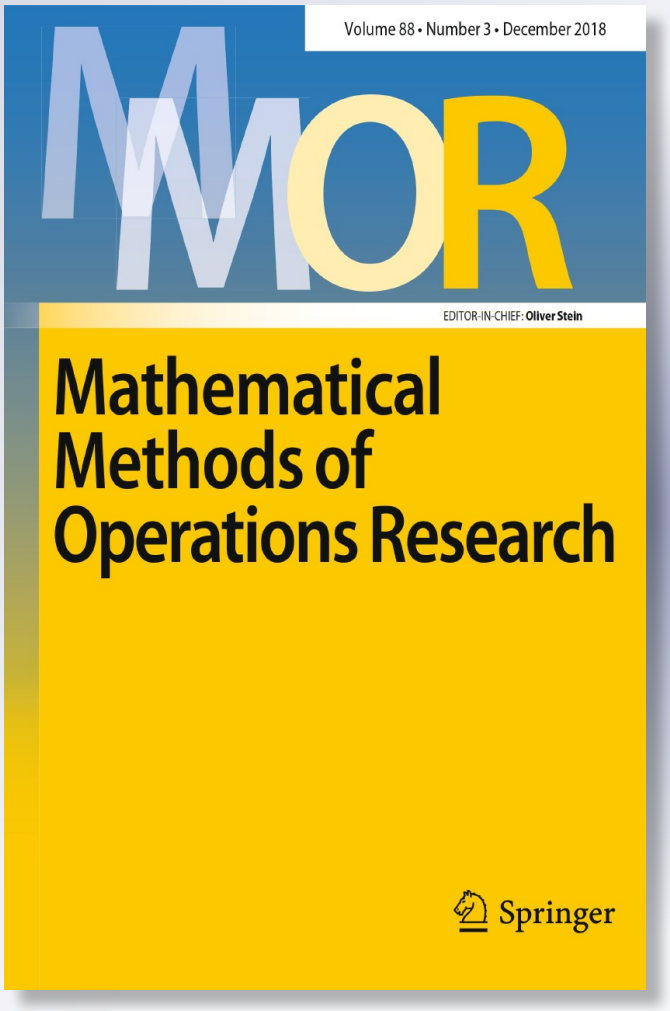
Asymptotics for the late arrivals problem

Carlo Lancia, Gianluca Guadagni, Sokol Ndreca & Benedetto Scoppola

Mathematical Methods of Operations Research

ISSN 1432-2994
Volume 88
Number 3

Math Meth Oper Res (2018) 88:475-493
DOI 10.1007/s00186-018-0643-3



Springer

Your article is published under the Creative Commons Attribution license which allows users to read, copy, distribute and make derivative works, as long as the author of the original work is cited. You may self-archive this article on your own website, an institutional repository or funder's repository and make it publicly available immediately.

Mathematical Methods of Operations Research (2018) 88:475–493
https://doi.org/10.1007/s00186-018-0643-3

ORIGINAL ARTICLE

Asymptotics for the late arrivals problem

Carlo Lancia[1] · **Gianluca Guadagni**[2] · **Sokol Ndreca**[3] · **Benedetto Scoppola**[4]

Received: 26 July 2017 / Accepted: 14 June 2018 / Published online: 26 June 2018
© The Author(s) 2018

Abstract
We study a discrete time queueing system where deterministic arrivals have i.i.d. exponential delays ξ_i. We describe the model as a bivariate Markov chain, prove its ergodicity and study the joint equilibrium distribution. We write a functional equation for the bivariate generating function, finding the solution on a subset of its domain. This solution allows us to prove that the equilibrium distribution of the chain decays super-exponentially fast in the quarter plane. We exploit the latter result and discuss the numerical computation of the solution through a simple yet effective approximation scheme in a wide region of the parameters. Finally, we compare the features of this queueing model with the standard $M/D/1$ system, showing that the congestion turns out to be very different when the traffic intensity is close to 1.

Keywords Late arrivals · Exponentially delayed arrivals · Pre-scheduled random arrivals · Queues with correlated arrivals · Bivariate generating function

Partially supported by PRIN 2012 "Problemi matematici in teoria cinetica ed applicazioni".

✉ Carlo Lancia
c.lancia@math.leidenuniv.nl
https://sites.google.com/view/clancia

Gianluca Guadagni
gg5d@virginia.edu

Sokol Ndreca
sokol@est.ufmg.br

Benedetto Scoppola
scoppola@mat.uniroma2.it

[1] Mathematical Institute Leiden University, Niels Bohrweg 1, 2333CA Leiden, The Netherlands

[2] Applied Mathematics, School of Engineering, University of Virginia, P.O. Box 400739, Charlottesville, VA 22904, USA

[3] Departamento de Estatística, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Pampulha, Belo Horizonte, Minas Gerais 31270-901, Brazil

[4] Dipartimento di Matematica, Università degli Studi di Roma Tor Vergata, Via Della Ricerca Scientifica 1, 00133 Rome, Italy

Mathematics Subject Classification 60J05 · 60K25

1 Introduction

In this paper we consider a single-server queue with deterministic service time, which is assumed of unitary length for the sake of simplicity. The ith customer arrives to the system at time

$$t_i = i + \xi_i\,, \qquad i \in \mathbb{N}\,, \tag{1}$$

where $\{\xi_i\}$ are i.i.d. exponential random variables with parameter β.

In the limit $\beta \to 0$ the point process (1) weakly converges to a Poisson process of parameter 1, whereas for fixed β the arrivals are negatively autocorrelated, see Collings (1976), Guadagni et al. (2011) and references therein.

We study the system described above for fixed β and we assume that arriving customers might balk with independent probability $1-\rho$. In other words, each customer can be deleted independently with probability $1-\rho$ before joining the queue. Besides being a mathematical expedient that ensures the existence of a stationary state,[1] balking is a way to model empty intervals in a constant stream of customers. Again, the point process (1) with balking weakly converges to a Poisson process, but with parameter ρ (Guadagni et al. 2011). In what follows, we refer to the balked version of (1) as Exponentially Delayed Arrivals (EDA).

Service can be delivered by the unique server only at discrete times. The length of the queue at time t is n_t; it represents the number of customers waiting to be served, including the customer that will be served precisely at time t, if any. Due to the balking, it is immediate to see that the traffic intensity of the system is ρ. Using Kendall's notation, we hereafter refer to the queue model described so far as $EDA/D/1$.

The $EDA/D/1$ model is motivated by the description of public and private transportation systems, including buses, trains, aircraft (Ball et al. 2001; Guadagni et al. 2011; Gwiggner and Nagaoka 2010; Iovanella et al. 2011) and vessels (Govier and Lewis 1963; Jagerman and Altiok 2003), appointment scheduling in outpatient services (a; Cayirli and Veral 2003; Mercer 1960, 1973) crane handling in dock operations (Daganzo 1990; Edmond 1975), and in general any system where scheduled arrivals are intrinsically subject to random variations. Preliminary results show that the model described above fits very well with actual data of inbound air traffic over a large hub (Caccavale et al. 2014; Lancia and Lulli 2017).

The appearance of the stochastic point process (1) can be traced back to Winsten's seminal paper (Winsten 1959). Winsten named such a queueing model *late process* and obtained results for the special case $\xi_i \in [0, 2]$ and service time exponentially distributed. At the end of Winsten's paper there is a discussion by Lindley, Wishart, Foster, and Takács, stating that Winsten's work can be considered as the first treatment of a queueing model with correlated arrivals (Winsten 1959, pp. 22–28).

