that this problem can be solved by using 'an infinite recursion'. Once this probability is known we can calculate the usual performance measures such as the mean queue length, the fraction of customers which abandons the system, the throughput, et cetera. We will only give an outline here. For a detailed discussion we refer to Nobel and Ster (2010).

Let us first describe the model in more detail. We consider a discrete-time queueing model with c servers and an unlimited waiting space. Customers who find all servers busy upon arrival join the queue, and wait for service. The arrival stream is modeled as in the previous sections, but the service times of the individual customers are now taken geometrically distributed with parameter β (taking generally distributed service times as in the previous sections makes the model intractable!). Again the 'late arrival' (LAS) setup is chosen with 'delayed access' (DA). Next, in every time slot every customer waiting in the queue decides to leave the queue (and so abandon the system) forever with probability θ and he will persist to stay in the queue with probability $1-\theta$. As said before, we give precedence to abandonments over arrivals. All the usual independencies are assumed. We will coin this model as LADS-DA/EAb, i.e. a *late-arrival system with delayed access and early abandonments*, with c servers.

To study the steady-state behavior of this LADS-DA/EAb model introduce

$$\begin{aligned} C_k &= \text{the the number of busy servers at time } k-, \\ Q_k &= \text{the number of customers in the queue at time } k-. \end{aligned}$$

Then with the usual interpretation for $k-$ the process $\{(C_k, Q_k) : k = 0, 1, 2, \ldots\}$ is an irreducible aperiodic DTMC. The state space is

$$\mathcal{S} = \{(0, 0), (1, 0), \ldots, (c - 1, 0)\} \cup \{(c, n)|n = 0, 1, 2, \ldots\}.$$

Due to the abandonments this DTMC is positive recurrent, and so we can define the following limiting joint distribution,

$$\pi(j, n) = \lim_{k\to\infty} \mathbb{P}(C_k = j; Q_k = n), \qquad (j, n) \in \mathcal{S}.$$

To study this limiting distribution we introduce the probability generating function (p.g.f.),

$$\Pi_c(z) = \sum_{n=0}^{\infty} \pi(c, n)z^n.$$

To find the probabilities $\pi(0, 0), \pi(1, 0), \ldots, \pi(c-1, 0)$ and the p.g.f. $\Pi_c(z)$ we write down the following system of balance equations,

$$\begin{aligned}\pi(j, 0) = &\sum_{k=0}^{c-1} \sum_{i=(k-j)^+}^{k} \binom{k}{i} \beta^i (1 - \beta)^{k-i} a_{j-k+i}\pi(k, 0) + \sum_{m=0}^{\infty} \sum_{i=c-j}^{c} \binom{c}{i} \beta^i (1 - \beta)^{c-i} \\ &\times \sum_{k=(m-j+c-i)^+}^{m} \binom{m}{k} \theta^k (1 - \theta)^{m-k} a_{j-c+i-m+k}\pi(c, m), \qquad (15)\end{aligned}$$

$$\begin{aligned}\pi(c, n) = &\sum_{k=0}^{c-1} \sum_{i=0}^{k} \binom{k}{i} \beta^i (1 - \beta)^{k-i} a_{n+c-k+i}\pi(k, 0) + \sum_{m=0}^{\infty} \sum_{i=0}^{c} \binom{c}{i} \beta^i (1 - \beta)^{c-i} \\ &\times \sum_{k=(m-n-i)^+}^{m} \binom{m}{k} \theta^k (1 - \theta)^{m-k} a_{n+i-m+k}\pi(c, m), \\ &j = 0, 1, 2, \ldots, c - 1; \quad n = 0, 1, 2, \ldots. \qquad (16)\end{aligned}$$

After tedious algebraic manipulations it turns out that the first type of balance Eq. (15) can be rewritten as

$$\begin{aligned}\pi(j, 0) = &\sum_{k=0}^{c-1} \sum_{i=(k-j)^+}^{k} \binom{k}{i} \beta^i (1 - \beta)^{k-i} a_{j-k+i}\pi(k, 0) \\ &+ \sum_{i=c-j}^{c} \binom{c}{i} \beta^i (1 - \beta)^{c-i} \sum_{r=0}^{j-c+i} \frac{(1 - \theta)^r}{r!} a_{j-c+i-r}\Pi_c^{(r)}(\theta), \\ &j = 0, 1, \ldots, c - 1, \qquad (17)\end{aligned}$$

where

$$\Pi_c^{(r)}(z) = \frac{d^r \Pi_c(z)}{dz^r}$$

is the r-th derivative of the p.g.f. $\Pi_c(z)$. So we have found a system of c linear equations with $2c$ unknowns,

$$\pi(0, 0), \pi(1, 0), \ldots, \pi(c - 1, 0);\ \Pi_c(\theta), \Pi_c'(\theta), \Pi_c''(\theta), \ldots, \Pi_c^{(c-1)}(\theta).$$

The second type of balance Eq. (16) can be put in the p.g.f. format

$$\Pi_c(z) = \sum_{k=0}^{c-1} \pi(k,0) \sum_{i=0}^{k} \binom{k}{i} \beta^i (1-\beta)^{k-i} z^{k-i-c} \left[\mathcal{A}(z) - \sum_{j=0}^{c-k+i-1} a_j z^j \right]$$
$$+ \sum_{i=0}^{c} \binom{c}{i} \left(\frac{\beta}{z}\right)^i (1-\beta)^{c-i}$$
$$\times \left[\mathcal{A}(z)\Pi_c(\theta + (1-\theta)z) - \sum_{r=0}^{i-1} \frac{[(1-\theta)z]^r}{r!} \sum_{j=0}^{i-r-1} a_j z^j \Pi_c^{(r)}(\theta) \right]. \quad (18)$$

Hence we have to calculate $2c$ unknown quantities: (i) the c probabilities

$$\pi(0,0), \pi(1,0), \ldots, \pi(c-1,0),$$

and (ii) the values of the c derivatives of the p.g.f. $\Pi_c(z)$ for $z = \theta$,

$$\Pi_c(\theta), \Pi_c'(\theta), \Pi_c''(\theta), \ldots, \Pi_c^{(c-1)}(\theta).$$

In the rest of this section we show how to calculate these unknown quantities for the special case of one server, i.e. $c = 1$ [for the general case we refer to Nobel and Ster (2010)]. For this simplified model Eqs. (17) and (18) become

$$\pi(0,0) = a_0\pi(0,0) + \beta a_0 \Pi_1(\theta). \quad (19)$$

$$\Pi_1(z) = \pi(0,0)\frac{\mathcal{A}(z) - a_0}{z} + \left(1 - \beta + \frac{\beta}{z}\right) \mathcal{A}(z)\Pi_1(\theta + (1-\theta)z) - \frac{\beta a_0}{z}\Pi_1(\theta). \quad (20)$$

