ON MULTISERVER RETRIAL QUEUES: HISTORY, OKUBO-TYPE HYPERGEOMETRIC SYSTEMS AND MATRIX CONTINUED-FRACTIONS

F. AVRAM

Département de Mathématiques, Avenue de l'Université
Pau, France
florin.avram@univ-pau.fr
http://web.univ-pau.fr/ avram/

D. MATEI

Institute of Mathematics "Simion Stoilow" of the Romanian Academy
Bucharest, P.O. Box 1-764, RO-014700, Romania
dmatei@imar.ro

Y. Q. ZHAO

School of Mathematics and Statistics, Carleton University
Ottawa, Canada
zhao@math.carleton.ca
http://people.math.carleton.ca/ zhao/

Received (17/07/2012)
Revised (4/08/2013)

We study two families of QBD processes with linear rates: (A) the multiserver retrial queue and its easier relative; and (B) the multiserver M/M/∞ Markov modulated queue.

The linear rates imply that the stationary probabilities satisfy a recurrence with linear coefficients; as known from previous work, they yield a "minimal/non-dominant" solution of this recurrence, which may be computed numerically by matrix continued-fraction methods.

Furthermore, the generating function of the stationary probabilities satisfies a linear differential system with polynomial coefficients, which calls for the venerable but still developing theory of holonomic (or D-finite) linear differential systems.

We provide a differential system for our generating function that unifies problems (A) and (B), and we also include some additional features and observe that in at least one particular case we get a special "Okubo-type hypergeometric system", a family that recently spurred considerable interest.

The differential system should allow further study of the Taylor coefficients of the expansion of the generating function at three points of interest: 1) the irregular singularity at 0; 2) the dominant regular singularity, which yields asymptotic series via classic methods like the Frobenius vector expansion; and 3) the point 1, whose Taylor series coefficients are the factorial moments.

Keywords: Retrial queue; stationary probabilities; Okubo hypergeometric system; minimal solution.

1. Introduction

Motivation: The practical questions of taking into account the blocking of queues, the dilemma of blocked customers either waiting or retrying later, and the eventual abandon of dissatisfied customers have been recognized by queueing theorists from the outstart – see for example Kosten (1947).

The modelling of these phenomena is more challenging mathematically than that of classical queues, since even the simplest M/M/1 + retrial case involves a bivariate modelling that tries to capture the interaction between two types of customers: those who "wait on the spot", and those who "postpone their requests for later".

A bit of history: the mystery of retrial queues with more than two servers. An impressive pioneering analysis of retrial queues stems from J.W. Cohen (1957), who obtained exact solutions in terms of contour integrals and Laguerre series expansions. For $s = 1$ and $s = 2$ servers, Cohen's results involve classic hypergeometric functions.

One of the striking features of the field is the "disappearing" of second-order classic hypergeometric equations for the generating function of the stationary probabilities when the number s of servers is bigger than 2. This has been noted several times in the literature – see, for example pp.25 of Kulkarni and Liang (1997) and pp.288 of Falin and Templeton (1997). Also, in the case of $s = 3$ servers, there is a considerable increase in complexity that forces one to turn to numerical methods as was done by Kim (1995) and Gómez-Corral and Ramalhoto (1999).

The non-Fuchsian singularity. Let us dwell further on this important type of singularity. The generating function of the stationary probability distribution satisfies the well-known Kolmogorov linear differential system. In our case, this has affine coefficients — see for example Riordan (1962), Keilson et al. (1968), Cohen (1957), Falin and Templeton (1997), Artalejo and Falin (2002, (2.2)), and our generalization (5.16).

With a low number of servers ($s \leq 2$), the system leads to a regular singularity at 0 (which explains the explicit hypergeometric solutions), while with a higher number of servers, it has a non-regular (non-Fuchsian) singularity.

Formal series solutions. The concept of classic (second order) hypergeometric functions has been fruitfully extended to that of functions satisfying holonomic (D-finite) equations, for example see Stanley (1980); Zeilberger (1990); Koepf (1997). The **general solution** of such equations falls straight under the blanket of the venerable Euler-Frobenius-Poincaré formal power series approach, which culminated recently in the developing of symbolic algebra tools like Maple's `DEtools` and `gfun` and Mathematica's `HolonomicFunctions`. Such programs easily identify the general formal series solutions of our special "Okubo-type" equations – see Section 7.

Non-dominant/minimal solutions. There is however one corner not covered well by the above mentioned blanket! Our generating function must be in the ergodic case the unique (up to a constant) analytic solution in the unit circle having positive Taylor coefficients around 0 (as ensured by ergodic theory), despite the fact that the equation under study has an irregular singularity at 0. This phenomenon seems less studied, and thus, while computer programs will provide the subspace of all (typically divergent) power series at 0, they will not provide direct information on the unique generating function/analytical solution with positive coefficients at 0 that is of interest in probability and combinatorics.

This difficulty has already been investigated by Pearce (1985) in a remarkable uncited paper, where he notes that at an irregular singular point there might be two "formal power series solutions that are divergent (have zero radius of convergence) but that nevertheless may be combined into a holomorphic solution with a positive radius of convergence in the vicinity of the singularity. This phenomenon does not appear to have been noted previously in the literature. The approach of expressing an arbitrary solution about an irregular singular point as a sum of solutions of known asymptotic behaviour is not very convenient for the investigation of regular solutions. In practice, the determination of regular solutions centers about the indicial equation and around the determination of the non-dominant/minimal solution for the recurrence satisfied by the formal series coefficients, for example see Pearce (1985, Ch 3).

The continued fraction construction of minimal solutions. For second order recurrences, the minimal solution may be obtained by a theorem of Pincherle, Jones and Thron (1980, Thm B.4, pg. 403), or Gautschi (1967), which constructs it by using continued fractions.

Phung-Duc et al. (2009; 2010b) showed that with three or four servers, the recurrence satisfied by the stationary distribution is of second order, and indeed it is a minimal solution of this recurrence. Furthermore, they constructed it as a linear combination of two explicit independent solutions of the recurrence which diverge (tend to infinity as the number of customers in the orbit tends to infinity), following an approach initiated by Choi et al. (1998).

Beyond second order recurrences, one needs a generalization of Pincherle's result to higher order recurrence relations, for example see Pearce (1989), and such a **matrix continued fractions generalization** was provided by Levrie and Bultheel (1996).

The matrix continued fractions approach to retrial queues (building on Pearce (1985, Thm. 1), Pearce (1989), and Levrie and Bultheel (1996)) was introduced by Hanschke (1999), who emphasized matrix-algorithmic aspects, similar to those of computing the R_n matrices of Neuts (the blocks of the RG-factorization in Li and Zhao (2004)).

Recently, Phung-Duc et al. (2010a, 2011) and Baumann and Sandmann (2010) have provided impressive applications of the matrix-continued approach to retrial queues and to general level dependent QBD's, respectively. These papers turn convergence results for matrix continued fractions into a practical tool (quite related to the R and G matrix algorithmic approach of Neuts).

Many other intriguing facts about retrial queues– exact results, asymptotics, approximations and probabilistic limit theorems – may be found in Falin and Templeton (1997); Artalejo and Falin (2002), and in the more recent literature. Let us end by listing a few of these intriguing facts:

(1) In the limit to zero retrial rate, Cohen (1957) discovered that certain retrial queues without a waiting buffer behave like a classic Erlang-loss system. However, the retrial buffer does not simply disappear, but gives rise to an increase in the arrival rate to the system (for a fine analysis of this phenomenon in the Halfin-Whitt regime, see Avram et al. (2013), Janssen (2012 (to appear))).

Note that the Cohen discontinuity from Cohen (1957) when the retrial rate converges to 0 has only been established for exponential transition times. The matrix approach advocated below might be useful for investigating this phenomenon under phase-type transition times.

(2) A fundamental result with one server is the "stochastic decomposition" of the generating function as a product, which contains the generating function of a limit model, for example see Artalejo and Falin (1994); Atencia and Moreno (2003); Krishnamoorthy et al. (2012 (online first: DOI: 10.1007/s11750-012-0256-6)).
(3) Especially interesting is the refining of the existing simple approximations like the RTA, or constant retrial rate approximation, and the Fredericks and Reisner approximation – see Wolff (1989, Ch 7), and the Grier et al. (1997) approximation.
(4) Finally, the role of censoring, which turned out to be crucial in the analysis of certain retrial models such as Liu and Zhao (2010); Liu et al. (2012), and Kim et al. (2012), deserves further investigation.
(5) The relation of the matrix continued approach to the Laguerre series expansions of Cohen (1957) is not understood.

Our motivation: The results mentioned above have typically been investigated for "simple" particular cases of retrial models. We feel that this sometimes obscures the underlying mathematical structure, and therefore decided to start again at the base: Kolmogorov's linear differential systems for the generating function of stationary probabilities for QBD's (5.16) and retrial queues (5.18), respectively. As a first bonus, we compute the dominant singularity.

As a second bonus, we hope to obtain in the future a full asymptotic expansion of the stationary probabilities, extending results of Liu et al. (2012), and Kim et al. (2012). In principle, this could be achieved by applying the vector Frobenius method, a new treatment that has been provided recently by Kim et al. (2012, Theorem 1), but this requires further work, as explained below.

Contents: Retrial queues are a particular case of quasi-birth-and-death processes (QBD's), and some background material on this family is presented in Section 2.

Our key definition, a general "advanced multiserver retrial model", is presented in Section 3. The generality adopted here helps us to organize various open questions (the study of which we hope to undertake in the future). Section 4 deals with the special Markovian case, a linear level-dependent QBD process, which subsumes several interesting queueing features like geometric acceptance, abandon, feedback etc.

Section 5 collects some simple but fundamental features of our differential system, namely:

(1) the matrices intervening in the Kolmogorov equations — see Lemmas 5.1, 5.2,
(2) a discussion of the corresponding singularities, and
(3) the stability condition ensuring analicity of the generating function in the unit circle — see Remark 6.2.

Section 6 contains results for persistent models.

Section 7 makes an interesting observation that persistent retrials with no feedback lead to a recently introduced type of "Okubo-type hypergeometric systems" (which always have an irregular singularity at 0 when $s \geq 3$).