The same problem was investigated also by Kendall (1964, page 11). Kendall remarked the great importance of systems with arrivals like (1): "*[...] perhaps too much attention has been paid to rather uninteresting variations on the fundamental*

[1] See Lemma 1 below.

Poisson stream. As soon as one considers variations dictated by the exigencies of the real world, rather than by the pursuit of mathematical elegance, severe difficulties are encountered; this is particularly well illustrated by the notoriously difficult problem of late arrivals." Kendall also provided the following elegant interpretation: if the random variables ξ_i are non-negative then the process defined by (1) is the output of the stationary $D/G/\infty$ queueing system (Kendall 1964). In particular, if the random variables ξ_i are exponentially distributed then $EDA/D/1$ can be viewed as a 2-stage tandem queueing network

$$D/M/\infty \to \cdot/D/1\,.$$

However, this is not the approach followed in this work.

Some years later, under the hypothesis that $\xi_i > 0$, Nelsen and Williams (1970) exactly characterised the distribution of the inter-arrival time intervals and their correlations. They also gave an explicit expression of these quantities in the particular case of ξ_i's exponentially distributed.

After the '70s only approximations of the arrival process (1) (Bloomfield and Cox 1972; Sabria and Daganzo 1989) or numerical studies of its output (Almaz and Altiok 2012; Ball et al. 2001; Nikoleris and Hansen 2012) seem to have appeared in the literature. In particular, Guadagni et al. (2011) presented a self-contained study of an arrival process like (1), assuming for ξ_i a compact-support distribution. They also proposed an approximation scheme that keeps the correlation of the arrivals and is able to compute in a quite accurate way the quantitative features of the queue. To the best of our knowledge, a queueing system with arrivals described by (1) still remains an open problem and the best results obtained so far are due to Winsten (1959).

$EDA/D/1$ is an example of a queueing system with correlated arrivals, a subject broadly studied in past years. There are many ways to impose a correlation to the arrival process. For instance, the parameters of the process may depend on their past realisation (Drezner 1999), or on some on/off sources (Wittevrongel and Bruneel 1999). Another relevant example of a queue model with correlated arrivals is the so-called Markov Modulated Queueing System. In Markov Modulated Queueing Systems the parameters are driven by an independent external Markovian process (Adan and Kulkarni 2003; Asmussen and Kella 2000; Combé and Boxma 1998; Lucantoni 1991; Pacheco et al. 2009). Our model shares with Markov Modulated Queueing Systems the property that one can define an external and independent Markovian process that drives the arrival rates. However, we see in Sect. 2 that the output of this external drive has also a deterministic effect on the queue length. More precisely, EDA can be interpreted as an independent drawing from an external pool of customers *late* at time t, see (3) below. Due to the memoryless property of the exponential delays, each customer late at time t will be still in the pool at time $t+1$ independently with probability $q \equiv e^{-\beta}$. This leads to binomial transitions in the number of late customers.

In Sect. 2 we show that $EDA/D/1$ can be described as a bivariate Markov chain representing the queue length and the number of late customers. We show that the distribution of the late customers is described by a INAR(1) process (Al-Osh and Alzaid 1987). Then we prove that the bivariate chain is ergodic and write the balance

equations of the stationary distribution, finding a functional equation for the bivariate generating function.

There exists an extensive literature about two-dimensional Markov models. Many methods for attacking the problem are available under two assumptions, namely, spatial homogeneity and finiteness of at least one marginal chain (Bini et al. 2005; Gail et al. 2000; Grassmann 2002; Latouche and Ramaswami 1999; Mitrani and Chakka 1995; Neuts 1989, 1995). Unfortunately, the Markov chain defined in Sect. 2 does not satisfy any of the mentioned requirements.

When both components of the Markov chain are infinite but space homogeneity is still ensured, the problem is typically attacked by reduction to a Riemann-Hilbert Boundary Value Problem. These may be solved, for example, by the uniformisation technique (Kingman 1961), conformal mappings (Cohen 1988; Cohen and Boxma 1983; Fayolle and Iasnogorodski 1979), the compensation method (Adan et al. 1993), or the Power Series Approximation (Blanc 1987a, b; Koole 1997; Hooghiemstra et al. 1988). The aforesaid binomial transitions are responsible for the lack of spatial homogeneity and are often encountered in Mathematical Biology (Artalejo et al. 2007; Brockwell et al. 1982; Economou 2004).

To the best of our knowledge, the functional equation (14) below has never previously appeared in the literature. Yet it is possible to mark some analogies with the functional equation studied by Economou and Kapodistria (2009), Economou et al. (2010),or Kapodistria (2011), the most important being that in both equations the right hand side exhibits the generating function computed in a convex combination in the parameter q. Adan et al. (2009), Altman and Yechiali (2006), Neuts (1994), and Yechiali (2007) give other examples of chains with binomial transitions.

In Sect. 3 we list some known results on INAR(1) process in order to describe the properties of the marginal distribution of the number of late customers and we write its exact analytical expression, which reveals the rich combinatorial structure of the problem. This intermediate result allows us to show that the stationary distribution of the $EDA/D/1$ queue has a super-exponential decay. Finally, in Sect. 4 we show that such a super-exponential behaviour enables a simple, yet very effective, numerical approximation scheme of the system balance equations. For a wide range of the system parameters, including typical values for real traffic applications of the model, we give a very good a priori estimate of the total-variation distance between the true and the approximate solution. Moreover, we compare the queue of $EDA/D/1$ with that of the standard $M/D/1$ system, i.e. with Poissonian arrivals. We show that the two queueing distributions may be very different, especially for heavy traffic load.

2 Stationary distribution: generating function and balance equations

Let us consider the process n_t, which describes the length of the queue at time t. This process is governed by the stochastic recursion

$$n_{t+1} = n_t + m_{(t,t+1]} - (1 - \delta_{n_t,0}) , \tag{2}$$

where $m_{(t,t+1]}$ is the number of arrivals in the interval $(t, t+1]$ according to the arrival process (1), and $\delta_{i,j}$ is Kronecker's delta. The term $1-\delta_{n_t,0}$ represents the action of the server in decreasing the queue length by one unit if this is non-empty at time t. Since the service time is deterministic, we focus on the so-called *embedded process* by observing the system at $t \in \mathbb{N}$, i.e. at departure instants.

The quantity $m_t \equiv m_{(t,t+1]}$ depends in general on the whole previous history of the system. Indeed, if for some large value of T, $m_s = 0$ for any $s \in \{t-T, t-T+1, ..., t-1\}$ then m_t is large with great probability. Conversely, if in the recent past the values of m_s have been large then m_t is expected to be small. This suggests that the arrival process is negatively autocorrelated, as proven by Guadagni et al. (2011). Hence, the recursion (2) does not depend only on the present value of n_t, and the memory of the process is infinite since T can be arbitrarily large.

Let now l_t denote the number of customers that are late at time t, i.e.

$$l_t \equiv \left|\{0 \le i \le t-1 \text{ such that } \xi_i > t-i\}\right|, \tag{3}$$

and let $p \equiv \int_0^1 f_\xi(t)dt = \int_0^1 \beta e^{-\beta t}dt = 1-e^{-\beta}$ and $q \equiv e^{-\beta}$. According to the memoryless property of the exponential delays $\{\xi_i\}$, independently on its scheduled arrival time, each customer late at time $t-1$ will continue to be late at time t with independent probabilty q, and it will arrive in queue with probabilty p. The process l_t is known in literature as INAR(1), also known as *Galton–Watson branching process with Bernoulli offspring with parameter q and Bernoulli immigration of parameter $q\rho$*. This is equivalently defined as

$$l_t = \sum_{i=1}^{l_{t-1}} \zeta_{t,i} + \eta_t\,,$$

where $\zeta_{t,i}$ and η_t are independent Bernoulli variables (of parameters $q\rho$ and q, respectively). The variables ζ are interpreted in terms of random *offsprings*, while the variables η represent random *immigrations*.