Substituting (19) in (20) gives

$$\Pi_1(z) = \left(1 - \beta + \frac{\beta}{z}\right) \mathcal{A}(z)\Pi_1(\theta + (1-\theta)z) + \pi(0,0)\frac{\mathcal{A}(z) - 1}{z}. \quad (21)$$

To find an expression for $\Pi_1(z)$ we introduce

$$z^{(k)} = 1 - (1-\theta)^k(1-z). \quad k = 0, 1, 2, \ldots.$$

So we can write for $k = 1, 2, \ldots,$

$$\Pi_1\left(z^{(k-1)}\right) = \left(1 - \beta + \frac{\beta}{z^{(k-1)}}\right) \mathcal{A}\left(z^{(k-1)}\right) \Pi_1\left(z^{(k)}\right) + \pi(0,0)\frac{\mathcal{A}\left(z^{(k-1)}\right) - 1}{z^{(k-1)}}. \quad (22)$$

Iterating Eq. (22) n times gives $[z = z^{(0)}]$

$$\Pi_1(z) = \prod_{k=0}^{n-1} \left(1 - \beta + \frac{\beta}{z^{(k)}}\right) \mathcal{A}\left(z^{(k)}\right) \Pi_1\left(z^{(n)}\right)$$
$$+ \pi(0,0) \sum_{k=0}^{n-1} \prod_{i=0}^{k-1} \left(1 - \beta + \frac{\beta}{z^{(i)}}\right) \mathcal{A}\left(z^{(i)}\right) \frac{\mathcal{A}\left(z^{(k)}\right) - 1}{z^{(k)}}. \quad (23)$$

Now we send n to infinity. Notice that

$$\lim_{n\to\infty} z^{(n)} = 1 \qquad \text{and} \qquad \Pi_1(1) = 1 - \pi(0,0).$$

So we get

$$\Pi_1(z) = \prod_{k=0}^{\infty}\left(1-\beta+\frac{\beta}{z^{(k)}}\right)\mathcal{A}\left(z^{(k)}\right)(1-\pi(0,0))$$
$$+\pi(0,0)\sum_{k=0}^{\infty}\prod_{i=0}^{k-1}\left(1-\beta+\frac{\beta}{z^{(i)}}\right)\mathcal{A}\left(z^{(i)}\right)\frac{\mathcal{A}\left(z^{(k)}\right)-1}{z^{(k)}}. \tag{24}$$

From (19) we have

$$\Pi_1(\theta) = \frac{1-a_0}{a_0\beta}\pi(0,0).$$

So taking $z=\theta$ in (24) we get an expression for $\pi(0,0)$, the probability that the system is empty, or, in other words, for

$$\pi(0,0) = \text{long-run fraction of time slots that the system is empty.}$$

We find after rearranging terms, [introduce $\theta^{(k)} := 1-(1-\theta)^{k+1}$]

$$\pi(0,0) = \frac{a_0\beta\prod_{k=0}^{\infty}\left(1-\beta+\frac{\beta}{\theta^{(k)}}\right)\mathcal{A}\left(\theta^{(k)}\right)}{1-a_0-a_0\beta\left[\begin{array}{c}\sum_{k=0}^{\infty}\prod_{i=0}^{k-1}\left(1-\beta+\frac{\beta}{\theta^{(i)}}\right)\mathcal{A}\left(\theta^{(i)}\right)\frac{\mathcal{A}\left(\theta^{(k)}\right)-1}{\theta^{(k)}}\\ -\prod_{k=0}^{\infty}\left(1-\beta+\frac{\beta}{\theta^{(k)}}\right)\mathcal{A}\left(\theta^{(k)}\right)\end{array}\right]}. \tag{25}$$

We can now find $\overline{Q}$, the *long-run average queue length*, by differentiating

$$\Pi_1(z) = \left(1-\beta+\frac{\beta}{z}\right)\mathcal{A}(z)\Pi_1(\theta+(1-\theta)z)+\pi(0,0)\frac{\mathcal{A}(z)-1}{z}.$$

After rearranging terms we get

$$\overline{Q} = \Pi_1'(1) = \frac{1}{\theta}\left[\mathcal{A}'(1)-\beta(1-\pi(0,0))\right].$$

This result is not surprising and has a clear-cut interpretation. Notice that

$$\begin{aligned}\mathcal{A}'(1) &= \text{the arrival rate}\\ \beta(1-\pi(0,0)) &= \text{the throughput}\\ \theta\overline{Q} &= \text{the abandonment rate}\end{aligned}$$

and of course

$$\text{abandonment rate} = \text{arrival rate} - \text{throughput}.$$

In other 'words'

$$\theta\overline{Q} = \mathcal{A}'(1)-\beta(1-\pi(0,0)).$$

For numerical results we refer again to Nobel and Ster (2010).

5.2 The LARS-DA model with abandonments

In this subsection we discuss the steady-state behavior of the one-server retrial model with abandonments, coined LARS-DA/EAb. This model has been studied in Wit (2010). As in Sect. 5.1 the service times are taken geometrically distributed with parameter β. Define

$$\begin{aligned} C_k &= \text{the the number of busy servers at time } k- \text{ (now only 0 or 1)},\\ Q_k &= \text{the number of customers in the orbit at time } k-. \end{aligned}$$

Then, $\{(C_k, Q_k) : k = 0, 1, 2, \ldots\}$ is an irreducible aperiodic DTMC.
The state space is

$$\mathcal{S} = \{(0, n)|n = 0, 1, 2, \ldots\} \cup \{(1, n)|n = 0, 1, 2, \ldots\}.$$

Define the following limiting joint distribution,

$$\pi(j, n) = \lim_{k\to\infty} \mathbb{P}(C_k = j; Q_k = n), \qquad (j, n) \in \mathcal{S}.$$

and introduce the two generating functions,

$$\Pi_j(z) = \sum_{n=0}^{\infty} \pi(j, n) z^n, \qquad j = 0, 1.$$

We want to find these two generating functions. First we give the system of balance equations,

$$\begin{aligned} \pi(0, n) = a_0 &\sum_{m=0}^{\infty} \binom{n+m}{m} \theta^m (1-\theta-r)^n \pi(0, n+m)\\ &+ \beta \sum_{k=0}^{n} a_k \sum_{m=0}^{\infty} \binom{n+m-k}{m} \theta^m (1-\theta)^{n-k} \pi(1, n+m-k), \end{aligned} \tag{26}$$

$$\begin{aligned} \pi(1, n) = (1-\beta) &\sum_{k=0}^{n} a_k \sum_{m=0}^{\infty} \binom{n+m-k}{m} \theta^m (1-\theta)^{n-k} \pi(1, n+m-k)\\ &+ \sum_{k=1}^{n+1} a_k \sum_{m=0}^{\infty} \binom{n+1+m-k}{m} \theta^m (1-\theta)^{n+1-k} \pi(0, n+1+m-k)\\ &+ a_0 \sum_{m=0}^{\infty} \binom{n+m+1}{m} \theta^m \left[(1-\theta)^{n+1} - (1-\theta-r)^{n+1}\right] \pi(0, n+1+m),\\ &n = 0, 1, 2, \ldots. \end{aligned} \tag{27}$$