Section 8 considers the asymptotic behavior of stationary probabilities (assuming a positive retrial rate). Note that the cases $s = 1, s = \infty$, and $1 < s < \infty$ correspond respectively to one variable scalar generating functions (extensively-studied — see for example Odlyzko (1995); Flajolet and Sedgewick (2009)), two variable scalar generating functions, essentially an open problem, whose study has only recently been tackled by Baryshnikov and Pemantle (2004); Pemantle and Wilson (2005); Pemantle (2010); Raichev and Wilson (2010), and Baryshnikov and Pemantle (2011), and to one variable vector generating functions. In this latter case, asymptotic expansions are available in principle by the vector Frobenius method — see Coddington and Levinson (1955); Wasow (2002), and Barkatou and Pflügel (1999). First terms of such expansions have already been obtained by Liu and Zhao (2010); Liu et al. (2012); Kim et al. (2012), and we are working currently on extending these to full asymptotic expansions.

2. Quasi-birth-and-death processes: a general framework for bivariate Markovian modelling

Introduction: Many important stochastic models involve multidimensional random walks with discrete state space, whose coordinates split naturally into:

(1) **one infinite valued coordinate** $N(t) \in \mathbb{N} = \{0, 1, 2, \ldots\}$, called *level*, and
(2) "the rest of the information" $I(t)$, called *phase or background*, which typically takes a **finite number of possible values**.

Partitioned according to the level, the infinitesimal generator Q of such a Markov process $(I(t), N(t))$ is a **block tridiagonal** matrix, called *level-dependent* QBD generator (LD-QBD):

$$Q = \begin{bmatrix} B_0 & A_0 & & & \\ C_1 & B_1 & A_1 & & \\ & C_2 & B_2 & A_2 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \tag{2.1}$$

(recall that Q is a matrix with nonnegative off-diagonal "rates/weights", and with row sums equal to 0). In the above and throughout the paper, all unspecified entries of a matrix are zero.

LD-QBD processes share the "skip free" structure of birth-and-death-processes; however, the "**level transition weights**" A_n, B_n, C_n are now **matrices**, inviting one to enter the noncommutative world.

Stationary distribution: One challenging problem of great interest for **positive recurrent** LD-QBD processes is the determination of the stationary distribution $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)$ partitioned by level, where $\boldsymbol{\pi}_n = (\pi_{0,n}, \pi_{1,n}, \pi_{2,n}, \ldots)$. The equilibrium equations

$$\boldsymbol{\pi} Q = 0 \tag{2.2}$$

in partitioned form yield the second degree vector recursion:

$$\begin{cases} \boldsymbol{\pi}_{n-1} A_{n-1} + \boldsymbol{\pi}_n B_n + \boldsymbol{\pi}_{n+1} C_{n+1} = 0, & n = 1, 2, \ldots \\ \boldsymbol{\pi}_0 B_0 + \boldsymbol{\pi}_1 C_1 = 0 \end{cases} \tag{2.3}$$

Note the absence of an initial value for $\boldsymbol{\pi}_0$! (Instead, the sequence is determined by the requirement of summability of $\boldsymbol{\pi}_n(i)$ in $n, \forall i$.)

LD-QBD processes: Level independent QBD processes have been intensively studied, but the level-dependent case is less understood due to the generality of the model. For some recent works discussing LD-QBD processes, see Bright and Taylor (1995); Ramaswami (1996); Li and Cao (2004); Li et al. (2006); Li (2010), and Avram and Fotso-Chedom (2011).

Despite recent theoretic and algorithmic progress in the study of formal series solutions, there is an intriguing scarcity of analytic works in the particular "holonomic" case of linear or polynomial dependence on the level, although such recurrences and the corresponding generating functions have been central in classic works of Euler, Gauss, Riemann, Liouville, Klein, Fuchs, Frobenius, Pincherle, Poincaré, Lie, Hilbert, Birkhoff, etc (see Gray (2008) for a delightful account), and despite recent theoretic and algorithmic progress in the study of formal series solutions. Narrowing the gap between this venerable mathematics and our intriguing applications has in fact been our motivation for this work.

We end by noting that very efficient numerical approaches for computing stationary probabilities for level-dependent QBD's have been provided recently by Baumann and Sandmann (2010); Phung-Duc et al. (2010a), and Phung-Duc et al. (2011).

3. The GI/G/s/K queue with retrial, Bernoulli acceptance, abandon and feedback

The inclusion of this general model in a paper which deals with a modified $M/M/s/K$ Markovian model requires some justification. We are motivated by the fact, as recognized in Hanschke (1999), that when the primary area has no free servers, first time arrivals face one of three choices: a) wait in the primary area (in the priority queue), b) abandon, and c) join the orbit for later retrials. The same three choices may be made after each service by the customer who has just completed his service (since feedback is allowed in our model), and retrial customers also face the first two choices. However, several recent fundamental papers still assume that one of these three options (which changes from paper to paper) may not happen.

One such example, and for us an especially important one, is the generalized Pollaczek-Khinchine formula of Atencia and Moreno (2003). We feel that their generating function formula Atencia and Moreno (2003, Thm 3.1) marks an important moment in the evolution of retrial queues, since a) it is general enough to include many particular cases studied previously, and b) it allows for general service times, so that it may be applied directly to empirical data.

Unfortunately, some of the three natural retrial options are still assumed not to happen in (2003); this motivated us to formulate the general three-choice model below, so that we may state an important open problem:

Question 1. Generalize the Atencia-Moreno formula for the one server, and eventually for the multi-server case of the model below.

Definition 1. The **overflow GI/G/s/K queue with retrial rate ν, orbit abandon probability p, Bernoulli acceptance probability p_a and feedback to orbit probabil-**

ity θ consists of two facilities: (1) a "primary service area" with s service places (or servers) and K extra waiting spaces (referred to as a priority queue), the only cases studied here being $K = 0$ and $K = \infty$. "Primary" customers arrive to the primary service area according to a renewal process with rate λ (the case of interest for us being Poisson arrivals); and (2) an additional infinite "overflow buffer" (referred to as an orbit).

The main "activities" of interest in the system are:

(1) Arrivals to the primary area with rate λ. If at least one server is free upon arrival, the customer is accepted with probability p_a for first time customers, sent to orbit with probability $\tilde{p}_a$, or forced to leave the system for some obscure reason with probability $\bar{p}_a = 1 - p_a - \tilde{p}_a$. However, if all servers are busy upon arrival, but $0 < K < \infty$ and there is still space in the priority queue, then these probabilities become $\beta_0, \tilde{\beta}_0, \bar{\beta}_0 = 1 - \beta_0 - \tilde{\beta}_0$ (acceptance means in this case that customers will wait in the priority queue of the primary area). To simplify notation, we will assimilate this case with the first, i.e. assume that $p_a = \beta_0, \tilde{p}_a = \tilde{\beta}_0, \ldots$ (since we consider only $K = 0$ below), keeping in mind that in certain applications a distinction might need to be made. Finally, if the servers and the priority queue are blocked, these probabilities become $\alpha_0 = 0, \tilde{\alpha}_0, \bar{\alpha}_0 = 1 - \tilde{\alpha}_0$. We denote the rates of the three possibilities, with available servers and after blocking of the primary area, respectively, by

$$\begin{cases} \lambda_q = \lambda p_a & \text{join the primary area} \\ \lambda_a = \lambda \bar{p}_a & \text{leave the system} \\ \lambda_o = \lambda \tilde{p}_a & \text{join the orbit} \end{cases} \qquad \begin{cases} \lambda_{qb} = 0 \\ \lambda_{ab} = \lambda \bar{\alpha}_0 \\ \lambda_{ob} = \lambda \tilde{\alpha}_0 \end{cases}$$

Remark 3.1. Our model is related to a model proposed by Hanschke (1999), essentially by adding feedback, and to the model of Atencia and Moreno (2003). To clarify this, we include here first Hanschke's, and then Atencia and Moreno's notations for the parameters (they correspond to certain probabilistic interpretations, which do not play a role in the mathematical analysis).

$$\begin{cases} \lambda_q = \lambda p \ (\mathcal{H}) = \lambda \ (\mathcal{AT}) & \text{join the primary area} \\ \lambda_a = \lambda(1-p)(1-r_1) \ (\mathcal{H}) = 0 \ (\mathcal{AT}) & \text{leave the system} \\ \lambda_o = \lambda(1-p) r_1 \ (\mathcal{H}) = 0 \ (\mathcal{AT}) & \text{join the orbit} \end{cases} \qquad \begin{cases} \lambda_{qb} = 0 \ (\mathcal{H}) = \lambda q \ (\mathcal{AT}) \\ \lambda_{ab} = \lambda(1-q_1) \ (\mathcal{H}) = 0 \ (\mathcal{AT}) \\ \lambda_{ob} = \lambda q_1 \ (\mathcal{H}) = \lambda p \ (\mathcal{AT}) \end{cases}$$

Clearly, adopting a unified notation would be beneficial.

(2) Leaving the orbit for retrial or abandon. The orbit may decrease by one at an "affine" rate $\nu(i) = \nu^{(0)} \, \mathbb{1}_{i>0} + i\nu$. The first term, which may be interpreted as the effect of service by one secondary server, will be assumed to be 0 throughout the paper. The second term is due to each customer reevaluating his position of remaining in the orbit at rate ν. At these evaluation times, he is informed first on the state of the primary area (free service, free waiting space, blocked primary area). Then, he abandons the system with probability $\bar{\alpha}$ or $\bar{p}$, depending on whether the primary area is blocked or not. In the second case the remaining proportion of $p = 1 - \bar{p}$ will go to the primary area. In the first, the remaining proportion of $\tilde{\alpha} = 1 - \bar{\alpha}$ will remain in the orbit. All customers in the orbit repeatedly reevaluate their position until a free place is secured in the primary area (either an idle server or in the priority queue), referred to as a successful retrial, or until abandon. We will denote the respective rates of going to the primary area and abandoning before and after blocking by

$$\begin{cases} \nu_q = \nu p & \text{join the primary area,} \\ \nu_a = \nu \bar{p} & \text{leave the system} \end{cases} \begin{cases} \nu_{qb} = 0 \\ \nu_{ab} = \nu \bar{\alpha} \end{cases}$$

When $\bar{\alpha} = \bar{p} = 0$, the retrials will be called persistent.

In Hanschke (1999), the respective rates are denoted by

$$\begin{cases} \nu_q = \nu p & \text{join the primary area} \\ \nu_a = \nu(1-p)(1-r_2) & \text{leave the system} \end{cases} \quad \begin{cases} \nu_{qb} = 0 \\ \nu_{ab} = \nu(1-q_2) \end{cases}$$

Remark 3.2. Note that we have not included a third possibility of rejoining the orbit, which could be appropriate in a discrete time model but not in a continuous time model, since staying put does not produce a transition rate in this case. For arrivals, it is also possible to similarly omit the abandoning arrivals in the specification of the model. This amounts mathematically to assuming that $\bar{p}_a = 1$, and letting λ denote the rate of nonrejected arrivals, as we will do from now on.