In our case, at time $(t-1)^-$ we have l_{t-1} late customers and each of them can independently arrive in queue in the time interval $[t-1, t)$ with probability $p = 1-q$. We interpret the fact that the customer arrives in queue in terms of having no offspring, while if the customer continues to be late it has an offspring. The immigration variable represents the new late customer, which is balked with probability $1-\rho$ and hence continues to be late at time t with probability ρq.

We can explicitly write the probability distribution of the quantity m_t conditioned to the value l_t: if the customer scheduled in the interval $(t, t+1]$ has balked, then

$$P\left(m_t = j \,|\, l_t = l\right) = \binom{l}{j} p^j q^{l-j} \equiv b_{j,l}\,, \tag{4}$$

otherwise

$$P\left(m_t = j \,|\, l_t = l\right) = \binom{l+1}{j} p^j q^{l+1-j} = b_{j,l+1}\,.$$

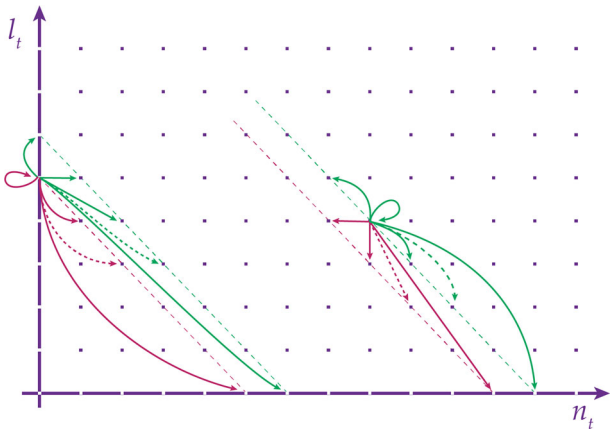


Fig. 1 Transitions of the $EDA/D/1$ queueing system in the quarter plane. They happen along the lines of Cartesian equation $x + y = n_t + l_t$ and $x + y = n_t + l_t - 1$ if $n_t \neq 0$, and along the lines of Cartesian equation $x + y = n_t + l_t$ and $x + y = n_t + l_t + 1$ if $n_t = 0$. Green transitions happen with probability ρ (no balking), red transitions happen with probability $1 - \rho$ (balking). (Color figure online)

All in all,

$$P\left(m_t = j \,|\, l_t = l\right) = b_{j,l}\,(1-\rho) + b_{j,l+1}\,\rho\,. \tag{5}$$

We describe the $EDA/D/1$ queue by the embedded process (n_t, l_t), $t \in \mathbb{N}$, a discrete-time Markov chain with the following transition probabilities:

$$\begin{aligned}
&\text{For } n > 0,\\
&\quad \mathcal{P}\left((n,l),(n+a-1,l-a+1)\right) = \rho\, b_{a,l+1}\,, && 0 \le a \le l+1\,,\\
&\quad \mathcal{P}\left((n,l),(n+a-1,l-a)\right) = (1-\rho)\, b_{a,l}\,, && 0 \le a \le l\,; \qquad (6)\\
&\text{For } n = 0,\\
&\quad \mathcal{P}\left((0,l),(a,l-a+1)\right) = \rho\, b_{a,l+1}\,, && 0 \le a \le l+1\,,\\
&\quad \mathcal{P}\left((0,l),(a,l-a)\right) = (1-\rho)\, b_{a,l}\,, && 0 \le a \le l\,. \qquad (7)
\end{aligned}$$

Figure 1 displays the transitions of the embedded chain (n_t, l_t) in the quarter plane.

Remark 1 The idea of describing the system as a bivariate Markov chain has been already used by Guadagni et al. (2011). However, they find more convenient to work with the number of customers that have already arrived at time t rather than the number of customers late at time t. This is a direct consequence of the assumption that delays have compact support.

From the theory of INAR(1) processes, l_t is an ergodic Markov chain for $q < 1$. The existence of the stationary state of the embedded chain (n_t, l_t) is guaranteed for $\rho, q < 1$ by the following

Lemma 1 *The bivariate chain (n_t, l_t) is ergodic if and only if $q < 1$ and $\rho < 1$.*

Proof If $q < 1$ and $\rho < 1$, the bivariate chain (n_t, l_t) is irreducible, see Fig. 1. Let us consider the process $\alpha_t = n_t + l_t$, which represent the diagonal in the quarter plane where the point (n_t, l_t) lies on, see Fig. 1. The bivariate chain has the property that $|\alpha_{t+1} - \alpha_t| \leq 1$. Equations (6)–(7) yield

$$P(\alpha_{t+1} = \alpha_t + 1 \,|\, n_t \neq 0) = 0\,, \tag{8}$$
$$P(\alpha_{t+1} = \alpha_t - 1 \,|\, n_t \neq 0) = (1-\rho)\,, \tag{9}$$
$$P(\alpha_{t+1} = \alpha_t + 1 \,|\, n_t = 0) = \rho\,, \tag{10}$$
$$P(\alpha_{t+1} = \alpha_t - 1 \,|\, n_t = 0) = 0\,. \tag{11}$$

In order to prove the positive recurrence of (n_t, l_t), we use Foster's criterion (Robert 2003, Cor. 8.7) setting $f(n_t, l_t) \equiv M\alpha_t + l_t + 1$ with $M = 2/(1-\rho)$ as the Lyapunov function. Thus, we need to show that there exist suitable positive constants K, γ such that

1. $\mathbb{E}[f(n_1, l_1) - f(n_0, l_0) \,|\, n_0 = N,\ l_0 = L] \leq -\gamma$ for $f(N, L) > K$;
2. $\mathbb{E}[f(n_1, l_1) \,|\, n_0 = N,\ l_0 = L] < \infty$ for $f(N, L) \leq K$;
3. the set $\{n, l \geq 0 \,:\, f(n, l) \leq K\}$ is finite.

Indeed

$$\begin{aligned}
&\mathbb{E}[f(n_1, l_1) - f(n_0, l_0) \,|\, n_0 = N, l_0 = L] \\
&\quad = M\mathbb{E}[\alpha_1 - \alpha_0 \,|\, n_0 = N, l_0 = L] + \mathbb{E}[l_1 - l_0 \,|\, l_0 = L]\,, \\
&\quad \leq M\rho\delta_{N,0} - M(1-\rho)(1-\delta_{N,0}) + \rho q - L(1-q)\,.
\end{aligned}$$

Then, using a little algebra, it can be shown that the first point of Foster's criterion is satisfied by $\gamma = 1$ and $K > 1 + \frac{(M+1)(1+\rho q+M\rho)}{1-\rho}$ (for example, $K = \frac{7}{(1-\rho)^2(1-q)}$). Point 2 of Foster's criterion holds because

$$\mathbb{E}[f(n_1, l_1) \,|\, n_0 = N, l_0 = L] \leq M(N + L + 1) + L + 2,$$

where we have used the property that α_t has only nearest-neighbour transitions. Point 3 is fulfilled by simply considering the definition of $f(n_t, l_t)$. For $\rho = 1$, the bivariate chain (n_t, l_t) is not ergodic because it is no longer irreducible[2]. □

For $\rho, q < 1$, Lemma 1 guarantees both the existence and the uniqueness of the stationary distribution

$$P_{n,l} \equiv \lim_{t\to\infty} P(n_t = n, l_t = l)\,. \tag{12}$$

Let us consider the following bivariate generating function:

$$P(z, y) \equiv \sum_{n,l\geq 0} z^n\, y^l\, P_{n,l}\,, \quad |z|, |y| \leq 1\,. \tag{13}$$

We are now ready to prove the main result of this section.