From the balance equations Eq. (26) we get

$$\begin{aligned} \Pi_0(z) = a_0 &\sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \binom{n+m}{m} \theta^m \pi(0, n+m)[(1-\theta-r)z]^n\\ &+ \beta \sum_{k=0}^{\infty} a_k z^k \sum_{m=0}^{\infty} \sum_{n=k}^{\infty} \binom{n+m-k}{m} \theta^m \pi(1, n+m-k)[(1-\theta)z]^{n-k}. \end{aligned} \tag{28}$$

Notice that for $j = 0, 1$ and $m = 0, 1, 2, \ldots,$

$$\sum_{n=0}^{\infty}(n+m)(n+m-1)\cdots(n+1)\pi(j, n+m)s^n$$
$$= \Pi_j^{(m)}(s), \text{ the } m\text{-th derivative of the p.g.f.} \Pi_j(\cdot).$$

Using the Taylor-expansion

$$\Pi_j(\theta+s) = \sum_{m=0}^{\infty}\frac{\theta^m}{m!}\Pi_j^{(m)}(s)$$

Eq. (28) can be rewritten as

$$\Pi_0(z) = a_0\Pi_0(\theta+(1-\theta-r)z) + \beta\mathcal{A}(z)\Pi_1(\theta+(1-\theta)z). \tag{29}$$

Similarly, from the balance equations Eq. (27) we get

$$\begin{aligned}\Pi_1(z) &= (1-\beta)\sum_{k=0}^{\infty}a_k z^k \sum_{m=0}^{\infty}\sum_{n=k}^{\infty}\binom{n+m-k}{m}\theta^m\pi(1, n+m-k)[(1-\theta)z]^{n-k}\\ &+\frac{1}{z}\sum_{k=1}^{\infty}a_k z^k \sum_{m=0}^{\infty}\sum_{n=k-1}^{\infty}\binom{n+1+m-k}{m}\theta^m\pi(0, n+1+m-k)\\ &\times[(1-\theta)z]^{n+1-k}\\ &+\frac{1}{z}a_0\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\binom{n+m+1}{m}\theta^m\pi(0, n+1+m)z^{n+1}\\ &\times\left[(1-\theta)^{n+1}-(1-\theta-r)^{n+1}\right],\end{aligned} \tag{30}$$

Using again the Taylor-expansion we get from Eq. (30)

$$\begin{aligned}\Pi_1(z) = (1-\beta)\mathcal{A}(z)\Pi_1(\theta+(1-\theta)z) + \frac{\mathcal{A}(z)}{z}\Pi_0(\theta+(1-\theta)z)\\ -\frac{a_0}{z}\Pi_0(\theta+(1-\theta-r)z).\end{aligned}$$

Using Eq. (29) we get $\Pi_1(z) =$

$$\left(1-\beta+\frac{\beta}{z}\right)\mathcal{A}(z)\Pi_1(\theta+(1-\theta)z) + \frac{1}{z}\left[\mathcal{A}(z)\Pi_0(\theta+(1-\theta)z) - \Pi_0(z)\right]. \tag{31}$$

Summarizing, we have found the following two equations for the p.g.f.'s $\Pi_0(\cdot)$ and $\Pi_1(\cdot)$,

$$\Pi_0(z) = a_0\Pi_0(\theta+(1-\theta-r)z) + \beta\mathcal{A}(z)\Pi_1(\theta+(1-\theta)z).$$
$$\Pi_1(z) = \left(1-\beta+\frac{\beta}{z}\right)\mathcal{A}(z)\Pi_1(\theta+(1-\theta)z) + \frac{1}{z}\left[\mathcal{A}(z)\Pi_0(\theta+(1-\theta)z) - \Pi_0(z)\right].$$

Notice that different arguments show up in the p.g.f.'s: z, $\theta+(1-\theta-r)z$ and $\theta+(1-\theta)z$. So, it will not be easy to find the functions $\Pi_0(\cdot)$ and $\Pi_1(\cdot)$.
Some form of iteration will be required as we have seen in Sect. 5.1. The problem has only been solved for the special cases $\beta = \theta$ and (in case $\beta \neq \theta$) for $\theta = 1 - r$. We refer to Wit (2010) for further details. The general case is still a topic for future research.

6 The LARS-DA model with a tolerant server

In this section we discuss a variant of the standard LARS-DA model in which all customers arriving in a slot are accepted for service when the server is idle. This in contrast with the standard model where only one of the arriving customers is chosen for service, while the others are (re)sent into the orbit. The accepted mixed batch of primary and secondary customers is subsequently served uninterruptedly in random order. Every customer has his own service time, and all service times of the different customers in the batch are assumed to be independent. Only after all customers of the mixed batch have been served the server stays idle for at least one time slot, as in the standard model.

We are again interested in the steady-state behavior of the number of customers in orbit and we take the residual service time of the mixed batch of customers currently in service as a supplementary variable. As in the previous sections we define a DTMC by observing the system at the epochs $k-$, that is at the start of the time slot k just after, possibly, a service of a mixed batch has started, but before the arrivals during time slot k have occurred. We introduce the following random variables,

$$\begin{aligned} H_k &= \text{the residual service time of the mixed batch in service at time } k-, \\ Q_k &= \text{the number of customers in orbit at time } k-. \end{aligned}$$

We define $H_k = 0$ when at epoch $k-$ the server is idle. Then the stochastic process $\{(H_k, Q_k) : k = 0, 1, 2, \ldots\}$ is again an irreducible aperiodic DTMC and under the stability condition that $\mathcal{A}'(1)\mathcal{B}'(1) < 1$ it is positive recurrent, as can be easily seen from the fact that $\mathcal{A}'(1)$ is the mean number of primary arrivals per time slot, and $\mathcal{B}'(1)$ is the mean service time. A formal proof of this stability condition can be given using Foster's criterion. Notice that the $+1$ added to the mean service time $\mathcal{B}'(1)$ in the stability condition of the standard LARS-DA model discussed in Sect. 2 is absent now. This is because the customers in an accepted mixed batch are served uninterruptedly. Using the same notation as in Sect. 2, to find $\Pi(w, z)$ we write down the system of balance equations,

$$\pi(0,n) = a_0(1-r)^n\pi(0,n) + \sum_{k=0}^{n} a_k\pi(1,n-k), \qquad n = 0,1,2,\ldots, \tag{32}$$

$$\begin{aligned} \pi(j,n) = {} & \sum_{k=0}^{n} a_k\pi(j+1,n-k) + \sum_{k=1}^{\infty} a_k \sum_{m=0}^{\infty} \binom{n+m}{m} r^m(1-r)^n b_j^{*(k+m)}\pi(0,n+m) \\ & + a_0 \sum_{m=1}^{\infty} \binom{n+m}{m} r^m(1-r)^n b_j^{*m}\pi(0,n+m), \quad j=1,2,\ldots; \quad n=0,1,2,\ldots. \end{aligned} \tag{33}$$