Remark 3.3. The parameter ν may be interpreted as the total activity rate per individual in the orbit (retrial + abandon due to impatience), and p and $\bar{p}$ as the respective probabilities of these activities; normally, one would expect $\bar{\alpha} \geq \bar{p}$ i.e. a bigger probability of leaving due to congestion.

Note also that from the mathematical modelling point of view, impatience is identical to having an additional infinite server dispatching the customers.

(3) Service and feedback. For each of the accepted first time or retrial customers, the service time follows a general distribution with affine service rate (=inverse of expected service time) of rate $\mu(i) = \mu^{(0)} \mathbb{I}_{i>0} + i\mu$. We will consider mostly the case when $\mu(i)$ is either constant or linear, with exponential service times. All service times are independent, and are also independent of the arrival, retrial and routing processes.

Feedback: leaving the system/joining the orbit after service. After service in the primary area, the customer may leave the system forever with probability $\bar{\theta}$, request an additional service in the primary area with probability θ, or join the orbit with probability $\tilde{\theta} = 1 - \theta - \bar{\theta}$ (the "no feedback" factor $\bar{\theta}$ is 1 in the standard model). We will denote the rates of these three possibilities by

$$\begin{cases} \mu_q = \mu\theta \\ \mu_a = \mu\bar{\theta} \\ \mu_o = \mu\tilde{\theta} \end{cases}$$

Here we have followed the notation of Atencia and Moreno (2003), while adding the third case with parameter $\tilde{\theta} \geq 0$ (note that since feedback is not affected by possible blocking of the primary area, we only need one set of rates in this case).

We included the above general model because we hope that it will encourage the emergence of generalizations of the Atencia-Moreno formula (2003, Thm 3.1). However, as is common in the multiserver literature, we will now go on to consider the M/M/s/K retrial model.

4. The Markovian model with exponential transition times

We now consider the case of exponential arrivals, services and retrials.

Definition 2. An **affine death QBD process** is a QBD process with affine dependence on the level:

$$\begin{aligned} A_n &= A, \quad C_n = nC + C^{(0)} 1_{\{n>0\}}, \\ B_n &= B - \tilde{A} - n\tilde{C} - \tilde{C}^{(0)} 1_{\{n>0\}} \end{aligned} \tag{4.8}$$

where B is a conservative generator, and $\tilde{A}$, $\tilde{C}^{(0)}$ and $\tilde{C}$ are diagonal matrices containing on their diagonals the sum of the rows of A, $C^{(0)}$ and C, respectively.

The two terms in C_n correspond to the cases of "**independent retrials**" and "**one retrial dispatcher**" in the orbit, respectively, and we here are mainly interested in the linear retrial case with $C_n = nC$ and $C^{(0)} = 0$, and in the constant retrial case $C = 0$, when we will denote $C^{(0)}$ simply by C.

Let $(N(t), I(t))$ denote respectively the numbers of customers in the orbit and in the service area at time t for a Markovian retrial system. Consider first the case with s servers and no extra waiting spaces ($K = 0$), assuming for now $C^{(0)} = 0$. Then, the retrial process $(N(t), I(t))$ with respective geometric losses $\bar{p}_0 = 1 - p \in [0, 1]$ and $\bar{\alpha}_0 = 1 - \alpha_0 \in [0, 1]$, orbit abandon probabilities $\bar{p} = 1 - p \in [0, 1]$ and $\bar{\alpha} = 1 - \alpha \in [0, 1]$, and feedback $\theta \in [0, 1]$ is a linear death QBD process with departure matrix $C_n = nC$, where:

$$C = \begin{bmatrix} \nu\bar{p} & \nu p & & & \\ & \nu\bar{p} & \nu p & & \\ & & \ddots & \ddots & \\ & & & \nu\bar{p} & \nu p \\ & & & & \nu\bar{\alpha} \end{bmatrix}$$

and with arrival and "loop" (transitions without level change) matrices:

$$A = \begin{bmatrix} \lambda\tilde{p}_a & & & & & \\ \mu\tilde{\theta} & \lambda\tilde{p}_a & & & & \\ & 2\mu\tilde{\theta} & \lambda\tilde{p}_a & & & \\ & & \ddots & \ddots & & \\ & & & (s-1)\mu\tilde{\theta} & \lambda\tilde{p}_a & \\ & & & & s\mu\tilde{\theta} & \lambda\tilde{\alpha}_0 \end{bmatrix} \tag{4.10}$$

$$B = \begin{bmatrix} -\lambda p_a & \lambda p_a & & & & \\ \mu\bar{\theta} & -(\lambda p_a + \mu\bar{\theta}) & \lambda p_a & & & \\ & 2\mu\bar{\theta} & -(\lambda p_a + 2\mu\bar{\theta}) & \lambda p_a & & \\ & & \ddots & \ddots & \ddots & \\ & & & (s-1)\mu\bar{\theta} & -(\lambda p_a + (s-1)\mu\bar{\theta}) & \lambda p_a \\ & & & & s\mu\bar{\theta} & -s\mu\bar{\theta} \end{bmatrix} \tag{4.11}$$

where $p + \bar{p} = 1$ and $\bar{p}_a + p_a = 1$.

Example 4.1. In the retrials model with $p_a = 1$ (all accepted when a server is available), the arrivals $A_n = A$ and "loops" are:

$$A = \begin{bmatrix} 0 & & & & & \\ \mu\tilde{\theta} & 0 & & & & \\ & 2\mu\tilde{\theta} & 0 & & & \\ & & \ddots & \ddots & & \\ & & & (s-1)\mu\tilde{\theta} & 0 & \\ & & & & s\mu\tilde{\theta} & \lambda_{ob} \end{bmatrix}$$

$$B = \begin{bmatrix} -\lambda & \lambda & & & & \\ \mu\bar{\theta} & -(\lambda + \mu\bar{\theta}) & \lambda & & & \\ & 2\mu\bar{\theta} & -(\lambda + 2\mu\bar{\theta}) & \lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & (s-1)\mu\theta & -(\lambda + (s-1)\mu\bar{\theta}) & \lambda \\ & & & & s\mu\bar{\theta} & -s\mu\bar{\theta} \end{bmatrix}$$

Fig. 1. States and transitions of the classic $M/M/3/3$ retrial queue.

The entries of A_n, B_n and C_n correspond respectively to upward, level and downward arrows in figure 1.

Remark 4.1. Our model includes three parameters which are less studied:

(1) the "abandon retrials if blocked" parameter $\bar{\alpha}$,
(2) the "impatience abandon retrials" parameter $\bar{p}$, which allows us to also include the easier case of Markov modulated M/M/∞ queues, characterized by $\bar{p} = 1$, which is considered for example in Economou (2005), and
(3) the feedback parameter θ.

5. The differential system for the generating functions

Let $\pi_{i,n}$ denote the stationary probabilities of having i customers in the primary area and n customers in the orbit, which satisfy the recursion (2.3). A classic approach for tackling this is via the **generating functions**

$$p_i(z) = \sum_{n=0}^{\infty} \pi_{i,n} z^n, i = 0, 1, 2 \ldots \iff \mathbf{p}(z) = \sum_{n=0}^{\infty} \boldsymbol{\pi}_n z^n, \tag{5.14}$$

where $\mathbf{p}(z) = (p_0(z), p_1(z), \ldots)$.

Lemma 5.1. a) For the affine death QBD process in (4.8) (with finite s and finite K), the recursion

$$\begin{cases} \boldsymbol{\pi}_{n-1} A + \boldsymbol{\pi}_n (B - \tilde{A} - n\tilde{C} - C^{(0)}) + \boldsymbol{\pi}_{n+1}((n+1)C + C^{(0)}) = 0, \ n = 1, 2, \ldots \\ \boldsymbol{\pi}_0 (B - \tilde{A}) + \boldsymbol{\pi}_1 (C + C^{(0)}) = 0 \end{cases} \tag{5.15}$$

yields a linear differential system

$$\mathbf{p}'(z) V(z) = \mathbf{p}(z) U(z) + \boldsymbol{\pi}_0 (\tilde{C}^{(0)} - z^{-1} C^{(0)}), \tag{5.16}$$

where $V(z)$ and $U(z)$ are square matrices of order $s + 1 + K$, given by:

$$U(z) = B + zA - \tilde{A} + z^{-1} C^{(0)} - \tilde{C}^{(0)}, \quad V(z) = z\tilde{C} - C \tag{5.17}$$

b) In the particular case of the model of Definition 1 with exponential transition times, $C^{(0)} = \tilde{C}^{(0)} = 0$, $K = 0$, and $1 \le s < \infty$, (5.16) holds with $V(z)$ and $U(z)$ given by:

(1) For $s = 1$,

$$V(z) = \nu \begin{bmatrix} z - \bar{p} & -p \\ 0 & \bar{\alpha}(z-1) \end{bmatrix}, \quad U(z) = \nu \begin{bmatrix} \tilde{\lambda}(z\tilde{p}_0 - 1) & \tilde{\lambda} p_0 \\ \tilde{\mu}(\bar{\theta} + \tilde{\theta} z) & \tilde{\lambda}_{ob}(z-1) - \tilde{\mu}(\bar{\theta} + \tilde{\theta}) \end{bmatrix}.$$

(2) For $2 \le s < \infty$:

$$V(z) = \nu^{-1}(z\tilde{C} - C) = \begin{bmatrix} z - \bar{p} & -p & & & \\ & z - \bar{p} & -p & & \\ & & \ddots & \ddots & \\ & & & z - \bar{p} & -p \\ & & & & \bar{\alpha}(z-1) \end{bmatrix} \tag{5.18}$$

$$U(z) = \nu^{-1}(B - \tilde{A} + zA) =$$
$$\begin{bmatrix} \Theta_0 & \tilde{\lambda} p_0 & & & & \\ \tilde{\mu}(\bar{\theta} + \tilde{\theta} z) & \Theta_1 & \tilde{\lambda} p_0 & & & \\ & 2\tilde{\mu}(\bar{\theta} + \tilde{\theta} z) & \Theta_2 & \tilde{\lambda} p_0 & & \\ & & \ddots & \ddots & \ddots & \\ & & & (s-1)\tilde{\mu}(\bar{\theta} + \tilde{\theta} z) & \Theta_{s-1} & \tilde{\lambda} p_0 \\ & & & & s\tilde{\mu}(\bar{\theta} + \tilde{\theta} z) & \tilde{\lambda}_{ob}(z-1) - s\tilde{\mu}(\bar{\theta} + \tilde{\theta}) \end{bmatrix}$$

where $\tilde{\lambda} = \frac{\lambda}{\nu}$, $\tilde{\mu} = \frac{\mu}{\nu}$ and

$$\Theta_k = \tilde{\lambda}(z\tilde{p}_a - 1) - k\tilde{\mu}(\bar{\theta} + \tilde{\theta}), \ k = 0, 1, \ldots, (s-1). \tag{5.19}$$

