Matrix-Analytic Methods in Stochastic Models

Proceedings of the Ninth International Conference on Matrix-Analytic Methods in Stochastic Models

Edited by: Qi-Ming He, Gábor Horváth, Miklós Telek

June 28 - 30, 2016
Budapest, Hungary

Contents

An interaction between queueing and change detection

[Opening address]

László Gerencsér
MTA SZTAKI
Institute of Computer Science and Control
Hungarian Academy of Sciences
Hungary
gerencser.laszlo@sztaki.mta.hu

ABSTRACT

A basic feature of a single server queue is waiting time recorded at the time of arrival of a new customer, following a simple non-linear dynamics. By a remarkable coincidence the same dynamics is obtained for what is called the the Page-Hinkley detector designed for real-time change detection of stochastic processes [1, 2]. According to this an alarm is given if the detector (an equivalent to waiting time) exceeds a prefixed threshold.

A result on the empirical tail-probabilities of the detector under very general technical conditions will be presented, assuming no change, thus providing an upper bound for the false alarm rate. These results translate to results on empirical large deviations for waiting times. The above mentioned technical condition can be verified in a variety of interesting special cases to be briefly presented.

Keywords

Waiting time; Page-Hinkley detector; Change detection.

1. REFERENCES

[1] L. Gerencsér and C. Prosdocimi. Input-output properties of the page-hinkley detector. *Systems and Control Letters*, 60:486–491, 2011.

[2] L. Gerencsér, C. Prosdocimi, and Zs. Vágó. Change detection for finite dimensional gaussian linear systems - a bound for the almost sure false alarm rate. In *Proc. of the 2013 European Control Conference (ECC)*, Zürich, Switzerland, July 2013.

Infinite Server Queueing-inventory Models

[Extended Abstract] *

Srinivas R. Chakravarthy
Department of Industrial and
Manufacturing Engineering
Kettering University
Flint, MI-48504, USA
schakrav@kettering.edu

Dhanya Shajin
Department of Mathematics
Cochin University of Science
and Technology
Cochin-682022, Kerala, India
dhanya.shajin@gmail.com

B. Lakshmy
Department of Mathematics
Cochin University of Science
and Technology
Cochin-682022, Kerala, India
lakshmykrishnaiyer
@yahoo.com

A. Krishnamoorthy
Department of Mathematics
Cochin University of Science
and Technology
Cochin-682022, Kerala, India
achyuthacusat
@gmail.com

ABSTRACT

In this paper, we study an $MAP/M/\infty$ queue associated with an inventory system. The inventory is replenished according to the (s, S)-policy. The (self) service and lead times are assumed to be exponentially distributed. No arriving customer is allowed to enter into the system (of infinite capacity) when there is no stock available for servicing the customer. Thus, every customer in service is attached with an inventory at the time of entering into the system. We employ algorithmic approach for the computation of various quantities of interest and derive some explicit expressions in some cases. An illustrative example is discussed.

Keywords

$MAP/M/\infty$ queue; inventory; lead time; (s, S) policy; phase type distribution; algorithmic probability

1. INTRODUCTION

This paper deals with a queueing-inventory system with an unlimited number of servers. One can think of this as a self-service system for the customers. Every customer needs to have an inventory item to be served. We believe this work to be the first in the literature involving an infinite number of servers in queueing-inventory process. Queueing-inventory models have been extensively discussed in the literature. However, most of these deal with a single server system and with servicing occurring one at a time. Papers dealing with with two or more servers are scarce in the literature; one paper in this direction is by Krishnamoorthy et al. [17]. We refer the reader to Sigman and Simchi-Levi [28], Berman and Kim [1], Berman and Sapna [2], Schwartz et al. ([26], [27]), Saffari et al. [25], Krishnamoorthy and Viswanath [15], Sivakumar and Arivarignan ([29], [30]), Krishnamoorthy and Anbazhagan [13], Krenzler and Daduna ([10], [11]), and Krishnamoorthy et al. ([14], [16]), for further details on single-server queueing-inventory systems. In the single server case explicit product form solution for the system state is obtained in ([10], [11], [14], [15], [16], [25], [26], [27]) under the condition that no customer joins the system when the inventory level is zero.

The case of a general bulk service rule in the context of single-server and with (s, S)-type replenishment policy was first studied in Chakravarthy et al. [5]. In this paper, the authors assume that at the beginning of a (batch) service the inventory level will be decreased by an amount equivalent to the size of the batch so that each customer will consume exactly one item from the inventory. Replenishment order is placed as soon as unattached inventory (see definition below) level falls to the set $\{0, 1, 2, ..., s\}$ for the first time after a replenishment. Thus, in effect the total inventory in the system can even be $2S$ in their paper. This is not the case with the rest of the papers mentioned above.

The infinite-server queueing systems have been investigated extensively in the literature. A sample of this is indicated below: Collings and Stoneman [6], Conolly [7], Eliazar [8], Foley [9], Keilson and Servi [12], Liu et al. [18], Mirasol [19], Newell [22], O'cinneide and Purdue [23], Ramalhoto [24], Stadje [31].

In the present paper we adopt the policy of attaching an inventory with every customer entering into the system for getting a service. We consider an infinite capacity queueing-inventory system with infinite number of servers to which customers arrive according to a Markovian arrival process with representation (D_0, D_1), the order of these matrices is m. The service time is exponentially distributed with parameter μ. Thus, when there are n customers in the system

*A full version of this paper will be submitted elsewhere for publication the rate of departures is given by $n\mu$. We adopt the policy of attaching inventory with all customers being served (which is same as all those who are present). Here inventory control is based on (s, S) policy as described in [5]: whenever the number of *unattached* inventory falls to s an order for replenishment is placed. Upon replenishment the inventory level will be brought to S. The crucial assumption that we make in this paper is that no customer joins the system during the time period with *unattached* inventory level equal to zero. This means that during such period of time if a customer walks in, he has to leave the system – lost case. The lead time for replenishment follows an exponential distribution with parameter θ. Since we are forbidding a customer entry during such period we are able to get explicit expression for the marginal probability of *unattached* inventory.

Using matrix-analytic methods we study the model as a level dependent quasi-birth and death process ($LDQBD$) of the form:

$$\mathcal{Q} = \begin{bmatrix} A_{00} & A_0 & & & & \\ A_{21} & A_{11} & A_0 & & & \\ & A_{22} & A_{12} & A_0 & & \\ & & A_{23} & A_{13} & A_0 & \\ & & & \ddots & \ddots & \ddots \end{bmatrix}, \qquad (1)$$

where A_{00}, A_0, A_{1i}, and A_{2i} are of order $m(S+1)$. We establish a number of results including deriving expressions for the marginal excess inventory.

We propose two truncation methods, one based on Neuts-Rao truncation and the other based on direct truncation, to arrive at the steady-state solution. A special case of the model dealing with $M/M/\infty$ is considered. We also provide an illustrative numerical example to highlight the behavior of the system under study and a few concluding remarks are given.

2. REFERENCES

[1] O. Berman, E. Kim: Dynamic inventory strategies for profit maximization in a service facility with stochastic service, demand and lead time, Math.Meth.Oper.Res, 60, 497 - 521 (2004).

[2] O. Berman, K. P. Sapna: Optimal service rates of a service facility with perishable inventory items, Naval Research Logistics, 49, 464 - 482 (2002).

[3] S.R. Chakravarthy: The batch Markovian arrival process: a review and future work, A. Krishnamoorthy et al. (Eds.), Advances in Probability Theory and Stochastic Process: Proc., Notable Publications, NJ, 21-49 (2001).

[4] S.R. Chakravarthy, A. Krishnamoorthy, V. C. Joshua: Analysis of a Multi-server Retrial Queue with Search of Customers from the Orbit, Performance Evaluation, 63, 776-798 (2006).

[5] S. R. Chakravarthy, Arunava Maity, U. C. Gupta: An (s, S) inventory in a queueing system with batch service facility, Ann Oper Res DOI 10.1007/s10479-015-2041-z, (2015).

[6] T. Collings, C. Stoneman: The $M/M/\infty$ queue with varying arrival and departure rates, Operations Research 24, 760-773 (1976).

[7] B. W. Conolly: The busy period for the infinite capacity server system $M/G/\infty$, In Saudii di probabilita, Statistica e Riceraca Operativa in Onore di G. Pompily, Institute di Calcolo delle Probabilita, Universita di Roma, Oderisi Gubbio Roma, 128 - 130 (1971).

[8] I. Eliazar: The $M/G/\infty$ system revisited: finiteness, summability, long range dependence, and reverse engineering, Queueing Syst. 55 ,71-82 (2007).

[9] R. D. Foley: The non-homogeneous $M/G/\infty$ queue, Opsearch 19, 40-48 (1982).

[10] R. Krenzler, H. Daduna: Loss systems in a random environment-embedded Markov chains analysis (2013) 1-54. http://preprint.math.unihamburg.de/public/papers/prst/prst2013-02.pdf.

[11] R. Krenzler, H. Daduna, Loss systems in a random environment steady-state analysis, Queueing Syst, DOI 10.1007/s11134-014-9426-6, (2014).

[12] J Keilson, L. D. Servi: The matrix $M/M/\infty$ system, Retrial models and Markov Modlated sources, Adv. Appl. Prob. 25(2), 453-471 (1993).

[13] A. Krishnamoorthy, N. Anbazhagan: Perishable Inventory System at Service Facilities with N Policy, Stochastic Analysis and Applications, Vol. 26, 120-135, (2007).

[14] A. Krishnamoorthy, R. Manikandan, B. Lakshmy: A revisit to queueing-inventory system with positive service time, Annals of Operations Research, Vol.207, DOI 10.1007/s10479-013-1437-x (2013).

[15] A. Krishnamoorthy, Narayanan C. Viswanath: Stochastic decomposition in production inventory with service time, EJOR, (2013). http://dx.doi.org/10.1016/j.ejor.2013.01.041

[16] A. Krishnamoorthy, Dhanya Shajin, B. Lakshmy: Product form solution for some queueing-inventory supply chain problem, OPSEARCH (Spriner), DOI 10.1007/s12597-015-0215-8, (2015).

[17] A. Krishnamoorthy, R. Manikandan, Dhanya Shajin: Analysis of multi-server queueing - inventory system, Advances in Operations Research, Vol. 2015, http://dx.doi.org/10.1155/2015/747328, (2015).

[18] L. Liu, B. R. K. Kashyap, J. G. C. Templeton: On the service system $M/M^R/\infty$ with impatient customers, Queueing Syst. 2, 363-372 (1987).

[19] Mirasol: The output of an $M/G/\infty$ queueing system is Poisson, Operations Research 11, 282-284 (1963).

[20] M.F. Neuts, Matrix-geometric solutions in stochastic models: an algorithmic approach. The Johns Hopkins University Press, Baltimore [1994 version is Dover Edition], (1981).

[21] M.F. Neuts, B. M. Rao: Numerical investigation of a multi-server retrial model. Queueing Syst 7:169Ű190, (1990).

[22] G. F. Newell: The $M/G/\infty$ Queue, SIAM Journal of Applied Mathematics 14, 86-88 (1966).

[23] C. A. O'cinneide, P. Purdue: The $M/M/\infty$ queue in a random environment, Journal of Applied Probability 23, 175-184 (1986).

[24] M. F. Ramalhoto: Bounds for the variance of the busy period of the $M/G/\infty$ queue, Adv. Appl. Prob. 16, 929 - 932 (1984).

[25] M. Saffari, S. Asmussen, R. Haji: The M/M/1 queue with inventory, lost sale and general lead times,, Queueing Syst, 75, 65-77 (2013).

[26] M. Schwarz, C. Sauer, H. Daduna, R. Kulik, R. Szekli: M/M/1 queueing systems with inventory, Queueing Syst 54:55-78 (2006).

[27] M. Schwarz, C. Wichelhaus, H. Daduna: Product form models for queueing networks with an inventory, Stochastic Models, 23(4), 627-663 (2007).

[28] K. Sigman, D. Simchi-Levi: Light traffic heuristic for an M/G/1 queue with limited inventory, Annals of OR, 40:371-380 (1992).

[29] B. Sivakumar, G. Arivarignan: A perishable inventory system with service facilities and negative customers, Advanced Modelling and Optimization, Vol.7,No.2, 193 - 210 (2005).

[30] B. Sivakumar, G. Arivarignan: A perishable inventory system at service facilities with negative customers, Information and Management Sciences, vol.17, no.2, 1 - 18 (2006).

[31] V. Stadje: The busy period of the queueing system $M/G/\infty$, Journal of Appl. Prob. 22, 697-704 (1985).

[32] Stewart W.J.(1994). Introduction to the Numerical Solution of Markov Chains. Princeton, NJ: Princeton University Press.

Performance Modeling of Delay-based Dynamic Speed Scaling

Caglar Tunc
Department of
Electrical and Electronics Engineering
Bilkent University
Ankara, Turkey
caglar@ee.bilkent.edu.tr

Nail Akar
Department of
Electrical and Electronics Engineering
Bilkent University
Ankara, Turkey
akar@ee.bilkent.edu.tr

ABSTRACT

Dynamic speed scaling refers to the continuous adjustment of the service speed of a server to balance performance and power consumption. In this paper, we study delay-based dynamic speed scaling in which the server selects a different service rate with a certain power consumption for the head-of-line (HOL) job according to its delay already experienced in the queue. As an extension, we also study the case where the jobs have deadlines. In this case, when the delay exceeds a certain threshold, the HOL job abandons the system without service in which case the job is said to be blocked. Under Poisson job arrivals and exponentially distributed job service times, we propose a multi-regime Markov fluid queue model to obtain the average power consumption, job blocking probabilities, and the distribution of delays experienced by served jobs. We validate the proposed model by simulations and evaluate the performance of a specific dynamic speed scaling scheme in terms of job blocking probabilities and attainable power gain in comparison with a fixed-rate server.

Keywords

dynmic speed scaling, multi-rate server, Markov fluid queue

1. INTRODUCTION

Speed scaling adapts the speed of a computer or communications system to tradeoff energy and performance [11]. In static speed scaling, a single speed is employed unless the system is idle and is put into a sleep mode when idle [11]. In dynamic speed scaling which is the focus of the current paper, the speed is adapted continuously based on the instantaneous state, i.e., number of jobs in the system, delay experienced by jobs, etc. Modern processors and computer systems allow dynamic speed scaling which leads the way to investigate its impact on fairness among jobs, amount of delay experienced by each job, power efficiency, etc. [3],[5],[6],[8]. The reference [2] studies policies for setting the speed of a processor towards optimization of the energy used and the maximum temperature attained. The speed of an Ethernet link is adapted in [7] by means of dynamic speed scaling for energy efficiency purposes.

In this paper, we study the performance of a single server which can be configured to serve a job using one of available service rates drawn from a *finite set*. Each service rate is associated with a distinct power consumption figure during the service. A service rate (or a power level) is then decided for the HOL job at the service start epoch according to the delay already experienced by this job in the queue. Furthermore, as an extension, we also allow the jobs to have strict delay deadlines as assumed in [12]. Therefore, for such delay intolerant systems, when the delay experienced by the HOL job exceeds a certain threshold, the job abandons the system without service, i.e., the job is blocked. This abandonment may also take place at the arrival epoch if service times of all jobs are known a-priori, or exactly when the queuing time hits the delay deadline while the job is waiting. However, as far as analysis of abandonment systems are concerned, the abandonment epoch is immaterial.

For analytical modeling, we assume Poisson job arrivals and exponentially distributed service times but the framework is amenable to more general distributions which is left for future research. A multi-regime Markov Fluid Queue (MRMFQ) model is proposed for this system, the steady-state solution of which provides expressions for the average power, job blocking probabilities, and the delays of the served jobs. The MRMFQ solver that we use is purely matrix-analytical and relies on ordered Schur decompositions and Sylvester equation solvers as its main engine. Once a numerical solution for these performance metrics through MRMFQs is in hand, one can use this analysis as an instrument to obtain sub-optimal dynamic speed scaling policies that attempt to minimize the power consumption while meeting delay or blocking performance requirements.

The paper is organized as follows. MRMFQs are briefly described in Section 2 along with the boundary conditions necessary to solve their steady-state distribution. The dynamic speed scaling model that we propose and its solution are given in Section 3. Numerical results are presented in Section 4. Finally, we conclude.

2. MARKOV FLUID QUEUES

In fluid queue models, a fluid acts as the input to and output of a buffer. In particular, Markov Fluid Queues (MFQ) are described by a joint Markovian process $(X(t), Z(t))$ where $X(t)$ represents the fluid level (or buffer content) and $Z(t)$ is an underlying finite state-space continuous-time Markov chain that determines the drift, i.e., the rate at which the buffer content $X(t)$ changes. The process $Z(t)$ is called the background (or modulating) process of the MFQ. MRMFQs are generalizations of single-regime MFQs in the sense that the buffer space in MRMFQs (also called multi-threshold, level-dependent, multi-layer, or feedback MFQs) is partitioned into a finite number of non-overlapping intervals which are called the regimes (or layers) of the MRMFQ [1],[4],[9],[10]. In MRMFQs, the infinitesimal generator of the background CTMC as well as the drift into the buffer depend on the regime at which the buffer level resides. The material below for the brief description of infinite-buffer MRMFQs and their notation is based on [9].

In an infinite-size MRMFQ, the buffer[1] is partitioned into $K > 1$ regimes with the boundaries $0 = T^{(0)} < T^{(1)} < \cdots < T^{(K-1)} < T^{(K)} = \infty$. If $T^{(k-1)} < X(t) < T^{(k)}$, the system is said to be in regime k at time t. Let $X(t) \in [0, \infty)$ and $Z(t) \in \{0, 1, \ldots, N-1\}$ denote the buffer content and the background process, respectively, at time t, as in usual MFQs. We denote the infinitesimal generator and drift matrices associated with regime k by $Q^{(k)}$ and $R^{(k)}$, respectively, for $1 \leq k \leq K$. The regime-k drift matrix $R^{(k)}$ is the diagonal matrix

$$R^{(k)} = \mathbf{diag}(r_0^{(k)}, r_1^{(k)}, \ldots, r_{N-1}^{(k)}),$$

where $r_i^{(k)}$ is the net drift of the buffer at state i and regime k. Note that $Q^{(k)}$ and $R^{(k)}$ are fixed within a given regime. Similar to $Q^{(k)}$ and $R^{(k)}$, we define $\tilde{Q}^{(k)}$ and $\tilde{R}^{(k)}$ as the infinitesimal generator and drift matrices associated with the boundary $T^{(k)}$ for $0 \leq k \leq K-1$, where the drift of state i at the boundary $T^{(k)}$ is denoted by $\tilde{r}_i^{(k)}$. We define the joint pdf vector $f^{(k)}(x)$ for regime k when $T^{(k-1)} < x < T^{(k)}$ as follows:

$$\begin{aligned} f_i^{(k)}(x) &= \lim_{t\to\infty} \frac{d}{dx} \Pr\{X(t) \leq x,\ Z(t) = i\}, & (1) \\ f^{(k)}(x) &= \begin{bmatrix} f_0^{(k)}(x) & f_1^{(k)}(x) & \ldots & f_{N-1}^{(k)}(x) \end{bmatrix}. & (2) \end{aligned}$$

Similarly, the steady-state probability mass accumulation vector $c^{(k)}$ is defined for each boundary point $T^{(k)}$ for $0 \leq k \leq K-1$ as follows:

$$\begin{aligned} c_i^{(k)} &= \lim_{t\to\infty} \Pr\{X(t) = T^{(k)},\ Z(t) = i\}, & (3) \\ c^{(k)} &= \begin{bmatrix} c_0^{(k)} & c_1^{(k)} & \ldots & c_{N-1}^{(k)} \end{bmatrix}. & (4) \end{aligned}$$

Note that probability mass accumulations cannot occur at $T^{(K)} = \infty$. Based on [9], the following set of differential equations holds for the joint pdf vector:

$$\frac{d}{dx} f^{(k)}(x) R^{(k)} = f^{(k)}(x) Q^{(k)}, \qquad (5)$$

with the following set of boundary conditions:

$$c_i^{(0)} = 0, \quad \forall i \in S_+^{(1)} \qquad (6)$$

$$c_i^{(k)} = 0, \quad \forall i \in \left(S_+^{(k)} \cap S_+^{(k+1)}\right) \cup \left(S_-^{(k)} \cap S_-^{(k+1)}\right) \qquad (7)$$

$$c_i^{(k)} = 0, \quad \forall i \in \left(S_-^{(k)} \cap S_+^{(k+1)}\right) \cap \left(\tilde{S}_+^{(k)} \cup S_-^{(k)}\right) \qquad (8)$$

$$f^{(1)}(0+) R^{(1)} = c^{(0)} \tilde{Q}^{(0)} \qquad (9)$$

$$f^{(k+1)}(T^{(k)}+) R^{(k+1)} - f^{(k)}(T^{(k)}-) R^{(k)} = c^{(k)} \tilde{Q}^{(k)} \qquad (10)$$

$$f_i^{(k)}(T^{(k)}-) = 0 \quad \forall i \in S_-^{(k)} \cup \left(\tilde{S}_0^{(k)} \cap \tilde{S}_+^{(k)}\right) \qquad (11)$$

$$f_i^{(k+1)}(T^{(k)}+) = 0 \quad \forall i \in \left(\tilde{S}_0^{(k)} \cap \tilde{S}_-^{(k)}\right) \cup S_+^{(k+1)} \qquad (12)$$

$$\left(\sum_{k=1}^{K} \int_{T^{(k-1)}+}^{T^{(k)}-} f^{(k)}(x) dx + \sum_{k=0}^{K-1} c^{(k)} \right) \mathbf{1} = 1 \qquad (13)$$

where $\mathbf{1}$ denotes a column vector of ones of appropriate size. Assuming invertibility of the per-regime drift matrices[2], the following similarity transformation is applied to the per-regime matrix $A^{(k)} = Q^{(k)} \left(R^{(k)}\right)^{-1}$ as follows:

$$A^{(k)} Y^{(k)} = Y^{(k)} \begin{bmatrix} \mathbf{0} & & \\ & A_-^{(k)} & \\ & & A_+^{(k)} \end{bmatrix}, \qquad (14)$$

where $A_-^{(k)}$ and $A_+^{(k)}$ have eigenvalues in the open left and open right half planes, respectively, and $Y^{(k)}$ is partitioned as:

$$\left(Y^{(k)}\right)^{-1} = \begin{bmatrix} L_0^{(k)} \\ L_-^{(k)} \\ L_+^{(k)} \end{bmatrix}. \qquad (15)$$

The matrix $Y^{(k)}$ can be obtained in a computationally stable and efficient way using the ordered Schur decomposition and a pair of Sylvester equations as shown in [9]. Subsequently, the joint pdf vector $f^{(k)}(x)$ for each regime k is given by the following matrix-exponential form:

$$f^{(k)}(x) = a^{(k)} \begin{bmatrix} L_0^{(k)} \\ e^{A_-^{(k)}(x - T^{(k-1)})} L_-^{(k)} \\ e^{-A_+^{(k)}(T^{(k)} - x)} L_+^{(k)} \end{bmatrix}, \qquad (16)$$

for $T^{(k-1)} < x < T^{(k)}$ and

$$a^{(k)} = \begin{bmatrix} a_0^{(k)} & a_-^{(k)} & a_+^{(k)} \end{bmatrix}$$

is the vector of unknown coefficients to be solved for. Since the regime K is of infinite size, the stability condition

$$\pi^{(K)} R^{(K)} \mathbf{1} < 0 \qquad (17)$$

should be satisfied where $\pi^{(K)}$ is the steady-state vector of $Q^{(K)}$. In addition to the boundary conditions (6)-(13) and the stability condition (17), $a_0^{(K)} = 0$ and $a_+^{(K)} = \mathbf{0}$ should hold since (16) must be bounded in regime K. Finally, one can solve for the unknowns $a^{(k)}$ and $c^{(k)}$ by re-writing equations (6)-(13) (using Eqn. (16)) in terms of the unknowns. This algorithm requires the solution of a linear matrix equation of at most size $N(2K+1)$. The computational complexity of the proposed algorithm can be reduced to $\mathcal{O}(N^3 K)$ on the basis of the observation that the linear matrix equation is in block tridiagonal form [13].