[2] For $\rho = 1$, the process α_t satisfies in fact $\alpha_{t+1} \geq \alpha_t$.

Theorem 1 *The bivariate generating function* (13) *satisfies*

$$P(z, y) = \frac{1 + \rho\,(\upsilon - 1)}{z}\,[(z - 1)\,P(0, \upsilon) + P(z, \upsilon)]\,, \tag{14}$$

where $\upsilon = \upsilon(z, y) \equiv z + q\,(y - z)$.

Proof For each $n, l \geq 0$, the balance equations of $EDA/D/1$ are the following:

$$P_{n,l} = (1 - \rho)\left(\sum_{j=0}^{n} P_{j+1,l+n-j}\,b_{n-j,l+n-j} + P_{0,l+n}\,b_{n,l+n}\right) + \rho\left(\sum_{j=0}^{n} P_{j+1,l+n-j-1}\,b_{n-j,l+n-j} + P_{0,l+n-1}\,b_{n,l+n}\right), \tag{15}$$

where $b_{j,l}$ are given by (4) and we agree that $P_{n,l} = 0$ whenever $n, l < 0$. The special cases $n = 0$ and $n = l = 0$ respectively lead to

$$P_{0,l} = (1 - \rho)\left(P_{1,l} + P_{0,l}\right)\,b_{0,l} + \rho\left(P_{1,l-1} + P_{0,l-1}\right)\,b_{0,l}\,, \tag{16}$$

$$P_{0,0} = (1 - \rho)(P_{1,0} + P_{0,0})\,. \tag{17}$$

To show that (15)–(17) hold, it suffices to write $P_{n_{t+1},l_{t+1}}$ in terms of P_{n_t,l_t} and then send t to infinity, in order to obtain the stationary state of the system.
Let us take (15) and multiply both sides by $z^n\,y^l$. Then take (16) and multiply both sides by y^l. Finally consider all such contributions, included the one in (17), and sum on n, l. The summation of all terms multiplied by $(1 - \rho)$ yields, after a bit of algebra,

$$(1 - \rho)\left[P(0, zp + yq) + \frac{1}{z}\,(P(z, zp + yq) - P(0, zp + yq))\right].$$

Analogously, the terms multiplied by ρ yield

$$\rho(zp + yq)\left[P(0, zp + yq) + \frac{1}{z}\,(P(z, zp + yq) - P(0, zp + yq))\right].$$

Summing up the two contributions, we get (14).

□

Remark 2 The functional equation (14) does not admit simple or immediate solutions. It is radically different from the functional equations typically studied in the literature (Cohen and Boxma 1983; Fayolle et al. 1999) and it is rather special in this respect. A simple solution can be found only in the particular case $z = 1$, see Sect. 3 below.

Remark 3 We end this section with a discussion of the special case $q = 0$. This means $\beta \to \infty$ or, in other words, no delay ξ. In this regime, the right-hand side of Eq. (14)

does not depend on y anymore, and $P(z, y) \equiv Q(z)$. The number of late customers is in fact $l_t = 0$ as the ith customer can not have a delay $\xi_i \geq 1$.

Then, Eq. (14) yields directly

$$Q(z) = \frac{1 + \rho(z-1)}{z}[(z-1)Q(0) + Q(z)], \tag{18}$$

where $Q(0) = 1 - \rho$ is the stationary probability of a void queue. Therefore, Eq. (18) is equivalent to

$$Q(z) = 1 + \rho(z-1),$$

which is the classical result of a $D/D/1$ queue with balking.

3 Asymptotic properties of $P_{n,l}$

In this section we first focus on P_l, the marginal distribution of late customers. Starting from the known expression of the generating function of INAR(1) process in the form of an infinite product, we find the exact analytical expression of P_l. Then, we derive the asymptotic behaviour of P_l and use it to infer asymptotics for $P_{n,l}$.
The marginal distribution of late customers is

$$P_l \equiv \sum_{n \geq 0} P_{n,l}$$

and its generating function is

$$\sum_{l \geq 0} P_l y^l = \sum_{n,l \geq 0} P_{n,l}\, y^l = P(1, y).$$

Setting $z = 1$ into Eq. (14) yields

$$P(1, y) = [1 + \rho q(y-1)]P(1, 1 + q(y-1)). \tag{19}$$

This is exactly the recurrence relation of the INAR(1) generating function, that can be written in terms of an infinite product as

$$P(1, y) = \prod_{k \geq 0} \left[1 + \rho q^{k+1}(y-1)\right]. \tag{20}$$

Remark 4 The infinite product (20) has an interesting combinatorial interpretation:

$$P(1, y) = \prod_{k \geq 0} \left[1 + \rho q^{k+1}(y-1)\right] = \frac{(\rho(1-y); q)_\infty}{1 + \rho(y-1)}, \tag{21}$$

where $(a;q)_\infty \equiv \prod_{k\geq 0}[1-aq^k]$ is the *infinite q-Pochhammer symbol*, also known as *infinite q-ascending factorial in a*. For $y = 1 - q/\rho$, $P(1, y) = \phi(q)\,(1-q)^{-1}$, where $\phi(q)$ is the well-known *Euler function*.

Remark 5 For $q < 1$, the right-hand side of (21) is analytic for each $y \in \mathbb{C}$. Therefore, the power series $P(1, y) = \sum_{l\geq 0} P_l\, y^l$, convergent for each $|y| \leq 1$, can be analytically continued in the whole complex plane. As such, we expect the marginal distribution P_l to decrease super-exponentially fast in l.

Following the insight given by Remark 5, we shift now the focus to the asymptotic behaviour of P_l and $P_{n,l}$. Expanding the product and rearranging it in powers of $\rho(y-1)$ yields

$$\begin{aligned} P(1,y) &= \prod_{k\geq 0}\Big(1+\rho\, q^{k+1}(y-1)\Big) \\ &= 1+\sum_{k\geq 1}\rho^k(y-1)^k\left[\sum_{m\geq\binom{k+1}{2}} d(m;k)q^m\right], \end{aligned} \tag{22}$$

where $d(m;k)$ is the number of partitions of m in k distinct parts.

Recall now the following two results from number theory.