The bivariate **generating function**

$$\varphi(y,z) := \sum_{i=0}^{s}\sum_{n=0}^{\infty}\pi_{i,n}y^i z^n$$

satisfies the equation:

$$\begin{aligned}\nu(z-\bar{p}-py)\varphi'_z(y,z)+\nu y^s(\bar{p}-\bar{\alpha}+z(\bar{\alpha}-1)-py)p'_s(z) = \\ \lambda\varphi(y,z)(\tilde{p}_0 z-1+p_0 y)+\mu\varphi'_y(y,z)(\bar{\theta}+\tilde{\theta}z-y(\bar{\theta}+\tilde{\theta})) \\ +y^s(z(\lambda_{ob}-\lambda\tilde{p}_0)-(\lambda_{ob}-\lambda)-\lambda p_0 y)p_s(z)\end{aligned} \tag{5.20}$$

c) In the infinite buffer case $K=\infty$, the differential system $\mathbf{p}'(z)V(z)=\mathbf{p}(z)U(z)$ formed from (5.22), (5.23) and (5.24) involves now infinite square matrices. For $s=1$ for example, the V and U matrices are given respectively by:

$$V(z) = \begin{bmatrix} z-\bar{p} & -p & & & \\ & (\bar{\alpha}+\alpha)z-\bar{\alpha} & -\alpha & & \\ & & (\bar{\alpha}+\alpha)z-\bar{\alpha} & -\alpha & \\ & & & \ddots & \ddots \\ & & & & \ddots \end{bmatrix}$$

and

$$U(z) = \begin{bmatrix} \tilde{\lambda}(z\tilde{p}_0-1) & \tilde{\lambda}p_0 & & & \\ \tilde{\mu}(\bar{\theta}+\tilde{\theta}z) & \tilde{\lambda}(\tilde{\alpha}_0 z-\tilde{\alpha}_0-\alpha_0)-\tilde{\mu}(\bar{\theta}+\tilde{\theta}) & \tilde{\lambda}\alpha_0 & & \\ & \tilde{\mu}(\bar{\theta}+\tilde{\theta}z) & \tilde{\lambda}(\tilde{\alpha}_0 z-\tilde{\alpha}_0-\alpha_0)-2\tilde{\mu}(\bar{\theta}+\tilde{\theta}) & \tilde{\lambda}\alpha_0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

Proof: a) Multiplying the equilibrium equations in (5.15) by z^n and summing up gives rise to a linear first order differential system:

$$\begin{aligned}&\sum_{n=0}^{\infty} z^n\boldsymbol{\pi}_{n-1}A+z^n\boldsymbol{\pi}_n(B-\tilde{A})-\boldsymbol{\pi}_n z^n(n\tilde{C}+\tilde{C}^{(0)}1_{\{n>0\}})+z^n\boldsymbol{\pi}_{n+1}((n+1)C+C^{(0)}) = \\ &z\boldsymbol{p}(z)A+\boldsymbol{p}(z)(B-\tilde{A}-\tilde{C}^{(0)})+\boldsymbol{\pi}_0\tilde{C}^{(0)}-z\boldsymbol{p}'(z)\tilde{C}+\boldsymbol{p}'(z)C+z^{-1}(\boldsymbol{p}(z)-\boldsymbol{\pi}_0)C^{(0)}=0\end{aligned}$$

which yields (5.16).

b,c) It is convenient to associate a number with each of the activities of the system. Let us denote by 1, 2, ..., 6, the inflow activities associated with the rates $\lambda_w,\lambda_{ob},\mu_o,\mu_a,\nu_a,\nu_w$, respectively, and by $\langle 12\rangle$, $\langle 34\rangle$ and $\langle 56\rangle$ the corresponding outflow activities (the distinction between the blocked and non-blocked cases is not made here).

It is easy to check that the **generating function terms associated with respective activities** are

1	$\lambda p_0 p_{i-1}(z)$	$\lambda \alpha_0 p_{i-1}(z)$	$i \geq s+1$
2	$\lambda \tilde{p}_0 z p_i(z)$	$\lambda \tilde{\alpha}_0 z p_i(z)$	$i \geq s$
$\langle 12 \rangle$	$\lambda(p_0 + \tilde{p}_0) p_i(z)$	$\lambda(\alpha_0 + \tilde{\alpha}_0) p_i(z)$	$i \geq s$
3	$(i+1)\mu\tilde{\theta} z p_{i+1}(z)$	$\dots$	$i \geq 0$
4	$(i+1)\mu\bar{\theta} p_{i+1}(z)$	$\dots$	$i \geq 0$
$\langle 34 \rangle$	$\mu i(\bar{\theta} + \tilde{\theta}) p_i(z)$	$\mu i(\bar{\theta} + \tilde{\theta}) p_i(z)$	$i \geq s$
5	$\nu \bar{p} p_i'(z)$	$\nu \bar{\alpha} p_i'(z)$	$i \geq s$
6	$\nu \alpha p_{i-1}'(z)$	$\nu \alpha p_{i-1}'(z)$	$i \geq s+1$
$\langle 56 \rangle$	$\nu(p + \bar{p}) z p_i'(z)$	$\nu(\alpha + \bar{\alpha}) z p_i'(z)$	$i \geq s$

The equilibrium equations may be written symbolically as

$$-6 + \langle 56 \rangle - 5 = 1 + 2 - \langle 12 \rangle - \langle 34 \rangle + 4 + 3 \tag{5.21}$$

In the interior, for $i = 1, \dots, s-1$, using the table this results in

$$\begin{aligned} &\nu\left(-p p_{i-1}'(z) + ((p+\bar{p})z - \bar{p}) p_i'(z)\right) = \\ &\lambda p_0 p_{i-1}(z) + \lambda(\tilde{p}_0 z - p_0 - \tilde{p}_0) - \mu i\left(\bar{\theta} + \tilde{\theta}\right) p_i(z) + (i+1)(\bar{\theta} + \tilde{\theta} z) p_{i+1}(z) \end{aligned} \tag{5.22}$$

as stated.

The exterior equilibrium equation for $i = s+1, s+2, \dots$ (also represented symbolically in (5.21)) is, due to the changes in rates:

$$\begin{aligned} &\nu\left(-\alpha p_{i-1}'(z) + ((\alpha+\bar{\alpha})z - \bar{\alpha}) p_i'(z)\right) = \\ &\lambda \alpha_0 p_{i-1}(z) + \left(\lambda(\tilde{\alpha}_0 z - \alpha_0 - \tilde{\alpha}_0) - \mu i(\bar{\theta} + \tilde{\theta})\right) p_i(z) + (i+1)(\bar{\theta} + \tilde{\theta} z) p_{i+1}(z) \end{aligned} \tag{5.23}$$

Finally, the boundary equilibrium equation for $i = s$ is

$$\begin{aligned} &\nu\left(-\alpha p_{i-1}'(z) + ((\alpha+\bar{\alpha})z - \bar{\alpha}) p_i'(z)\right) = \\ &\lambda p_0 p_{i-1}(z) + \left(\lambda(\tilde{\alpha}_0 z - \alpha_0 - \tilde{\alpha}_0) - \mu i(\bar{\theta} + \tilde{\theta})\right) p_i(z) + (i+1)(\bar{\theta} + \tilde{\theta} z) p_{i+1}(z) \end{aligned} \tag{5.24}$$

Remark 5.1. The bivariate generating function equation (5.20) generalizes the equation Falin and Templeton (1997, (2.21)):

$$\nu(z-y)\varphi_z'(y,z) - \nu y^s (z-y) p_s'(z) = \lambda \varphi(y,z)(-1+y) + \mu \varphi_y'(y,z)(1-y) + \lambda y^s (z-y) \varphi_s(z).$$

Remark 5.2. The Markov modulated M/M/∞ case. In the particular case of a Markov modulated M/M/∞ queue, with environment transitions specified by B, the arrival and departure matrices A and C are diagonal, $C^{(0)} = 0$, and the differential system (5.16) for the generating function becomes

$$(1-z)\mathbf{p}'(z)C = \mathbf{p}(z)((1-z)A - B), \tag{5.25}$$

or

$$\mathbf{p}'(z) = \mathbf{p}(z)(\hat{A} - (1-z)^{-1}\hat{B}), \quad \hat{A} = AC^{-1}, \hat{B} = BC^{-1}, \tag{5.26}$$

when C is invertible.

In this case, O'Cinneide and Purdue (1986, Thm 3.1) expand (5.25) in power series around $z = 1$, thus getting a recurrence for the conditional factorial moments:

$$\mathbf{p}^{(k)}(1)(kC - B) = k\mathbf{p}^{(k-1)}(1)A, \quad k = 1, 2, \dots \tag{5.27}$$

$$\mathbf{p}(1)B = 0 \tag{5.28}$$

By the last equation, the phase stationary probabilities $\mathbf{p}(1)$ are the stationary vector of B.

Note that the factorial moments yield efficient approximations for the stationary probabilities, simply by "shifting the expansion" to around $z = 0$.

This resolved an open problem signalled by Neuts. Citing Neuts (1995), page 274: "We note that the infinite-server M/M/∞ queue in a Markovian environment is surprisingly resistent to analytic solution ... Brute force numerical solution ... seems necessary because of the lack of a mathematically elegant solution." It is intriguing to investigate whether a similar approach based on a series expansion around the regular point 1 could produce the factorial moments for retrial queues as well.

Remark 5.3. When $K = 0$, the singularities of the system (5.16) are the roots of $Det(V(z)) = \bar{\alpha}(z-\bar{p})^s(z-1)$. However, it turns out that the singularity at 1 may be simplified by replacing the last equation by the sum of the equations, and dividing by $z - 1$.