[1] Note that the buffer may be of finite or infinite size in the more general case, the latter of which will be of interest in this study.

[2] When there are states with zero drifts, one can follow the procedure described in [9] to handle the case of singular per-regime drift matrices.

3. SYSTEM MODEL

We will now describe the system model for the delay-based dynamic speed scaling system. For this system, we assume that the jobs arrive at the server according to a Poisson process with rate λ and service times are exponentially distributed with parameter μ_k for $1 \leq k \leq K+1$ where $\{\mu_1, \mu_2, \ldots, \mu_{K+1}\}$ is the set of service rates (typically in ascending order), selected from the interval $[\mu_{min}, \mu_{max}]$. Regime boundaries are ordered as $0 = T^{(0)} < T^{(1)} < \cdots < T^{(K)} < T^{(K+1)} = \infty$. When $T^{(k-1)} \leq D(t) < T^{(k)}$, then the HOL job is served with rate μ_k where $D(t)$ denotes the delay already experienced by the HOL job at service start time t. We let $S(t)$ (sojourn time process) denote the sojourn time of the job being served by the server. If there are no jobs being served at time t, then $S(t) = 0$. Moreover, let $A(t)$ (unfinished work process) denote the unfinished work in the system at time t. It is clear that a job arriving at the system at time t with $T^{(k-1)} \leq A(t) < T^{(k)}$ is to be eventually served at rate μ_k.

The sample paths for the two processes $S(t)$ and $A(t)$ are given in figures 1a and 1b, respectively, for an example scenario with two thresholds $T^{(1)} = 2$ and $T^{(2)} = 4$ and for the case of job arrivals occurring at $t = 0, 2, 3, 4, 8, 13$. For the sake of convenience, the service times in regimes 1, 2, and 3, are deterministically set to 3, 2, and 1, respectively, in this example. Due to abrupt jumps in both processes, neither of the two processes can be represented as an MFQ with finite drifts. Therefore, we propose a $(K+1)$-regime MFQ, namely the joint process $(X(t), Z(t))$ where the fluid level $X(t)$ is obtained by replacing abrupt downward jumps in $S(t)$ by linear decrements corresponding to a drift of minus unity. Fig. 1c depicts $X(t)$ for the same example.

The sample path followed by $X(t)$ can indeed be modeled by an MRMFQ. Moreover, it is clear from sample path arguments that the steady-state distribution of the process $S(t)$ ($A(t)$) can be derived from that of $(X(t), Z(t))$ by censoring out the states corresponding to negative (positive) drifts. Therefore, we will first focus on the MRMFQ model for $X(t)$. For this purpose, I_k is defined as the *service state* in regime k for $k = 1, \ldots, K+1$ and in this state, the job is being served with rate μ_k and $X(t)$ is increased with a drift of 1. When the service of the current job completes in state I_k, the system transits into state $\mathcal{D}$ during which $X(t)$ is decreased with a drift of 1 for an exponentially distributed amount of time with mean $1/\lambda$ so that the delay of the new HOL job is reduced by an amount corresponding to its inter-arrival time. If $T^{(k-1)} \leq X(t) < T^{(k)}$ for some $k \leq K$, the system transits into state I_k and so on. However, if $X(t) \geq T^{(K)}$, either of the following cases may follow: i) the HOL job is served with some finite rate μ_{K+1} if the system is delay-tolerant, or ii) the HOL job is blocked because of the excessive delay. For both cases, we will refer to $T^{(K)}$ as the *delay threshold*. Moreover, $X(t)$ may hit zero in state $\mathcal{D}$ meaning that there are no jobs waiting in the queue. When $X(t) = 0$, once a job arrives at the system, the server selects a service rate of μ_1 for this new job. Hence, the only transition at the boundary $X(t) = 0$ occurs out of state $\mathcal{D}$ into state I_1 with rate λ. With states $\mathcal{D}$ and I_i for $i = 1, \ldots, K+1$, the background process, denoted by $Z(t)$, has $K+2$ states in total. State transitions for the possible cases are illustrated in Figure 2. Moreover, with the states ordered as $I_{K+1}, I_K, \ldots, I_1, \mathcal{D}$, the infinitesimal generator matrix of regime-j, denoted by $Q^{(j)}$, for $j = 1, \ldots, K+1$ can be written as follows: $Q^{(j)} =$

$$
\begin{array}{c|cccccccc|}
 & I_{K+1} & \cdots & I_{j+1} & I_j & I_{j-1} & \cdots & I_1 & \mathcal{D} \\
I_{K+1} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
I_{j+1} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\
I_j & 0 & \cdots & 0 & -\mu_j & 0 & \cdots & 0 & \mu_j \\
I_{j-1} & 0 & \cdots & 0 & 0 & -\mu_{j-1} & \cdots & 0 & \mu_{j-1} \\
\vdots & \vdots & & \vdots & \vdots & & \ddots & \vdots & \vdots \\
I_1 & 0 & \cdots & 0 & 0 & 0 & \cdots & -\mu_1 & \mu_1 \\
\mathcal{D} & 0 & \cdots & 0 & \lambda & 0 & \cdots & 0 & -\lambda
\end{array}
$$

Note that since $X(t)$ increases in the service state I_k for $k = 1, \ldots, j-1$, there may be transitions from state I_k to state $\mathcal{D}$ in regime j for $k \leq j$. We set $\tilde{Q}^{(j)} = Q^{(j+1)}$ as the generator at boundary-j for $j = 1, \ldots, K$. $\tilde{Q}^{(0)}$ is similar to $Q^{(1)}$ except that there is no transition from state I_1 to state $\mathcal{D}$ at boundary-0 and only transition is from state $\mathcal{D}$ to state I_1. Moreover, drift matrices at regime-k and boundary-k, denoted by $R^{(k)}$ and $\tilde{R}^{(k)}$, respectively, are written as follows:

$$R^{(k)} = \mathbf{diag}(\mathbf{I}, -1), \quad 1 \leq k \leq K+1, \tag{18}$$

$$\tilde{R}^{(k)} = \begin{cases} R^{(k+1)}, & 1 \leq k \leq K, \\ \mathbf{max}(0, R^{(1)}), & k = 0, \end{cases} \tag{19}$$

where $\mathbf{max}$ is the element-wise operator and $\mathbf{I}$ denotes an identity matrix of appropriate size.

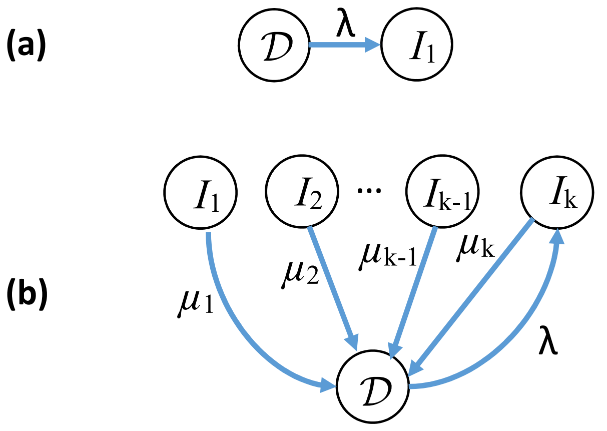


Figure 2: State transitions (a) for $X(t) = 0$ and (b) in regime k for $k = 1, ..., K+1$.

3.1 Steady-state Solution

Since the unfinished work process $A(t)$ determines the amount of delay that newly arriving jobs (which arrive to the system according to a Poisson process) will experience, the steady-state probability distribution of state $\mathcal{D}$ can be used to obtain the quantities of interest including the average system power, blocking probability, and the delay distribution, by a direct consequence of the *PASTA* property. Therefore, in order to obtain the steady-state distribution of $A(t)$ from that of the fluid process $(X(t), Z(t))$, we censor out all the service states and subsequently normalize the steady-state distributions. In mathematical terms, we calculate the steady-state distribution of $A(t)$ from that of $(X(t), Z(t))$ as

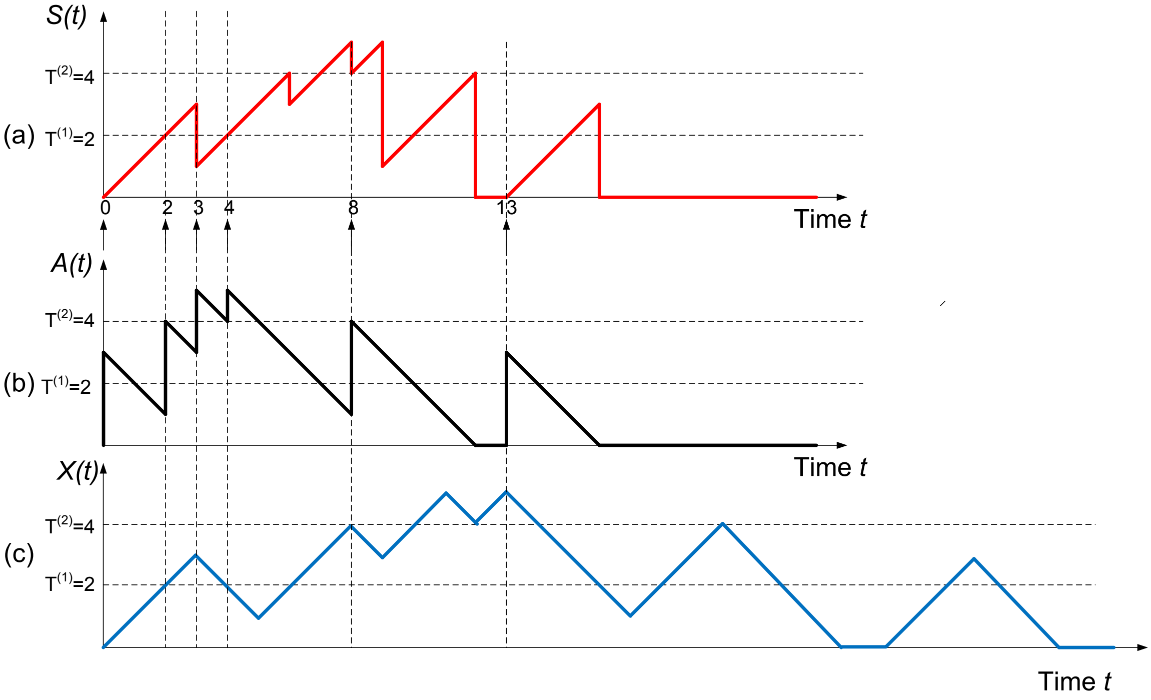


Figure 1: Sample paths of (a) $S(t)$, (b) $A(t)$ and (c) $X(t)$.

follows:

$$\lim_{t\to\infty} \Pr\{A(t) \le x\} = \lim_{t\to\infty} \frac{\Pr\{Z(t) = \mathcal{D}, X(t) \le x\}}{\Pr\{Z(t) = \mathcal{D}\}}. \quad (20)$$

We denote the probability that a newly arriving job finds the system in regime k by p_k for $k = 1, \ldots, K+1$. Mathematically,

$$p_k = \lim_{t\to\infty} \Pr\{T^{(k-1)} < A(t) < T^{(k)}\}, 1 \le k \le K+1. \quad (21)$$

Moreover, we denote the probability that a newly arriving job finds the queue empty by p_0, i.e., $p_0 = \lim_{t\to\infty} \Pr\{A(t) = 0\}$. Similarly, we denote the probability that a job is served with rate μ_k by q_k for $k = 1, \ldots, K+1$. Note that $q_k = p_k$ for $k \ge 2$, with the only exception that $q_1 = p_0 + p_1$; because even if a job arrives to the system at boundary-0, it will be served with a rate of μ_1. With these definitions, the average system power P_{avg} can be written as:

$$P_{avg} = p_0 P_I + (1-p_0) \sum_{k=1}^{K+1} \frac{\frac{q_k}{\mu_k}}{\sum\limits_{i=1}^{K+1} \frac{q_i}{\mu_i}} P_k, \quad (22)$$

where P_k denotes the operating power associated with rate μ_k and P_I is the power consumed when the server is idle. The distribution of the delay that the arriving jobs experience can also be computed via the steady-state distribution of $A(t)$. The cumulative distribution function of the steady-state queuing delay $D(t)$, denoted by $F_D(x)$, can be written as:

$$F_D(x) = \lim_{t\to\infty} \Pr\{D(t) \le x\} = \lim_{t\to\infty} \Pr\{A(t) \le x\}, \quad (23)$$

which can directly be obtained from the steady-state solution of $A(t)$.

3.2 The Case of Abandonments

For delay intolerant systems, the HOL jobs which have experienced a delay larger than $T^{(K)}$ will abandon the system without service. In that case, such jobs are said to be blocked. In order to model blocking, we let $\mu_{K+1} \to \infty$ and assume that the server does not consume any energy during this process. Then, we calculate the blocking probability, denoted by p_b, as follows:

$$p_b = \lim_{\mu_{K+1}\to\infty} p_{K+1} = \lim_{t\to\infty} \lim_{\mu_{K+1}\to\infty} \Pr\{A(t) \ge T^{(K)}\}. \quad (24)$$

Note that as $\mu_{K+1} \to \infty$, q_{K+1}/μ_{K+1} terms in (22) approach zero. However, in the numerical examples, we set μ_{K+1} to some large value compared to μ_{max} (such as 1e6) and show that such approximations do not give rise to adverse effects on the numerical accuracy of the results. In this case, the identity (22) is slightly modified in order to prevent numerical errors in the computation of average power:

$$P_{avg} = p_0 P_I + (1-p_0) \sum_{k=1}^{K} \frac{\frac{q_k}{\mu_k}}{\sum\limits_{i=1}^{K} \frac{q_i}{\mu_i}} P_k. \quad (25)$$

Since the solution to the dynamic speed scaling problem with or without abandonments has been reduced to the steady-state solution of an MRMFQ with $K+2$ states and $K+1$ regimes, the computational complexity of the overall algorithm is $O(K^4)$.

4. NUMERICAL EXAMPLES

In the numerical examples, we evaluate the performance of the dynamic speed scaling system in terms of the average system power and blocking probability for various values of the parameter K, the delay threshold $T^{(K)}$ and the job arrival rate λ. In average power calculations, we set $P_I = 0$. Similar to the studies in [2],[3],[5], and [6], we assume that $P_k = c\mu_k^{\alpha}$ for $k = 1, \ldots, K+1$, where α is a scaling factor and c is a constant. Moreover, as in [3],[6], we fix the power scaling factor $\alpha = 2$, and the constant c is set to unity.

4.1 Example 1

In the first example, we consider abandonments. For this purpose, we fix $K = 2$, $T^{(1)} = 10$, $T^{(2)} = 20$, $\mu_1 = 0.5$, $\mu_2 = 1$ and compare the blocking probability p_b and average system power P_{avg} with simulation results as μ_3 is increased beyond $\mu_{max} = 1$, indicating that HOL jobs with delays greater than $T^{(2)}$ abandon the system without service. In the simulations, we run a single instance until 1e7 jobs are blocked. Four values of μ_3, namely 1e2, 1e4, 1e6, and 1e8, are tested. We define $\eta = \lambda/\mu_{max}$ as the load on the system and for $\eta = 0.4, 0.8$, which correspond to light and heavy loads, respectively, the results are tabulated in Table 1. As it can be observed from Table 1, for both heavy and light load cases, the blocking probability converges beyond $\mu_3 = $ 1e6 to values which have negligible differences from the simulation results. Hence, for the remaining examples with blocking, we propose in our MFQ model that μ_{K+1} is set to 1e6.

4.2 Example 2

In the second example, we propose and investigate a specific rate adjustment policy, which will be referred to as the *Piecewise Linear Rate Adjustment Policy (PiLRAP)*. In particular, PiLRAP selects service rates from piecewise linear functions of the unfinished work process $A(t)$ (or delay $D(t)$ depending on when the service rate selection is to be made) from the interval $[\mu_{min}, \mu_{max}]$. We assume that the service rate of regime K is μ_{max} in order to serve the jobs in this regime more aggressively, and jobs with delays larger than $T^{(K)}$ are blocked.

We first define the point (x_0, y_0) such that the rate function is piecewise linear in the intervals $[(0,0),(x_0,y_0)]$ and $[(x_0,y_0),(T^{(K)},\mu_{max})]$. For a given $K < \infty$, we let $x_0 = lT^{(K)}/K$ and $y_0 = m\mu_{max}/K$ for $0 \le l \le K$, $1 \le m \le K$. An example of the service rate function of interest is depicted in Figure 3 when $\mu_{min} = 0$, $\mu_{max} = 1$, $T^{(K)} = 10$, $K = 10$ and three different (x_0, y_0) points.

The point (x_0, y_0) determines the service rates assigned for different unfinished work values and K identifies the number of service rates, or power levels, that the server will use. Consequently, the PiLRAP system parameters K, x_0, and y_0 have significant impact on the performance of the system in terms of average system power P_{avg} and blocking probability p_b. In the current example, we fix K to 20 and study how P_{avg} and p_b change as a function of x_0 and y_0 in Figure 4 for $\lambda = 0.6$, $\mu_{max} = 1$ and $T^{(K)} = 20$. In general, lower blocking probabilities can be obtained by trading off the average system power. For a given parameter set (including K), one can obtain from Figure 4 the optimal point (x_0, y_0) such that the average power consumption is minimized while the blocking probability is kept under a certain desired value.

Table 1: Blocking probability p_b and average system power P_{avg} compared with simulation results for two values of $\eta = 0.4, 0.8$.

μ_3	p_b (%)		P_{avg}	
	$\eta = 0.4$	$\eta = 0.8$	$\eta = 0.4$	$\eta = 0.8$
1e2	0.1123	3.0429	0.2238	0.6662
1e4	0.1118	3.0196	0.2238	0.6664
1e6	0.1118	3.0193	0.2238	0.6664
1e8	0.1118	3.0193	0.2238	0.6664
Sim	0.1118	3.0185	0.2238	0.6664

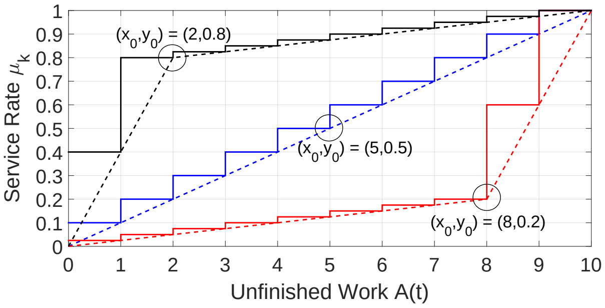


Figure 3: Service rate function (dashed lines) and actual service rate μ_k (straight lines) as functions of the unfinished work $A(t)$ for $\mu_{min} = 0$, $\mu_{max} = 1$, $T^{(K)} = 10$, $K = 10$ and $(x_0, y_0) \in \{(2, 0.8), (5, 0.5), (8, 0.2)\}$.

4.3 Example 3

In the final example, we investigate how the number of available service rates K and load η impact the location of the optimal point (x_0, y_0). In order to evaluate the performance of PiLRAP in terms of power savings, we define the percentage power gain G as follows:

$$G = 100\frac{(P_f - P_{avg})}{P_f} \tag{26}$$

where P_f is the average system power of the benchmark policy which has a fixed service rate, i.e., $K = 1$ and $\mu_1 = \mu_{max}$, which is an M/M/1 queue with a service rate of μ_{max}. The average power consumption of the M/M/1 server with a load of $\rho = \lambda/\mu_{max}$ can be written as follows:

$$P_f = (1-\rho)P_I + \rho P_1 \tag{27}$$

since the operating power of the server is either P_1 or zero depending on whether the server is busy or not. We also define (x_0^*, y_0^*) as the optimal (x_0, y_0) point that minimizes P_{avg} while keeping p_b below a threshold set to 0.01 in this example. Similarly, we define the attainable power gain, denoted by G^*, as the power gain obtained by choosing $(x_0, y_0) = (x_0^*, y_0^*)$. We demonstrate how x_0^*, y_0^* and G^* change as functions of K in figures 5, 6 and 7, respectively. It can be observed from figures 5 and 6 that as K increases, the optimal point (x_0^*, y_0^*) favors a more aggressive rate adjustment policy by reducing x_0^* and assigning high service rates even for very low unfinished work values. Moreover, we observe that as the load η increases, y_0^* also increases so that the server can handle the increased load. However, there are some exceptional noisy points, which is due to the fact that the service rates can only by selected from integer multiples of μ_{max}/K which causes a certain non-monotonic behavior.

We also observe from Figure 7 that the relative change in the attainable power gain is only marginal beyond $K =$

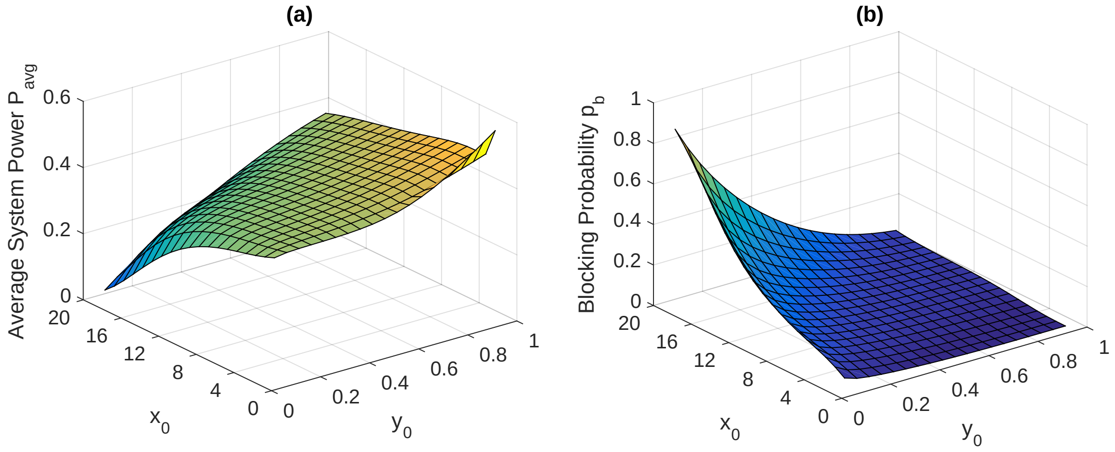


Figure 4: (a) Average system power P_{avg} and (b) blocking probability p_b as functions of parameters x_0 and y_0 for $K = 20$.