Lemma 2 *(Yaglom and Yaglom 1964) If $m > \binom{k+1}{2}$ then the number of partitions of m in k distinct parts equals the number of partitions of $m - \binom{k+1}{2}$ into at most k parts (not necessarily distinct).*

Lemma 3 *(Andrews 1998; Hardy and Wright 1979) Let $p_{\leq k}(m)$ be the number of partitions of m in parts that do not exceed k. Then $p_{\leq k}(m)$ equals the number of partitions of m into at most k parts and*

$$P_{\leq k}(q) \equiv \sum_{m\geq 0} p_{\leq k}(m)\, q^m = \prod_{i=1}^{k}\frac{1}{1-q^i}\,.$$

Using Lemmas 2 and 3 we can recast after some algebra (22) as

$$P(1,y) = \sum_{k\geq 0}\frac{\rho^k q^{\binom{k+1}{2}}(y-1)^k}{\prod_{i=1}^{k}[1-q^i]}\,, \tag{23}$$

and by differentiation

$$P_l = \sum_{k\geq l}\frac{(-1)^{k-l}\rho^k q^{\binom{k+1}{2}}\binom{k}{l}}{\prod_{i=1}^{k}[1-q^i]}\,. \tag{24}$$

Remark 6 To the best of our knowledge, the explicit expression (24) has been proved in literature in relatively recent times, see Uchimura and Saitô (2011, 2015). Our proof, based solely on the combinatorial structure of the generating function and on the two Lemmas above, is slightly shorter.

Now we use (24) to obtain an upper bound on P_l. Let $(q;q)_l = \prod_{i=1}^{l}\left[1-q^i\right]$ be the q-Pochammer symbol of the pair (q,q). The following inequalities hold:

$$\binom{m+l}{l} = \prod_{k=1}^{m}\left(1+\frac{l}{k}\right) \le (1+l)^m\,,$$

$$\prod_{i=l+1}^{l+m}\left[1-q^i\right] \ge \prod_{i=1}^{m}\left[1-q^i\right] = (q;q)_m\,,$$

$$(q;q)_l \ge (q;q)_\infty\,.$$

Thus,

$$P_l = \frac{\rho^l\, q^{\binom{l+1}{2}}}{(q;q)_l}\sum_{m\ge 0}\frac{(-\rho)^m q^{(l+1)m} q^{\binom{m}{2}}\binom{m+l}{l}}{\prod_{i=l+1}^{l+m}[1-q^i]}\,,$$

$$\le \frac{\rho^l\, q^{\binom{l+1}{2}}}{(q;q)_\infty}\sum_{m\ge 0}\frac{q^{\binom{m}{2}}[\rho\, q^{l+1}\,(l+1)]^m}{(q;q)_m}\,, \tag{25}$$

$$= \rho^l\, q^{\binom{l+1}{2}}\frac{\prod_{k\ge 0}\left[1+q^{k+l+1}\rho(l+1)\right]}{(q;q)_\infty}\,, \tag{26}$$

where from (25) to (26) we have used the properties of q-ascending factorials and q-binomial coefficients (Gasper and Rahman 2004). If l is sufficiently large then $q^l\rho(l+1)\le 1$, and (26) yields

$$P_l \le \rho^l\, q^{\binom{l+1}{2}}\frac{(-q;q)_\infty}{(q;q)_\infty}. \tag{27}$$

Remark 7 Equations (24) and (27) show that, asymptotically in l, the leading order of P_l is $\rho^l q^{\binom{l+1}{2}}$. This fact can be directly implied from the arrival process (1). In fact, the most likely way to have l late customers (l large) is that each of the customers originally scheduled in the interval $[t-l,t)$ do not balk and are late at time t, an event of probability $\rho q^l\,\rho q^{l-1}\cdots\rho q = \rho^l q^{\binom{l+1}{2}}$.

Since $P_{n,l}\le P_l$, we have just obtained the following asymptotic result: uniformly in n, the equilibrium distribution $P_{n,l}$ decays super-exponentially fast in l. More precisely,

$$P_{n,l} = O\left(\rho^l q^{\binom{l+1}{2}}\right)\quad \text{for } l\to\infty\,. \tag{28}$$

As a matter of fact, the super-exponential decay of $P_{n,l}$ may be proved asymptotically in n too. Considering the auxiliary process $\alpha_t = n_t + l_t$, which we have already encountered in the proof of Lemma 1, we can write

$$p_a \equiv P(\alpha_t = a) = \sum_{\substack{n,l \geq 0 \\ n+l=a}} P_{n,l}\,, \quad a \geq 0\,.$$

It is straightforward to prove that the generating function of p_a is

$$P(z, z) = \frac{1 + \rho(z-1)}{1-\rho} P(0, z)\,. \tag{29}$$

From (29), it follows immediately that

$$p_a = P_{0,a} + \frac{\rho}{1-\rho} P_{0,a-1}\,. \tag{30}$$

For $a = n + l$, formulas (28) and (30) then yield

$$P_{n,l} \leq p_a = P_{0,a} + \frac{\rho}{1-\rho} P_{0,a-1} = O(\rho^a q^{\binom{a}{2}})\,. \tag{31}$$

Therefore, the following asymptotic result holds:

Theorem 2 *The equilibrium distribution $P_{n,l}$ decays super-exponentially fast as either $n \to \infty$ or $l \to \infty$. More precisely,*

$$P_{n,l} = O\left(\rho^{n+l} q^{\binom{n+l}{2}}\right) \quad \textit{for } n, l \to \infty\,. \tag{32}$$

Remark 8 Equation (29) gives an interesting connection between the equilibrium distribution of the quantity $\alpha_t = n_t + l_t$ and the stationary probability of having l late customers given that the queue is void. Figure 1 shows that the latter event drives the dynamic of α_t through an independent Bernoulli random variable with parameter ρ, which explains the factor $1 + \rho(z-1)$.

4 Approximate computation of $P_{n,l}$

As we have just shown that the joint stationary probability $P_{n,l}$ decreases super-exponentially fast in the limit of either $n, l \to \infty$, the natural question arising is whether a bare truncation of the infinite linear system (15)–(17) is sufficient to obtain a satisfactory numerical expression of $P_{n,l}$. As we will see, in this case the answer is positive due to (32), and this may prove crucial in contexts where practical solutions are needed.

Another natural question, once the numerical expressions of the probability distribution of our process are obtained, is whether the average performance of this queueing system is comparable with a classical memoryless system, say an $M/D/1$ with the same traffic intensity ρ, or if it is definitely different. This question is natural also because we know, see above, that for q very close to 1 the arrival process converges to a Poisson process, and we may ask what are the actual values of q such that, with respect

to the average congestion, the two processes, Poisson and EDA, become indistinguishable. As it will be clear, the difference between the average queue lengths remains large also for q very close to 1, due to the negative autocorrelation of the EDA arrivals. This can be interesting also in many applications, for instance in the air traffic case.

In order to compute numerically the probability distribution, we truncate the infinite system of balance equations (15)–(17) by fixing an integer $\alpha_{\max}$ and imposing $P_{n,l} = 0$ for $n + l > \alpha_{\max}$. Then we map the quarter plane $\{n, l \in \mathbb{N} \times \mathbb{N}\}$ onto the set of non-negative integers and relabel the unknowns $P_{n,l}$ as π_i, recasting (15)–(17) as $\pi = \pi Q$. For the details of both mapping and relabeling, see "Appendix A"

Next, we map the set $\{n, l \in \mathbb{N} \times \mathbb{N}$ such that $n + l \leq \alpha_{\max}\}$ onto the set of non-negative integers $\{0, 1, \ldots, k_{\max}\}$, where $k_{\max} \equiv \binom{\alpha_{\max}+1}{2}$. We want to consider the truncated system

$$\begin{cases} \tilde{\pi}_i = \sum_{j=0}^{k_{\max}} \tilde{\pi}_j Q_{j,i} \quad \text{for} \quad i = 1, 2, \ldots, k_{\max} - 1 \,, \\ \sum_{j=0}^{k_{\max}} \tilde{\pi}_j = 1 \,. \end{cases} \tag{33}$$