Lemma 5.2. (Pearce's Lemma) With $K = 0$ and $p + \bar{p} = 1$, the system (5.16) is equivalent to the **simplified system** $\mathbf{p}'(z)V(z) = \mathbf{p}(z)U(z)$, with $V(z)$ and $U(z)$ for $s \geq 2$ given by

$$V(z) = \begin{bmatrix} z-\bar{p} & -p & & & \cdots & 1 \\ & z-\bar{p} & -p & & & 1 \\ & & \ddots & \ddots & & \vdots \\ & & & z-\bar{p} & -p & 1 \\ & & & & z-\bar{p} & 1 \\ & & & & & \bar{\alpha} \end{bmatrix} \tag{5.29}$$

$$U(z) = \begin{bmatrix} \Theta_0 & \tilde{\lambda}p_0 & & & & \tilde{\lambda}\tilde{p}_0 \\ \tilde{\mu}(\bar{\theta}+\tilde{\theta}z) & \Theta_1 & \tilde{\lambda}p_0 & & & \tilde{\lambda}\tilde{p}_0 + \tilde{\theta}\tilde{\mu} \\ & 2\tilde{\mu}(\bar{\theta}+\tilde{\theta}z) & \Theta_2 & \tilde{\lambda}p_0 & & \tilde{\lambda}\tilde{p}_0 + 2\tilde{\theta}\tilde{\mu} \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & (s-1)\tilde{\mu}(\bar{\theta}+\tilde{\theta}z) & \Theta_{s-1} & \tilde{\lambda}\tilde{p}_0 + (s-1)\tilde{\theta}\tilde{\mu} \\ & & & & s\tilde{\mu}(\bar{\theta}+\tilde{\theta}z) & \tilde{\lambda}_{ob} + s\tilde{\theta}\tilde{\mu} \end{bmatrix} \tag{5.30}$$

For $s = 1$, the matrices $V(z)$ and $U(z)$ are given by:

$$V(z) = \nu \begin{bmatrix} z-\bar{p} & 1 \\ 0 & \bar{\alpha} \end{bmatrix}, \quad U(z) = \nu \begin{bmatrix} \tilde{\lambda}(z\tilde{p}_0 - 1) & \tilde{\lambda}p_0 \\ \tilde{\mu}(\bar{\theta}+\tilde{\theta}z) & \tilde{\lambda}_{ob} + \tilde{\theta}\tilde{\mu} \end{bmatrix}$$

Proof: Let us replace the last equation by the sum of the equations, i.e. replace the last columns of V and U by the sum of the columns. Since B is a conservative generator, this sum equation is $\boldsymbol{p}'(z)(z-1)\tilde{C}\mathbf{1} = \boldsymbol{p}(z)(z-1)\tilde{A}\mathbf{1}$, and $z-1$ may be simplified, yielding:

$$\sum_{i=0}^{s-1} p_i'(z) + \bar{\alpha} p_s'(z) = \tilde{\lambda}\Big(\tilde{p}_0 \sum_{i=0}^{s-1} p_i(z) + \alpha_0 p_s(z)\Big) + \tilde{\theta}\tilde{\mu} \sum_{i=1}^{s} i p_i(z) \tag{5.31}$$

Remark 5.4. The "simplification" by $z-1$ of the original system seems to have been first noticed in Pearce (1989, (4.6)), who used the first s equations of (5.29) to obtain formulas expressing $p_i(z), i = 1, \ldots, s$ in terms of $p_0(z)$, and then used (5.31) to get a scalar ODE for $p_0(z)$ (and recursion for the respective probabilities).

Remark 5.5. The system (5.29) has been solved in particular cases when its degree is at most two in terms of classic hypergeometric functions, basically by "table look up" of the solution — see for example Hanschke (1987); Falin and Templeton (1997), and especially Choi et al. (1998), who assumes $\bar{\alpha} > 0$ and includes feedback.

One must further distinguish the two cases of:

(1) "Non-persistent retrials" with $\bar{\alpha} > 0$, for which the system is always ergodic.
(2) "Persistent retrials" with $\bar{\alpha} = 0$, when the determinant is 0 identically, and a reduction of the dimension by 1 is possible.

Below, we consider only the second case.

6. Persistent retrials ($\bar{\alpha} = 0$) and dimension reduction

A special treatment is necessary in the case of "persistent retrials" with $\bar{\alpha} = 0$, when our system becomes $\mathbf{p}'(z)V(z) = \mathbf{p}(z)U(z)$ with $V(z)$ given by

$$\begin{bmatrix} z-\bar{p} & -p & & & & 1 \\ & z-\bar{p} & -p & & & 1 \\ & & \ddots & \ddots & & \vdots \\ & & & z-\bar{p} & -p & 1 \\ & & & & z-\bar{p} & 1 \\ & & & & & 0 \end{bmatrix}$$

Since $V(z)$ is not invertible when $\bar{\alpha} = 0$, it is convenient to eliminate the last component $p_s(z)$ from the $s+1$-st equation provided by the last columns in (5.29):

$$(\tilde{\lambda}_o + s\tilde{\theta}\tilde{\mu}) p_s(z) = \sum_{i=0}^{s-1} p_i'(z) - \sum_{i=0}^{s-1} (\tilde{\lambda}\tilde{p}_a + i\tilde{\theta}\tilde{\mu}) p_i(z) \tag{6.32}$$

The, because only the last of the first s equations contains $p_s(z)$, in the form

$$\begin{aligned}(z-\bar{p})p_{s-1}'(z) - p p_{s-2}'(z) &= s\tilde{\mu}(\bar{\theta} + \tilde{\theta} z) p_s(z) + \cdots \\ &= \frac{s\mu(\bar{\theta} + \tilde{\theta} z)}{\lambda_{ob} + s\tilde{\theta}\mu} \left(\sum_{i=0}^{s-1} p_i'(z) - \sum_{i=0}^{s-1} (\tilde{\lambda}\bar{p} + i\tilde{\theta}\tilde{\mu}) p_i(z) \right) + \cdots\end{aligned}$$

we arrive, letting

$$\widetilde{\boldsymbol{p}}(z) = (p_0(z), \dots, p_{s-1}(z)) \tag{6.33}$$

denote the first s unknowns, to the simplified system:

$$\widetilde{\boldsymbol{p}}'(z)(V_{s-1}(z) - \kappa \boldsymbol{L}) = \widetilde{\boldsymbol{p}}(z)(U_{s-1}(z) - \kappa(\tilde{\lambda}\tilde{p}_a \boldsymbol{L} + \tilde{\theta}\tilde{\mu}\boldsymbol{l}_1) \tag{6.34}$$

where V_{s-1} and U_{s-1} are projections of $V(z)$ and $U(z)$ on the first s coordinates, where $\boldsymbol{L}$ denotes a matrix with ones on the last column and 0 else, $\boldsymbol{l}_1$ denotes a matrix with $0, 1, 2, \dots$ on the last column and 0 else, and

$$\kappa = \kappa(z) = \frac{s\mu(\bar{\theta} + \tilde{\theta}z)}{\lambda_{ob} + s\tilde{\theta}\mu} = \frac{\bar{\theta} + \tilde{\theta}z}{\rho + \tilde{\theta}} \tag{6.35}$$

Explicitly, we have:

$$V(z) = \begin{bmatrix} z - \bar{p} & -p & & & -\kappa \\ & z - \bar{p} & -p & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & z - \bar{p} & -p - \kappa \\ & & & & z - \bar{p} - \kappa \end{bmatrix} \tag{6.36}$$

$$U(z) = \begin{bmatrix} \Theta_0 & \tilde{\lambda}p_a & & & -\kappa\tilde{\lambda}\tilde{p}_a \\ \tilde{\mu}(\bar{\theta} + \tilde{\theta}z) & \Theta_1 & \tilde{\lambda}p_a & & -\kappa(\tilde{\lambda}\tilde{p}_a + \tilde{\theta}\tilde{\mu}) \\ & 2\tilde{\mu}(\bar{\theta} + \tilde{\theta}z) & \Theta_2 & \tilde{\lambda}p_a & -\kappa(\tilde{\lambda}\tilde{p}_a + 2\tilde{\theta}\tilde{\mu}) \\ & & \ddots & \ddots & \tilde{\lambda}p_a - \kappa(\tilde{\lambda}\tilde{p}_a + (s-2)\tilde{\theta}\tilde{\mu}) \\ & & & (s-1)\tilde{\mu}(\bar{\theta} + \tilde{\theta}z) & \tilde{\lambda}(z\tilde{p}_a - 1) - (s-1)\tilde{\mu}\bar{\theta} - (s-1)\tilde{\mu}\tilde{\theta}(\kappa + 1) - \kappa\tilde{\lambda}\tilde{p}_a \end{bmatrix}$$

Remark 6.1. The determinant of the reduced system is

$$Det(V(z)) = (z - \bar{p})^{s-1}(z - \bar{p} - \kappa(z)) = (z - \bar{p})^{s-1}\frac{\rho}{\rho + \tilde{\theta}}(z - \bar{p} - \tilde{\rho}(1 + \frac{\tilde{\theta}}{\bar{\theta}}\bar{p})) = 0$$

where $\tilde{\rho} = \frac{s\mu\bar{\theta}}{\lambda_{ob}}$. **Therefore, for $s \geq 3$, $\bar{p}$ is an irregular singularity.**

Remark 6.2. We consider here only the stable case, when stationary probabilities exist. As is well known, when $K = 0$, the system is stable if $\bar{\alpha} > 0$ (nonpersistent retrials). When $\bar{\alpha} = 0$, we conjecture that ergodicity holds precisely when the dominant regular singularity is outside the unit circle

$$z_r := \bar{p} + \frac{1 + \frac{\tilde{\theta}}{\bar{\theta}}\bar{p}}{\rho} > 1 \Longleftrightarrow p\rho < 1 + \frac{\tilde{\theta}}{\bar{\theta}}\bar{p} \tag{6.37}$$

where

$$\rho := \frac{\lambda_{ob}}{\mu\bar{\theta}s} = \tilde{\rho}^{-1}. \tag{6.38}$$

Equivalently,

$$\xi := p(\rho\bar{\theta} + \tilde{\theta}) + \theta < 1$$

This reduces to Hanschke's condition $p\rho < 1$ when $\tilde{\theta} = 0$, and to the Atencia-Moreno condition $\rho < 1 - \theta = \bar{\theta}$ when $\tilde{\theta} = 0$ and $p = 1$.