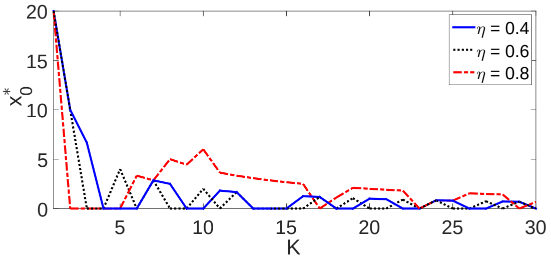


Figure 5: x_0^* as a function of K for $\eta = 0.4, 0.6, 0.8$.

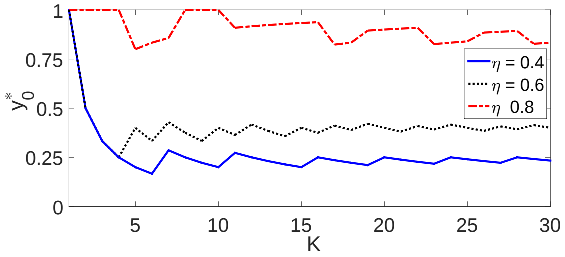


Figure 6: y_0^* as a function of K for $\eta = 0.4, 0.6, 0.8$.

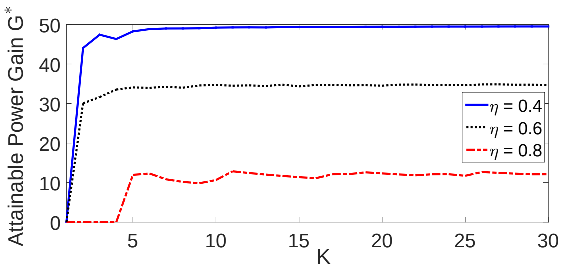


Figure 7: G^* as a function of K for $\eta = 0.4, 0.6, 0.8$.

20 for all of the three load values tried. Since this is the case, we now fix $K = 20$ and plot the attainable power gain as a function of load η in Figure 8. The attainable power gain G^* decreases monotonically as the load η approaches 1. This is due to the fact that as η gets closer to 1, the server tends to select higher service rates which approach μ_{max} in order to handle high arrival rates and to keep the blocking probability p_b below the threshold. Therefore, we can conclude that the gain in terms of average system power is significant especially for lightly and moderately loaded systems.

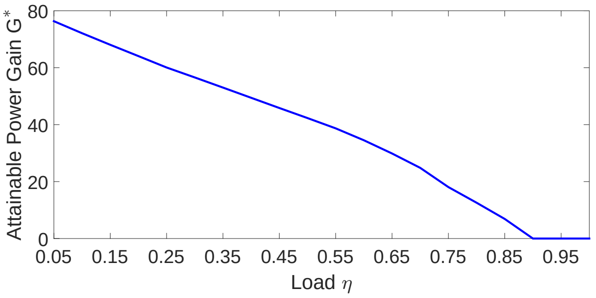


Figure 8: Attainable power gain G^* as a function of the load η for $K = 20$.

5. CONCLUSIONS

In this study, we propose a methodology to model a dynamic speed scaling system in which the service rate is to be chosen from a finite set of values depending on the queue waiting time of the HOL job at the epoch of service start. We envision two possible scenarios: HOL jobs with delays greater than a certain threshold may either be served with a finite service rate or they abandon the system. In the numerical examples, we focused on the latter case. We proposed a multi-regime Markov fluid queue model, the steady-state solution of which is used to calculate the average power consumption, the blocking probability, and the distribution of queue waiting times. Using the MRMFQ solver as an instrument, a specific dynamic speed scaling policy, namely the Piecewise Linear Rate Adjustment Policy (PiLRAP) is investigated to find the optimum parameters of this policy giving rise to minimized power consumption under blocking probability constraints. Future work will consist of development of stochastic models for more general arrival processes and service time distributions as well as the investigation of more general dynamic speed scaling policies (as extensions of PiLRAP). The case of zero drifts in MRMFQs will also be investigated for studying dynamic speed scaling systems with abandonment for dealing with the case $\mu_{K+1} \to \infty$ exactly, and not approximately as in Eqn. (24).

6. ACKNOWLEDGMENTS

The authors would like to thank the Science and Research Council of Turkey (TUBITAK) for supporting this study under project grant EEEAG-115E360.

7. REFERENCES

[1] A. Badescu, S. Drekic, and D. Landriault. On the analysis of a multi-threshold Markovian risk model. *Scandinavian Actuarial Journal*, 2007(4):248–260, 2007.

[2] N. Bansal, T. Kimbrel, and K. Pruhs. Speed scaling to manage energy and temperature. *J. ACM*, 54(1):3:1–3:39, March 2007.

[3] L. Chen and N. Li. On the interaction between load balancing and speed scaling. *Selected Areas in Communications, IEEE Journal on*, 33(12):2567–2578, Dec 2015.

[4] A. da Silva Soares and G. Latouche. Fluid queues with level dependent evolution. *European Journal of Operational Research*, 196(3):1041 – 1048, 2009.

[5] T. V. Dinh, L. L. H. Andrew, and Y. Nazarathy. Architecture and robustness tradeoffs in speed-scaled queues with application to energy management. *Intern. J. Syst. Sci.*, 45(8):1728–1739, aug 2014.

[6] M. Elahi, C. Williamson, and P. Woelfel. Decoupled speed scaling: Analysis and evaluation. *Performance Evaluation*, 73:3 – 17, 2014. Special Issue on the 9th International Conference on Quantitative Evaluation of Systems.

[7] C. Gunaratne, K. Christensen, B. Nordman, and S. Suen. Reducing the Energy Consumption of Ethernet with Adaptive Link Rate (ALR). *Computers, IEEE Transactions on*, 57(4):448–461, April 2008.

[8] J. M. H. Jennifer M. George. Dynamic control of a queue with adjustable service rate. *Operations Research*, 49(5):720–731, 2001.

[9] H. E. Kankaya and N. Akar. Solving multi-regime feedback fluid queues. *Stochastic Models*, 24(3):425–450, 2008.

[10] M. Mandjes, D. Mitra, and W. Scheinhardt. Models of network access using feedback fluid queues. *Queueing Syst. Theory Appl.*, 44(4):2989–3002, 2003.

[11] A. Wierman, L. L. H. Andrew, and M. Lin. Speed scaling: An algorithmic perspective. In *Handbook of Energy-Aware and Green Computing*, pages 385–406. CRC Press, 2 Jan 2012.

[12] F. Yao, A. Demers, and S. Shenker. A Scheduling Model for Reduced CPU Energy. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, FOCS '95, pages 374–, Washington, DC, USA, 1995. IEEE Computer Society.

[13] M. A. Yazici and N. Akar. The finite/infinite horizon ruin problem with multi-threshold premiums: a Markov fluid queue approach. To appear at Annals of Operations Research, 2016.

Shift techniques for Quasi–Birth-and-Death processes: Canonical factorizations and matrix equations

Dario A. Bini
Università di Pisa, Italy
bini@dm.unipi.it

Guy Latouche
Université Libre de Bruxelles, Belgium
latouche@ulb.ac.be

Beatrice Meini
Università di Pisa, Italy
meini@dm.unipi.it

ABSTRACT

We revisit the shift technique applied to Quasi–Birth-and-Death (QBD) processes (He, Meini, and Rhee, SIAM J. Matrix Anal. Appl., 2001) by bringing the attention to the existence and properties of canonical factorizations. To this regard, we prove new results concerning the solutions of the quadratic matrix equations associated with the QBD. These results find applications to the solution of the Poisson equation for QBDs.

Keywords

Quasi-Birth-and-Death processes; Shift technique; Canonical factorizations; Quadratic matrix equations

1. OUTLINE OF THE MAIN RESULTS

Consider a discrete time Quasi–Birth-and-Death process (QBD) defined by the transition probability matrix

$$P = \begin{bmatrix} B & A_1 & & \\ A_{-1} & A_0 & A_1 & \\ & A_{-1} & A_0 & \ddots \\ & & \ddots & \ddots \end{bmatrix} \qquad (1)$$

where B, A_{-1}, A_0 and A_1 are square matrices of order $m < \infty$. Assume that P and $A_{-1} + A_0 + A_1$ are irreducible.

Define the matrix G as the minimal nonnegative solution of the equation

$$A_{-1} + (A_0 - I)X + A_1 X^2 = 0, \qquad (2)$$

and the matrix $\widehat{G}$ as the minimal nonnegative solution of the dual equation [13, 2]

$$A_1 + (A_0 - I)X + A_{-1} X^2 = 0. \qquad (3)$$

For this class of problems, together with (2) and (3), the reversed equations

$$X^2 A_{-1} + X(A_0 - I) + A_1 = 0,$$
$$A_{-1} + X(A_0 - I) + X^2 A_1 = 0,$$

have a relevant interest. These equations have a minimal nonnegative solution R and $\widehat{R}$, respectively, which can be explicitly related to G and $\widehat{G}$ [13, 15]. These solutions have an interesting probabilistic interpretation and their computation is a fundamental task in the analysis of QBD processes. Moreover they provide the factorization

$$\varphi(z) = (I - zR)K(I - z^{-1}G)$$

of the Laurent polynomial $\varphi(z) = z^{-1}A_{-1} + A_0 - I + zA_1$, where K is a nonsingular matrix. A factorization of this kind is canonical if $\rho(R) < 1$ and $\rho(G) < 1$, where ρ denotes the spectral radius. It is weak canonical if $\rho(R) \le 1$ and $\rho(G) \le 1$.

We introduce the matrix polynomial

$$B(z) = A_{-1} + z(A_0 - I) + z^2 A_1 = z\varphi(z)$$

and define the roots of $B(z)$ as the zeros of the polynomial $\det B(z)$. If ξ is a root of $B(z)$ we say that v is an eigenvector associated with ξ if $v \neq 0$ and $B(\xi)v = 0$. The location of the roots of $B(z)$ determines the classification of the QBD as positive, null recurrent or transient, and governs the convergence and the efficiency of the available numerical algorithms for approximating G and R [2]. In particular, $B(z)$ has always a root on the unit circle, namely, the root $\xi = 1$, and the corresponding eigenvector is the vector e of all ones, i.e., $B(1)e = 0$.

If the QBD is recurrent, the root $\xi = 1$ is the eigenvalue of largest modulus of the matrix G and $Ge = e$. In the transient case, that root is the eigenvalue of largest modulus of R. These facts have been used to improve convergence properties of numerical methods for computing the matrix G. The idea, introduced in [11] and based on the results of [5], is to “shift” the root $\xi = 1$ of $B(z)$ to zero or to infinity, and to construct a new quadratic matrix polynomial

$$B_{\boldsymbol{s}}(z) = A_{-1}^{\boldsymbol{s}} + z(A_0^{\boldsymbol{s}} - I) + z^2 A_1^{\boldsymbol{s}}$$

having the same roots as $B(z)$, except for the root equal to 1, which is replaced with 0 or infinity (the super- or subscript $\boldsymbol{s}$ means “shifted”). This idea has been subsequently developed and applied in [4, 8, 9, 10, 12, 14].

In this talk we revisit the shift technique, and we focus on the properties of the canonical factorizations. In particular, we prove new results concerning the existence and properties of the solutions of the quadratic matrix equations obtained after the shift [3].

By following [2], we recall that in the positive recurrent case the root $\xi = 1$ can be shifted to zero by multiplying $B(z)$ to the right by a suitable function (right shift), while in the transient case the root $\xi = 1$ can be shifted to infinity by multiplying $B(z)$ to the left by another suitable function (left shift). In the null recurrent case, where $\xi = 1$ is a root of multiplicity 2, shift is applied both to the left and to the right so that one root 1 is shifted to zero and the other root 1 is shifted to infinity (double shift). In all the cases, the new Laurent matrix polynomial $\varphi_{\boldsymbol{s}}(z) = z^{-1}B_{\boldsymbol{s}}(z)$ is invertible on an annulus containing the unit circle in the complex plane and we prove that it admits a canonical factorization which is related to the weak canonical factorization of $\varphi(z)$. As a consequence, we relate G and R with the solutions $G_{\boldsymbol{s}}$ and $R_{\boldsymbol{s}}$ of minimal spectral radius of the matrix equations

$$A_{-1}^{\boldsymbol{s}} + (A_0^{\boldsymbol{s}} - I)X + A_1^{\boldsymbol{s}}X^2 = 0,$$
$$X^2A_{-1}^{\boldsymbol{s}} + X(A_0^{\boldsymbol{s}} - I) + A_1^{\boldsymbol{s}} = 0,$$

respectively.

A less trivial issue is the existence of the canonical factorization of $\varphi_{\boldsymbol{s}}(z^{-1})$. We show that such factorization exists and we provide an explicit expression for it, for the three different kinds of shifts. The existence of such factorization allows us to express the minimal nonnegative solutions $\widehat{G}$ and $\widehat{R}$ of the matrix equations $A_{-1}X^2 + (A_0 - I)X + A_1 = 0$ and $A_{-1} + X(A_0 - I) + X^2A_1 = 0$, in terms of the solutions of minimal spectral radius $\widehat{G}_{\boldsymbol{s}}$ and $\widehat{R}_{\boldsymbol{s}}$ of the equations

$$A_{-1}^{\boldsymbol{s}}X^2 + (A_0^{\boldsymbol{s}} - I)X + A_1^{\boldsymbol{s}} = 0,$$
$$A_{-1}^{\boldsymbol{s}} + X(A_0^{\boldsymbol{s}} - I) + X^2A_1^{\boldsymbol{s}} = 0,$$

respectively.

The existence of the canonical factorizations of $\varphi_{\boldsymbol{s}}(z)$ and $\varphi_{\boldsymbol{s}}(z^{-1})$ has interesting consequences. Besides providing computational advantages in the numerical solution of matrix equations, it may improve the numerical conditioning of the problem. In fact, while null recurrent problems are ill-conditioned, the shifted counterparts are not. A convenient computational strategy to solve a null recurrent problem consists in transforming it into a new one, say by means of the double shift; solve the latter by using a quadratic convergent algorithm like cyclic reduction or logarithmic reduction [2]; then recover the solution of the original problem from the one of the shifted problem. For this conversion, the expressions relating the solutions of the shifted equations to those of the original equations are fundamental.

Another useful application of these results is the solution of the Poisson problem for QBDs where we are looking for a vector u such that

$$(I - P)u = g$$

where g is a given vector [1], [6]. In fact, using the theory of matrix difference equations and of resolvent triples of [7], we can characterize all the solutions in the positive recurrent and in the transient cases. This theory cannot be applied directly to null recurrent QBDs. However, by performing the shift of eigenvalues, we transform the original Poisson problem into a new one where, unlike in the null recurrent case, the eigenvalue 1 is simple. For the new problem, the theory of matrix difference equations can be applied and the solution of the original problem can be easily related to the solution of the modified problem.

2. REFERENCES

[1] D. A. Bini, S. Dendievel, G. Latouche, and B. Meini. General solution of the Poisson equation for QBDs. In preparation, 2016.

[2] D. A. Bini, G. Latouche, and B. Meini. *Numerical methods for structured Markov chains.* Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2005. Oxford Science Publications.

[3] D. A. Bini, G. Latouche, and B. Meini. Shift techniques for quasi-birth and death processes: canonical factorizations and matrix equations. arXiv:1601.07717, 2016.

[4] D. A. Bini, B. Meini, and I. M. Spitkovsky. Shift techniques and canonical factorizations in the solution of M/G/1-type Markov chains. *Stochastic Models*, 21(2-3):279–302, 2005.

[5] A. Brauer. Limits for the characteristic roots of a matrix. IV. Applications to stochastic matrices. *Duke Math. J.*, 19:75–91, 1952.

[6] S. Dendievel, G. Latouche, and Y. Liu. Poisson's equation for discrete-time quasi-birth-and-death processes. *Performance Evaluation*, 70:564–577, 2013.

[7] I. Gohberg, P. Lancaster, and L. Rodman. *Matrix Polynomials.* Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1982. Computer Science and Applied Mathematics.

[8] C.-H. Guo. Comments on a shifted cyclic reduction algorithm for quasi-birth-death problems. *SIAM J. Matrix Anal. Appl.*, 24(4):1161–1166 (electronic), 2003.

[9] C.-H. Guo and N. J. Higham. Iterative solution of a nonsymmetric algebraic Riccati equation. *SIAM J. Matrix Anal. Appl.*, 29(2):396–412, 2007.

[10] C.-H. Guo, B. Iannazzo, and B. Meini. On the doubling algorithm for a (shifted) nonsymmetric algebraic Riccati equation. *SIAM J. Matrix Anal. Appl.*, 29(4):1083–1100, 2007.

[11] C. He, B. Meini, and N. H. Rhee. A shifted cyclic reduction algorithm for quasi-birth-death problems. *SIAM J. Matrix Anal. Appl.*, 23(3):673–691 (electronic), 2001/02.

[12] B. Iannazzo and F. Poloni. A subspace shift technique for nonsymmetric algebraic Riccati equations associated with an M-matrix. *Numer. Linear Algebra Appl.*, 20(3):440–452, 2013.

[13] G. Latouche and V. Ramaswami. *Introduction to matrix analytic methods in stochastic modeling.* ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.

[14] B. Meini. A "shift-and-deflate" technique for quadratic matrix polynomials. *Linear Algebra and its Applications*, 438(4):1946 – 1961, 2013. 16th ILAS Conference Proceedings, Pisa 2010.

[15] M. F. Neuts. *Matrix-geometric solutions in stochastic models*, volume 2 of *Johns Hopkins Series in the Mathematical Sciences.* Johns Hopkins University Press, Baltimore, Md., 1981.

Algorithms for Stationary Distributions of Fluid Queues: Interpretations and Re-interpretations[*]

Nigel Bean
School of Mathematical Sciences
The University of Adelaide
Adelaide, Australia
nigel.bean@adelaide.edu.au

Giang T. Nguyen
School of Mathematical Sciences
The University of Adelaide
Adelaide, Australia
giang.nguyen@adelaide.edu.au

Federico Poloni
Dipartimento di Informatica
Università di Pisa
Pisa, Italy
federico.poloni@unipi.it

ABSTRACT

We give a probabilistic interpretation of the family of algorithms known as *doubling*, which are the most effective algorithms for computing the return probability matrix Ψ, the key ingredient in the stationary distribution of a stochastic fluid queue.

To this end, we first revisit the links presented in [1, 8] between fluid queues and quasi-birth-death processes. In particular, we give alternative interpretations for these connections, and then generalize this framework of understanding to give a probabilistic meaning for the initial step of doubling algorithms. Then, we give an interpretation for the iterative step of these algorithms.

CCS Concepts

•**Mathematics of computing** → **Probabilistic algorithms;** *Markov processes;*

Keywords

doubling algorithms; stochastic fluid flows; quasi-birth-death processes; stationary distribution

1. INTRODUCTION

Stochastic fluid queues are two-dimensional Markov processes frequently used for modeling real-life applications. In a fluid queue $\{X_t, \varphi_t\}_{t\geq 0}$, the *phase* φ_t is a continuous-time Markov chain on a finite state space $\mathcal{S}$, and the *level* X_t varies linearly at rate c_i whenever $\varphi_t = i$, $i \in \mathcal{S}$. We assume $X_t \in [0, \infty)$, that is, there exists a regulated boundary at 0. The joint stationary distribution of the level and the phase has been well-analyzed. Asmussen [2], Rogers [15], and Karandikar and Kulkarni [11] independently derived this distribution using different approaches: time reversal, the theory of generators of Markov processes, Wiener-Hopf factorization, and partial differential equations. More recently, Ramaswami [14] and da Silva Soares and Latouche [8] obtained new representations using matrix-analytic methods.

The key component for obtaining the stationary distribution is the probability matrix Ψ, of which each entry Ψ_{ij} is the probability of the fluid returning, from above, to the initial level x in phase j, after starting in phase i and avoiding all levels below x. This matrix Ψ is also the minimal nonnegative solution to a *nonsymmetric algebraic Riccati equation* (NARE) of the form

$$B - AX - XD + XCX = 0. \tag{1}$$

These probabilistic and algebraic characterizations of Ψ have led to considerable efforts in developing algorithms for computing the matrix efficiently. Asmussen [2] presented three iterative schemes, while Guo [9] analyzed fixed-point iterations and Newton's method. Following a different path, Ramaswami [14] and da Silva Soares and Latouche [8] proved that one can approximate fluid processes using quasi-birth-death (QBD) processes, thus allowing quadratically convergent algorithms originally developed for QBDs—such as Logarithmic Reduction [12] and Cyclic Reduction [6]—to be used for solving for Ψ. Bean *et al.* [4] proposed First-Exit and Last-Entrance and gave probabilistic interpretations for these two algorithms, as well as for the Newton's method, one of Asmussen's iterative schemes, and the Logarithmic Reduction applied to the QBD version of fluid processes.

Here, we focus on a family of algorithms for solving (1), known as *doubling*, including structure-preserving doubling algorithm (SDA) [10], SDA shrink-and-shift (SDA-ss) [7], and alternating-directional doubling algorithm (ADDA) [16]. These algorithms are known to be more computationally efficient than all the ones treated in [2, 14, 4]. We give these doubling algorithms a probabilistic interpretation, which to the best of our knowledge is the first one available. Probabilistic interpretations are useful as they give an intuitive explanation of how a numerical algorithm works, which in turn allows for shorter and more elegant proofs, as well as improvements and generalizations of the algorithm. The combination of purely linear-algebraic manipulations and probabilistic understanding has paved ways for many significant theoretical developments, as already seen for QBDs [13, 5] and fluid queues [14, 8, 4].

Thus, the aim of this work is twofold. One is to understand doubling algorithms more thoroughly from a proba-

[*]N. Bean and G. Nguyen acknowledge the financial support of the Australian Research Council through the Discovery Grant DP110101663, and of the ARC Centre of Excellence for Mathematical and Statistical Frontiers. F. Poloni is partially supported by a PRA research project of the University of Pisa.

bilistic point of view, with an eye to possible future generalizations. The other is to make the algorithms—which were developed purely from a linear algebra perspective—more accessible to probabilists who work on stochastic fluid flows.

In Section 2, we present the doubling algorithms. In Section 3, we give alternative interpretations for the links in [1, 8] between fluid queues and QBD processes, and then generalize these to give a probabilistic meaning for the initial starting point of doubling algorithms; we also offer an interpretation for the iterations of these algorithms.

2. DOUBLING ALGORITHMS

Let $\mathcal{S}_+ := \{i \in \mathcal{S} : c_i > 0\}$, $\mathcal{S}_- := \{i \in \mathcal{S} : c_i < 0\}$, $n := |\mathcal{S}|$, $n_+ := |\mathcal{S}_+|$, and $n_- := |\mathcal{S}_-|$. We denote by C the diagonal rate matrix $\operatorname{diag}(c_i)_{i\in\mathcal{S}}$ for the level X_t, and by T the generator of the phase process φ_t.