The idea we present is not new and has been already discussed, for instance by Tijms (1994) for stationary distributions with geometric tail. "Appendix B" shows that there exists a sequence $\{\varepsilon_j\}$ such that

$$\sum_{i > k_{\max}} \pi_i \, Q_{i,j} \leq \varepsilon_j \,, \qquad j = 0, 1, \ldots, k_{\max} - 1 \,,$$
$$\sum_{i > k_{\max}} \pi_i \leq \varepsilon_{k_{\max}} \,.$$

An a priori estimate of the error introduced by the truncation can be obtained from *perturbation theory* (Golub and Loan 2012, § 2.6.2):

$$\sum_{j=0}^{k_{\max}} |\pi_j - \tilde{\pi}_j| \leq \kappa(A) \sum_{i=0}^{k_{\max}} \varepsilon_i \,, \tag{34}$$

where A is the $k_{\max} \times k_{\max}$ matrix

$$A_{ij} \equiv \begin{cases} \delta_{i,j} - Q_{i,j} \,, & i = 0, 1, \ldots, k_{\max} - 1 \,, \\ 1 & i = k_{\max} \,, \end{cases} \tag{35}$$

$\delta_{i,j}$ is the usual Kronecker's delta, and

$$\kappa(A) \equiv \|A\|_1 \|A^{-1}\|_1$$

is the norm-1 condition number of the matrix A.

From "Appendix B",

$$\sum_{j=0}^{k_{\max}} \varepsilon_j \leq 2\alpha_{\max} \frac{(-q;q)_\infty}{(q;q)_\infty} \rho^{\alpha_{\max}+1} q^{\binom{\alpha_{\max}+1}{2}} \leq 2\alpha_{\max} \frac{(-q;q)_\infty}{(q;q)_\infty} q^{\binom{\alpha_{\max}+1}{2}} \,. \tag{36}$$

Numerical evaluations show that $\frac{(-q;q)_\infty}{(q;q)_\infty} q^{\binom{\alpha_{\max}}{2}}$ remains smaller than 10^{-16} for $\alpha_{\max} = 100$ and $q \leq .98$.

Estimating the condition number of a matrix is a notably difficult problem and a vast literature exists on this topic (Golub and Kahan 1965; Qi 1984; Johnson 1989). Since one of the aims of the present section is showing that a bare truncation of the balance equations (6)–(7) may be sufficient for an approximate computation of $P_{n,l}$, we evaluate $\kappa(A)$ numerically in the ρq-plane when $\alpha_{\max} = 100$, finding that the condition number is not larger than 10^5 for $\rho, q \leq 0.99$. Combining these two results we obtain that, uniformly in $\rho \leq 0.99$, the a priori norm-1 approximation error is less than 10^{-12} for q up to 0.98.

Remark 9 For air traffic applications, $q = 0.98$ corresponds to typical delays of the order of one hour. The same value of q for other transport systems, e.g. trains or buses, correspond to even higher delays. Consider also that typical values of traffic intensity ρ in extremely congested systems, e.g. London Heathrow Airport, do not exceed 0.98, see Caccavale et al. (2014). Therefore, the approximation scheme presented in this section is very fit for real life applications.

We conclude this section by analysing the performance of EDA/D/1 under congestion, i.e. ρ, q close to 1. First, we offer a comparison of the marginal distribution of the queue according to $EDA/D/1$ and the classical $M/D/1$ system. Figure 2 shows the distribution of the queue length in 25 different scenarios, with ρ and q increasing from 0.9 to 0.98 and 0.95, respectively. $EDA/D/1$ and $M/D/1$ are compared varying q keeping ρ fixed. We see that, even for very large values of parameter q, the super-exponential decay of $EDA/D/1$ yields both a much thinner tail and a much shorter average queue length. The plot also suggests that the distribution of the queue approaches the $M/D/1$ cases for $q \to 1$, which is expected in view of the weak convergece of 1 to a Poisson process. Yet this phenomenon is much more evident for low values of ρ – not reported in Fig. 2.

Figure 3 shows the average queue length. We note that the mean value of the queue never exceed 6, while for $\rho = .98$ the average queue of $M/D/1$ is equal to 25. This shows, as outlined above, that the negative autocorrelation of the arrivals 1 has an heavy impact on the congestion when the traffic intensity is high. As mentioned, such difference tends to vanish for low traffic intensity. A deeper comparison, extended to a wider range of values ρ, q is available at the address https://github.com/clancia/EDA/blob/master/Asymptotics.ipynb.

5 Conclusions

In this paper we have addressed a single-server queueing system with deterministic service time and *exponentially delayed arrivals*. The point process describing these arrivals dates back to the '50s of the past century and was studied by Kendall and others.

We have described the model as a bivariate Markov chain, proved that the latter is ergodic, wrote the balance equations of the stationary distribution, and found a functional equation for the bivariate generating function. Then we have focused on the

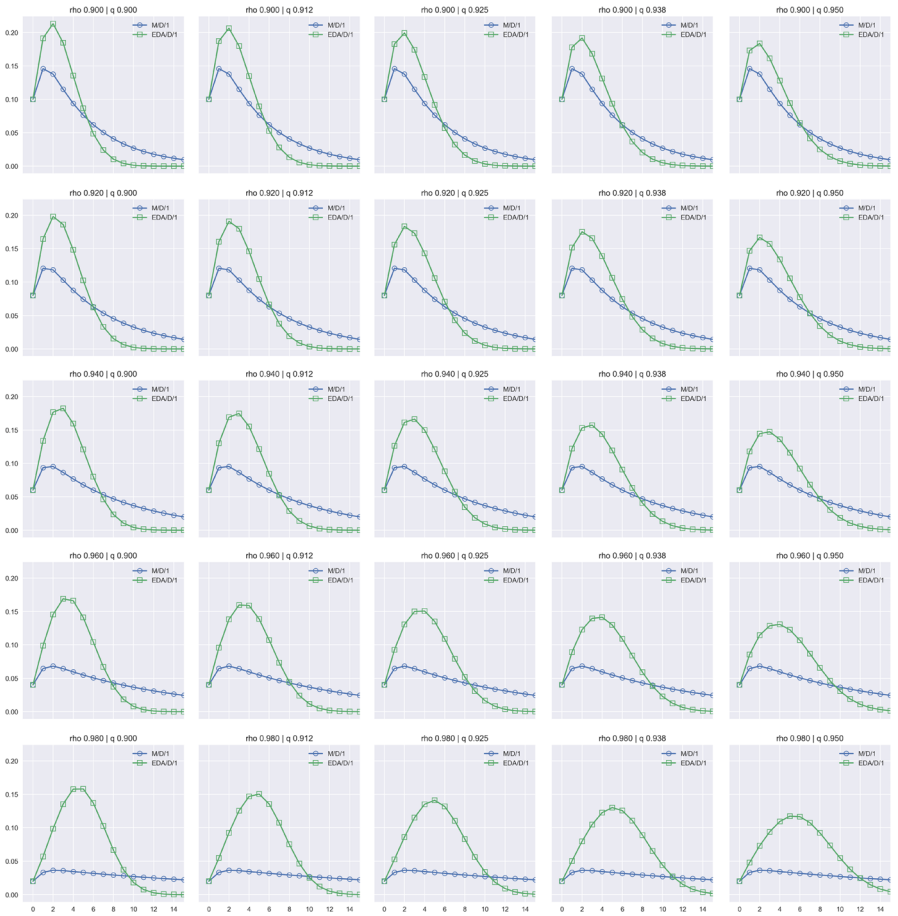


Fig. 2 Comparison of the queue distribution of EDA/D/1 and M/D/1 models for $\rho = 0.90, 0.92, 0.94, 0.96, 0.98$ and $q = 0.90, 0.9125, 0.925, 0.9375, 0.95$. (Color figure online)

marginal distribution of the number of late customers and found its exact expression. This intermediate step has enabled the fundamental result on the super-exponential decay of the joint stationary distribution. The characterisation of the asymptotic behaviour has finally led us to show that the solution to the balance equations can be approximately computed in a simple yet very accurate way. We have than compared this queueing system with a classical memoryless system, and we have found that the two systems are deeply different for high traffic intensity