Here $(b_j^{*(n+m)})_{j=n+m}^{\infty}$ denotes the $(n+m)$-fold convolution of the service-time distribution $(b_j)_{j=1}^{\infty}$. From Eqs. (32) and (33) we get immediately

$$\Pi_0(z) = a_0\Pi_0((1-r)z) + \mathcal{A}(z)\Pi_1(z), \tag{34}$$

$$\begin{aligned} \Pi_j(z) = {} & \mathcal{A}(z)\Pi_{j+1}(z) + \sum_{k=1}^{\infty} a_k \sum_{m=0}^{\infty} \frac{r^m}{m!} b_j^{*(k+m)} \Pi_0^{(m)}((1-r)z) \\ & + a_0 \sum_{m=1}^{\infty} \frac{r^m}{m!} b_j^{*m} \Pi_0^{(m)}((1-r)z). \end{aligned} \tag{35}$$

Here $\Pi_0^{(m)}(\cdot)$ is the m-th derivative of the partial generating function $\Pi_0(\cdot)$. Next, multiplying (35) by w^j and summing over $j = 1, 2, \ldots$ gives

$$\begin{aligned}\Pi(w,z) - \Pi_0(z) = {} & \frac{\mathcal{A}(z)}{w}\left(\Pi(w,z) - w\Pi_1(z) - \Pi_0(z)\right) \\ & + \sum_{k=1}^{\infty} a_k \sum_{m=0}^{\infty} \frac{r^m}{m!}[\mathcal{B}(w)]^{k+m}\Pi_0^{(m)}((1-r)z) \\ & + a_0 \sum_{m=1}^{\infty} \frac{r^m}{m!}[\mathcal{B}(w)]^{m}\Pi_0^{(m)}((1-r)z).\end{aligned}$$

Using the Taylor-expansion and after canceling terms we get

$$\begin{aligned}\Pi(w,z) - \Pi_0(z) = {} & \frac{\mathcal{A}(z)}{w}\left(\Pi(w,z) - w\Pi_1(z) - \Pi_0(z)\right) + \mathcal{A}(\mathcal{B}(w))\Pi_0(r\mathcal{B}(w) + (1-r)z) \\ & - a_0\Pi_0((1-r)z).\end{aligned}$$

Rearranging terms and using (34) gives

$$(w - \mathcal{A}(z))\Pi(w,z) = -\mathcal{A}(z)\Pi_0(z) + w\mathcal{A}(\mathcal{B}(w))\Pi_0(r\mathcal{B}(w) + (1-r)z)). \tag{36}$$

So, as in the previous sections, the problem is to find the unknown partial generating function $\Pi_0(z)$. The standard trick is again to take $w = \mathcal{A}(z)$ in (36). This gives

$$\Pi_0(z) = \mathcal{A}(\mathcal{B}(\mathcal{A}(z)))\Pi_0(r\mathcal{B}(\mathcal{A}(z)) + (1-r)z). \tag{37}$$

This equation asks for iteration. Introduce for any z, $|z| \leq 1$,

$$z^{(0)} := z \quad \text{and} \quad z^{(k)} := r\mathcal{B}\left(\mathcal{A}\left(z^{(k-1)}\right)\right) + (1-r)z^{(k-1)}, \quad k = 1, 2, \ldots.$$

Now we get by iteration from (37)

$$\begin{aligned}\Pi_0(z) &= \mathcal{A}(\mathcal{B}(\mathcal{A}(z)))\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(1)}\right)\right)\right)\Pi_0\left(r\mathcal{B}\left(\mathcal{A}\left(z^{(1)}\right) + (1-r)z^{(1)}\right)\right) = \cdots = \\ &= \prod_{k=0}^{n-1}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right)\Pi_0\left(z^{(n)}\right).\end{aligned}$$

It is not difficult to see that under the stability condition $\mathcal{A}'(1)\mathcal{B}'(1) < 1$, $\lim_{k\to\infty} z^{(k)} = 1$ for all $|z| \leq 1$ and further that

$$\lim_{n\to\infty}\prod_{k=0}^{n-1}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right)$$

exists.

So, sending n to infinity we get the following result for the partial generating function $\Pi_0(\cdot)$

$$\Pi_0(z) = \prod_{k=0}^{\infty}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right)\Pi_0(1). \tag{38}$$

Because we know that $\Pi_0(1)$ is the long-run fraction of time slots that the server is idle, we have $\Pi_0(1) = 1 - \mathcal{A}'(1)\mathcal{B}'(1)$ and then we get

$$\Pi_0(z) = (1 - \mathcal{A}'(1)\mathcal{B}'(1))\prod_{k=0}^{\infty}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right). \tag{39}$$

Now, from (36) and (39) we find

$$\Pi(w,z)=\frac{w\mathcal{A}(\mathcal{B}(w))\Pi_0(r\mathcal{B}(w)+(1-r)z)-\mathcal{A}(z)\Pi_0(z)}{w-\mathcal{A}(z)}$$
$$=(1-\mathcal{A}'(1)\mathcal{B}'(1))\frac{w\mathcal{A}(\mathcal{B}(w))\prod_{k=0}^{\infty}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(\phi^{(k)}(w,z)\right)\right)\right)-\mathcal{A}(z)\prod_{k=0}^{\infty}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right)}{w-\mathcal{A}(z)}, \tag{40}$$

where

$$\phi^{(0)}(w,z):=r\mathcal{B}(w)+(1-r)z \quad \text{and} \quad \phi^{(k)}(w,z):=r\mathcal{B}\left(\mathcal{A}\left(\phi^{(k-1)}(w,z)\right)\right) + (1-r)\phi^{(k-1)}(w,z).$$

Next, we consider the marginal distribution of the orbit size, say $\{q_n\}_{n=0}^{\infty}$ with p.g.f. $\mathcal{Q}(z)=\Pi(1,z)=\sum_{n=0}^{\infty}q_nz^n$. Then we get from (40)

$$\mathcal{Q}(z)=\frac{\Pi_0(r+(1-r)z)-\mathcal{A}(z)\Pi_0(z)}{1-\mathcal{A}(z)}$$
$$=(1-\mathcal{A}'(1)\mathcal{B}'(1))\frac{\prod_{k=0}^{\infty}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(\phi^{(k)}(1,z)\right)\right)\right)-\mathcal{A}(z)\prod_{k=0}^{\infty}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right)}{1-\mathcal{A}(z)}. \tag{41}$$