Given a matrix $P \in \mathbb{R}^{n\times n}$, we partition P as follows

$$P = \begin{bmatrix} E & G \\ H & F \end{bmatrix}, \text{ with } E \in \mathbb{R}^{n_+\times n_+}, F \in \mathbb{R}^{n_-\times n_-}. \tag{2}$$

We also partition C and T in a similar manner, into sub-blocks C_+, C_-, T_{++}, etc. Define

$$\widehat{E} := E(I - GH)^{-1}E, \tag{3a}$$

$$\widehat{F} := F(I - HG)^{-1}F, \tag{3b}$$

$$\widehat{G} := G + E(I - GH)^{-1}GF, \tag{3c}$$

$$\widehat{H} := H + F(I - HG)^{-1}HE. \tag{3d}$$

Then the function

$$\mathcal{F}(P) := \begin{bmatrix} \widehat{E} & \widehat{G} \\ \widehat{H} & \widehat{F} \end{bmatrix} \tag{4}$$

is known as the *doubling* map, which is well-defined when $I - GH$ and $I - HG$ are nonsingular.

Let

$$\alpha_{\text{opt}} := \min_{i\in\mathcal{S}_-}\left|\frac{C_{ii}}{T_{ii}}\right|, \quad \beta_{\text{opt}} := \min_{i\in\mathcal{S}_+}\left|\frac{C_{ii}}{T_{ii}}\right|, \tag{5}$$

Next, choose two real constants $0 \le \alpha \le \alpha_{\text{opt}}$, $0 \le \beta \le \beta_{\text{opt}}$, not both being zero, and define

$$P_0 := Q^{-1}R, \tag{6}$$

where

$$Q := \begin{bmatrix} C_+ - \alpha T_{++} & -\beta T_{+-} \\ -\alpha T_{-+} & |C_-| - \beta T_{--} \end{bmatrix}, \tag{7}$$

$$R := \begin{bmatrix} C_+ + \beta T_{++} & \alpha T_{+-} \\ \beta T_{-+} & |C_-| + \alpha T_{--} \end{bmatrix}. \tag{8}$$

Applying the doubling map to P_0 iteratively gives rise to the following sequence

$$P_k := \begin{bmatrix} E_k & G_k \\ H_k & F_k \end{bmatrix} := \mathcal{F}^k(P_0) \quad \text{for } k \ge 0. \tag{9}$$

where $\mathcal{F}^k$ denotes the composition of $\mathcal{F}$ with itself k times. By [16, Theorem 3.3] when $\alpha, \beta \ne 0$, and by [7, Theorem 8] otherwise, we have $P_k > 0$, $G_k < G_{k+1}$, $H_k < H_{k+1}$, and the limiting behaviour of the sub-blocks of P_k is

$$\lim_{k\to\infty} E_k = 0, \quad \lim_{k\to\infty} F_k =: F_\infty, \text{ for some } F_\infty$$

$$\lim_{k\to\infty} G_k = \Psi, \quad \text{and } \lim_{k\to\infty} H_k = \widehat{\Psi},$$

where Ψ and $\widehat{\Psi}$ are the minimal nonnegative solutions to

$$C_+^{-1}T_{+-} + \Psi|C_-|^{-1}T_{--} + C_+^{-1}T_{++}\Psi + \Psi|C_-|^{-1}T_{-+}\Psi = 0, \tag{10}$$

$$|C_-|^{-1}T_{-+} + \widehat{\Psi}C_+^{-1}T_{++} + |C_-|^{-1}T_{--}\widehat{\Psi} + \widehat{\Psi}C_+^{-1}T_{+-}\widehat{\Psi} = 0. \tag{11}$$

The convergence rate in these limits is known to be quadratic.

The algorithm described above is ADDA. The algorithms SDA-ss and SDA correspond to letting $\alpha := 0$, and letting $\alpha = \beta := \min(\alpha_{\text{opt}}, \beta_{\text{opt}})$, respectively.

3. (RE-)INTERPRETATIONS

3.1 The Return Probability Matrix Ψ

For simplicity, assume the fluid has unit rates, and the process starts at level 0 in a phase in $\mathcal{S}_+$. If we condition on the time y of the first transition to a phase in $\mathcal{S}_-$ (which is also the level where the fluid first changes direction from going upward to going downward), a standard argument gives

$$\Psi = \int_0^\infty \exp(T_{++}y)T_{+-}\exp(Uy)\mathrm{d}y, \tag{12}$$

where $U := T_{--} + T_{-+}\Psi$ is the generator of the downward record process.

In [8], the authors uniformized the first upward part of the path, and then the remaining part, with the same rate. Here, we give a slightly more general version, in which there are two different uniformization rates λ and μ. In particular, we uniformize the first upward part of the path with rate λ, and the remainder with rate μ, obtaining

$$\Psi = \int_0^\infty [\sum_{k=0}^\infty \mathrm{e}^{-\lambda y}\frac{(\lambda y)^k}{k!}P_{\lambda++}]\lambda P_{\lambda+-}[\sum_{n=0}^\infty \mathrm{e}^{-\mu y}\frac{(\mu y)^n}{n!}V_\mu]\mathrm{d}y.$$

with $P_\lambda := I + \lambda^{-1}T$, $P_\mu := I + \mu^{-1}T$, and $V_\mu := I + \mu^{-1}U$. Swapping the order of summation and integration gives

$$\Psi = \sum_{k,n=0}^\infty \gamma_{k,n}P_{\lambda++}^k P_{\lambda+-}V_\mu^n, \tag{13}$$

where

$$\gamma_{kn} = \frac{(k+n)!}{k!n!}\frac{\lambda^{k+1}\mu^n}{(\lambda+\mu)^{k+n+1}}.$$

Da Silva Soares and Latouche [8] gave a probabilistic interpretation for (13): sample paths of the uniformized process are decomposed into disjoint sets $\mathcal{A}_{k,n}$, corresponding to paths with

- k uniformization events in the time it takes to increase from 0 to y (which is y, as we assume unit rates),
- an additional event at time y, and then
- n uniformization events in the time it takes the downward record process to decrease from y to 0.

Using probabilistic arguments, we can rearrange the sum in (13) in different ways, each involving only one summation variable. Let

$$W := (I + \mu^{-1}U)(I - \lambda^{-1}U)^{-1}.$$

Theorem 3.1.

$$\Psi = \sum_{k=0}^{\infty} P_{\lambda++}^{k} P_{\lambda+-} (I - \lambda^{-1}U)^{-k-1} \tag{14}$$

$$= \sum_{n=0}^{\infty} (I - \mu^{-1}T_{++})^{-n-1} P_{\mu+-} V_{\mu}^{n}, \tag{15}$$

$$= \sum_{m=0}^{\infty} (I + \lambda^{-1}T_{++})^{m} (I - \mu^{-1}T_{++})^{-m-1} \times (P_{\lambda+-}W + P_{\mu+-})W^{m}. \tag{16}$$

The proof of Theorem 3.1, and that of another theorem stated later, are included in our forthcoming paper [3]. Theorem 3.1 presents three ways of categorizing returning-to-zero sample paths of a fluid process, using two Poisson processes $\mathcal{U}_\lambda$ and $\mathcal{D}_\mu$ with rate λ and μ, respectively.

1. For the sum in (14):

We group the sample paths according to the number k of events of the Poisson process $\mathcal{U}_\lambda$ in $[0, y)$, which provide k uniformizing steps represented by $P_{\lambda++}^{k}$ in the sum. We denote by u_i, $0 < u_1 < \cdots < u_k < y$, the sequence of events. There is also an event at $y =: u_{k+1}$, represented by $P_{\lambda+-}$.

Let $\tau_i := \inf\{t \geq y : X_t = u_i\}$, $i = 0, \ldots, k+1$, be the first time the fluid reaches level u_i on the way down after the turning point at y, with $u_0 := 0$. Then, $y = \tau_{k+1} < \tau_k < \cdots < \tau_0$.

For each interval (τ_{i+1}, τ_i), we observe the fluid process once it reaches a new downward record u_i at time τ_i—this is indicated by the matrix $(1 - \lambda^{-1}U)^{-1}$, of which each element (j_1, j_2) is the probability that the fluid is in state j_2 at time τ_i, after starting in state j_1 at time τ_{i+1}. There are $k+1$ intervals, which explains the term $(1 - \lambda^{-1}U)^{-k-1}$.

2. For the sum in (15):

We group the sample paths in an opposite manner: according to the number of n events of the process $\mathcal{D}_\mu$ in an interval of length y, which provides n uniformizing steps of the downward record process (with generator U), represented by V_{μ}^{n}. We denote by d_i, $i = 1, \ldots, n$, the sequence of events, and let $d_0 := y$, and $d_{n+1} := 0$.

Then, for the initial upward journey from level 0 to level y, we simply observe the fluid process at the end of each exponentially distributed interval (d_{i+1}, d_i). There are $n+1$ intervals, which explains the term $(1 - \mu^{-1}T_{++})^{-n-1}$. Once at level y, there is a switch in direction with matrix $P_{\mu+-}$.

3. For the sum in (16):

The interpretation of this sum is based on a combination of the previous two constructions. Let $\{u_i\}_{i\geq 1}$ and $\{d_i\}_{i\geq 1}$ denote the sequences of events of $\mathcal{U}_\lambda$ and of $\mathcal{D}_\mu$ in $[0, y)$, respectively. We define a new sequence, based on $\{u_i\}$ and $\{d_i\}$, as follows

$$c_0 = 0,$$
$$c_{2i+1} = \min\{d_i : d_i > c_{2i}\},$$
$$c_{2i+2} = \min\{u_i : u_i > c_{2i+1}\},$$

for $i \geq 0$.

In this construction, each interval $[c_{2i}, c_{2i+1}]$ contains only one point of $\mathcal{D}_\mu$, which lies at its right endpoint, and has exponentially distributed length with parameter μ; on the other hand, each interval $[c_{2i+1}, c_{2i+2}]$ contains only one point of $\mathcal{U}_\lambda$, which lies at its right endpoint, and has exponentially distributed length with parameter λ.

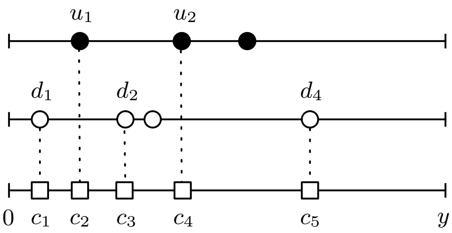


Figure 1: An illustration of the construction for c_i

Let $N := \operatorname{argmax}\{c_j : c_j < y\}$, then N is the number of events c_j in the interval $[0, y)$. There are two cases: N is even, and N is odd. For brevity, we consider only the case when N is odd in this abstract, and analyse both cases in [3].

If $N = 2m + 1$ for some integer $m \geq 0$, let $c_{2m+2} := y$. Thus, there are $2m + 2$ exponentially distributed intervals, with parameters alternating between μ and λ, starting with μ and ending with λ. These sample paths contribute to the following term of the sum (16):

$$\sum_{m=0}^{\infty} (I + \lambda^{-1}T_{++})^{m} (I - \mu^{-1}T_{++})^{-m-1} \times P_{\lambda+-} W^{m+1}.$$

During the initial upward journey from 0 to y, for the first $2m$ intervals, we alternate between observing the fluid at the end of an interval, represented by $(I - \mu^{-1}T_{++})^{-1}$, and doing a uniformization step with $I + \lambda^{-1}T_{++}$.

We have another observation at the end of the $(2m+1)$th interval, and then a uniformization step with $P_{\lambda+-}$, signaling a switch from the upward direction to downward for the fluid, at $c_{2m+2} = y$.

Let $\rho_i := \inf\{t \geq y : X_t = c_i\}$, $i = 1, \ldots, 2m+2$, and let $\rho_0 := c_0 = 0$. Then, ρ_i is the sequence of first hitting times to level c_i from above, and $\rho_{i+1} < \rho_i$.

For the $2m + 2$ time intervals (ρ_{i+1}, ρ_i), we alternate between observing the fluid at the end, represented by $(I - \lambda^{-1}U)^{-1}$, and doing a uniformization step with $I + \mu^{-1}U$.

This implies, that if we observe the fluid over the interval at the end of which it increases from level c_i to level c_{i+1} for the first time, then we do a uniformization step for the path where the fluid decreases from c_{i+1} to c_i for the first time, and vice versa. This construction for the downward journey over the time interval $[y, c_0]$, explains the term $W^{m+1} = (I + \mu^{-1}U)^{m+1}(I - \lambda^{-1}U)^{-m-1}$.

3.2 QBDs and Fluid Queues

Based on the aforementioned interpretations, we construct three QBD processes for which the first downward return matrices G contains the Ψ as one of its blocks.

Theorem 3.2. *Consider three quasi-birth-death processes* *with corresponding groups of probability matrices:*

$$A_{-1} = \frac{1}{2}\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix},\ A_0 = \frac{1}{2}\begin{bmatrix} I & P_{\lambda+-} \\ 0 & P_{\lambda--} \end{bmatrix},\ A_1 = \frac{1}{2}\begin{bmatrix} P_{\lambda++} & 0 \\ P_{\lambda-+} & 0 \end{bmatrix};$$

$$B_{-1} = \frac{1}{2}\begin{bmatrix} 0 & P_{\mu+-} \\ 0 & P_{\mu--} \end{bmatrix},\ B_0 = \frac{1}{2}\begin{bmatrix} P_{\mu++} & 0 \\ P_{\mu-+} & I \end{bmatrix},\ B_1 = \frac{1}{2}\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix};$$

$$C_{-1} = \frac{1}{2}\begin{bmatrix} 0 & P_{\mu+-} \\ 0 & P_{\mu--} \end{bmatrix},\ C_0 = \frac{1}{2}\begin{bmatrix} P_{\mu++} & P_{\lambda+-} \\ P_{\mu-+} & P_{\lambda--} \end{bmatrix},$$

$$C_1 = \frac{1}{2}\begin{bmatrix} P_{\lambda++} & 0 \\ P_{\lambda-+} & 0 \end{bmatrix}.$$

Then, the matrices G recording the first passage probabilities to a lower level of these QBDs, respectively, are

$$G_A = \begin{bmatrix} 0 & \Psi \\ 0 & (I - \lambda^{-1}U)^{-1} \end{bmatrix},\ G_B = \begin{bmatrix} 0 & \Psi \\ 0 & V_\mu \end{bmatrix},\ G_C = \begin{bmatrix} 0 & \Psi \\ 0 & W \end{bmatrix}.$$

We note that the QBD associated with A_{-1}, A_0, A_1 is based on our interpretation of the sum (14), and is a modification of the QBD given in [1]. The QBD associated with B_{-1}, B_0, B_1 is based on our interpretation of (15), and is a modification of the QBD given in [8]. The third process, associated with C_{-1}, C_0, C_1, is the most general QBD of the three, and is based on our interpretation of (16).

3.3 Interpretation of Doubling Algorithms

The interpretation for doubling algorithms involves two parts:

(a) First, we show that the initial values P_0, defined in (6), are associated to QBDs of the form

$$D_{-1} = \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix},\ D_0 = \begin{bmatrix} 0 & G \\ H & 0 \end{bmatrix},\ D_1 = \begin{bmatrix} E & 0 \\ 0 & 0 \end{bmatrix},\quad (17)$$

which are directly linked to the processes constructed in Theorem 3.2. (We set $\alpha = 1/\mu$ and $\beta = 1/\lambda$.) In particular, the QBDs (17) are obtained by censoring out intervals during which the QBDs in Theorem 3.2 oscillate between two consecutive levels.

(b) Then, we give an interpretation for the iteration $P_{k+1} = \mathcal{F}(P_k)$. A doubling algorithm is equivalent to applying Cyclic Reduction to the QBD process (17) (as proved in [7]), so this probabilistic interpretation is similar to that of Cyclic Reduction.

Note that Cyclic Reduction preserves the zero structure of (17). This is the underlying reason doubling-based algorithms are more computationally efficient than applying Cycling Reduction directly to the QBDs given in Theorem 3.2.

4. REFERENCES

[1] S. Ahn and V. Ramaswami. Fluid flow models and queues—a connection by stochastic coupling. *Stoch. Models*, 19(3):325–348, 2003.

[2] S. Asmussen. Stationary distributions for fluid flow models with or without Brownian noise. *Communications in Statistics: Stochastic Models*, 11(1):21–49, 1995.

[3] N. Bean, G. T. Nguyen, and F. Poloni. Algorithms for stationary distributions of fluid queues: Interpretations and re-interpretations. *In preparation*, 2016.

[4] N. Bean, M. M. O'Reilly, and P. G. Taylor. Algorithms for return probabilities for stochastic fluid flows. *Stoch. Models*, 21(1):149–184, 2005.

[5] D. A. Bini, G. Latouche, and B. Meini. *Numerical Methods for Structured Markov Chains.* Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2005.

[6] D. A. Bini and B. Meini. The cyclic reduction algorithm: from Poisson equation to stochastic processes and beyond. *Numer. Algorithms*, 51:23–60, 2009.

[7] D. A. Bini, B. Meini, and F. Poloni. Transforming algebraic Riccati equations into unilateral quadratic matrix equations. *Numer. Math.*, 116(4):553–578, 2010.

[8] A. da Silva Soares and G. Latouche. Further results on the similarity between fluid queues and QBDs. In G. Latouche and P. Taylor, editors, *Matrix-analytic methods. Theory and applications*, pages 89–106. World Scientific, Singapore, 2002.

[9] C.-H. Guo. Nonsymmetric algebraic Riccati equations and Wiener-Hopf factorization for M-matrices. *SIAM Journal on Matrix Analysis and Applications*, 23(1):225–242, 2001.

[10] X.-X. Guo, W.-W. Lin, and S.-F. Xu. A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation. *Numer. Math.*, 103:393–412, 2006.

[11] R. L. Karandikar and V. Kulkarni. Second-order fluid flow models: Reflected Brownian motion in a random environment. *Oper. Res*, 43:77–88, 1995.

[12] G. Latouche and V. Ramaswami. A Logarithmic Reduction algorithm for Quasi-Birth-Death processes. *J. Appl. Prob.*, 30:650–674, 1993.

[13] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia PA, 1999.

[14] V. Ramaswami. Matrix analytic methods for stochastic fluid flows. In D. Smith and P. Hey, editors, *Teletraffic Engineering in a Competitive World (Proceedings of the 16th International Teletraffic Congress)*, pages 1019–1030. Elsevier Science B.V., Edinburgh, UK, 1999.

[15] L. C. G. Rogers. Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. *Ann. Appl. Probab.*, 4:390–413, 1994.

[16] W.-G. Wang, W.-C. Wang, and R.-C. Li. Alternating-directional doubling algorithm for M-matrix algebraic Riccati equations. *SIAM J. Matrix Anal. Appl.*, 33(1):170–194, 2012.

A numerical framework for computing the limiting distribution of a stochastic fluid-fluid process [*]

Nigel Bean
School of Mathematical Sciences
The University of Adelaide
Adelaide, Australia
nigel.bean@adelaide.edu.au

Giang T. Nguyen
School of Mathematical Sciences
The University of Adelaide
Adelaide, Australia
giang.nguyen@adelaide.edu.au

Malgorzata O'Reilly
Faculty of Science, Engineering,
and Technology
The University of Tasmania
Hobart, Australia
Malgorzata.OReilly@utas.edu.au

Vikram Sunkara
Department of Mathematics and
Computer Science
Freie Universität Berlin
Berlin, Germany
sunkara@mi.fu-berlin.de

ABSTRACT

We present a numerical framework for computing the limiting distribution of a so-called *stochastic fluid-fluid* model. Introduced by Bean and O'Reilly (2014), a stochastic fluid-fluid process is a Markov processes $\{X_t, Y_t, \varphi_t\}_{t\geq 0}$, where the first level X_t is driven by the Markov chain φ_t, and the second level Y_t is driven by φ_t as well as by X_t. That paper gave a closed-form expression for the stationary distribution given in terms of operators acting on measures, which does not lend itself easily to numerical computations.

In our work, we apply the discontinuous Galerkin method to numerically obtain the stationary distribution of a stochastic fluid-fluid model, and illustrate it using a specific example of a stochastic fluid-fluid.

CCS Concepts

•**Mathematics of computing** → **Probabilistic algorithms;** *Markov processes;*

Keywords

stochastic fluid–fluid processes; stationary distribution; discontinuous Galerkin method

1. INTRODUCTION

A stochastic fluid process $\{X_t, \varphi_t\}_{t\geq 0}$ is a two-dimensional Markov process, where the phase φ_t is a continuous-time Markov chain on a finite state space $\mathcal{S}$, and the level X_t varies linearly at rate c_i whenever $\varphi_t = i$, $i \in \mathcal{S}$. A subset of Markov additive processes, stochastic fluids have been well-analysed in the past two decades. There are two recent generalizations of stochastic fluid processes to a higher dimension: Miyazawa and Zwart [3] analyzed discrete-time multidimensional Markov additive processes, and Bean and O'Reilly [1] studied the so-called *stochastic fluid-fluid process*, the latter is our focus in this paper.

A stochastic fluid-fluid is a Markov process $\{X_t, Y_t, \varphi_t\}_{t\geq 0}$, where the phase φ_t is still a Markov chain on a finite state space $\mathcal{S}$; $X_t \in (-\infty, \infty)$ is the first fluid, which varies linearly at rate c_i whenever $\varphi_t = i$, $i \in \mathcal{S}$; and Y_t represents the second fluid, which varies linearly at rate $r_i(x)$ whenever $X_t = x$ and $\varphi_t = i$. Thus, Y_t is a Markov process on $\mathbb{R} \times \mathcal{S}$.

While the analyses in [3] and [1] are markedly different, both papers drew inspiration from Neuts' matrix-analytic approach to obtain limiting behavior of these processes, working with operators on function spaces instead of matrices. Thus, their closed-form expressions for the limiting distributions ([3, Theorem 4.1], [1, Theorem 2]) are given in terms of operators and measures, which are not immediately amenable to numerical computations for real-life applications. One standard way to numerically handle operators on function spaces is to construct approximations of the operators. To this end, there are existing numerical methods, such as *finite differences*, *finite volume*, *finite elements*, and *the discontinuous Galerkin method.*

In this paper, we apply the discontinuous Galerkin method to compute the stationary distribution of a stochastic fluid-fluid and illustrate this using a specific example, which is an on-off bandwidth-sharing system of two processors [2]. Inputs into the processors are turned on and off by a Markov chain, and the combined output capacity is fixed and allocated according to the workload of the first, more important, processor. Latouche *et al.* [2] computed the bounds for the marginal limiting distribution of the workload of the second processor. In our work, we numerically obtain this distribution, and verify the results by using it to compute the marginal limiting distribution of the workload of the first

[*] N. Bean, G. Nguyen, and V. Sunkara acknowledge the financial support of the Australian Research Council through the Discovery Grant DP110101663. The first two authors also acknowledge the support of the ARC Centre of Excellence for Mathematical and Statistical Frontiers.

processor, which we then compare with the analytical solution obtained via a classical analysis of stochastic fluid models. As an additional verification, we also compare our numerical results against simulations.