In spite of the big efforts we have put forward to find the solution to the functional equation (14), the complete solution of the problem is still out of reach. An expansion in powers of q seems to be a promising approach to obtain (at least) an approximate expression of the bivariate generating function. This method allows to set up a recursive scheme to compute the coefficients of the power series (Lancia 2013). Unfortunately,

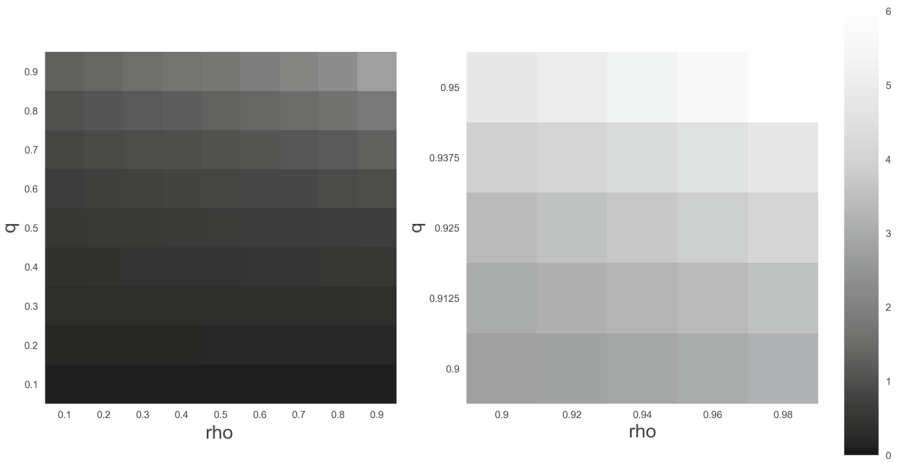


Fig. 3 Mean queue length of EDA/D/1 for different pairs (ρ, q)

the computations are quite involved and need some additional work to be refined. This is the subject of ongoing research.

Acknowledgements G.G. and S.N. appreciate the Mathematics Department of the University of Rome 'Tor Vergata' for the kind support. We thank the anonymous referees for their work.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix A: Map of the quarter plane onto non-negative integers

Define the map F of the quarter plane onto the set of the non-negative integers and its inverse G:

$$F(n,l) \equiv \binom{n+l}{2} + l\,,$$

$$g(m) \equiv \max\left\{ j \geq 0 \text{ such that } \binom{j+1}{2} \leq m \right\} = \left\lfloor \frac{-1+\sqrt{1+8m}}{2} \right\rfloor\,,$$

$$G(m) \equiv (g(m), m - g(m))\,,$$

where $\lfloor \cdot \rfloor$ denotes the lower integer part, i.e. the *floor* operation. Recall that $k_{\max} = \binom{\alpha_{\max}+1}{2}$ for a fixed positive integer $\alpha_{\max}$. Define the $k_{\max} \times k_{\max}$ matrix

$$A_{ij} \equiv \begin{cases} \delta_{i,j} - \mathcal{P}(G(i), G(j))\,, & i = 0, 1, \ldots, k_{\max} - 1\,, \\ 1 & i = k_{\max}\,, \end{cases} \tag{37}$$

where $\mathcal{P}(\cdot,\cdot)$ is given by (6)–(7), and

$$\pi_j \equiv P_{g(j),j-g(j)}\,, \qquad j=0,1,\ldots,k_{\max}\,. \tag{38}$$

Appendix B: Truncated system

By product of (32) and direct inspection of (6)–(7), for $i+j=\alpha_{\max}$,

$$\sum_{\substack{n,l\geq 0\\ n+l>\alpha_{\max}}} P_{n,l}\,\mathcal{P}((n,l),(i,j)) \leq (1-\rho)\frac{(-q;q)_\infty}{(q;q)_\infty}\rho^{\alpha_{\max}+1}q^{\binom{\alpha_{\max}+1}{2}}\,, \tag{39}$$

while for $i+j<\alpha_{\max}$,

$$\sum_{\substack{n,l\geq 0\\ n+l>\alpha_{\max}}} P_{n,l}\,\mathcal{P}((n,l),(i,j)) = 0\,, \tag{40}$$

where $P_{n,l}$ and $\mathcal{P}(\cdot,\cdot)$ are defined by (6)–(7) and (12), respectively. Also,

$$\sum_{\substack{n,l\geq 0\\ n+l>\alpha_{\max}}} P_{n,l} \leq 2\frac{(-q;q)_\infty}{(q;q)_\infty}\rho^{\alpha_{\max}+1}q^{\binom{\alpha_{\max}+1}{2}}\,. \tag{41}$$

References

Adan I, Kulkarni VG (2003) Single-server queue with Markov-dependent inter-arrival and service times. Queueing Syst 45(2):113–134

Adan I, Wessels J, Zijm WHM (1993) A compensation approach for two-dimensional Markov processes. Adv Appl Probab 25(4):783–817

Adan I, Economou A, Kapodistria S (2009) Synchronized reneging in queueing systems with vacations. Queueing Syst 62(1):1–33

Al-Osh MA, Alzaid AA (1987) First-order integer-valued autoregressive [inar (1)] process. J Time Ser Anal 8(3):261–275

Almaz OA, Altiok T (2012) Simulation modeling of the vessel traffic in Delaware River: impact of deepening on port performance. Simul Model Pract Theory 22:146–165

Altman E, Yechiali U (2006) Analysis of customers' impatience in queues with server vacations. Queueing Syst 52(4):261–279

Andrews GE (1998) The theory of partitions. Cambridge University Press, Cambridge

Artalejo JR, Economou A, Lopez-Herrero MJ (2007) Evaluating growth measures in an immigration process subject to binomial and geometric catastrophes. Math Biosci Eng MBE 4(4):573

Asmussen S, Kella O (2000) A multi-dimensional martingale for Markov additive processes and its applications. Adv Appl Probab 32(2):376–393

Bailey NTJ (1952) A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. J R Stat Soc Ser B (Methodol) 14(2):185–199

Ball MO, Vossen T, Hoffman R (2001) Analysis of demand uncertainty effects in ground delay programs. In: 4th USA/Europe air traffic management R&D seminar, pp 51–60

Bini D, Latouche G, Meini B (2005) Numerical methods for structured Markov chains. Oxford University Press, Oxford