It is illustrative to calculate $\Pi_0'(z)$,

$$\Pi_0'(z)=(1-\mathcal{A}'(1)\mathcal{B}'(1))\sum_{k=0}^{\infty}\mathcal{A}'\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right)\mathcal{B}'\left(\mathcal{A}\left(z^{(k)}\right)\right)\mathcal{A}'\left(z^{(k)}\right)\left(z^{(k)}\right)' \times\prod_{i\neq k}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(i)}\right)\right)\right).$$

To evaluate this expression we calculate $\left(z^{(k)}\right)'$. Using the recursion for $z^{(k)}$ we get by iteration

$$\left(z^{(k)}\right)'=\left(r\mathcal{B}'\left(\mathcal{A}\left(z^{(k-1)}\right)\right)\mathcal{A}'\left(z^{(k-1)}\right)+1-r\right)\left(z^{(k-1)}\right)'$$
$$=\cdots=\prod_{i=0}^{k-1}\left(r\mathcal{B}'\left(\mathcal{A}\left(z^{(i)}\right)\right)\mathcal{A}'\left(z^{(i)}\right)+1-r\right)$$

because $\left(z^{(0)}\right)'=z'=1$. So we find

$$\Pi_0'(z)=\sum_{k=0}^{\infty}\mathcal{A}'\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right)\mathcal{B}'\left(\mathcal{A}\left(z^{(k)}\right)\right)\mathcal{A}'\left(z^{(k)}\right) \times\prod_{j=0}^{k-1}\left(r\mathcal{B}'\left(\mathcal{A}\left(z^{(j)}\right)\right)\mathcal{A}'\left(z^{(j)}\right)+1-r\right) \times\prod_{i\neq k}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(i)}\right)\right)\right)(1-\mathcal{A}'(1)\mathcal{B}'(1)). \tag{42}$$

From (42) we see immediately that $\Pi_0'(1)=\left(\mathcal{A}'(1)\right)^2\mathcal{B}'(1)/r$.

Now we can also calculate the *long-run average number of customers in the orbit*, say $\overline{Q}$ (seen at time epochs $k-$)

$$\overline{Q} = \mathcal{Q}'(1) = \lim_{z\to 1} \frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{\Pi_0(r+(1-r)z) - \mathcal{A}(z)\Pi_0(z)}{1-\mathcal{A}(z)}\right].$$

We first calculate $\mathcal{Q}'(z)$ in terms of $\Pi_0(z)$ and $\Pi_0'(z)$,

$$\begin{aligned}&\mathcal{Q}'(z)\\ &\quad= \frac{(1-\mathcal{A}(z))\left[\Pi_0'(r+(1-r)z)(1-r)-\mathcal{A}(z)\Pi_0'(z)\right]+\mathcal{A}'(z)\left[\Pi_0(r+(1-r)z)-\Pi_0(z)\right]}{(1-\mathcal{A}(z))^2}.\end{aligned}$$

Using L'Hôpital we get

$$\begin{aligned}\overline{Q} = \mathcal{Q}'(1) = \lim_{z\to 1}&\left\{\Pi_0'(z) - \frac{\Pi_0''(r+(1-r)z)(1-r)^2 - \mathcal{A}(z)\Pi_0''(z)}{2\mathcal{A}'(z)}\right.\\ &\left.-\frac{\mathcal{A}''(z)\left[\Pi_0(r+(1-r)z)-\Pi_0(z)\right]}{2\mathcal{A}'(z)(1-\mathcal{A}(z))}\right\}.\end{aligned}$$

Next using L'Hôpital in the third term we see that

$$\overline{\mathcal{Q}} = \left[1 - \frac{r\mathcal{A}''(1)}{2\left(\mathcal{A}'(1)\right)^2}\right]\Pi_0'(1) + \frac{r(2-r)}{2\mathcal{A}'(1)}\Pi_0''(1). \tag{43}$$

Because we know already that $\Pi_0'(1) = \left(\mathcal{A}'(1)\right)^2\mathcal{B}'(1)/r$, to calculate $\overline{\mathcal{Q}}$ we still have to evaluate $\Pi_0''(1)$. This requires some tedious algebra,

$$\begin{aligned}\Pi_0''(z) &= (1-\mathcal{A}'(1)\mathcal{B}'(1))\\ &\quad\times\left(\sum_{k=0}^{\infty}\mathcal{A}'\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right)\mathcal{B}'\left(\mathcal{A}\left(z^{(k)}\right)\right)\mathcal{A}'\left(z^{(k)}\right)\left(z^{(k)}\right)'\prod_{i\neq k}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(i)}\right)\right)\right)\right)'.\end{aligned}$$

To do the job we introduce the following abbreviations

$$\begin{aligned}\Gamma_k(z) &= \mathcal{A}'\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right)\mathcal{B}'\left(\mathcal{A}\left(z^{(k)}\right)\right)\mathcal{A}'\left(z^{(k)}\right),\\ \Delta_k(z) &= \left(z^{(k)}\right)' = \prod_{i=0}^{k-1}\left(r\mathcal{B}'\left(\mathcal{A}\left(z^{(i)}\right)\right)\mathcal{A}'\left(z^{(i)}\right)+1-r\right),\\ \Omega_k(z) &= \prod_{i\neq k}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(i)}\right)\right)\right).\end{aligned}$$

Then we have that

$$\begin{aligned}\Pi_0''(z) = (1-\mathcal{A}'(1)\mathcal{B}'(1))&\left\{\sum_{k=0}^{\infty}\Gamma_k'(z)\Delta(z)\Omega(z) + \sum_{k=0}^{\infty}\Gamma_k(z)\Delta'(z)\Omega(z)\right.\\ &\left.+\sum_{k=0}^{\infty}\Gamma_k(z)\Delta(z)\Omega'(z)\right\}.\end{aligned} \tag{44}$$