To relate to Hanschke (1999), note that he assumes $p + \bar{p} \le 1$ and $\alpha + \bar{\alpha} \le 1$, while we assume $p + \bar{p} = 1$ (which allows us to use Pearce's Lemma). Our assumption is w.l.o.g., however to compare with Hanschke (1999, Thm 1), we must rewrite his ergodicity condition in our notation as

$$\frac{\lambda_{ob}}{s\mu}\nu_q < \nu_q + \nu_a \tag{6.39}$$

which shows that our ergodicity condition (6.37) generalizes his.

Unfortunately, we do not have a reference for the rather fundamental and plausible conjecture that the obviously necessary ergodicity condition that the dominant singularity lies outside the unit circle is also sufficient [§] .

Remark 6.3. The asymptotic behavior may be determined by expanding the function around the (regular) singularity z_r — see Section 8.

Note that the parameters p, θ and α_0 play a crucial role in the location of the singularities, unlike p_a.

Example 6.1. For $K = 0$ and $s = 1$ or $s = 2$ servers, the matrices V and U are given, respectively, by $V = z - \bar{p} - \kappa(z)$ and $U = \tilde{\lambda}(z\tilde{p}_a - 1 - \kappa(z)\tilde{p}_a)$, and

$$V(z) = \begin{bmatrix} z - \bar{p} & -p - \kappa(z) \\ 0 & z - \bar{p} - \kappa(z) \end{bmatrix}$$

$$U(z) = \begin{bmatrix} \tilde{\lambda}(z\tilde{p}_a - 1) & \tilde{\lambda}(p_a - \kappa(z)\tilde{p}_a) \\ \tilde{\mu}(\bar{\theta} + \tilde{\theta}z) & \tilde{\lambda}(z\tilde{p}_a - 1 - \kappa(z)\tilde{p}_a) - \tilde{\mu}(\bar{\theta} + \tilde{\theta} + \kappa(z)\tilde{\theta}) \end{bmatrix}$$

Example 6.2. The solution for $s = 1$ is given, after putting

$$\xi := p(\rho + \tilde{\theta}) + \theta,\ u = 1 - \xi + \rho - \rho z = \bar{\theta} + \bar{p}\tilde{\theta} - \rho(z - \bar{p}), \bar{\lambda} = \tilde{\lambda}(1 - \bar{p}\tilde{p}_a)(1 + \frac{\tilde{\theta}}{\rho})$$

by

$$\begin{aligned} q(z) &= cu^{-\bar{\lambda}} e^{-\frac{\tilde{\lambda}\tilde{p}_a}{\rho}u} \\ p(z) &= \rho(1 - \bar{p}\tilde{p}_a)u^{-1}q(z) \end{aligned} \tag{6.40}$$

with singularity at $z_r > 1$ when $\xi < 1$, on the branch cut $[z_r, \infty)$. Using $q(z) + p(z)|_{z=1} = 1$ yields easily the proportionality constant c.

From (6.40) it may be shown that asymptotically

$$(p_{0,n}, p_{1,n}) = c(1 - \xi)^{\bar{\lambda}+1} \frac{(n+1)^{(\bar{\lambda}-1)}\xi^n}{\Gamma(\bar{\lambda})} \left(1, \frac{\xi(\bar{\lambda} + n)}{\bar{\lambda}}\right) \sim_{n\to\infty} c(1 - \xi)^{\bar{\lambda}+1} \frac{n^{\bar{\lambda}-1}}{\Gamma(\bar{\lambda})}\xi^n(1, \frac{n}{\tilde{\mu}})$$

[§] We believe this since Kolmogorov's system encodes in principle all the necessary information on our problem, but we have been unable to check this beyond the particular cases studied by Hanschke (1999) and Atencia and Moreno (2003).

where $\bar\lambda = \tilde\lambda(1+\frac{\tilde\theta}{\rho})$, and $(n+1)^{(\bar\lambda-1)} = (n+1)(n+2)\cdots$ is an ascending Pochammer symbol, in line with Kim et al. (2012) (who worked in the classic case with $\xi = \rho$).

With $\tilde p_a = 0$, the exact solution is:

$$(q(z), p(z)) = \frac{1-\xi}{1-\xi+\rho}\left(\frac{1-\xi}{1-\xi+\rho-\rho z}\right)^{\tilde\lambda(1+\frac{\tilde\theta}{\rho})}(1, (\bar\theta+\bar p\tilde\theta)\tilde\rho-(z-\bar p))^{-1})$$

a negative binomial distribution.

Furthermore, if $\bar\theta = 1$ and $p = 1$, this yields the classic

$$(q(z), p(z)) = (1-\rho)\left(\frac{1-\rho}{1-\rho z}\right)^{\tilde\lambda}\left(1, \frac{\rho}{1-\rho z}\right)$$

and $(p_{0,n}, p_{1,n}) = (1-\rho)^{\tilde\lambda+1}(\frac{\rho^n\tilde\lambda_{(n)}}{n!}, \frac{\rho^{n+1}(\tilde\lambda+1)_{(n)}}{n!})$ see — Falin and Templeton (1997) pg. 101.

7. The classic persistent retrial model ($\bar\alpha = 0$) with full acceptance ($p_a = 1$) and no feedback to orbit ($\tilde\theta = 0$): a generalized Okubo system

In this section we consider the case of $\bar\alpha = \tilde\theta = \tilde p_a = 0$, for which the system (6.34) becomes:

$$\tilde{\boldsymbol p}'(z)(V_{s-1}(z) - \tilde\rho \boldsymbol L) = \tilde{\boldsymbol p}(z)U_{s-1} \tag{7.41}$$

where $\tilde\rho = \frac{s\mu}{\lambda_{ob}}$, and $V_{s-1}(z)$ and U_{s-1} are given, respectively, by:

$$\begin{bmatrix} z-\bar p & -p & & & & -\tilde\rho \\ & z-\bar p & -p & & & -\tilde\rho \\ & & \ddots & \ddots & & \vdots \\ & & & z-\bar p & -p & -\tilde\rho \\ & & & & z-\bar p & -p-\tilde\rho \\ & & & & & z-\bar p-\tilde\rho \end{bmatrix}$$

$$\begin{bmatrix} -\tilde\lambda & \tilde\lambda & & & & \\ \bar\mu & -\tilde\lambda-\bar\mu & \tilde\lambda & & & \\ & 2\bar\mu & -\tilde\lambda-2\bar\mu & \tilde\lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & (s-2)\bar\mu & -\tilde\lambda-(s-2)\bar\mu & \tilde\lambda \\ & & & & (s-1)\bar\mu & -\tilde\lambda-(s-1)\bar\mu \end{bmatrix} \tag{7.42}$$

where $\bar\mu = \tilde\mu\bar\theta$.

Remark 7.1. For any number s of servers, $U_{s-1} = B_{s-1} - \tilde\lambda E$, where E is a matrix with a 1 in the lower right corner and 0 else, is the generator of the time until blocking of an Erlang loss system with s servers.

Since $\boldsymbol{U} = U_{s-1}$ is now a constant matrix, the reduced system $\widetilde{\boldsymbol{p}}'(z)V_{s-1}(z) = \widetilde{\boldsymbol{p}}(z)\boldsymbol{U}$ is of "generalized Okubo type" defined by a canonical form [§] :

$$\widetilde{\boldsymbol{p}}'(z)\big(z\boldsymbol{I} - \boldsymbol{T}\big) = \widetilde{\boldsymbol{p}}(z)\boldsymbol{U} \tag{7.43}$$

with $\boldsymbol{T}$ given by

$$\begin{bmatrix} \bar{p} & p & & & & \tilde{\rho} \\ & \bar{p} & p & & & \tilde{\rho} \\ & & \ddots & \ddots & & \vdots \\ & & & \bar{p} & p & \tilde{\rho} \\ & & & & \bar{p} & p+\tilde{\rho} \\ & & & & & \bar{p}+\tilde{\rho} \end{bmatrix} \tag{7.44}$$

Remark 7.2. The recently introduced class of generalized Okubo-type systems constitutes a powerful generalization of classic hypergeometric equations (see for example Balser et al. (2006); Hiroe (2009), and Oshima (2011) for an introduction). The Jordan form of $\boldsymbol{T}$ turns out to be useful for classifying differential systems.

Note that we have only two Jordan blocks (for any number s of servers), for which the eigenvalues $\bar{p}$ and $\bar{p}+\tilde{\rho}$ of $\boldsymbol{T}$ yield the singular points of our system (irregular and regular, respectively), and that the Jordan form of our $\boldsymbol{T}$ is not diagonal.

According to Kohno (1999), Ch4, a Fuchsian equation may be written as a system of type (7.43) iff T is diagonal (this case includes several important special functions, like the generalized hypergeometric, and the Pochammer system).

Therefore, the case of persistent retrial queues with $s \geq 3$ is qualitatively different from the cases $s = 1$ and $s = 2$, since the point $z = \bar{p}$ is now a non-Fuchsian irregular singular point, with non-diagonal $\boldsymbol{T}$, a situation that has been much less studied in the literature.

Remark 7.3. When $p > 0$, putting $y = \frac{z-\bar{p}}{p}$ reduces the problem effectively to the **pure retrial** case with $p = 1$ and $\bar{p} = 0$, after replacing $\tilde{\rho}$ and ρ by $\bar{\rho} = \tilde{\rho}/p$ and $\xi = p\rho$, respectively. We will call this case the **standardized Okubo problem**.

Recall that we suppose throughout that the dominant regular singularity $\bar{p}+\tilde{\rho}$ is outside the unit circle, i.e. $p < \tilde{\rho}$, or simply $\tilde{\rho} > 1$ in the standardized case.

In the following subsections, we will revisit the cases of $s = 2$ and $s = 3$, respectively.

[§] There is also an equivalent formulation

$$\boldsymbol{q}'(z) = -\boldsymbol{q}(z)\Big(\boldsymbol{T} + \frac{\boldsymbol{U}+\boldsymbol{I}}{z}\Big)$$

from which (7.43) is obtained by Laplace transform.