2. NUMERICAL FRAMEWORK

2.1 A one-sided stochastic fluid-fluid

We assume there is a regulated boundary at level 0 for the first fluid, and thus $X_t \in [0, \infty)$. Furthermore, we assume positive recurrence, for the existence of the joint stationary distribution $\boldsymbol{\pi}(y) = (\pi_i(y))_{i \in \mathcal{S}}$ for $\{X_t, Y_t, \varphi_t\}$, where for a given set $\mathcal{A}$

$$\pi_i(y)(\mathcal{A}) = \lim_{t \to \infty} \frac{\partial \mathbb{P}(X_t \in \mathcal{A}, Y_t \leq y, \varphi_t = i)}{\partial y}.$$

An expression for $\boldsymbol{\pi}(y)$ is given in [1, Thm 2], in terms of operators acting on measures, with the two most important ones being B and Ψ: for a set $\mathcal{A}$ and a measure vector $\boldsymbol{\mu} = (\mu_i)_{i \in \mathcal{S}}$, $\boldsymbol{\mu}\mathrm{e}^{Bt}(\mathcal{A})$ gives the probability of $X_t \in \mathcal{A}$, after starting at time zero according to $\boldsymbol{\mu}$; $\boldsymbol{\mu}\Psi(\mathcal{A})$ gives the probability of Y_t returning to zero and doing so when $X_t \in \mathcal{A}$, given that the initial distribution is $\boldsymbol{\mu}$.

2.2 Discontinuous Galerkin (DG) method

Discontinuous Galerkin methods are used to approximate the solution to a system of partial differential equations, and work as follows. Consider a sequence of points called *nodal points*, each interval between two consecutive nodal points is referred to as a *mesh*.

Within each mesh, we have a *finite element* approximation. This method constructs a finite-dimensional smooth Sobolev space, by choosing appropriate piecewise polynomial basis functions, and then projecting the partial differential equations onto this space. This projection leads to a new system of equations, which are referred to as the *weak form* of the original PDEs.

There is a flux operator, which moves probability from one mesh to another, in a manner similar to the underlying principle of a *finite volume* approximation: integrating the PDEs over each mesh, and then constructing a new system of ordinary differential equations, which describe the change in the integral over the mesh. This method conserves probability, and can handle discontinuities like jumps and point masses.

The discontinuous Galerkin method leads to a global approximation in the space of piecewise functions. Intuitively, we relax the continuity between meshes to gain the conservation of probability. The local approximations within a mesh are as smooth as desired, by appropriate choice of the basis functions.

2.3 DG applied to a stochastic fluid-fluid

Here, we give a brief description of how to apply the discontinuous Galerkin method to approximate the operator B. Let $f_i(x,t)$ be the probability density function of X_t taking value x at time t, and φ_t being $i \in \mathcal{S}$. Then, the functions $f_i(x,t)$ satisfy the system of partial differential equations

$$\frac{\partial}{\partial t} f_i(x,t) = \sum_{j \in \mathcal{S}} T_{ji} f_j(x,t) - c_i \frac{\partial}{\partial x} f_i(x,t), \qquad (1)$$

subject to suitable boundary conditions, where T is the generator of the Markov chain φ_t.

Denote by $\{x_k\}_{k=1,\dots,K}$ a sequence of nodal points, and by $\{\mathcal{D}_k\}_{k=1,\dots,K-1}$ the sequence of corresponding meshes. Let $\phi_n^k : \mathcal{D}_k \mapsto \mathbb{R}_+$, $n = 0, \dots, N$, be the basis functions on the kth mesh. The span of our basis functions ϕ_n^k forms our approximation space, $V_k := \oplus_{k=1}^{K-1} \{\phi_0^k, \dots, \phi_N^k\}$. Then, a function $u_i \in V_K$ has the form:

$$u_i(x,t) = \sum_{k=1}^{K-1} \sum_{n=0}^{N} \alpha_n^{i,k}(t) \phi_n^k(x) \quad \text{for } x \in \Omega, t \in \mathbb{R}_+.$$

We can show that the weak form of the evolution of the probability density $f_i(x,t)$ of $\{X_t, \varphi_t\}$ is the following system of ordinary differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{\alpha}^i(t) = \sum_{j=1}^{|\mathcal{S}|} T_{ji} I \boldsymbol{\alpha}^j(t) + M^{-1}(c_i G + F)\boldsymbol{\alpha}^i(t). \qquad (2)$$

where $\boldsymbol{\alpha}^{i,k}(t) = (\alpha_n^{i,k}(t))_{n=0,1,\dots,N}$, I is an appropriately-sized identity matrix, and for the k^{th} mesh,

$$M_{m,n}^k = \int_{\mathcal{D}_k} \phi_n^k(x) \phi_m^k(x) \mathrm{d}x,$$

$$G_{m,n}^k = \int_{\mathcal{D}_k} \phi_n^k(x) \frac{\partial}{\partial x} \phi_m^k(x) \mathrm{d}x,$$

$$F_{m,\bullet}^k = f^*(u, x_k^\ell) \phi_m^k(x_k^\ell) - f^*(u, x_k^r) \phi_m^k(x_k^r),$$

with x_k^ℓ and x_k^r being the respective left and right endpoints of $\mathcal{D}_k$, $n \in \{0, \dots, N\}, m \in \{0, \dots, n\}$, and f^* represents the *numerical flux* of probability going from $\mathcal{D}_{k-1}$ into $\mathcal{D}_k$.

Using this weak form, we can construct an approximation to the operator B, and then the exponential operator e^{Bt}, and so on.

We perform numerical experiments on the on-off system described briefly in the Introduction. For the first buffer X_t, which is driven by only the Markov chain φ_t, we find that our piecewise linear DG approximation is close to the results of one-million Monte Carlo simulations, and to the analytical results obtained in [2]. For the second buffer Y_t, driven by φ_t and X_t, we compare the DG approximation of the return-probability operator Ψ against Monte Carlo simulations, and find reasonable agreement.

Furthermore, we analyse different choices for the level of spatial discretisation and the degree of polynomial basis functions, with respect to the order of convergence in relevant error terms.

3. REFERENCES

[1] N. G. Bean and M. O'Reilly, Malgorzata. The stochastic fluid-fluid model: A stochastic fluid model driven by an uncountable-state process, which is a stochastic fluid itself. *Stochastic Processes and their Applications*, 124:1741–1772, 2014.

[2] G. Latouche, G. T. Nguyen, and Z. Palmowski. Two-dimensional fluid queues with temporary assistance. volume 27 of *Springer Proceedings in Mathematics & Statistics*, chapter 9, pages 187–207. Springer Science, New York, NY, to be published.

[3] M. Miyazawa and B. Zwart. Wiener-Hopt factorizations for a multidimensional markov additive process and their applications to reflected processes. *Stochastic Systems*, 2:67–114, 2012.

Efficient cyclic reduction for QBDs with rank structured blocks: algorithm and analysis

[Extended Abstract]

Dario A. Bini
University of Pisa
Pisa, Italy
bini@dm.unipi.it

Stefano Massei
Scuola Normale Superiore
Pisa, Italy
stefano.massei@sns.it

Leonardo Robol
KU Leuven
Leuven, Belgium
leonardo.robol@cs.kuleuven.be

ABSTRACT

The computation of the steady state distribution of Quasi Birth-and-Death Markov chains involves the solution of a matrix equation of the kind $X = A_{-1} + A_0X + A_1X^2$ where the blocks A_i are $m \times m$ nonnegative matrices such that $A_{-1} + A_0 + A_1$ is irreducible and stochastic. We provide an effective algorithm for solving quadratic matrix equations with quasi-separable coefficients. This case comprises all the applications in which the blocks of the transition matrix are banded, for example the doubly QBD processes [8].

The algorithm is based on a suitable implementation of Cyclic Reduction (CR) which relies on the technology of rank-structured matrices [11]. In fact, we prove that all the matrices generated by CR have off-diagonal submatrices with small numerical rank so that they are well approximated by rank-structured matrices. This property allows us to implement CR with a substantially lower cost with respect to the general case. The results of the numerical experiments confirm a significant speed up over the general algorithms, already starting with the moderately small size $m \approx 10^2$.

Keywords

Markov chains, QBD processes, Cyclic reduction, Matrix equhtations, Hierarchical matrices, Quasi-separable matrices

1. INTRODUCTION

Cyclic reduction (CR) is an effective tool that can be used for solving several problems in linear algebra and in polynomial computations [4]. One of its important applications concerns the computation of the minimal nonnegative solution of the matrix equation

$$X = A_{-1} + A_0X + A_1X^2, \qquad (1)$$

encountered in Quasi Birth-Death (QBD) Markov chains, where A_{-1}, A_0, and A_1 are given $m \times m$ nonnegative matrices such that $A_{-1}+A_0+A_1$ is irreducible and stochastic and where X is the matrix unknown [2, 4]. The computation of the minimal solution allows one to recover the steady state vector π of the Markov chain [9].

CR computes four sequences of matrices, $A_i^{(h)}$, $i = -1, 0, 1$ and $\widehat{A}_0^{(h)}$, according to the following equations

$$\begin{aligned}
&A_1^{(h+1)} = -A_1^{(h)}\, S^{(h)}\, A_1^{(h)}, \qquad S^{(h)} = (A_0^{(h)} - I)^{-1}\\
&A_0^{(h+1)} = A_0^{(h)} - A_1^{(h)}\, S^{(h)}\, A_{-1}^{(h)} \;-\; A_{-1}^{(h)}\, S^{(h)}\, A_1^{(h)},\\
&A_{-1}^{(h+1)} = -A_{-1}^{(h)}\, S^{(h)}\, A_{-1}^{(h)},\\
&\widehat{A}_0^{(h+1)} = \widehat{A}_0^{(h)} - A_1^{(h)}\, S^{(h)}\, A_{-1}^{(h)},
\end{aligned} \qquad (2)$$

for $h = 0, 1, \ldots$, with $A_i^{(0)} = A_i$, $i = -1, 0, 1$ and $\widehat{A}_0^{(0)} = A_0 - I$. It can be proved [2] that the sequence $-(\widehat{A}_0^{(h)})^{-1}A_{-1}$ converges to the minimal nonnegative solution G of the matrix equation (1) as $h \to \infty$.

Each step of CR requires one matrix inversion and a small number of matrix multiplications which, without any further assumption on the structure of the blocks A_i, costs of $O(m^3)$ arithmetic operations (ops).

There are several models from the applications where the blocks A_i exhibit special structures. These specific features of the blocks can be used to decrease the computational complexity of the iterations by means of *ad hoc* adaptations of CR, some examples are treated in [3], [10], [1].

Here, we are interested in analysing the case where the blocks A_i are quasi-separable matrices. That is, the *off-diagonal submatrices* of A_{-1}, A_0 and A_1, contained in the upper or in the lower triangular part, have low rank with respect to m. The maximum of the ranks of the off-diagonal submatrices is called *quasi-separable rank*. It is well-known that the quasi-separable rank is invariant under matrix inversion and is sub-additive with respect to matrix addition and multiplication [11]. Moreover, the basic operations like matrix inversion and matrix multiplication have a low cost if performed with quasi-separable matrices [11].

A matrix with quasi-separable rank k is said k*-quasi-separable*. Observe that k-quasi-separable matrices include banded matrices. In particular, for doubly QBD processes as well as for bidimensional random walks [8], the matrices A_i are tridiagonal and in particular are 1-quasi-separable.

Our goal is to exploit the quasi-separable structure of the blocks A_i. Unfortunately, the matrices $A_i^{(h)}$ generated by CR through (2), do not preserve in general the quasi-separable structure. However, they remain k-quasi-separable in an approximate sense.

2. NUMERICAL PRESERVATION OF THE QUASI-SEPARABLE RANK

We start with an empirical observation: plotting the singular values of the off-diagonal blocks of the matrices $A_i^{(h)}$ shows an interesting behaviour. The singular values of the off-diagonal submatrices of any $A_i^{(h)}$ generated by CR have an exponential decay. This fact is shown in Figure 1 where A_i, $i = -1, 0, 1$ are randomly generated tridiagonal matrices of size $m = 1600$ and we plot the singular values of the largest south-western off-diagonal submatrix of $A_0^{(h)}$ for $h = 1, 2, \dots, 20$.

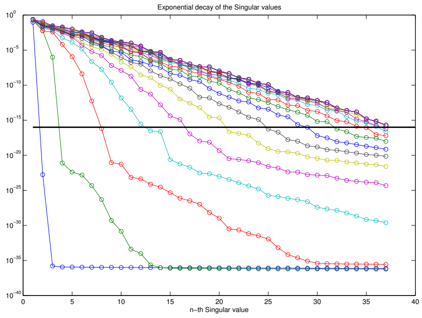


Figure 1: Log-scale plot of the singular values of the largest south-western submatrix of $A_0^{(h)}$ contained in the lower triangular part, for $m = 1600$ and $h = 1, \dots, 20$. The horizontal line denotes the machine precision threshold. Matrices are randomly generated so that $A_i \geq 0$ are tridiagonal matrices and $A_{-1} + A_0 + A_1$ is stochastic.

It is evident that, even though the number of nonzero singular values grows at each step of CR, the number of singular values above the machine precision – denoted by a horizontal line in Figure 1 – is bounded by a moderate constant. This fact implies that even if the rank of the off-diagonal submatrices grows up to saturation, their numerical rank remains small, that is, these submatrices are well approximated by low rank matrices.

Moreover, the singular values seem to stay below a straight-line which constitutes an asymptotic bound. That is, they get closer to this line as $h \to \infty$. The logarithm scale suggests that the computed singular values $\sigma_i^{(h)}$, ordered so that $\sigma_i^{(h)} \geq \sigma_{i+1}^{(h)}$, decay exponentially with i and the basis of the exponential grows with h but has a limit less than 1.

We will prove this asymptotic property relating this limit to the spectrum of the solutions of certain quadratic matrix equations strictly related to (1). Then we exploit this property to design an effective implementation of CR.

2.1 Main properties of CR

Associate the matrices $A_i^{(h)}$, $i = -1, 0, 1$ defined in (2) with the matrix Laurent polynomial $\varphi^{(h)}(z) := -z^{-1}A_{-1}^{(h)} + (I - A_0^{(h)}) - zA_1^{(h)}$, where $\varphi^{(0)}(z) = \varphi(z) = z^{-1}A_{-1} + (I - A_0) - zA_1$, and define the matrix rational function $\psi^{(h)}(z) = \varphi^{(h)}(z)^{-1}$. The following property holds [2]

$$\begin{cases} \psi^{(0)}(z) := \psi(z), \\ \psi^{(h+1)}(z^2) := \frac{1}{2}(\psi^{(h)}(z) + \psi^{(h)}(-z)), \end{cases}$$

thus, we deduce that $\psi^{(h)}(z^{2^h}) = \frac{1}{2^h} \sum_{j=0}^{2^h-1} \psi^{(0)}(\omega_N^j z)$, where $\omega_N = e^{\frac{2\pi}{N}\mathbf{i}}$ is a principal N-th root of the unity for $N = 2^h$, and $\mathbf{i}$ denotes the imaginary unit, so that

$$\varphi^{(h)}(z^{2^h}) = \left(\frac{1}{2^h} \sum_{j=0}^{2^h-1} \psi^{(0)}(\omega_N^j z) \right)^{-1}. \qquad (3)$$

Observe that in the case where A_{-1}, A_0 and A_1 are tridiagonal, then $\varphi(z)$ is tridiagonal as well, so that for any value of z such that $\det \varphi(z) \neq 0$, the matrix $\psi(z)$ is quasi-separable, that is, $\operatorname{tril}(\psi(z)) = \operatorname{tril}(L)$, $\operatorname{triu}(\psi(z)) = \operatorname{triu}(U)$, where L and U are matrices of rank 1 [11] and $\operatorname{tril}(A)$, $\operatorname{triu}(A)$ denote the lower and upper triangular matrix, respectively, formed by the entries of A.

Another thing to highlight is that, by performing an interpolation on the unit circle, we can retrieve the blocks $A_i^{(h)}$ starting from $\varphi^{(h)}(z)$. In fact, if ξ is a primitive 6-th root of the unity, then one can verify that

$$A_{-1}^{(h)} = \frac{1}{3}\left(\xi\varphi^{(h)}(\xi) + \xi^5\varphi^{(h)}(\xi^5) - \varphi^{(h)}(-1) \right), \qquad (4)$$

$$A_0^{(h)} = \frac{1}{2}\left(\varphi^{(h)}(z) + \varphi^{(h)}(-z) \right), \qquad (5)$$

$$A_1^{(h)} = \frac{1}{3}\left(\xi^5\varphi^{(h)}(\xi) + \xi\varphi^{(h)}(\xi^5) - \varphi^{(h)}(-1) \right). \qquad (6)$$

2.1.1 Laurent expansion of $\psi(z)$

In order to prove the exponential decay of the singular values of the off-diagonal submatrices, we rely on the Laurent expansion of $\psi(z)$. We recall the following result [2, Theorem 3.20].

THEOREM 2.1. *Let $\varphi(z) = -z^{-1}A_{-1} + I - A_0 - zA_1$ with $A_i \in \mathbb{R}^{m \times m}$, $i = -1, 0, 1$ and assume that $\det(z\varphi(z))$ has zeros ξ_i, $i = 1, \dots, 2m$ which satisfy*

$$|\xi_1| \leq \dots \leq |\xi_m| < \delta < 1 < \delta^{-1} < |\xi_{m+1}| \leq \dots \leq |\xi_{2m}|. \qquad (7)$$

Moreover, suppose that there exist matrices R and $\widehat{R}$ with spectral radius less than 1 which solve the following equations

$$A_1 + X(A_0 - I) + X^2 A_{-1} = 0,$$
$$X^2 A_1 + X(A_0 - I) + A_{-1} = 0,$$

respectively. Then expanding $\varphi(z)^{-1} = \psi(z) = \sum_{j=-\infty}^{+\infty} z^j H_j$ we have

$$H_j = \begin{cases} H_0 \widehat{R}^{-j} & j \leq 0 \\ H_0 R^j & j \geq 0 \end{cases}. \qquad (8)$$

In the Markovian case, except for the null recurrent scenario where $\xi_m = 1 = \xi_{m+1}$, the hypothesis of the previous theorem are satisfied by suitably scaling the variable of $\varphi(z)$.

2.2 Exponential decay of the singular values

We now draw a sketch of the proof of the exponential decay property in the off-diagonal blocks of the sequences generated by the CR.

(i) Exploiting the low quasiseparable rank of the starting blocks, the recurrence relation (3) and the Laurent expansion (8) we prove the property for $\psi^{(h)}(z)$ with z on the unit circle.

(ii) With linear algebra techniques we show how to retrieve similar bounds for $\varphi^{(h)}(z)$.

(iii) By performing interpolations by means of (4),(5) and (6) we get bounds for the sequences A_i $i = -1, 0, 1$.

2.2.1 The property of $\psi^{(h)}(z)$

Define $\mathcal{X}_{C,M,L,k,\rho,\delta}$ the set of Laurent matrix polynomials $\varphi(z) = -z^{-1}A_{-1} + I - A_0 - zA_1$ such that

- the splitting property (7) with radius of the split δ holds for $\varphi(z)$ and for each matrix polynomial obtained as the lower right principal submatrix of $\varphi(z)$;
- the coefficients A_j are k-quasiseparable and $\|A_j\|_2 \leq L$, $j = -1, 0, 1$, for a given constant L;
- denoting $\varphi_i(z) = -z^{-1}A_{-1,i} + I - A_{0,i} - zA_{1,i}$ the matrix function obtained from $\varphi(z)$ by taking the rows and columns from i to m, assume that the matrix equations
$$\begin{aligned} A_1 + (A_0 - I)X + A_{-1}X^2 &= 0, \\ A_1 + X(A_0 - I) + X^2 A_{-1} &= 0, \\ A_{1,i} + (A_{0,i} - I)X + A_{-1,i}X^2 &= 0, \\ A_{1,i} + X(A_{0,i} - I) + X^2 A_{-1,i} &= 0 \end{aligned}$$
have minimal nonnegative solutions R, $\widehat{R}$, R_i and $\widehat{R}_i$ respectively, $i = 1, \ldots, m$;
- the numerical ranges of the matrices $\delta^{-1}R$, $\delta^{-1}\widehat{R}$, $\delta^{-1}R_i$ and $\delta^{-1}\widehat{R}_i$ are included in a compact and connected set of logarithmic capacity smaller than ρ.

We have the following

Theorem 2.2. *For any $\varphi(z) \in \mathcal{X}_{C,M,L,k,\rho,\delta}$, for any $\epsilon > 0$ such that $\rho + \epsilon < 1$, for any $h \in \mathbb{N}$, and $z \in \mathbb{C}$, $|z| = 1$, every off-diagonal block $\tilde{C}(z)$ of $\psi^{(h)}(z)$, is such that*
$$\sigma_j(\tilde{C}(z)) \leq \gamma \|\psi^{(h)}(z)\|_2 \cdot e^{-\alpha j},$$
where $\alpha = \frac{1}{12k}|\log(\rho+\epsilon)|$ and $\sigma_j(\cdot)$ indicate the j-th singular value of the argument.

3. AN ALGORITHM USING $\mathcal{H}$-MATRICES

We have provided an implementation of CR, which applies to matrix functions $\varphi(z)$ having quasi-separable blocks, and relies on the approximate quasi-separable structure induced by the decay of the singular values. We relied on the $\mathcal{H}$-matrix representation of [5, 6, 7].

3.1 $\mathcal{H}$-matrix representation

Here, we give a brief and informal description of the $\mathcal{H}$-matrix representation that we have implemented. For full details we refer to [5] where an overview of the definition and use of hierarchical matrices is given.

Let $A \in \mathbb{R}^{n \times n}$ be a k-quasiseparable matrix such that $A = \left[\begin{smallmatrix} A_{11} & A_{22} \\ A_{21} & A_{22} \end{smallmatrix}\right]$, $A_{11} \in \mathbb{R}^{n_1 \times n_1}$, $A_{22} \in \mathbb{R}^{n_2 \times n_2}$, with $n_1 := \lfloor \frac{n}{2} \rfloor$ and $n_2 := \lceil \frac{n}{2} \rceil$. Observe that the antidiagonal blocks A_{12} and A_{21} do not involve any element of the main diagonal of A, hence they are representable as a sum of at most k rank-1 matrices. Moreover the diagonal blocks A_{11} and A_{22} are square matrices with the same rank structure of A. Therefore these diagonal blocks are recursively represented with a similar partitioning. If blocks become small enough, they are stored as full matrices.

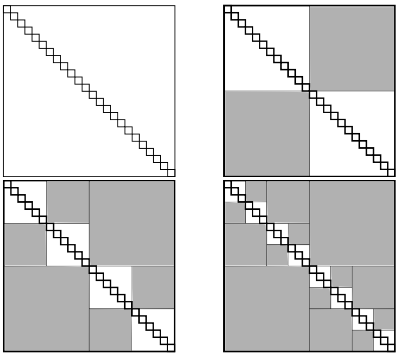


Figure 2: The $\mathcal{H}$-matrix representation. The blocks filled with grey are represented as sum of a few rank-1 matrices, the diagonal blocks in the last step are stored as full matrices.