So we need the derivatives of $\Gamma(z)$, $\Delta(z)$ and $\Omega(z)$,

$$\begin{aligned}
\Gamma_k'(z) = \Big\{ & \mathcal{A}''\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right)\left[\mathcal{B}'\left(\mathcal{A}\left(z^{(k)}\right)\right)\mathcal{A}'\left(z^{(k)}\right)\right]^2 \\
& + \mathcal{A}'\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right)\mathcal{B}'\left(\mathcal{A}\left(z^{(k)}\right)\right)\mathcal{A}''\left(z^{(k)}\right) \\
& + \mathcal{A}'\left(\mathcal{B}\left(\mathcal{A}\left(z^{(k)}\right)\right)\right)\mathcal{B}''\left(\mathcal{A}\left(z^{(k)}\right)\right)\left[\mathcal{A}'\left(z^{(k)}\right)\right]^2\Big\}\left(z^{(k)}\right)',
\end{aligned}$$

$$\begin{aligned}
\Delta_k'(z) = \left(z^{(k)}\right)'' &= \left(\prod_{i=0}^{k-1}\left(r\mathcal{B}'\left(\mathcal{A}\left(z^{(i)}\right)\right)\mathcal{A}'\left(z^{(i)}\right)+1-r\right)\right)' \\
&= r\sum_{i=0}^{k-1}\left[\mathcal{B}''\left(\mathcal{A}\left(z^{(i)}\right)\right)\left[\mathcal{A}'\left(z^{(i)}\right)\right]^2+\mathcal{B}'\left(\mathcal{A}\left(z^{(i)}\right)\right)\mathcal{A}''\left(z^{(i)}\right)\right]\left(z^{(i)}\right)' \\
&\quad \times\prod_{j\neq i}\left(r\mathcal{B}'\left(\mathcal{A}\left(z^{(j)}\right)\right)\mathcal{A}'\left(z^{(j)}\right)+1-r\right),
\end{aligned}$$

$$\Omega_k'(z) = \left(\prod_{i\neq k}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(i)}\right)\right)\right)\right)' = \sum_{i\neq k}\Gamma_i(z)\left(z^{(i)}\right)'\prod_{j\neq i,k}\mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(z^{(j)}\right)\right)\right).$$

Next we evaluate all these functions and their derivatives at $z=1$,

$$\begin{aligned}
&\Gamma_k(1) = \left[\mathcal{A}'(1)\right]^2\mathcal{B}'(1), \qquad \Delta_k(1) = \left(r\mathcal{A}'(1)\mathcal{B}'(1)+1-r\right)^k, \qquad \Omega_k(1)=1, \\
&\Gamma_k'(1) = \left[\mathcal{A}''(1)\mathcal{A}'(1)\mathcal{B}'(1)\left\{\mathcal{A}'(1)\mathcal{B}'(1)+1\right\}+\left(\mathcal{A}'(1)\right)^3\mathcal{B}''(1)\right]\left(r\mathcal{A}'(1)\mathcal{B}'(1)+1-r\right)^k, \\
&\Delta_k'(1) = \left\{\left[\mathcal{A}'(1)\right]^2\mathcal{B}''(1)+\mathcal{A}''(1)\mathcal{B}'(1)\right\} \\
&\qquad\qquad \times\left(r\mathcal{A}'(1)\mathcal{B}'(1)+1-r\right)^{k-1}\frac{1-\left(r\mathcal{A}'(1)\mathcal{B}'(1)+1-r\right)^k}{1-\mathcal{A}'(1)\mathcal{B}'(1)}, \\
&\Omega_k'(1) = \left[\mathcal{A}'(1)\right]^2\mathcal{B}'(1)\left[\frac{1}{r(1-\mathcal{A}'(1)\mathcal{B}'(1))}-\left(r\mathcal{A}'(1)\mathcal{B}'(1)+1-r\right)^k\right].
\end{aligned}$$

Notice that $\Delta_0'(1)=0$. Now we can evaluate the series in (44) for $z=1$,

$$\begin{aligned}
\sum_{k=0}^{\infty}\Gamma_k'(1)\Delta_k(1)\Omega_k(1) &= \frac{\mathcal{A}''(1)\mathcal{A}'(1)\mathcal{B}'(1)\left\{\mathcal{A}'(1)\mathcal{B}'(1)+1\right\}+\left(\mathcal{A}'(1)\right)^3\mathcal{B}''(1)}{1-(r\mathcal{A}'(1)\mathcal{B}'(1)+1-r)^2}, \\
\sum_{k=1}^{\infty}\Gamma_k(1)\Delta_k'(1)\Omega_k(1) &= \frac{\left[\mathcal{A}'(1)\right]^2\mathcal{B}'(1)\left\{\left[\mathcal{A}'(1)\right]^2\mathcal{B}''(1)+\mathcal{A}''(1)\mathcal{B}'(1)\right\}}{(1-\mathcal{A}'(1)\mathcal{B}'(1))\left(1-(r\mathcal{A}'(1)\mathcal{B}'(1)+1-r)^2\right)}, \\
\sum_{k=0}^{\infty}\Gamma_k(1)\Delta_k(1)\Omega_k'(1) &= \left[\mathcal{A}'(1)\right]^4\left[\mathcal{B}'(1)\right]^2 \\
&\quad \times\left[\frac{1}{r^2(1-\mathcal{A}'(1)\mathcal{B}'(1))^2}-\frac{1}{1-(r\mathcal{A}'(1)\mathcal{B}'(1)+1-r)^2}\right].
\end{aligned}$$

Plugging in these expressions in (44), taking $z=1$, gives an expression for $\Pi_0''(1)$ and using (43) we get a *closed form expression* for $\overline{Q}$. Notice that for the standard LARS-DA model we did *not* get a closed form expression for the mean obit size $\overline{Q}$.

Next we look at several other performance measures, such as the size of a served batch, the number of customers in the orbit at a service completion epoch, and the number of customers in the orbit at the start of a service, all in steady-state.

Define $\mathcal{D}(z) = \sum_{m=0}^{\infty} d_m z^m$ as the p.g.f. of the *number of customers in the orbit at the end of a (batch) service*, in other words the *limiting departure distribution*. Then we have that $d_m = \sum_{k=0}^{m} a_k \pi(1, m-k)/\Pi_1(1)$, because there is a one to one correspondence between the departure epochs and the epochs with a residual [batch] service time equal to 1 [notice that $\Pi_1(1)$ is the steady-state probability that an epoch is a departure epoch]. So from (34) we get for the limiting departure distribution

$$\mathcal{D}(z) = \frac{\Pi_0(z) - a_0 \Pi_0((1-r)z)}{1 - \mathcal{A}'(1)\mathcal{B}'(1) - a_0 \Pi_0(1-r)}. \tag{45}$$

Now it is easy to see that the p.g.f. of the orbit size distribution at the start of a service is given by $\mathcal{D}((1-r)z + r)$ and the p.g.f. of the distribution of the batch taken into service is given by $\mathcal{A}(z)\mathcal{D}(rz + 1 - r)$. Notice that the orbit size seen at a service completion epoch is randomly split into a part going into service and a part staying in the orbit. These two parts are dependent. Also the orbit size distribution at arbitrary epochs $k-$ differs from the orbit size distribution at the start of a service, $\mathcal{Q}(z) \neq \mathcal{D}((1-r)z+r)$ [compare (41) and (45)]. Of course this is in accordance with the observation that, although the primary arrivals follow a Bernoulli distribution, due to possible arrival from the orbit the start of service epochs *do not*, i.e. BASTA does not hold.