Blanc JPC (1987a) A numerical study of a coupled processor model. University of Limburg, Maastricht
Blanc JPC (1987b) On a numerical method for calculating state probabilities for queueing systems with more than one waiting line. J Comput Appl Math 20:119–125
Bloomfield P, Cox DR (1972) A low traffic approximation for queues. J Appl Probab 9(4):832–840
Brockwell PJ, Gani J, Resnick SI (1982) Birth, immigration and catastrophe processes. Adv Appl Probab 14(4):709–731
Caccavale MV, Iovanella A, Lancia C, Lulli G, Scoppola B (2014) A model of inbound air traffic: the application to Heathrow airport. J Air Transp Manag 34:116–122
Cayirli T, Veral E (2003) Outpatient scheduling in health care: a review of literature. Prod Oper Manag 12(4):519–549
Cohen JW (1988) Boundary value problems in queueing theory. Queueing Syst 3(2):97–128
Cohen JW, Boxma OJ (1983) Boundary value problems in queueing system analysis, vol 79. North Holland, Amsterdam
Collings PS (1976) Limit theorems for processes of randomly displaced regular events. J Appl Probab 13(3):530–537
Combé MB, Boxma OJ (1998) BMAP modelling of a correlated queue. In: Walrand J, Bagchi K, Zobrist G (eds) Network performance modeling and simulation, chap 7. Gordon and Breach Science Publishers, Amsterdam, pp 177–196
Daganzo CF (1990) The productivity of multipurpose seaport terminals. Transp Sci 24(3):205–216
Drezner Z (1999) On a queue with correlated arrivals. J Appl Math Decis Sci 3(1):75–84
Economou A (2004) The compound poisson immigration process subject to binomial catastrophes. J Appl Probab 41(2):508–523
Economou A, Kapodistria S (2009) q-series in Markov chains with binomial transitions. Probab Eng Inf Sci 23(1):75
Economou A, Kapodistria S, Resing J (2010) The single server queue with synchronized services. Stoch Models 26(4):617–648
Edmond E (1975) Operation capacity of container berths for scheduled service by queuing theory. Dock Harbor Auth 56:230–234
Fayolle G, Iasnogorodski R (1979) Two coupled processors: the reduction to a Riemann–Hilbert problem. Probab Theory Relat Fields 47(3):325–351
Fayolle G, Iasnogorodski R, Malyshev V (1999) Random walks in the quarter-plane: algebraic methods, boundary value problems and applications, vol 40. Springer, Berlin
Gail HR, Hantler SL, Taylor BA (2000) Use of characteristic roots for solving infinite state Markov chains. Comput Probab 24:205–255
Gasper G, Rahman M (2004) Basic hypergeometric series, vol 96. Cambridge University Press, Cambridge
Golub G, Kahan W (1965) Calculating the singular values and pseudo-inverse of a matrix. J Soc Ind Appl Math Ser B Numer Anal 2(2):205–224
Golub G H, Van Loan C F (2012) Matrix computations, 4th edn. John Hopkins University Press, Baltimore
Govier LJ, Lewis T (1963) Stock levels generated by a controlled-variability arrival process. Oper Res 11(5):693–701
Grassmann WK (2002) Real eigenvalues of certain tridiagonal matrix polynomials, with queueing applications. Linear Algebra Appl 342(1):93–106
Guadagni G, Ndreca S, Scoppola B (2011) Queueing systems with pre-scheduled random arrivals. Math Methods Oper Res 73(1):1–18
Gwiggner C, Nagaoka S (2010) Analysis of fuel efficiency in highly congested arrival flows. In: Electronic Navigation Research Institute, Asia-Pacific international symposium on aerospace technology, Tokyo-Japan
Hardy GGH, Wright EM (1979) An introduction to the theory of numbers. Oxford University Press, Oxford
Hooghiemstra G, Keane M, Van De Ree S (1988) Power series for stationary distributions of coupled processor models. SIAM J Appl Math 48(5):1159–1166
Iovanella A, Scoppola B, Pozzi S, Tedeschi A (2011) The impact of 4D trajectories on arrival delays in mixed traffic scenarios. In: Schaefer D (ed) Proceedings of the SESAR innovation days
Jagerman D, Altiok T (2003) Vessel arrival process and queueing in marine ports handling bulk materials. Queueing Syst 45(3):223–243
Johnson CR (1989) A Gersgorin-type lower bound for the smallest singular value. Linear Algebra Appl 112:1–7
Kapodistria S (2011) The M/M/1 queue with synchronized abandonments. Queueing Syst 68(1):79–109

Kendall DG (1964) Some recent work and further problems in the theory of queues. Theory Probab Appl 9(1):1–13
Kingman JFC (1961) Two similar queues in parallel. Ann Math Stat 32(4):1314–1323
Koole G (1997) On the use of the power series algorithm for general Markov processes, with an application to a Petri net. INFORMS J Comput 9(1):51–56
Lancia C (2013) Looking through the cutoff window. Ph.D. Thesis, Technische Universiteit Eindhoven
Lancia C, Lulli G (2017) Data-driven modelling and validation of aircraft inbound-flow at some major European airports. *ArXiv preprint at* arXiv:1708.02486
Latouche G, Ramaswami V (1999) Introduction to matrix geometric methods in stochastic modeling. ASA-SIAM series on statistics and applied probability. SIAM, Philadelphia PA
Lucantoni DM (1991) New results on the single server queue with a batch Markovian arrival process. Stoch Models 7(1):1–46
Mercer A (1960) A queueing problem in which arrival times of the customers are scheduled. J R Stat Soc Ser B (Methodol) 22:108–113
Mercer A (1973) Queues with scheduled arrivals: a correction, simplification and extension. J R Stat Soc Ser B (Methodol) 35(1):104–116
Mitrani I, Chakka R (1995) Spectral expansion solution for a class of Markov models: application and comparison with the matrix-geometric method. Perform Eval 23(3):241–260
Nelsen RB, Williams T (1970) Random displacements of regularly spaced events. J Appl Probab 7(1):183–195
Neuts MF (1989) Structured stochastic matrices of M/G/1 type and their applications, vol 701. Marcel Dekker, New York
Neuts MF (1994) An interesting random walk on the non-negative integers. J Appl Probab 31(1):48–58
Neuts MF (1995) Matrix-geometric solutions in stochastic models: an algorithmic approach. Dover Publications, Mineola
Nikoleris T, Hansen M (2012) Queueing models for trajectory-based aircraft operations. Transp Sci 46(4):501–511
Pacheco A, Prabhu NU, Tang LC (2009) Markov-modulated processes and semiregenerative phenomena. World Scientific, Singapore
Qi L (1984) Some simple estimates for singular values of a matrix. Linear Algebra Appl 56:105–119
Robert P (2003) Stochastic networks and queues. Springer, Berlin
Sabria F, Daganzo CF (1989) Approximate expressions for queueing systems with scheduled arrivals and established service order. Transp Sci 23(3):159–165
Tijms HC (1994) Stochastic models: an algorithmic approach, vol 303. Wiley, Hoboken
Uchimura Y, Saitô K (2011) Stationary distributions of the bernoulli type Galton–Watson branching process with immigration. Commun Stoch Anal 5(3):457–480
Uchimura Y, Saitô K (2015) Asymptotic behavior of the Bernoulli type Galton–Watson branching process with immigration. Random Oper Stoch Equ 23(1):1–10
Winsten CB (1959) Geometric distributions in the theory of queues. J R Stat Soc Ser B (Methodol) 21(1):1–35
Wittevrongel S, Bruneel H (1999) Discrete-time queues with correlated arrivals and constant service times. Comput Oper Res 26(2):93–108
Yaglom AM, Yaglom IM (1964) Challenging mathematical problems with elementary solutions: combinatorial analysis and probability theory. Dover Publications, Mineola
Yechiali U (2007) Queues with system disasters and impatient customers when system is down. Queueing Syst 56(3):195–202