7.1. *Two servers*

We note first that for $\mathbf{s} = \mathbf{2}$, it is also possible to treat the general Pearce system (5.29):

$$\mathbf{p}'(z)\begin{bmatrix} z-\bar{p} & -p & 1 \\ 0 & z-\bar{p} & 1 \\ 0 & 0 & \bar{\alpha} \end{bmatrix} = \mathbf{p}(z)\begin{bmatrix} \tilde{\lambda}(\tilde{p}_a z-1) & \tilde{\lambda}p_a & \tilde{\lambda}\tilde{p}_a \\ \tilde{\mu}(\bar{\theta}+\tilde{\theta}z) & \tilde{\lambda}(\tilde{p}_a z-1)-\tilde{\mu}\bar{\theta} & \tilde{\lambda}\tilde{p}_a+\tilde{\mu}\tilde{\theta} \\ 0 & 2\tilde{\mu}(\bar{\theta}+\tilde{\theta}z) & \tilde{\lambda}_{ob}+2\tilde{\mu}\tilde{\theta} \end{bmatrix}$$

However, we consider only the Okubo case (with $\bar{\alpha} = 0$, with full acceptance $\tilde{p}_a = 0$ and no feedback $\tilde{\theta} = 0$), when the system is:

$$\mathbf{p}'(z)\begin{bmatrix} z-\bar{p} & -p & 1 \\ 0 & z-\bar{p} & 1 \\ 0 & 0 & 0 \end{bmatrix} = \mathbf{p}(z)\begin{bmatrix} -\tilde{\lambda} & \tilde{\lambda} & 0 \\ \bar{\mu} & -\tilde{\lambda}-\bar{\mu} & 0 \\ 0 & 2\bar{\mu} & \tilde{\lambda}_{ob} \end{bmatrix}$$

(compare with Hanschke (1987, (4.3),(4.4)), who assumes $\bar{p} = 0$, $\bar{\theta} = 1$, and $\lambda_{ob} = \lambda$).

After eliminating $p_2(z)$ using (6.32), the reduced Okubo system (7.43) and (7.44) for $\widetilde{\boldsymbol{p}}(z) = (p_0(z), p_1(z))$ becomes: §

$$\widetilde{\boldsymbol{p}}'(z)\begin{bmatrix} z-\bar{p} & -p-\rho \\ 0 & z-\bar{p}-\rho \end{bmatrix} = \widetilde{\boldsymbol{p}}(z)\begin{bmatrix} -\tilde{\lambda} & \tilde{\lambda} \\ \tilde{\mu} & -\tilde{\lambda}-\tilde{\mu} \end{bmatrix} \qquad (7.45)$$

8. Shifting to the dominant regular singularity and asymptotics for the pure retrial case

Consider now the standardized Okubo problem $\widetilde{\boldsymbol{p}}'(y)\big(y\boldsymbol{I}-\boldsymbol{T}\big) = \widetilde{\boldsymbol{p}}(y)\boldsymbol{U}$ with U defined in (7.42), and $\boldsymbol{T} = T_+ + \bar{\rho}\boldsymbol{L}$, where

$$T_+ = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & 1 & \\ & & & \ddots & \ddots \\ & & & & \ddots \end{pmatrix}, \quad \boldsymbol{L} = \begin{pmatrix} 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

§Hanschke (1987) obtains when $p = 1$ classic hypergeometric solutions. He considers the second order scalar equation $Lq(z) = 0$ for $q(z) = p_0(z)$, which is:

$$L = (\tilde{\lambda}+\bar{\mu}+(z-\tilde{\rho})D)(\tilde{\lambda}+zD) - \bar{\mu}(\tilde{\lambda}+(1+\tilde{\rho})D) =$$
$$\tilde{\lambda}^2 + (z(2\tilde{\lambda}+\ \bar{\mu}+1) - \tilde{\rho}(3\tilde{\lambda}/2+\bar{\mu}+1))D + z(z-\tilde{\rho})D^2$$

(using $DzD = D + zD^2$), where $\tilde{\rho} = \frac{\mu s}{\lambda_{ob}}$ was defined in (6.38). Hanschke (1987, (4.8)) notes that putting $\rho z = x$ shifts the singularity $z = \rho^{-1}$, yielding the canonic Gauss hypergeometric equation

$$x(x-1)q''(x) + [x(2\tilde{\lambda}+\bar{\mu}+1) - (\frac{3\tilde{\lambda}}{2}+\bar{\mu}+1)]q'(x) + \tilde{\lambda}^2 q(x) = 0$$

whose only analytic solution in the unit disk is the Gauss hypergeometric series. This determines all unknowns up to a proportionality constant, obtained using $q(z) + p_1(z) + p_2(z)|_{z=1} = 1$. The explicit formulas for $p_i(z)$ are given in Falin and Templeton (1997, Thm. 2.1), which lead to some explicit formulas for classic performance measures, such as $E[N] = \sum_{i=0}^{2} p_i'(1)$, given in Falin and Templeton (1997, Thm. 2.2) (see also Falin (1990, (91-93))). Further factorial moments $\sum_{i=0}^{2} p_i''(1) = E[N(N-1)]$, $\sum_{i=0}^{2} p_i'''(1) = E[N(N-1)(N-2)]$, ... are given in the last section of Hanschke (1987).

in which the singularities of the system are the eigenvalues 0, which is irregular, and $\bar{\rho}$, which is regular.

Since $\boldsymbol{T}^s = \bar{\rho}\boldsymbol{T}^{s-1}$, the resolvent $\big(y\boldsymbol{I} - \boldsymbol{T}\big)^{-1}$ may be easily computed, yielding:

$$(yI - \boldsymbol{T})^{-1} = y^{-1}\left(\boldsymbol{I} + \frac{1}{y}\boldsymbol{T} + \cdots + \frac{1}{y^{s-2}}\boldsymbol{T}^{s-2} + \frac{1}{y^{s-2}(y-\bar{\rho})}\boldsymbol{T}^{s-1}\right)$$

For $s = 3$ for example, we find:

$$\begin{aligned}(yI - \boldsymbol{T})^{-1} &= y^{-1}\left(\boldsymbol{I} + \frac{1}{y}\boldsymbol{T} + \frac{1}{y(y-\bar{\rho})}\boldsymbol{T}^2\right) = y^{-1}\left[\boldsymbol{I} + \frac{1}{y}\boldsymbol{T} + \frac{1}{\bar{\rho}}\left(\frac{1}{y-\bar{\rho}} - \frac{1}{y}\right)\boldsymbol{T}^2\right] \\ &= y^{-1}\left[\boldsymbol{I} + \frac{1}{y}\left(\boldsymbol{T} - \frac{1}{\bar{\rho}}\boldsymbol{T}^2\right) + \frac{1}{\bar{\rho}}\frac{1}{y-\bar{\rho}}\boldsymbol{T}^2\right] \\ &= y^{-1}\left[\boldsymbol{I} + \frac{1}{y}\begin{pmatrix}0 & 1 & -(1+\xi) \\ 0 & 0 & 0 \\ 0 & 0 & 0\end{pmatrix} + \frac{1}{y-\bar{\rho}}\begin{pmatrix}0 & 0 & \xi + 1 + \xi^{-1} \\ 0 & 0 & 1+\xi^{-1} \\ 0 & 0 & \xi^{-1}\end{pmatrix}\right]\end{aligned}$$

where we used $\boldsymbol{L}^2 = \boldsymbol{L}$, $\boldsymbol{L}T_+ = \boldsymbol{0}$, $T_+\boldsymbol{L} := \boldsymbol{L}_1 = \begin{pmatrix}0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0\end{pmatrix}$, and $\boldsymbol{T}^2 = \begin{pmatrix}0 & 0 & \xi + 1 + \xi^{-1} \\ 0 & 0 & 1+\xi^{-1} \\ 0 & 0 & \xi^{-1}\end{pmatrix}$.

Multiplying the resolvent on the RHS, we get $\forall s \geq 3$ a differential system

$$\widetilde{\boldsymbol{p}}'(y) = \widetilde{\boldsymbol{p}}(y)\boldsymbol{R} = \widetilde{\boldsymbol{p}}(y)\left((y-\bar{\rho})^{-1}\boldsymbol{D} + y^{-1}\boldsymbol{D}_1 + y^{-2}\boldsymbol{D}_2\right)$$

of Poincaré rank 1 at $y = 0$ (thus this singularity is nonregular) and of rank 0 at $y = \bar{\rho}$, with

$$\boldsymbol{D} = \begin{pmatrix} 0 & \vdots \\ & 0 \\ 0 & \lambda(1+\rho^{-1}) \\ 0 & -\lambda-(s-1)\mu \end{pmatrix}, \tag{8.47}$$

$$\boldsymbol{D}_1 = \begin{pmatrix} -\lambda & 0 & \ldots & \\ 0 & -\lambda-\mu & 0 & \\ & & \ddots & \\ & & -\lambda-(s-2)\mu & -\lambda(1+\rho^{-1}) \\ 0 & 0 & 0 & 0 \end{pmatrix}, \boldsymbol{D}_2 = \begin{pmatrix} 0 & \lambda & 0 & \\ 0 & 0 & \lambda & 0 \\ & & 0 & 0 \\ & & & \end{pmatrix}.$$

The asymptotics of the stationary probabilities, and in fact a full asymptotic expansion, may be obtained by looking for logarithmic free regular solutions $\widetilde{\boldsymbol{p}}(y) = (\bar{\rho} - y)^w \boldsymbol{g}(y)$, where $\boldsymbol{g}(y)$ is analytic at the regular singularity $y = \bar{\rho}$. The eigenvalue w, as well as the first term of such an expansion, are obtained from the spectral decomposition of the matrix $\boldsymbol{D}$.

In principle, this expansion requires an application of the vector version of Frobenius's formal series expansions around singular points — see for example Coddington and Levinson (1955, Ch 4), Wasow (2002, Ch 4), and, for recent developments, Barkatou and Pflügel (1999).

Indeed, the first term of such an expansion was recently provided by Liu et al. (2012) and Kim et al. (2012) (who use a different dimensionality reduction from $s + 1$ to s, by the important probabilistic technique of "censoring").

Note however that the vector Frobenius method is available in several flavors (which typically require $\boldsymbol{D}$ to be non-singular, which is not the case in our differential system, nor in Liu et al. (2012) or Kim et al. (2012, Thm 1)), and therefore obtaining a full asymptotic expansion requires further work.

Acknowledgements: We thank Jesus Artalejo, JL. Lopez, A. Quadrat, T. Cluzeau, F. Beukers, C. Koutschan and the anonymous referees for very useful advice.

The first two authors were partially supported by the French-Romanian Programme LEA Math-Mode.

F. Avram is a full professor in the Mathematics department at Université de Pau, France. He received his Ph.D. degree in Mathematics from Cornell University in 1986. He has published over 60 research papers in internationally-refereed journals, in the field of applied probability and stochastic processes (with particular emphasis on queueing networks, risk theory and mathematical finance).