Now we can present formulae for the mean size of a served batch, say $\overline{C}$ and for the mean orbit size at the start of a service, say $\overline{O_s}$,

$$\begin{aligned}\overline{C} &= \frac{\mathrm{d}}{\mathrm{d}z}\left[\mathcal{A}(z)\mathcal{D}(rz + 1 - r)\right]_{z=1} = \mathcal{A}'(1) + r\frac{\Pi_0'(1) - a_0(1-r)\Pi_0'(1-r)}{1 - \mathcal{A}'(1)\mathcal{B}'(1) - a_0\Pi_0(1-r)},\\ \overline{O_s} &= \frac{\mathrm{d}}{\mathrm{d}z}\left[\mathcal{D}((1-r)z + r)\right]_{z=1} = (1-r)\frac{\Pi_0'(1) - a_0(1-r)\Pi_0'(1-r)}{1 - \mathcal{A}'(1)\mathcal{B}'(1) - a_0\Pi_0(1-r)}.\end{aligned}$$

Using (39) and (42) we can calculate $\Pi_0(1-r)$ and $\Pi_0'(1-r)$, although we cannot evaluate the infinite product and the infinite series for $z = 1 - r$. So to calculate $\Pi_0'(1-r)$ numerically, truncation will be unavoidable here. Remember that we have seen earlier that $\Pi_0'(1) = \left(\mathcal{A}'(1)\right)^2 \mathcal{B}'(1)/r$.

To conclude, we will look at the mean busy period $\overline{L}$, defined this time as the time lapse which starts when the server begins the service of a newly arrived batch and no customer is in the orbit and ends when the server completes the service of a batch leaving behind an empty orbit. From (39) we get taking $z = 0$,

$$\pi(0,0) = (1 - \mathcal{A}'(1)\mathcal{B}'(1)) \prod_{k=0}^{\infty} \mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(0^{(k)}\right)\right)\right).$$

Notice that $0^{(0)} = 0$ and $0^{(k)} = r\mathcal{B}\left(\mathcal{A}\left(0^{(k-1)}\right)\right) + (1-r)0^{(k-1)}, \quad k = 1, 2, \ldots$. Then we get, using the same reasoning as in the earlier sections,

$$\overline{L} = \frac{1}{1 - a_0}\left(\frac{1}{(1 - \mathcal{A}'(1)\mathcal{B}'(1)) \prod_{k=0}^{\infty} \mathcal{A}\left(\mathcal{B}\left(\mathcal{A}\left(0^{(k)}\right)\right)\right)} - 1\right).$$

For numerical results we refer to Nobel (2013).

7 Final remarks

In this paper we have given a brief overview of some discrete-time late arrival retrial models. In all models the PGM has been used. For the standard LARS-DA model we have seen that the p.g.f. of the steady-state orbit size distribution can be found, but not in closed form because an infinite product shows up, and also the formula for the mean orbit size is not in closed form. The same remarks can be made about the models discussed in Sect. 3 and in Sect. 4. The LARS-DA/EAb model with non-persistent customers turned out to be more complicated. We have discussed a delay model with abandonments in Sect. 5.1, but we only discussed a model with geometrically distributed service times. Even under this restriction closed form expressions for the p.g.f. of the queue length cannot be found. For the retrial model with abandonments we were faced with the 'curse of intractability': only under further restrictions some results can be obtained. We have made the general observation that models turn out to become intractable when the choices for the customers become more versatile: for loss and delay models where after arrival the customers must accept passively their future, analytical results are parsimonious, but for retrial models the situation is already more difficult, and once you give the customers also the option to leave the system, intractability is not far away. With that respect the model with the tolerant server, discussed in Sect. 6, is interesting. For this model the p.g.f. of the orbit-size distribution contains an infinite product, but the mean orbit size has a closed form solution. In the light of our earlier observation, one can say that in the model with the tolerant server, the retrial aspect is suppressed by the tolerance of the server, giving the model a flavor of a delay model as well. Apparently, this mixed retrial/delay characteristic makes the model more susceptible for mathematical analysis, in the sense that closed form expressions exist for performance measures which do not have closed form expressions in pure retrial models.

So, our next step is to discuss other models with a mixed delay/retrial character. In Sharkawy (2014) the retrial model of Sect. 6 is enriched with a second arrival stream of low-priority customers who are served only, one by one, when no (mixed batch of) original customers has arrived in a slot in which the server is idle. The low-priority customers who cannot be served upon arrival are placed in a queue which is served in first-come-first-served order. Unfortunately this model has not been fully solved yet. More recently, a successful analysis has been given in Nobel (2015) for a mixed delay/retrial model in which the high-priority customers are delayed and served one by one on a first-come-first-served basis and the low-priority customers act as classical retrial customers, i.e. the server always accepts only one customer at a time. We mentioned this model already in Sect. 4, and the analysis illustrates the potential of extending the versatility of this type of retrial models beyond the models discussed in this paper without being trapped by 'the curse of intractability'. In general we can conclude that finding a good balance of more active versus more passive roles for the customers seems to be a prerequisite for success in the field of discrete-time retrial queues.

References

Aboul-Hassan, A., Rabia, S. I., & Taboly, F. A. (2008). A discrete-time Geo/G/1 retrial queue with general retrial times and balking customers. *Journal of the Korean Statistical Society*, *37*, 335–348.

Aboul-Hassan, A., Rabia, S. I., & Al-Mujahid, A. (2010). A discrete-time Geo/G/1 retrial queue with starting failures and impatient customers. In M. L. Gavrilova & C. J. K. Tan (Eds.), *Transactions on computer Science* (pp. 22–50). New York: Springer.

Alfa, A. S. (2002). Discrete time queues and matrix-analytic methods. *Top*, *10*, 147–210.

Alfa, A. S. (2006). Discrete-time analysis of the $GI/G/1$ system with Bernoulli retrials: An algorithmic approach. *Annals of Operations Research*, *141*, 51–66.

Amador, J., & Moreno, P. (2011). Analysis of the successful and blocked events in the Geo/Geo/c retrial queue. *Computers and Mathematics with Applications*, *61*, 2667–2682.

Artalejo, J. R., Atencia, I., & Moreno, P. (2005). A discrete-time $Geo^{[X]}/G/1$ retrial queue with control of admission. *Applied Mathematical Modelling*, *29*, 1100–1120.

Artalajo, J. R., Economou, A., & Gómez-Corral, A. (2008). Algorithmic analysis of the Geo/Geo/c retrial queue. *European Journal of Operational Research*, *189*, 1042–1056.

Artalejo, J. R., & Gómez-Corral, A. (2008). *Retrial queueing systems*. Berlin: Springer.

Artalejo, J. R., & Li, Q. L. (2010). Performance analysis of a block-structured discrete-time retrial queue with state-dependent arrivals. *Discrete Event Dynamic Systems*, *20*, 325–347.