3.2 Quasiseparable CR

If the quasiseparable rank of the $\mathcal{H}$-matrices is small if compared to the dimension m, then the algorithms which perform the arithmetic operations have almost linear complexity [5]. In particular, we can achieve asymptotic complexity $O(m \log m)$ for matrix addition and $O(m \log^2 m)$ for matrix multiplication and inversion. This is almost optimal, provided that the rank remains sufficiently low.

In order to fully exploit the numerical quasiseparable structure we perform the arithmetic operations (2) of CR adaptively with respect to the rank of the blocks. This means that the result of an arithmetic operation (eg. matrix multiplication) will be an $\mathcal{H}$-matrix with the same partitioning, where each low rank block is a truncated reduced SVD of the corresponding block of the exact result. Hence the rank is not a priori fixed but depends on a threshold ϵ at which the truncation is done. The parameter ϵ can be regarded as the desired accuracy (for us, it is close to the machine precision 2.22×10^{-16}) and can be crucial for the performance of the algorithm.

In Figure 3 we report the CPU time, in seconds, required by our implementation and by the general implementation of CR for several values of the size m and for the values $\epsilon = 10^{-8}, 10^{-12}, 10^{-16}$ in the case of tridiagonal blocks, and in the case of banded blocks with variable bandwidth.

4. CONCLUSIONS

We have experimentally observed the exponential decay of the singular values of certain off-diagonal submatrices generated by cyclic reduction applied to certain QBD stochastic processes of practical interest. We have formally related the rate of decay to geometric features of sets regarding the matrix function $\varphi(z)$ associated with the QBD.

We have provided a software implementation of CR, for QBD with banded blocks encountered in the analysis of bidimensional random walks, which relies on this decay property. The speed up that we get with respect to standard CR is substantial even with moderately large size of the blocks.

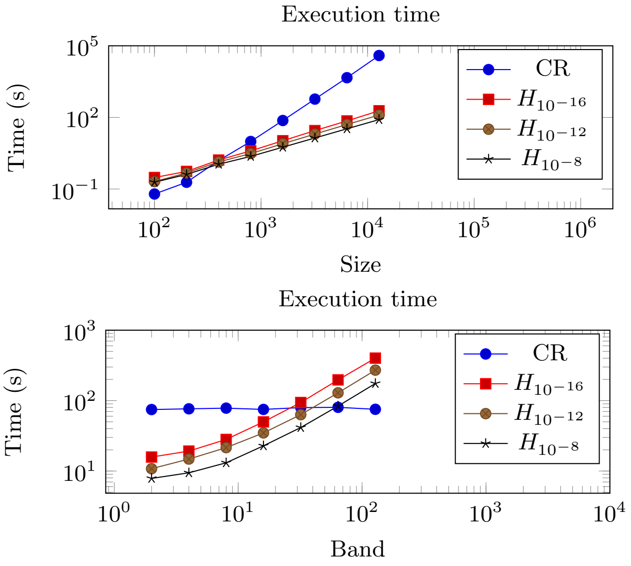


Figure 3: Timings of CR. Above, CR is applied to tridiagonal blocks with increasing size. Below, CR is applied to band blocks with increasing band and size fixed to 1600.

5. REFERENCES

[1] D. A. Bini, P. Favati, and B. Meini. A Compressed Cyclic Reduction for QBD processes with Low-Rank Upper and Lower Transitions. In *Matrix-Analytic Methods in Stochastic Models*, pages 25–40. Springer, 2013.

[2] D. A. Bini, G. Latouche, and B. Meini. *Numerical methods for structured Markov chains*. Oxford University Press, 2005.

[3] D. A. Bini and B. Meini. Effective methods for solving banded Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 20(3):700–719, 1999.

[4] D. A. Bini and B. Meini. The cyclic reduction algorithm: from Poisson equation to stochastic processes and beyond. *Numerical Algorithms*, 51(1):23–60, 2009.

[5] S. Börm, L. Grasedyck, and W. Hackbusch. Hierarchical matrices. *Lecture notes*, 21:2003, 2003.

[6] S. Börm, L. Grasedyck, and W. Hackbusch. Introduction to hierarchical matrices with applications. *Engineering Analysis with Boundary Elements*, 27(5):405–422, 2003.

[7] L. Grasedyck and W. Hackbusch. Construction and arithmetics of h-matrices. *Computing*, 70(4):295–334, 2003.

[8] M. Miyazawa. Tail decay rates in double qbd processes and related reflected random walks. *Mathematics of Operations Research*, 34(3):547–575, 2009.

[9] M. F. Neuts. *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Courier Corporation, 1981.

[10] J. F. Pérez and B. Van Houdt. Quasi-birth-and-death processes with restricted transitions and its applications. *Performance Evaluation*, 68(2):126–141, 2011.

[11] R. Vandebril, M. Van Barel, and N. Mastronardi. *Matrix computations and semiseparable matrices: linear systems*, volume 1. JHU Press, 2007.

Generalised reward generator for stochastic fluid models.

Aviva Samuelson
School of Physical Sciences
University of Tasmania
TAS 7001, Australia
aviva.samuelson@utas.edu.au

Małgorzata M. O'Reilly[*]
School of Physical Sciences
University of Tasmania
TAS 7001, Australia

Nigel G. Bean
School of Mathematical Sciences,
University of Adelaide and
ARC Centre of Excellence for
Mathematical and Statistical
Frontiers

ABSTRACT

We construct a generalised reward matrix $\mathbf{Z}(\underline{s})$, which is an extension of the fluid generator $\mathbf{Q}(s)$ of a stochastic fluid model (SFM). We classify the generators that are projections of $\mathbf{Z}(\underline{s})$, including the generator $\mathbf{Q}(s)$, and discuss the application of the resulting generators in different contexts.

As one application example, for the case with nonzero mean drift, we derive a new Riccati equation for the key matrix $\mathbf{\Psi}$, which records the probabilities of the first return to the original level.

The Riccati equation has the form $\mathbf{\Psi}+\mathbf{\Psi}\mathbf{M}_{-+}\mathbf{\Psi} = \mathbf{M}_{+-}$, where parameters $\mathbf{M}_{+-}$ and $\mathbf{M}_{-+}$ are block matrices in the matrix $\mathbf{M}$, which records the expected number of visits to the original level, before the unbounded fluid drifts to $\pm\infty$.

Finally, we derive the explicit form $\mathbf{\Psi} = \mathbf{M}_{+-}(\mathbf{I}+\mathbf{M}_{--})^{-1}$.

1. INTRODUCTION

Consider the stochastic fluid model (SFM) [2, 3, 4, 10, 11, 12, 13, 16, 17], denoted $\{(\varphi(t), X(t)) : t \geq 0\}$, with phase variable $\varphi(t) \in \mathcal{S} = \{1, \ldots, n\}$ and level variable $X(t) \geq 0$ with a lower boundary at zero, such that:

- $\{\varphi(t) : t \geq 0\}$ is an irreducible continuous-time Markov chain (CTMC) with state space $\mathcal{S}$ and generator $\mathbf{T} = [\mathcal{T}_{ij}]_{i,j\in\mathcal{S}}$;
- $X(t)$ changes at rate $c_{\varphi(t)} = dX(t)/dt$ at time t whenever $X(t) > 0$, and at rate $\max\{c_{\varphi(t)}, 0\}$ whenever $X(t) = 0$.

These models have been used in the analysis of a variety of real-life situations, including telecommunications systems [16], risk assessment [6], power generation systems [9] and congestion control [15]. A classical application example is modelling a telecommunications buffer, using phase $\varphi(t)$ to represent a switch active at time t, and the fluid level $X(t)$ to represent the amount of data in the buffer at time t. The rate of change c_i of the level in the buffer will depend on whether the switch i that is active at time t, lets the data into or out of the buffer, which corresponds to a positive or negative rate, respectively. The analytical expressions for the stationary and transient analysis of this model have been derived in the literature, and powerful algorithms exist for the numerical evaluations of various performance measures [2, 3, 4, 5, 10, 11, 12, 17].

A key matrix in the theory of SFMs is the fluid generator $\mathbf{Q}(s)$ introduced in [11], where s is some complex number. This matrix appears in the expressions for a variety of quantities, including the probability matrices $\mathbf{\Psi}$ and $\mathbf{\Xi}$ defined in [11].

In this paper, we consider the following generalisation of $\mathbf{Q}(s)$. Suppose that while in phase i, a reward is accumulated at some constant real-valued rate r_i per unit of time spent in i. Further, suppose that we wish to track the accumulation of the reward for different phases individually. In order to model this situation, we construct the *generalised reward generator*, denoted $\mathbf{Z}(\underline{s})$, where $\underline{s}$ is some complex vector. The details of the construction are given in Section 2.

We note that $\mathbf{Q}(s)$ is a one-dimensional *projection* of the multi-dimensional LST generator $\mathbf{Z}(\underline{s})$, corresponding to the case when the reward is simply time, with $r_i = 1$ for all i. The details of this and other projections of $\mathbf{Z}(\underline{s})$ are given in Section 3.

In Section 4 we classify the generators that are derived from $\mathbf{Z}(\underline{s})$, and discuss their physical interpretations and applications.

In Section 5, using a particular projection of $\mathbf{Z}(\underline{s})$, denoted $\mathbf{Z}^{+}(s)$, we derive a new Riccati equation for $\mathbf{\Psi}$ and an explicit form of $\mathbf{\Psi}$ with the parameters, for both, being derived from the blocks of $\mathbf{Z}^{+}(0)$. Conclusions follow in Section 6.

2. GENERATOR $\mathbf{Z}(\underline{s})$

In this section we introduce the matrix $\mathbf{Z}(\underline{s})$ and derive the results that describe its physical interpretation as a generator of the SFM. Matrix $\mathbf{Z}(\underline{s})$ is a generalisation of the fluid generators $\mathbf{Q}(s)$ introduced in [11] and $\mathbf{W}(s)$ introduced in [8].

The set $\mathcal{S}$ and generator $\mathbf{T}$ are partitioned depending on the sign of the rates c_i. That is, we partition the set $\mathcal{S}$ as $\mathcal{S} = \mathcal{S}_{+} \cup \mathcal{S}_{-} \cup \mathcal{S}_{0}$, where $\mathcal{S}_{+} = \{i \in \mathcal{S} : c_i > 0\}$, $\mathcal{S}_{-} = \{i \in \mathcal{S} : c_i < 0\}$, and $\mathcal{S}_{0} = \{i \in \mathcal{S} : c_i = 0\}$. Further, we partition

[*] The second author would like to thank the Australian Research Council for funding this research through Linkage Project LP140100152.

the generator $\mathbf{T}$ according to the partition of $\mathcal{S}$, as

$$\mathbf{T} = \begin{array}{c} \\ \mathcal{S}_+ \\ \mathcal{S}_- \\ \mathcal{S}_0 \end{array} \begin{array}{c} \begin{array}{ccc} \mathcal{S}_+ & \mathcal{S}_- & \mathcal{S}_0 \end{array} \\ \left[\begin{array}{ccc} \mathbf{T}_{++} & \mathbf{T}_{+-} & \mathbf{T}_{+0} \\ \mathbf{T}_{-+} & \mathbf{T}_{--} & \mathbf{T}_{-0} \\ \mathbf{T}_{0+} & \mathbf{T}_{0-} & \mathbf{T}_{00} \end{array}\right] \end{array}. \tag{1}$$

Also, we define diagonal matrices $\mathbf{C}_+ = \text{diag}(c_i)_{i\in\mathcal{S}_+}$ and $\mathbf{C}_- = \text{diag}(|c_i|)_{i\in\mathcal{S}_-}$. This notation has been adopted since the various quantities calculated in the analysis of the SFMs appear in a similar block matrix form.

Let $I(\cdot)$ denote the indicator function. For any $i \in \mathcal{S}$, let $r_i \in \mathbb{R}$ be some fixed real constant, and define the random variable $W_i(z,t)$ such that

$$W_i(z,t) = \int_{u=z}^{t} r_i I(\varphi(u) = i)du, \tag{2}$$

which we interpret as the total *i-type reward* earned during the time interval $[z,t]$, assuming the reward is earned at a rate r_i per each unit time spent in i.

Also, define diagonal matrices to collect the rates r_i, $\mathbf{R}_0 = \text{diag}(r_i)_{i\in\mathcal{S}_0}$, $\mathbf{R}_+ = \text{diag}(r_i)_{i\in\mathcal{S}_+}$, and $\mathbf{R}_- = \text{diag}(r_i)_{i\in\mathcal{S}_-}$.

Let $\underline{\mathbf{s}} = [s_i]$ be a (row) vector with $s_i \in \mathbb{C}$, and denote $\mathbf{D}_+ = \text{diag}(s_i)_{i\in\mathcal{S}_+}$, $\mathbf{D}_- = \text{diag}(s_i)_{i\in\mathcal{S}_-}$, and $\mathbf{D}_0 = \text{diag}(s_i)_{i\in\mathcal{S}_0}$.

Assume that $s_1, s_2, \ldots, s_n$ are such that

$$\chi(\mathbf{T}_{00} - \mathbf{D}_0\mathbf{R}_0) < 0, \tag{3}$$

where $\chi(\mathbf{A})$ denotes the eigenvalue with maximum real part of the matrix $\mathbf{A}$. This condition guarantees that the integral $\int_{y=0}^{\infty} e^{(\mathbf{T}_{00}-\mathbf{D}_0\mathbf{R}_0)y}dy$ is well defined and then equal to the inverse $(\mathbf{T}_{00} - \mathbf{D}_0\mathbf{R}_0)^{-1}$. We define the matrix $\mathbf{Z}(\underline{\mathbf{s}})$ as

$$\mathbf{Z}(\underline{\mathbf{s}}) = \left[\begin{array}{cc} \mathbf{Z}_{++}(\underline{\mathbf{s}}) & \mathbf{Z}_{+-}(\underline{\mathbf{s}}) \\ \mathbf{Z}_{-+}(\underline{\mathbf{s}}) & \mathbf{Z}_{--}(\underline{\mathbf{s}}) \end{array}\right], \tag{4}$$

where

$$\begin{aligned} \mathbf{Z}_{++}(\underline{\mathbf{s}}) &= \mathbf{C}_+^{-1}[\mathbf{T}_{++} - \mathbf{D}_+\mathbf{R}_+ - \mathbf{T}_{+0}(\mathbf{T}_{00} - \mathbf{D}_0\mathbf{R}_0)^{-1}\mathbf{T}_{0+}], \\ \mathbf{Z}_{--}(\underline{\mathbf{s}}) &= \mathbf{C}_-^{-1}[\mathbf{T}_{--} - \mathbf{D}_-\mathbf{R}_- - \mathbf{T}_{-0}(\mathbf{T}_{00} - \mathbf{D}_0\mathbf{R}_0)^{-1}\mathbf{T}_{0-}], \\ \mathbf{Z}_{+-}(\underline{\mathbf{s}}) &= \mathbf{C}_+^{-1}[\mathbf{T}_{+-} - \mathbf{T}_{+0}(\mathbf{T}_{00} - \mathbf{D}_0\mathbf{R}_0)^{-1}\mathbf{T}_{0-}], \\ \mathbf{Z}_{-+}(\underline{\mathbf{s}}) &= \mathbf{C}_-^{-1}[\mathbf{T}_{-+} - \mathbf{T}_{-0}(\mathbf{T}_{00} - \mathbf{D}_0\mathbf{R}_0)^{-1}\mathbf{T}_{0+}]. \end{aligned} \tag{5}$$

Now, define the random variable

$$h(t) = \int_{u=0}^{t} |c_{\varphi(u)}|du, \tag{6}$$

interpreted as the total amount of fluid that has entered or left the buffer $X(\cdot)$ during the time interval $[0,t]$, and referred to as the *in-out fluid* [11] of the process $X(\cdot)$. Also, define the random variable, for $y > 0$,

$$\omega(y) = \inf\{t > 0 : h(t) = y\}, \tag{7}$$

interpreted as the first time at which the in-out fluid of the process $X(\cdot)$ reaches level y.

Next, as a generalisation of a similar quantity in [8], for any $i,j \in \mathcal{S}_+\cup\mathcal{S}_-$, any $y > 0$, $t > 0$, and $w_1, w_2, \ldots, w_n \geq 0$, we define

$$\begin{aligned} &\delta_i^y(j,t;w_1,w_2,\ldots,w_n) = P(\varphi(\omega(y)) = j, \omega(y) \leq t, \\ &\qquad W_k(0,\omega(y)) \leq w_k, k = 1,\ldots,n \mid X(0) = 0, \varphi(0) = i), \end{aligned} \tag{8}$$

which we interpret as the joint probability mass/distribution function that, given the process $\{(\varphi(t), X(t)) : t \geq 0\}$ starts from level 0 in phase i, the in-out fluid of the process $X(\cdot)$ reaches level y for the first time at the time $\omega(y) \leq t$, does so in phase $\varphi(\omega(y)) = j$, and the k-type rewards at time $\omega(y)$ satisfy $W_k(0,\omega(y)) \leq w_k$ for all $k = 1,\ldots,n$.

Also, define the corresponding multi-dimensional Laplace-Stieltjes transform (LST) matrix $\tilde{\mathbf{\Delta}}^y(\underline{\mathbf{s}})$ such that, for any $y > 0$, any vector $\underline{\mathbf{s}} = [s_i]$ satisfying condition (3), and any $i,j \in \mathcal{S}_+\cup\mathcal{S}_-$,

$$\begin{aligned} [\tilde{\mathbf{\Delta}}^y(\underline{\mathbf{s}})]_{ij} &= E\big(e^{-(s_1W_1(0,\omega(y))+\ldots+s_nW_n(0,\omega(y)))} \\ &\quad \times I(\varphi(\omega(y)) = j) \mid \varphi(0) = i\big), \qquad (9) \\ &= \int_{t=0}^{\infty}\int_{w_1=0}^{r_1t}\ldots\int_{w_n=0}^{r_nt} e^{-(s_1w_1+\ldots+s_nw_n)} \\ &\quad \times d\delta_i^y(j,t;w_1,w_2,\ldots,w_n). \qquad (10) \end{aligned}$$

is the LST of the distribution of $(W_1(0,\omega(y)),\ldots,W_n(0,\omega(y)))$ and $\varphi(\omega(y)) = j$, given $\varphi(0) = i$.

Further, for $i,j \in \mathcal{S}_0$, $t > 0$, and $w_1, w_2, \ldots, w_n \geq 0$, let

$$\begin{aligned} &\beta_i^t(j;w_1,w_2,\ldots,w_n) = P(\varphi(t) = j, \varphi(u) \in \mathcal{S}_0, 0 \leq u \leq t, \\ &\qquad W_k(0,t) \leq w_k, k = 1,2,\ldots,n \mid \varphi(0) = i), \end{aligned} \tag{11}$$

which we interpret as the joint probability mass/distribution function that the phase remains in the set $\mathcal{S}_0$ at least for the duration of time t, the phase at time t is $\varphi(t) = j$, and the k-type rewards at time t satisfy $W_k(0,t) \leq w_k$ for all $k = 1,\ldots,n$, given the process $\{(\varphi(t), X(t)) : t \geq 0\}$ starts from level 0 in phase i.

Also, define matrix $\tilde{\mathbf{B}}^t(\underline{\mathbf{s}})$ such that

$$\begin{aligned} [\tilde{\mathbf{B}}^t(\underline{\mathbf{s}})]_{ij} &= \int_{w_1=0}^{r_1t}\ldots\int_{w_n=0}^{r_nt} e^{-(s_1w_1+\ldots+s_nw_n)} \\ &\quad \times d\beta_i^t(j,x;w_1,w_2,\ldots,w_n), \end{aligned} \tag{12}$$

is the LST of the distribution of $(W_1(0,t),\ldots,W_n(0,t))$, $\varphi(t) = j$ and the phase process remains in the set $\mathcal{S}_0$ at least for the duration of time t, given $\varphi(0) = i$.

Theorem 1 proves that the matrix $(\mathbf{T}_{00} - \mathbf{D}_0\mathbf{R}_0)$ is the generator of $\tilde{\mathbf{B}}^t(\underline{\mathbf{s}})$.

THEOREM 1. *$\tilde{\mathbf{B}}^t(\underline{\mathbf{s}})$ is well defined for all $t > 0$, and all $s_i \in \mathbb{C}$, and*

$$\tilde{\mathbf{B}}^t(\underline{\mathbf{s}}) = e^{(\mathbf{T}_{00}-\mathbf{D}_0\mathbf{R}_0)t}. \tag{13}$$

Proof: To prove this result we closely follow the methodology developed in [8, Theorem 1]. The difference is that here, we track the i-type rewards of the possible states $i \in \mathcal{S}$ separately.

First, note for all $a, b \geq 0$, and $k = 1,\ldots,n$, that

$$W_k(0,a+b) = W_k(0,a) + W_k(a,a+b),$$

and that by conditioning on the state of all the random variables at time a the behaviour of the process during the interval $[0,a)$ is independent of the behaviour of the process during the interval $(a,a+b]$. Therefore, by the law of total probability,

$$\begin{aligned}
[\tilde{\mathbf{B}}^{(a+b)}(\underline{s})]_{ij} &= \int_{w_1=0}^{r_1(a+b)} \cdots \int_{w_n=0}^{r_n(a+b)} e^{-(s_1 w_1 + \ldots + s_n w_n)} \\
&\quad \times d\beta_i^{(a+b)}(j; w_1, \ldots, w_n) \\
&= \sum_{k \in \mathcal{S}_0} \int_{u_1=0}^{r_1 a} \cdots \int_{u_n=0}^{r_n a} e^{-(s_1 u_1 + \ldots + s_n u_n)} \\
&\quad \times d\beta_i^{(a)}(k; u_1, \ldots, u_n) \\
&\quad \times \int_{h_1=0}^{r_1 b} \cdots \int_{h_n=0}^{r_n b} e^{-(s_1 h_1 + \ldots + s_n h_n)} \\
&\quad \times d\beta_k^{(b)}(j; h_1, \ldots, h_n) \\
&= \sum_{k \in \mathcal{S}_0} [\tilde{\mathbf{B}}^{(a)}(\underline{s})]_{ik} [\tilde{\mathbf{B}}^{(b)}(\underline{s})]_{kj},
\end{aligned}$$

and so

$$\tilde{\mathbf{B}}^{(a+b)}(\underline{s}) = \tilde{\mathbf{B}}^{(a)}(\underline{s}) \tilde{\mathbf{B}}^{(b)}(\underline{s}). \tag{14}$$

Also, denoting $\tilde{\mathbf{B}}^0(\underline{s}) = \lim_{t \to 0^+} \tilde{\mathbf{B}}^y(\underline{s})$, we have

$$\tilde{\mathbf{B}}^0(\underline{s}) = \mathbf{I}, \quad \lim_{y \to 0^+} ||\tilde{\mathbf{B}}^y(\underline{s}) - \mathbf{I}|| = \mathbf{0}. \tag{15}$$

By (14) and (15), $\{\tilde{\mathbf{B}}^t(\underline{s}), t > 0\}$ is a strongly continuous semigroup and so $\tilde{\mathbf{B}}^t(\underline{s})$ must be of the form $e^{\mathbf{V}(\underline{s})t}$, where

$$\mathbf{V}(\underline{s}) = \frac{d}{dh} \tilde{\mathbf{B}}^h(\underline{s}) \bigg|_{h=0^+}. \tag{16}$$

Consider the case when the process starts in phase $i \in \mathcal{S}_0$ at time zero and is observed at some time $h > 0$ in phase $j \in \mathcal{S}_0$. Given h is small, there are only two possible events that could occur with probability greater than $o(h)$.