D. Matei is a senior researcher in the Geometry and Topology group at the "Simion Stoilow" Institute of Mathematics of the Romanian Academy in Bucharest. He received his Ph.D. degree in Mathematics from Northeastern University, Boston in 1999. He has published research papers in algebraic geometry and topology. His research interests include hypergeometric differential equations, birth-and-death processes and random walks.

Y.Q. Zhao received his Ph.D. degree at the University of Saskatchewan. He has been a full professor at Carleton University, Ottawa, since 2003, and was the Director of the School of Mathematics and Statistics from 2004 to 2007 and from 2010 to 2011. Dr. Zhao's research interests are in applied probability and stochastic processes, with particular emphasis on computer and telecommunication network applications. He has published over 70 papers in internationally-refereed journals.

References

J.R. Artalejo and G.I. Falin. Standard and retrial queuesing systems: a comparative analysis. *Revista Matemática Complutense*, 15(1):101–130, 2002.

J.R. Artalejo and J.I. Falin. Stochastic decomposition for retrial queues. *Top*, 2(2):329–342, 1994.

I. Atencia and P. Moreno. A queueing system with linear repeated attempts, bernoulli schedule and feedback. *Top*, 11(2):285–310, 2003.

F. Avram and D. Fotso-Chedom. On symbolic rg factorization of quasi-birth-and-death processes. *Top*, 19(2):317–335, 2011.

F. Avram, A.J.E.M. Janssen, and J.S.H. van Leeuwaarden. Loss systems with slow retrials in the Halfin-Whitt regime. *Adv. in Appl. Probab*, 45(1):274–294, 2013.

W. Balser, C. Röscheisen, and F. Steiner. Systems of linear ordinary differential equations—a review of three solution methods. *Ulmer Seminare ber Funktionalanalysis und Differentialgleichungen*, (11):53–62, 2006.

M. Barkatou and E. Pflügel. An algorithm computing the regular formal solutions of a system of linear differential equations. *Journal of Symbolic Computation*, 28(4):569–587, 1999.

Y. Baryshnikov and R. Pemantle. Convolutions of inverse linear functions via multivariate residues. *Preprint*, 2004.

Y. Baryshnikov and R. Pemantle. Asymptotics of multivariate sequences, part III: quadratic points. *Advances in Mathematics*, 228(6):3127–3206, 2011.

H. Baumann and W. Sandmann. Numerical solution of level dependent quasi-birth-and-death processes. *Procedia Computer Science*, 1(1):1561–1569, 2010.

L. Bright and P.G. Taylor. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stochastic Models*, 11(3):497–525, 1995.

B.D. Choi, Y.C. Kim, and Y.W. Lee. The M/M/c retrial queue with geometric loss and feedback. *Computers & Mathematics with Applications*, 36(6):41–52, 1998.

E.A. Coddington and N. Levinson. *Theory of ordinary differential equations*. Tata McGraw-Hill Education, 1955.

J.W. Cohen. Basic problems of telephone traffic theory and the influence of repeated calls. *Philips Telecommunication Review*, 18(2):49–100, 1957.

A. Economou. Generalized product-form stationary distributions for Markov chains in random environments with queueing applications. *Advances in Applied Probability*, 37(1):185–211, 2005.

G. Falin. A survey of retrial queues. *Queueing Systems*, 7(2):127–167, 1990.

G.I. Falin and J.G.C. Templeton. *Retrial Queues*. Chapman & Hall/CRC, 1997.

P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge Univ Pr, 2009. ISBN 0521898064.

W. Gautschi. Computational aspects of three-term recurrence relations. *SIAM review*, 9(1): 24–82, 1967.

A. Gomez-Corral and M.F. Ramalhoto. The stationary distribution of a Markovian process arising in the theory of multiserver retrial queueing systems. *Mathematical and Computer Modelling*, 30(3):141–158, 1999.

J. Gray. *Linear Differential Equations and Group Theory from Riemann to Poincaré*. Birkhäuser Boston, 2008.

N. Grier, W.A. Massey, T. McKoy, and W. Whitt. The time-dependent Erlang loss model with retrials. *Telecommunication Systems*, 7(1):253–265, 1997.

T. Hanschke. Explicit formulas for the characteristics of the M/M/2/2 queue with repeated attempts. *Journal of Applied Probability*, 24:486–494, 1987.

T. Hanschke. A matrix continued fraction algorithm for the multiserver repeated order queue. *Mathematical and Computer Modelling*, 30(3):159–170, 1999.

K. Hiroe. Twisted Euler transform of differential equations with an irregular singular point. *Arxiv preprint arXiv:0912.5124*, 2009.

A. Janssen. Asymptotic analysis of Cohen's equation for retrial queues in the Halfin-Whitt regime. *Asymptotic Analysis*, 2012 (to appear).

W. B. Jones and W. J. Thron. Continued fractions: Analytic theory and applications (encyclopedia of mathematics and its applications, vol. 11) author. 1980.

J. Keilson, J. Cozzolino, and H. Young. A service system with unfilled requests repeated. *Operations Research*, 16:1126–1137, 1968.

J. Kim, J. Kim, and B. Kim. Tail asymptotics of the queue size distribution in the M/M/m retrial queue. *Journal of Computational and Applied Mathematics*, 236:3445–3460, 2012.

Y.C. Kim. On M/M/3/3 retrial queueing system. *Honam Mathematical Journal*, 17:141–147, 1995.

W. Koepf. The algebra of holonomic equations. *Mathematische Semesterberichte*, 44(2): 173–194, 1997.

M. Kohno. *Global Analysis in Linear Differential Equations*. Kluwer Academic Publishers, 1999.

L. Kosten. On the influence of repeated calls in the theory of probabilities of blocking. *De Ingenieur*, 59:1–25, 1947.

A. Krishnamoorthy, P.K. Pramod, and S.R. Chakravarthy. Queues with interruptions: a survey. *TOP*, 2012 (online first: DOI: 10.1007/s11750-012-0256-6).

V.G. Kulkarni and H.M. Liang. Retrial queues revisited. *Frontiers in Queueing: Models and Applications in Science and Engineering*, page 19, 1997.

P. Levrie and A. Bultheel. Matrix continued fractions related to first-order linear recurrence systems. *Electronic Transactions on Numerical Analysis*, 4:46–63, 1996.

Q.-L. Li and J. Cao. Two types of rg-factorizations of quasi-birth-and-death processes and their applications to stochastic integral functionals. *Stochastic models*, 20(3):299–340, 2004.

Q.-L. Li and Y.Q. Zhao. *The RG-factorization in block-structured Markov renewal processes with applications*. Observation, Theory and Modeling of Atmospheric Variability, ed: Xun Zhu, World Scientific, 2004.

Q.-L. Li, Y. Ying, and Y. Q. Zhao. A bmap/g/1 retrial queue with a server subject to breakdowns and repairs. *Annals of Operations research*, 141(1):233–270, 2006.

Q.L. Li. *Constructive computation in stochastic models with applications: the RG-Factorizations*. Springer Verlag, 2010.

B. Liu and Y.Q. Zhao. Analyzing retrial queues by censoring. *Queueing Systems*, 64(3): 203–225, 2010.

B. Liu, X. Wang, and Y.Q. Zhao. Tail asymptotics for M/M/c retrial queues with non-persistent customers. *Operational Research*, 12(2):173–188, 2012.

M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach*. Dover Publications, 1995.

C.A. O'Cinneide and P. Purdue. The M/M/∞ queue in a random environment. *Journal of Applied Probability*, 23:175–184, 1986.

A.M. Odlyzko. Asymptotic enumeration methods. *Handbook of Combinatorics*, 2:1063–1229, 1995.

T. Oshima. Fractional calculus of Weyl algebra and Fuchsian differential equations. *Arxiv preprint arXiv:1102.2792*, 2011.

C.E.M. Pearce. Holomorphic solutions about an irregular singular point of an ordinary linear differential equation. *Journal of the Australian Mathematical Society (Series A)*, 39(02): 178–186, 1985.

C.E.M. Pearce. Extended continued fractions, recurrence relations and two-dimensional Markov processes. *Advances in Applied Probability*, 21:357–375, 1989.

R. Pemantle. Analytic combinatorics in d variables: an overview. *Algorithmic probability and combinatorics: AMS Special Sessions on Algorithmic Probability and Combinatorics,* *October 5-6, 2007, DePaul University, Chicago, Illinois: AMS Special Session, October 4-5, 2008, University of British Columbia, Vancouver, BC, Canada*, 520:195, 2010.

R. Pemantle and M.C. Wilson. Twenty combinatorial examples of asymptotics derived from multivariate generating functions. *Arxiv preprint math/0512548*, 2005.

T. Phung-Duc, H. Masuyama, S. Kasahara, and Y. Takahashi. M/m/3/3 and m/m/4/4 retrial queues. *Journal of Industrial and Management Optimization*, 5(3):431, 2009.

T. Phung-Duc, H. Masuyama, S. Kasahara, and Y. Takahashi. A simple algorithm for the rate matrices of level-dependent qbd processes. In *Proceedings of the 5th International Conference on Queueing Theory and Network Applications*, pages 46–52. ACM, 2010a.

T. Phung-Duc, H. Masuyama, S. Kasahara, and Y. Takahashi. State-dependent M/M/c/c+r retrial queues with Bernoulli abandonment. *Journal of Industrial and Management Optimization*, 6:517–540, 2010b.

T. Phung-Duc, H. Masuyama, S. Kasahara, and Y. Takahashi. A matrix continued fraction approach to multiserver retrial queues. *Annals of Operations Research*, pages 1–23, 2011.

A. Raichev and M.C. Wilson. Asymptotics of coefficients of multivariate generating functions: improvements for multiple points. *Arxiv preprint arXiv:1009.5715*, 2010.

V. Ramaswami. Matrix analytic methods: A tutorial overview with some extensions and new results. *Lecture Notes in Pure and Applied Mathematics*, pages 261–296, 1996.

J. Riordan. *Stochastic Service Systems*. John, Wiley, 1962.

R. P. Stanley. Differentiably finite power series. *European J. Combin*, 1(2):175–188, 1980.

W. Wasow. *Asymptotic expansions for ordinary differential equations*. Courier Dover Publications, 2002.

R.W. Wolff. *Stochastic modeling and the theory of queues*. Prentice hall Englewood Cliffs, NJ, 1989.

D. Zeilberger. Three recitations on holonomic systems and hypergeometric series. *Séminaire Lotharingien de Combinatoire*, 24:28, 1990.