Atencia, I., Fortes, I., Nishimura, S., & Sanchez, S. (2010). A discrete-time retrial queueing system with recurrent customers. *Computers and Operations Research*, *37*(7), 1167–1173.

Atencia, I., & Moreno, P. (2004). A discrete-time $Geo/G/1$ retrial queue with general retrial times. *Queueing Systems*, *48*, 5–21.

Atencia, I., & Moreno, P. (2004). Discrete-time $Geo^{[X]}/G_H/1$ retrial queue with Bernoulli feedback. *Computers and Mathematics with Applications*, *47*, 1273–1294.

Atencia, I., & Moreno, P. (2006). A discrete-time $Geo/G/1$ retrial queue with server breakdowns. *Asia-Pacific Journal of Operational Research*, *23*, 247–271.

Atencia, I., & Moreno, P. (2006). A discrete-time $Geo/G/1$ retrial queue with the server subject to starting failures. *Annals of Operations Research*, *141*, 85–107.

Atencia, I., & Moreno, P. (2006). $Geo/G/1$ retrial queue with 2nd optional service. *International Journal of Operational Research*, *1*, 340–362.

Bruneel, H., & Kim, B. G. (1993). *Discrete-time models for communication systems including ATM*. Dordrecht: Kluwer Academic Publishers.

Chaudhry, M. L. (1993). Exact and approximate numerical solutions of steady-state single-server bulk-arrival discrete-time queues: $Geom^X/G/1$. *International Journal of Mathematical and Statistical Sciences*, *62*, 133–185.

Chaudhry, M. L., & Gupta, U. C. (1997). Queue-length and waiting-time distributions of discrete-time $GI^X/Geom/1$ queueing systems with early and late arrivals. *Queueing Systems*, *25*, 307–324.

Choi, B. D., & Kim, J. W. (1997). Discrete-time Geo_1, $Geo_2/G/1$ retrial queueing systems with two types of calls. *Computers and Mathematics with Applications*, *33*, 79–88.

Falin, G. I., & Templeton, J. G. C. (1997). *Retrial queues*. London: Chapman & Hall.

Garnett, O., Mandelbaum, A., & Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, *4*, 208–227.

Jain, J. T., Mohanty, S. G., & Böhm, W. (2007). *A course on queueing models*. Boca Raton: Chapman & Hall/CRC.

Kim, B., & Kim, J. (2010). Queue size distribution in a discrete-time D-BMAP/G/1 retrial queue. *Computers and Operations Research*, *37*(7), 1220–1227.

Li, H., & Yang, T. (1998). $Geo/G/1$ discrete time retrial queue with Bernoulli schedule. *European Journal of Operational Research*, *111*, 629–649.

Li, H., & Yang, T. (1999). Steady-state queue size distribution of discrete-time $PH/Geo/1$ retrial queues. *Mathematical and Computer Modelling*, *30*, 51–63.

Mandelbaum, A., Massey, W., Reiman, M., Stolyar, A., & Rider, B. (2002). Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems*, *21*, 149–171.

Mandelbaum, A.,& Zeltyn, S. (2007). Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. *Advances in Services Innovations* (pp. 17–45). Berlin: Springer.

Miyazawa, T., & Takagi, H. (1994). Advances in discrete-time queues. *Queueing Systems*, *18*, 1–3.

Nobel, R. D. (2013). A queueing model in discrete time with retrials and a tolerant server. In *Proceedings of the 8th international conference on queueing theory and network applications*, 30 July - 2 Aug, Taichung, Taiwan.

Nobel, R. D. (2015). A mixed discrete-time delay/retrial queueing model for handover calls and new calls competing for a target channel. *Under Review*.

Nobel, R. D., & Moreno, P. (2005). A discrete-time retrial queueing model with Bernoulli feedback, *International conference on operations research applications in infrastructure development in conjunction with 38th Annual Convention of operation research society of India (ICORAID-2005-ORSI)*, Bangalore, 27–29 Dec 2005.

Nobel, R. D., & Moreno, P. (2005a). A discrete-time priority loss/retrial queueing model with two types of traffic. In B.D. Choi (Ed.) *Proceedings of the Korea-Netherlands joint conference on queueing theory and its applications to telecommunication systems*, (pp. 189–207) Seoul.

Nobel, R. D., & Moreno, P. (2008). A discrete-time retrial queueing model with one server. *European Journal of Operational Research*, *189*, 1088–1103.

Nobel, R. D., van der Ster, S. (2010). A discrete-time queueing model with abandonments, *Proceedings of the 5th international conference on queueing theory and network applications*, (pp. 29–34).

Palm, C. (1953). Methods of judging the annoyance caused by congestion. *Tele*, *4*, 189–208.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes*. Cambridge: Cambridge University Press.

el Sharkawy, S. (2014). Een discrete tijd retrial/delay wachtrijmodel met klanten van hoge en lage prioriteit, Bachelor's thesis. Amsterdam (in Dutch): Vrije Universiteit.

Takagi, H. (1993). *Queueing Analysis: A Foundation of Performance Evaluation, Vol. 3, Discrete-Time Systems*, North-Holland, New York.

Takahashi, M., Osawa, H., & Fujisawa, T. (1999). $Geo^{[X]}/G/1$ retrial queue with non-preemptive priority. *Asia-Pacific Journal of Operational Research*, *16*, 215–234.

Tijms, H. C. (2003). *A first course on stochastic models*. New York: Wiley.

Wang, J., & Zhang, P. (2009). A single-server discrete-time retrial G-queue with server breakdowns and repairs. *Acta Mathematica Applicatae Sinnica, English Series*, *25*(4), 675–684.

Wang, J., & Zhang, P. (2009a). A discrete-time retrial queue with negative customers and unreliable server. *Computers & Industrial Engineering*, *56*, 1216–1222.

Wang, J., & Zhao, Q. (2007). Discrete-time Geo$/G/1$ retrial queue with general retrial times and starting failures. *Mathematical and Computer Modelling*, *45*, 853–863.

de Wit, W. (2010). *Discrete-time retrial model with abandonments*, Master's thesis, , Amsterdam: Vrije Universiteit.

Wu, J., Liu, Z., & Peng, Y. (2011). A discrete-time Geo$/G/1$ retrial queue with preemptive resume and collisions. *Applied Mathematical Modelling*, *35*, 837–847.

Wu, J., Wang, J., & Liu, Z. (2013). A discrete-time Geo$/G/1$ retrial queue with preferred and impatiant customers. *Applied Mathematical Modelling*, *37*, 2552–2561.

Yang, T., Posner, M. J. M., Templeton, J. G. C., & Li, H. (1994). An approximation method for the $M/G/1$ retrial queue with general retrial times. *European Journal of Operational Research*, *76*, 552–562.

Yang, T., & Li, H. (1995). On the steady-state queue size distribution of the discrete-time $Geo/G/1$ queue with repeated customers. *Queueing Systems*, *21*, 199–215.