1. The phase process remains in phase i until time h. This event occurs with probability $e^{-\lambda_i h}$ with $\lambda_i = -\mathcal{T}_{ii}$. The corresponding i-type reward is $r_i h$, k-type reward for $k \neq i$ is zero, and so the LST of the rewards conditional on this event occurring is $e^{s_i(r_i h)}$. We multiply the probability by the conditional LST, and store the result in a diagonal matrix $\tilde{\mathbf{B}}_1^t(\underline{s})$ with the (i,i)-th entry given by

   $$[\tilde{\mathbf{B}}_1^h(\underline{s})]_{ii} = e^{-(s_i r_i + \lambda_i)h}.$$

   It follows that,

   $$\frac{d}{dh} [\tilde{\mathbf{B}}_1^h(\underline{s})]_{ii} \bigg|_{h=0^+} = -(s_i r_i + \lambda_i),$$

   and so

   $$\frac{d}{dh} \tilde{\mathbf{B}}_1^h(\underline{s}) \bigg|_{h=0^+} = -(\mathbf{D}_0 \mathbf{R}_0 + \mathbf{\Lambda}_0), \tag{17}$$

   where $\mathbf{\Lambda}_0$ is a diagonal matrix with $[\mathbf{\Lambda}_0]_{ii} = \lambda_i$ for all $i \in \mathcal{S}_0$.

2. The phase process makes a single transition from i to phase $j \neq i \in \mathcal{S}_0$ at some time $u \in (0, h]$ and remains there until time h. The process undergoes the following set of steps:

   - First, the process leaves phase i at some time $u \in (0, h]$ and does so with probability density $(\lambda_i) e^{-\lambda_i u}$.
   - Next, the process makes a transition from phase i to phase j with probability $\mathcal{T}_{ij}/\lambda_i$.
   - Further, the process remains in phase j for the remaining $(h-u)$ time with probability $e^{-\lambda_j(h-u)}$.

   Therefore, the probability of this type of event occurring is $\lambda_i e^{-\lambda_i u} (\mathcal{T}_{ij}/\lambda_i) e^{\lambda_j (h-u)}$. The corresponding i-type reward is $r_i u$, j-type reward is $r_j(h-u)$, k-type reward for $k \neq i, j$ is zero, and so the LST of the rewards conditional on this event occurring is $e^{-s_i r_i u - s_j r_j (h-u)}$. We multiply the probability by the conditional LST and integrate over all $u \in (0, h]$, before finally storing the result in a matrix $\tilde{\mathbf{B}}_2^t(\underline{s})$ with the (i,j)-th entry, for $i \neq j$, given by

   $$[\tilde{\mathbf{B}}_2^h(\underline{s})]_{ij} = \int_{u=0}^{h} e^{-s_i r_i u - s_j r_j (h-u)} \lambda_i e^{-\lambda_i u} \frac{\mathcal{T}_{ij}}{\lambda_i} e^{-\lambda_j (h-u)} du.$$

   It follows that,

   $$\begin{aligned}
   &\frac{d}{dh} [\tilde{\mathbf{B}}_2^h(\underline{s})]_{ij} \bigg|_{h=0^+} \qquad\qquad (18) \\
   &= \frac{d}{dh} \left[\mathcal{T}_{ij} e^{-h(s_j r_j + \lambda_j)} \int_{u=0}^{h} e^{-u(\lambda_i + s_i r_i - \lambda_j - s_j r_j)} du \right]_{h=0^+} \\
   &= \mathcal{T}_{ij},
   \end{aligned}$$

   and so

   $$\frac{d}{dh} \tilde{\mathbf{B}}_2^h(\underline{s}) \bigg|_{h=0^+} = \mathbf{T}_{00} + \mathbf{\Lambda}_0. \tag{19}$$

Consequently, for all $i, j \in \mathcal{S}_0$, we have

$$\begin{aligned}
[\mathbf{V}(\underline{s})]_{ij} &= \left[\frac{d}{dh} \left(\tilde{\mathbf{B}}_1^h(\underline{s}) + \tilde{\mathbf{B}}_2^h(\underline{s}) \right) \bigg|_{h=0^+} \right]_{ij} \\
&= [\mathbf{T}_{00} - \mathbf{D}_0 \mathbf{R}_0]_{ij}.
\end{aligned} \tag{20}$$

The result follows. ■

The theorem below proves that $\mathbf{Z}(\underline{s})$ is the generator of $\tilde{\mathbf{\Delta}}^y(\underline{s})$.

Theorem 2. *For any $y > 0$, $\tilde{\mathbf{\Delta}}^y(\underline{s})$ exists and*

$$\tilde{\mathbf{\Delta}}^y(\underline{s}) = e^{\mathbf{Z}(\underline{s})y}. \tag{21}$$

Proof: To prove this result we closely follow the methodology developed in [8, Theorem 2]. The difference is that here, we track the i-type rewards of the possible states $i \in \mathcal{S}$ separately.

Suppose that the process starts in level 0 from some phase $i \in \mathcal{S}_+ \cup \mathcal{S}_-$ at time 0, and the in-out fluid hits level $y + v$ for the first time at some time $\omega(y+v)$, and does so in phase $j \in \mathcal{S}_+ \cup \mathcal{S}_-$. In order for this to occur,

- first the in-out fluid of the process $X(\cdot)$ must hit level y at time $\omega(y)$, in some phase $\ell \in \mathcal{S}_+ \cup \mathcal{S}_-$, and
- next, the in-out fluid of the process $X(\cdot)$ must hit level $y + v$ at time $\omega(y+v)$, in phase j.

Since $\omega(y) \leq \omega(y+v)$ for $y, v > 0$, it follows by (2) that for all $k = 1, \ldots, n$,

$$W_k(0, \omega(y+v)) = W_k(0, \omega(y)) + W_k(\omega(y), \omega(y+v)). \tag{22}$$

As in the previous proof, by conditioning on the state of all random variables at the time $\omega(y)$, the behaviour of the process during the interval $[0,\omega(y))$ is independent of the behaviour of the process during the interval $(\omega(y),\omega(y+v)]$, by the memoryless property of the Markov chain, for all $y,v>0$. By (9), (22), and the law of total probability,

$$\begin{aligned}
[\tilde{\boldsymbol{\Delta}}^{y+v}(\underline{\mathbf{s}})]_{ij} &= E\big(e^{-(s_1W_1(0,\omega(y+v))+\ldots+s_nW_n(0,\omega(y+v)))} \\
&\quad \times I(\varphi(\omega(y+v))=j)|\varphi(0)=i\big) \\
&= \int_{t=0}^{\infty}\int_{w_1=0}^{r_1t}\cdots\int_{w_n=0}^{r_nt} \\
&\quad \times e^{-(s_1w_1+\ldots+s_nw_n)}d\delta_i^{y+v}(j,t;w_1,\ldots,w_n) \\
&= \sum_{\ell\in\mathcal{S}_+\cup\mathcal{S}_-}\Big[\int_{z=0}^{\infty}\int_{v_1=0}^{r_1z}\cdots\int_{v_n=0}^{r_nz} \\
&\quad \times e^{-(s_1v_1+\ldots+s_nv_n)}d\delta_i^{y}(\ell,z;v_1,\ldots,v_n)\Big] \\
&\quad \times\Big[\int_{u=0}^{\infty}\int_{h_1=0}^{r_1u}\cdots\int_{h_n=0}^{r_nu} \\
&\quad \times e^{-(s_1h_1+\ldots+s_nh_n)}d\delta_\ell^{v}(j,u;h_1,\ldots,h_n)\Big] \\
&= \sum_{\ell\in\mathcal{S}_+\cup\mathcal{S}_-}[\tilde{\boldsymbol{\Delta}}^{y}(\underline{\mathbf{s}})]_{i\ell}[\tilde{\boldsymbol{\Delta}}^{v}(\underline{\mathbf{s}})]_{\ell j}.
\end{aligned}$$

Therefore, for all $y,v>0$,

$$\tilde{\boldsymbol{\Delta}}^{y+v}(\underline{\mathbf{s}})=\tilde{\boldsymbol{\Delta}}^{y}(\underline{\mathbf{s}})\tilde{\boldsymbol{\Delta}}^{v}(\underline{\mathbf{s}}). \tag{23}$$

Also, denoting $\tilde{\boldsymbol{\Delta}}^{0}(\underline{\mathbf{s}})=\lim_{y\to0^+}\tilde{\boldsymbol{\Delta}}^{y}(\underline{\mathbf{s}})$, when $j=i$,

$$\begin{aligned}
[\tilde{\boldsymbol{\Delta}}^{0}(\underline{\mathbf{s}})]_{ii} &= \int_{t=0}^{\infty}\int_{w_1=0}^{r_1t}\cdots\int_{w_n=0}^{r_nt} \\
&\quad \times e^{-(s_1w_1+\ldots+s_nw_n)}d\delta_i^{0}(i,t;w_1,w_2,\ldots,w_n) \\
&= 1,
\end{aligned}$$

and when $j\neq i$,

$$\begin{aligned}
[\tilde{\boldsymbol{\Delta}}^{0}(\underline{\mathbf{s}})]_{ij} &= \int_{t=0}^{\infty}\int_{w_1=0}^{r_1t}\cdots\int_{w_n=0}^{r_nt} \\
&\quad \times e^{-(s_1w_1+\ldots+s_nw_n)}d\delta_i^{0}(j,t;w_1,w_2,\ldots,w_n) \\
&= 0,
\end{aligned}$$

and so

$$\tilde{\boldsymbol{\Delta}}^{0}(\underline{\mathbf{s}})=\mathbf{I},\qquad \lim_{y\to0^+}||\tilde{\boldsymbol{\Delta}}^{y}(\underline{\mathbf{s}})-\mathbf{I}||=\mathbf{0}. \tag{24}$$

By (23) and (24), $\{\tilde{\boldsymbol{\Delta}}^{y}(\underline{\mathbf{s}}),y>0\}$ is a strongly continuous semi-group, and so $\tilde{\boldsymbol{\Delta}}^{y}(\underline{\mathbf{s}})=e^{\mathbf{U}(\underline{\mathbf{s}})y}$, where

$$\mathbf{U}(\underline{\mathbf{s}})=\frac{d}{dh}\tilde{\boldsymbol{\Delta}}^{h}(\underline{\mathbf{s}})\Big|_{h=0}. \tag{25}$$

We consider the case $i,j\in\mathcal{S}_+$. The proof for the other cases follow by an analogous argument. We will show that

$$[\mathbf{U}(\underline{\mathbf{s}})]_{ij}=[\mathbf{Z}_{++}(\underline{\mathbf{s}})]_{ij}.$$

Assume the process starts from level $X(0)=0$ in phase $i\in\mathcal{S}_+$ and the in-out fluid hits level h, for some small $h\geq0$, at time $\omega(h)$, and does so in phase $j\in\mathcal{S}_+$. Since h is small, there are only three possible events that could occur with a probability greater than $o(h)$.

1. The process remains in phase $i\in\mathcal{S}_+$ until the in-out fluid reaches level h. In this case, $\omega(h)=h/|c_i|$ and the probability of this occurring is $e^{-\lambda_i(h/|c_i|)}$. The corresponding i-type reward is $r_ih/|c_i|$, k-type reward for $k\neq i$ is zero, and so the LST of the rewards conditional on this event occurring is $e^{-s_i(r_ih/|c_i|)}$. We multiply the probability by the conditional LST, the result of which is stored in a diagonal matrix $\tilde{\boldsymbol{\Delta}}_1^h$ with the (i,i)-th entry given by

   $$[\tilde{\boldsymbol{\Delta}}_1^h(\underline{\mathbf{s}})]_{ii}=e^{-s_i(r_ih/|c_i|)}e^{-\lambda_i(h/|c_i|)}.$$

   It follows that

   $$\frac{d}{dh}[\tilde{\boldsymbol{\Delta}}_1^h(\underline{\mathbf{s}}_i)]_{ii}\Big|_{h=0^+}=-\frac{s_ir_i+\lambda_i}{|c_i|},$$

   and so

   $$\frac{d}{dh}[\tilde{\boldsymbol{\Delta}}_1^h(\underline{\mathbf{s}}_i)]_{ii}\Big|_{h=0^+}=-\mathbf{C}_+^{-1}(\mathbf{D}_+\mathbf{R}_++\boldsymbol{\Lambda}_+), \tag{26}$$

   where $\boldsymbol{\Lambda}_+=diag(\lambda_i)_{i\in\mathcal{S}_+}$.

2. The process makes a single transition from state i to state $j\neq i\in\mathcal{S}_+$ when the in-out fluid hits some level $u\in(0,h]$ (at time $u/|c_i|$) and remains there until time $\omega(h)$, that is for a further $(h-u)/|c_j|$. Therefore, the probability of this type of event occurring is

   $$(\frac{\lambda_i}{|c_i|})e^{-\lambda_i(u/|c_i|)}\frac{\mathcal{T}_{ij}}{\lambda_i}e^{-\lambda_j((h-u)/|c_j|)}. \tag{27}$$

   It follows that $\omega(h)=r_iu/|c_i|+r_j(h-u)/|c_j|$, and the corresponding LST of the rewards conditional on this event occurring is

   $$e^{-s_ir_iu/|c_i|+s_jr_j(h-u)/|c_j|}. \tag{28}$$

   By multiplying (27) and (28), and integrating over all $u\in(0,h]$, we obtain the (i,j)th entry of $\tilde{\boldsymbol{\Delta}}_2^h$, given by,

   $$\begin{aligned}
   [\tilde{\boldsymbol{\Delta}}_2^h(\underline{\mathbf{s}})]_{ij} &= \int_{u=0}^{h}e^{-s_ir_iu/|c_i|-s_jr_j(h-u)/|c_j|}(\frac{1}{|c_i|}) \\
   &\quad \times e^{-\lambda_i(u/|c_i|)}\mathcal{T}_{ij}e^{-\lambda_j((h-u)/|c_j|)}du.
   \end{aligned} \tag{29}$$

   Therefore, for $i\neq j$,

   $$\begin{aligned}
   &\frac{d}{dh}[\tilde{\boldsymbol{\Delta}}_2^h(\underline{\mathbf{s}})]_{ij}\Big|_{h=0^+} \\
   &= \frac{d}{dh}\Big[\int_{u=0}^{h}e^{-s_ir_iu/|c_i|-s_jr_j(h-u)/|c_j|}(\frac{1}{|c_i|}) \\
   &\quad \times e^{-\lambda_i(u/|c_i|)}\mathcal{T}_{ij}e^{-\lambda_j((h-u)/|c_j|)}du\Big]_{h=0^+} \\
   &= \frac{\mathcal{T}_{ij}}{|c_i|},
   \end{aligned} \tag{30}$$

   and so

   $$\frac{d}{dh}\tilde{\boldsymbol{\Delta}}_2^h(\underline{\mathbf{s}})\Big|_{h=0^+}=\mathbf{C}_+^{-1}(\mathbf{T}_{++}+\boldsymbol{\Lambda}_+). \tag{31}$$

3. The process transitions from state i into some $\ell\in\mathcal{S}_0$, and then, after spending some time t in $\mathcal{S}_0$, transitions into $j\in\mathcal{S}_+$ and remains there until time $\omega(h)$.

   - First, the process leaves phase i when the in-out fluid hits level u at time $u/|c_i|$, with probability density $(\lambda_i/|c_i|)e^{-\lambda_i(u/|c_i|)}$.

- Next, the process makes a transition from $i \in \mathcal{S}_+$ to $\ell \in \mathcal{S}_0$ with probability $[\mathbf{T}_{+0}]_{i\ell}/\lambda_i$.
- Further, the process remains in set $\mathcal{S}_0$ for the duration of time t and then makes a transition to phase $j \in \mathcal{S}_+$, with probability density $[e^{\mathbf{T}_{00}t}\mathbf{T}_{0+}]_{ij}$.
- Finally, the process remains in phase j until the in-out fluid reaches level h, with probability $e^{-\lambda_j((h-u)/|c_j|)}$.

It follows that $\omega(h) = u/|c_i| + t + (h-u)/|c_j|$, the probability of this occurring is

$$\frac{1}{|c_i|}e^{-\lambda_i(\frac{u}{|c_i|})}[\mathbf{T}_{+0}e^{\mathbf{T}_{00}t}\mathbf{T}_{0+}]_{ij}e^{-\lambda_j(\frac{(h-u)}{|c_j|})}, \tag{32}$$

and the corresponding LST components of the rewards conditional on this event occurring is

$$e^{-(s_i\frac{u}{|c_i|})}, \quad e^{\mathbf{D}_0\mathbf{R}_0 t}, \quad e^{(s_j\frac{h-u}{|c_j|})}. \tag{33}$$

By multiplying the terms in (32) and (33) in an appropriate order, and integrating over all $u \in (0,h]$, and $t \in (0,\infty)$, we obtain the (i,j)-th entry of $\tilde{\boldsymbol{\Delta}}_3$, given by

$$\begin{aligned}[\tilde{\boldsymbol{\Delta}}_3^h(\underline{\mathbf{s}})]_{ij} \quad = \quad & \int_{u=0}^{h} \frac{1}{|c_i|}e^{-(s_i\frac{r_i u}{|c_i|}+s_j\frac{r_j(h-u)}{|c_j|})}e^{-\lambda_i(\frac{u}{|c_i|})} \\ & \times [\mathbf{T}_{+0}\int_{t=0}^{\infty} e^{\mathbf{D}_0\mathbf{R}_0 t}e^{\mathbf{T}_{00}t}dt\mathbf{T}_{0+}]_{ij} \\ & \times e^{-\lambda_j(\frac{(h-u)}{|c_j|})}du. \end{aligned} \tag{34}$$

Since $\underline{\mathbf{s}}$ satisfies the condition (3), the inner integral exists and is given by $-(\mathbf{T}_{00} - \mathbf{D}_0\mathbf{R}_0)^{-1}$, and so

$$\frac{d}{dh}[\tilde{\boldsymbol{\Delta}}_3^h(\underline{\mathbf{s}})]_{ij}\bigg|_{h=0^+} = -\frac{1}{|c_i|}[\mathbf{T}_{+0}(\mathbf{T}_{00}-\mathbf{D}_0\mathbf{R}_0)^{-1}\mathbf{T}_{0+}]_{ij},$$

and

$$\frac{d}{dh}\tilde{\boldsymbol{\Delta}}_3^h(\underline{\mathbf{s}})\bigg|_{h=0^+} = -\mathbf{C}_+^{-1}\mathbf{T}_{+0}(\mathbf{T}_{00} - \mathbf{D}_0\mathbf{R}_0)^{-1}\mathbf{T}_{0+}. \tag{35}$$

By above,

$$\begin{aligned} & [\mathbf{U}(\underline{\mathbf{s}})]_{ij} \\ = \ & \left[\frac{d}{dh}\tilde{\boldsymbol{\Delta}}^h(\underline{\mathbf{s}})\bigg|_{h=0^+}\right]_{ij} \\ = \ & \left[\frac{d}{dh}(\tilde{\boldsymbol{\Delta}}_1^h(\underline{\mathbf{s}}) + \tilde{\boldsymbol{\Delta}}_2^h(\underline{\mathbf{s}}) + \tilde{\boldsymbol{\Delta}}_3^h(\underline{\mathbf{s}}))\bigg|_{h=0^+}\right]_{ij} \\ = \ & [\mathbf{C}_+^{-1}\{(\mathbf{T}_{++} - \mathbf{D}_+\mathbf{R}_+ - \mathbf{T}_{+0}(\mathbf{T}_{00} - \mathbf{D}_0\mathbf{R}_0)^{-1}\mathbf{T}_{0+}\}]_{ij} \\ = \ & [\mathbf{Z}_{++}(\underline{\mathbf{s}})]_{ij}. \end{aligned} \tag{36}$$

In a manner analogous to the argument above, we prove the expression for $\mathbf{Z}_{+-}(\underline{\mathbf{s}})$ (which clearly must have a zero contribution from Case 1). By symmetry, the expressions for $\mathbf{Z}_{-+}(\underline{\mathbf{s}})$ and $\mathbf{Z}_{--}(\underline{\mathbf{s}})$ also follow. ■

3. PROJECTIONS OF $\mathbf{Z}(\underline{\mathbf{s}})$

Below, we study a range of *projections* of the generalised reward matrix $\mathbf{Z}(\underline{\mathbf{s}})$ to generators of one-dimensional LSTs, and discuss their applications.

To do this, we define the random variable

$$W(z,t) = \sum_{i\in\mathcal{S}} W_i(z,t), \tag{37}$$

interpreted as the *total reward* earned in all states during the time interval $[z,t]$, and replace $e^{-(s_1W_1(z,t),\ldots,s_nW_n(z,t))}$ with a function of the scalar $s \in \mathbb{C}$, $e^{-sW(z,t)}$, in the definitions of the multi-dimensional LSTs. The resulting projections are generators of the one-dimensional LSTs of the distribution of the total reward (37). For example, we replace (9) with

$$[\tilde{\boldsymbol{\Delta}}^y(s)]_{ij} \quad = \quad E\big(e^{-sW(0,\omega(y))}I(\varphi(\omega(y)) = j) \mid \varphi(0) = i\big), \tag{38}$$

and make associated changes accordingly.

The first example is the fluid generator $\mathbf{Q}(s)$ defined in [11],

$$\mathbf{Q}(s) = \begin{bmatrix} \mathbf{Q}_{++}(s) & \mathbf{Q}_{+-}(s) \\ \mathbf{Q}_{-+}(s) & \mathbf{Q}_{--}(s) \end{bmatrix}, \tag{39}$$

where

$$\begin{aligned} \mathbf{Q}_{++}(s) &= \mathbf{C}_+^{-1}[\mathbf{T}_{++} - s\mathbf{I} - \mathbf{T}_{+0}(\mathbf{T}_{00} - s\mathbf{I})^{-1}\mathbf{T}_{0+}], \\ \mathbf{Q}_{--}(s) &= \mathbf{C}_-^{-1}[\mathbf{T}_{--} - s\mathbf{I} - \mathbf{T}_{-0}(\mathbf{T}_{00} - s\mathbf{I})^{-1}\mathbf{T}_{0-}], \\ \mathbf{Q}_{+-}(s) &= \mathbf{C}_+^{-1}[\mathbf{T}_{+-} - \mathbf{T}_{+0}(\mathbf{T}_{00} - s\mathbf{I})^{-1}\mathbf{T}_{0-}], \\ \mathbf{Q}_{-+}(s) &= \mathbf{C}_-^{-1}[\mathbf{T}_{-+} - \mathbf{T}_{-0}(\mathbf{T}_{00} - s\mathbf{I})^{-1}\mathbf{T}_{0+}]. \end{aligned} \tag{40}$$

This generator is a projection of $\mathbf{Z}(\underline{\mathbf{s}})$ obtained by setting all reward rates in (37) to $r_i = 1$, so that the amount of the reward earned in i is equal to the time spent in i.

The physical interpretation of $\mathbf{Q}(s)$ established in [11] is that $[e^{\mathbf{Q}(s)y}]_{ij}$ records the LST of the distribution of time for the process to first reach in-out fluid level y and do so in phase j, assuming the process starts from level 0 in phase i. The generator $\mathbf{Q}(s)$ was used in [10, 11, 12] to evaluate the key matrices $\boldsymbol{\Psi}(s)$, $\boldsymbol{\Xi}(s)$, and related quantities of the SFM $\{(\varphi(t), X(t)) : t \geq 0\}$.

A projection of $\mathbf{Z}(\underline{\mathbf{s}})$ with any real reward rates r_i is the generator $\mathbf{W}(s)$ derived in [8],

$$\mathbf{W}(s) = \begin{bmatrix} \mathbf{W}_{++}(s) & \mathbf{W}_{+-}(s) \\ \mathbf{W}_{-+}(s) & \mathbf{W}_{--}(s) \end{bmatrix},$$

with

$$\begin{aligned} \mathbf{W}_{++}(s) &= \mathbf{C}_+^{-1}[(\mathbf{T}_{++} - s\mathbf{R}_+) - \mathbf{T}_{+0}(\mathbf{T}_{00} - s\mathbf{R}_0)^{-1}\mathbf{T}_{0+}], \\ \mathbf{W}_{--}(s) &= \mathbf{C}_-^{-1}[(\mathbf{T}_{--} - s\mathbf{R}_-) - \mathbf{T}_{-0}(\mathbf{T}_{00} - s\mathbf{R}_0)^{-1}\mathbf{T}_{0-}], \\ \mathbf{W}_{+-}(s) &= \mathbf{C}_+^{-1}[\mathbf{T}_{+-} - \mathbf{T}_{+0}(\mathbf{T}_{00} - s\mathbf{R}_0)^{-1}\mathbf{T}_{0-}], \\ \mathbf{W}_{-+}(s) &= \mathbf{C}_-^{-1}[\mathbf{T}_{-+} - \mathbf{T}_{-0}(\mathbf{T}_{00} - s\mathbf{R}_0)^{-1}\mathbf{T}_{0+}]. \end{aligned} \tag{41}$$

Generator $\mathbf{W}(s)$ was used to analyse the coupled evolution of two fluids: the fluid $\{(\varphi(t), X(t)) : t \geq 0\}$ with a lower boundary zero, as defined in Section 1, and the unbounded fluid $\{(\varphi(t), Y(t)) : t \geq 0\}$ with rates r_i. As shown in [8], $[e^{\mathbf{W}(s)y}]_{ij}$ is the LST of the distribution of the total shift in the fluid $Y(\cdot)$, expressed as $Y(\omega(y)) - Y(0)$, accumulated at the time $\omega(y)$ when the *in-out fluid of the process* $X(\cdot)$ first reaches level y, and $\varphi(\omega(y)) = j$, given $\varphi(0) = i$. In this sense, the amount of the reward earned in i is equal to the shift in $Y(\cdot)$ accumulated while in i.

In this paper, of particular interest is the following projection of $\mathbf{Z}(\underline{\mathbf{s}})$, denoted $\mathbf{Z}^+(s)$. For a stochastic fluid model $\{(\varphi(t), X(t)), t \geq 0\}$ with level variable $X(t)$ unbounded above, we define

$$h_+(t) = \int_{u=0}^{t} c_{\varphi(u)} I(i \in \mathcal{S}_+) du, \qquad (42)$$

interpreted as the total amount of fluid that has flowed *into* the buffer $X(\cdot)$ during the time interval $[0, t]$, referred to as the *upward shift*, since it records the total shift in the fluid during the times the fluid level was increasing.

Suppose that we want to track only this upward shift in the fluid $\{(\varphi(t), X(t)) : t \geq 0\}$. To achieve this, we consider the projection with $W(0, t) = h_+(t)$, let $\mathbf{R}_+ = \mathbf{C}_+$, $\mathbf{R}_- = \mathbf{0}$, and $\mathbf{R}_0 = \mathbf{0}$ in (4)-(5), resulting in the matrix

$$\mathbf{Z}^+(s) = \begin{bmatrix} \mathbf{Z}^+_{++}(s) & \mathbf{Z}^+_{+-}(s) \\ \mathbf{Z}^+_{-+}(s) & \mathbf{Z}^+_{--}(s) \end{bmatrix}, \qquad (43)$$

where $\mathbf{Z}^+_{++}(s) = \mathbf{Q}_{++}(0) - s\mathbf{I}$, $\mathbf{Z}^+_{--}(s) = \mathbf{Q}_{--}(0)$, $\mathbf{Z}^+_{+-}(s) = \mathbf{Q}_{+-}(0)$, and $\mathbf{Z}^+_{-+}(s) = \mathbf{Q}_{-+}(0)$.

Here, we establish the following physical interpretation of $\mathbf{Z}^+(s)$. By Theorem 2, for any $i, j \in \mathcal{S}$ and $y > 0$,

$$[e^{\mathbf{Z}^+(s)y}]_{ij} = E(e^{-sh_+(\omega(y))} I(\varphi(\omega(y)) = j) \mid \varphi(0) = i) \quad (44)$$

is the LST of the distribution of the total upward shift in $X(\cdot)$ accumulated by the time the in-out fluid of the process $X(\cdot)$ first reaches level y and does so in phase j, given that the process starts in phase i at time zero.

In [16], $\mathbf{Z}^+(s)$ is used to calculate various loss rates for a doubly-bounded SFM, corresponding to the fluid lost during periods of congestion when the buffer collecting the fluid is full.

By symmetry, for a stochastic fluid model $\{(\varphi(t), X(t)), t \geq 0\}$ with level variable $X(t)$ unbounded below, we define

$$h_-(t) = \int_{u=0}^{t} |c_{\varphi(u)}| I(i \in \mathcal{S}_-) du, \qquad (45)$$

interpreted as the total amount of fluid that has flowed out of the buffer $X(\cdot)$ during the time interval $[0, t]$, referred to as the *downward shift*, since it records the total shift in the fluid during the times the fluid level was decreasing.

In order to track the downward shift in the fluid, we define matrix $\mathbf{Z}^-$, given by

$$\mathbf{Z}^-(s) = \begin{bmatrix} \mathbf{Z}^-_{++}(s) & \mathbf{Z}^-_{+-}(s) \\ \mathbf{Z}^-_{-+}(s) & \mathbf{Z}^-_{--}(s) \end{bmatrix}, \qquad (46)$$

where $\mathbf{Z}^-_{++}(s) = \mathbf{Q}_{++}(0)$, $\mathbf{Z}^-_{--}(s) = \mathbf{Q}_{--}(0) - s\mathbf{I}$, $\mathbf{Z}^-_{+-}(s) = \mathbf{Q}_{+-}(0)$, and $\mathbf{Z}^-_{-+}(s) = \mathbf{Q}_{-+}(0)$.

By Theorem 2, for any $i, j \in \mathcal{S}$,

$$[e^{\mathbf{Z}^-(s)y}]_{ij} = E(e^{-sh_-(\omega(y))} I(\varphi(\omega(y)) = j) \mid \varphi(0) = i) \quad (47)$$

is the LST of the distribution of the total downward shift in $X(\cdot)$ accumulated by the time the in-out fluid of the process $X(\cdot)$ first reaches level y and does so in phase j, given that the process starts in phase i at time zero.

Note that, in a stochastic fluid model $\{(\varphi(t), X(t)), t \geq 0\}$ with unbounded level variable $X(t) \in (-\infty, +\infty)$, we have $h(t) = h_+(t) + h_-(t)$ and $X(t) = X(0) + h_+(t) - h_-(t)$.

4. GENERATORS DERIVED FROM $\mathbf{Z}(\underline{\mathbf{s}})$

In this section we discuss useful applications of generators that are expressed in terms of $\mathbf{Z}(\underline{\mathbf{s}})$.

First, consider the following generalisation of matrices $\mathbf{\Psi}(s)$, $\mathbf{\Xi}(s)$, $\mathbf{G}^{(x,y)}(s)$, and $\mathbf{H}^{(x,y)}(s)$ discussed in [7, 11], denoted $\mathbf{\Psi}(\underline{\mathbf{s}})$, $\mathbf{\Xi}(\underline{\mathbf{s}})$, $\mathbf{G}^{(x,y)}(\underline{\mathbf{s}})$, and $\mathbf{H}^{(x,y)}(\underline{\mathbf{s}})$, respectively. Let $\theta(x) = \inf\{t > 0 : X(t) = x\}$ be the first time the process hits level x.

Matrix $\mathbf{\Psi}(\underline{\mathbf{s}}) = [\Psi(\underline{\mathbf{s}})_{ij}]_{i \in \mathcal{S}_+, j \in \mathcal{S}_-}$ is such that, for all $i \in \mathcal{S}_+$ and $j \in \mathcal{S}_-$,

$$\begin{aligned} \Psi(\underline{\mathbf{s}})_{ij} &= E(e^{-(s_1 W_1(0,\theta(0)) + \ldots + s_n W_n(0,\theta(0)))} I(\varphi(\theta(0)) = j) \\ &\quad \mid \varphi(0) = i, X(0) = 0) \qquad (48) \end{aligned}$$

is the LST of the distribution of $(W_1(0, \theta(0)), \ldots, W_n(0, \theta(0)))$ and $\varphi(\theta(0)) = j$, given $\varphi(0) = i$ and $X(0) = 0$.

Matrix $\mathbf{\Xi}(\underline{\mathbf{s}}) = [\Xi(\underline{\mathbf{s}})_{ij}]_{i \in \mathcal{S}_-, j \in \mathcal{S}_+}$ is symmetrical to $\mathbf{\Psi}(\underline{\mathbf{s}})$ for an unbounded fluid, in which $c_{\varphi(t)} = dX(t)/dt$ always, so that $X(t) \in (-\infty, +\infty)$. That is, for all $i \in \mathcal{S}_-$ and $j \in \mathcal{S}_+$, $\Xi(\underline{\mathbf{s}})_{ij}$ is defined by the right-hand side of (48).

Note that, with $\underline{\mathbf{0}}$ denoting a vector of zeros of appropriate size, matrices $\hat{\mathbf{\Psi}}(\underline{\mathbf{0}})$ and $\hat{\mathbf{\Xi}}(\underline{\mathbf{0}})$ are equivalent to matrices $\mathbf{\Psi}$ and $\mathbf{\Xi}$, respectively, defined in [11].

Matrix $\mathbf{G}^{(x,y)}(\underline{\mathbf{s}}) = [G^{(x,y)}(\underline{\mathbf{s}})_{ij}]_{i,j \in \mathcal{S}_+ \cup \mathcal{S}_-}$ is such that, for all $i, j \in \mathcal{S}_+ \cup \mathcal{S}_-$ and $0 < x < y$,

$$\begin{aligned} G^{(x,y)}(\underline{\mathbf{s}})_{ij} &= E(e^{-(s_1 W_1(0,\theta(0)) + \ldots + s_n W_n(0,\theta(0)))} \\ &\quad \times I(\varphi(\theta(0)) = j, \theta(0) < \theta(y)) \\ &\quad \mid \varphi(0) = i, X(0) = x) \qquad (49) \end{aligned}$$

is the LST of the distribution of $(W_1(0, \theta(0)), \ldots, W_n(0, \theta(0)))$ and $\varphi(\theta(0)) = j$ under the taboo $\theta(0) < \theta(y)$, given $\varphi(0) = i$ and $X(0) = x$. Denote $\mathbf{G}^{(0,y)}(\underline{\mathbf{s}}) = \lim_{x \to 0^+} \mathbf{G}^{(x,y)}(\underline{\mathbf{s}})$.

Conversely, matrix $\mathbf{H}^{(x,y)}(\underline{\mathbf{s}}) = [H^{(x,y)}(\underline{\mathbf{s}})_{ij}]_{i,j \in \mathcal{S}_+ \cup \mathcal{S}_-}$ is such that, for all $i, j \in \mathcal{S}_+ \cup \mathcal{S}_-$ and $0 < x < y$,

$$\begin{aligned} H^{(x,y)}(\underline{\mathbf{s}})_{ij} &= E(e^{-(s_1 W_1(0,\theta(y)) + \ldots + s_n W_n(0,\theta(y)))} \\ &\quad \times I(\varphi(\theta(y)) = j, \theta(y) < \theta(0)) \\ &\quad \mid \varphi(0) = i, X(0) = x) \qquad (50) \end{aligned}$$

is the LST of the distribution of $(W_1(0, \theta(y)), \ldots, W_n(0, \theta(y)))$ and $\varphi(\theta(y)) = j$ under the taboo $\theta(y) < \theta(0)$, given $\varphi(0) = i$ and $X(0) = x$. Denote $\mathbf{H}^{(x,x)}(\underline{\mathbf{s}}) = \lim_{y \to x^+} \mathbf{H}^{(x,y)}(\underline{\mathbf{s}})$.

We can show by using techniques similar to [11] that when $s_i \geq 0$ for all $i \in \mathcal{S}$, $\mathbf{\Psi}(\underline{\mathbf{s}})$ is the minimum nonnegative solution of the Riccati equation

$$\mathbf{Z}_{+-}(\underline{\mathbf{s}}) + \hat{\mathbf{\Psi}}(\underline{\mathbf{s}}) \mathbf{Z}_{-+}(\underline{\mathbf{s}}) \hat{\mathbf{\Psi}}(\underline{\mathbf{s}}) + \mathbf{Z}_{++}(\underline{\mathbf{s}}) \hat{\mathbf{\Psi}}(\underline{\mathbf{s}}) + \hat{\mathbf{\Psi}}(\underline{\mathbf{s}}) \mathbf{Z}_{--}(\underline{\mathbf{s}}) = 0,$$

with similar results for $\mathbf{\Xi}(\underline{\mathbf{s}})$, and algorithms in [12] can be used for finding these solutions. Further, we can derive expressions for $\mathbf{G}^{(x,y)}(\underline{\mathbf{s}})$ and $\mathbf{H}^{(x,y)}(\underline{\mathbf{s}})$ using methodology similar to [7].

Now, we introduce generators $\mathbf{J}_1(\underline{\mathbf{s}})$ and $\mathbf{J}_2(\underline{\mathbf{s}})$, which are generalisations of similar quantities in [10, 11, 12],

$$\begin{aligned} \mathbf{J}_1(\underline{\mathbf{s}}) &= \mathbf{Z}_{--}(\underline{\mathbf{s}}) + \mathbf{Z}_{-+}(\underline{\mathbf{s}}) \mathbf{\Psi}(\underline{\mathbf{s}}), \qquad (51) \\ \mathbf{J}_2(\underline{\mathbf{s}}) &= \mathbf{Z}_{++}(\underline{\mathbf{s}}) + \mathbf{Z}_{+-}(\underline{\mathbf{s}}) \mathbf{\Xi}(\underline{\mathbf{s}}). \qquad (52) \end{aligned}$$

These are useful in constructing algorithms for the numerical evaluation of $\mathbf{\Psi}(\underline{\mathbf{s}})$ and $\mathbf{\Xi}(\underline{\mathbf{s}})$, as shown in [12].

The physical interpretation of $\mathbf{J}_1(\underline{\mathbf{s}})$ is that, for $i, j \in \mathcal{S}_-$,

$$\begin{aligned} [e^{\mathbf{J}_1(\underline{\mathbf{s}})y}]_{ij} &= E(e^{-(s_1 W_1(0,\theta(0)) + \ldots + s_n W_n(0,\theta(0)))} I(\varphi(\theta(0)) = j) \\ &\quad \mid \varphi(0) = i, X(0) = y) \qquad (53) \end{aligned}$$

is the LST of the distribution of $(W_1(0, \theta(0)), \ldots, W_n(0, \theta(0)))$ and $\varphi(\theta(0)) = j$, given $\varphi(0) = i$ and $X(0) = y$.

The physical interpretation of $\mathbf{J}_2(\underline{s})$ follows by symmetry for the unbounded fluid $X(t) \in (-\infty, +\infty)$.

Further, we define generators $\mathbf{J}_3(\underline{s})$ and $\mathbf{J}_4(\underline{s})$, which are generalisations of $\mathbf{K}(s)$ in [18],

$$\begin{aligned} \mathbf{J}_3(\underline{s}) &= \mathbf{Z}_{++}(\underline{s}) + \mathbf{\Psi}(\underline{s})\mathbf{Z}_{-+}(\underline{s}), && (54)\\ \mathbf{J}_4(\underline{s}) &= \mathbf{Z}_{--}(\underline{s}) + \mathbf{\Xi}(\underline{s})\mathbf{Z}_{+-}(\underline{s}). && (55) \end{aligned}$$

The physical interpretation of $\mathbf{J}_3(\underline{s})$ is that for $i, j \in \mathcal{S}_+$,

$$\begin{aligned} [e^{\mathbf{J}_3(\underline{s})y}]_{ij} &= \int_{w_1=0}^{\infty} \cdots \int_{w_n=0}^{\infty} e^{-(s_1w_1+\ldots+s_nw_n)} \\ &\qquad \gamma_i(y, j; w_1, \ldots, w_n)dw_1 \cdots dw_n \qquad (56) \end{aligned}$$

is the *Laplace transform* of the density $\gamma_i(y, j; w_1, \ldots, w_n)$ with respect to the rewards, that the process crosses level y in phase j, while avoiding level zero, when the rewards are $(w_1, \ldots, w_n)$, and given $\varphi(0) = i$.

The physical interpretation of $\mathbf{J}_4(\underline{s})$ follows by symmetry for the unbounded fluid $X \in (-\infty, +\infty)$.

We also define generators $\mathbf{J}_5(\underline{s})$ and $\mathbf{J}_6(\underline{s})$, which are generalisations of $\left(\mathbf{Q}_{++} + \mathbf{Q}_{+-}\mathbf{H}^{(b,b)}(0)\right)$ used in [14],

$$\begin{aligned} \mathbf{J}_5(\underline{s}) &= \mathbf{Z}_{++}(\underline{s}) + \mathbf{Z}_{+-}(\underline{s})\mathbf{H}^{(b,b)}(\underline{0}), && (57)\\ \mathbf{J}_6(\underline{s}) &= \mathbf{Z}_{--}(\underline{s}) + \mathbf{Z}_{-+}(\underline{s})\mathbf{G}^{(0,b)}(\underline{0}). && (58) \end{aligned}$$

To establish the physical interpretation of $\mathbf{J}_5(\underline{s})$, define, for a doubly-bounded fluid $X(t) \in [0, b]$,

$$\begin{aligned} \widehat{h}(t) &= \int_{u=0}^{t} |c_{\varphi(u)}| I(X(u) = b)du, && (59)\\ \widehat{W}_i(z, t) &= \int_{u=z}^{t} r_i I(\varphi(u) = i, X(u) = b)du, && (60) \end{aligned}$$

interpreted as *censored* in-out fluid and i-type rewards, respectively, accumulated only during periods of congestion when the buffer is full.

Let $\delta(y) = \inf\{t > 0 : \widehat{h}(t) = y\}$ be the first time the censored in-out fluid reaches level y. Then, for all $i, j \in \mathcal{S}_+$,

$$\begin{aligned} [e^{\mathbf{J}_5(\underline{s})y}]_{ij} &= E(e^{-(s_1\widehat{W}_1(0,\delta(y))+\ldots+s_n\widehat{W}_n(0,\delta(y)))} \\ &\qquad \times I(\varphi(\delta(y)) = j, \delta(y) < \theta(0)) \\ &\qquad \mid \varphi(0) = i, X(0) = b) \qquad (61) \end{aligned}$$

is the LST of the distribution of $(\widehat{W}_1(0, \delta(y)), \ldots, \widehat{W}_n(0, \delta(y)))$ and $\varphi(\delta(y)) = j$ under the taboo $\delta(y) < \theta(0)$, given $\varphi(0) = i$ and $X(0) = b$.

The physical interpretation of $\mathbf{J}_6(\underline{s})$ follows by symmetry, for the rewards earned only when the buffer is empty.

5. NEW RICCATI EQUATION FOR $\mathbf{\Psi}$

In this section, we derive a new Riccati equation for $\mathbf{\Psi}$ and an explicit expression for $\mathbf{\Psi}$ in terms of appropriately defined matrix $\mathbf{M}$. We are currently investigating whether this new matrix $\mathbf{M}$ can be computed efficiently.

Consider a stochastic fluid model $\{(\varphi(t), X(t)), t \geq 0\}$ with unbounded level variable $X(t) \in (-\infty, +\infty)$. Throughout this section, *we assume that the process is transient.* (Note that the only other alternative is null-recurrence).

For $0 \leq x \leq y$, define $\mathbf{f}_y(x) = [f_y(x)_{ij}]_{i,j\in\mathcal{S}_+\cup\mathcal{S}_-}$ such that, for all $i, j \in \mathcal{S}_+ \cup \mathcal{S}_-$, $f_y(x)_{ij}$ is the inverse of the LST $[e^{\mathbf{Z}^+(s)y}]_{ij}$ so that

$$\begin{aligned} f_y(x)_{ij} &= \frac{d}{dx} P(h_+(\omega(y)) \leq x, \varphi(\omega(y)) = j \\ &\qquad \mid X(0) = 0, \varphi(0) = i), \qquad (62) \end{aligned}$$

is the probability density that the total upward shift in $X(\cdot)$ at the time $\omega(y)$ is $h_+(\omega(y)) = x$ and the phase is $\varphi(\omega(y)) = j$, given that the process starts in phase i at time zero. It follows that $\sum_j \int_{x=0}^{y} f_y(x)_{ij} dx = 1$.

Note that by using the method of Abate and Whitt [1], we can obtain $\mathbf{f}_y(x)$ by numerically inverting $e^{\mathbf{Z}^+(s)y}$.

Theorem 3. *Let $\mathbf{M} = [M_{ij}]$ be a matrix defined by*

$$M_{ij} = \int_{y=0}^{\infty} f_y(y/2)_{ij} dy, \qquad (63)$$

and partitioned according to $\mathcal{S}_+ \cup \mathcal{S}_-$ as

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{++} & \mathbf{M}_{+-} \\ \mathbf{M}_{-+} & \mathbf{M}_{--} \end{bmatrix}. \qquad (64)$$

Then $\mathbf{M}$ has the form

$$\mathbf{M} = \begin{bmatrix} \mathbf{\Psi}\mathbf{M}_{-+} & (\mathbf{I} - \mathbf{\Psi}\mathbf{\Xi})^{-1}\mathbf{\Psi} \\ \mathbf{\Xi}(\mathbf{I} - \mathbf{\Psi}\mathbf{\Xi})^{-1} & \mathbf{\Xi}\mathbf{M}_{+-} \end{bmatrix}. \qquad (65)$$

Proof: First, note that by (63), the quantity M_{ij} is the expected number of visits to state $(j, 0)$ given that the process starts in state $(i, 0)$, for all $i, j \in \mathcal{S}_+ \cup \mathcal{S}_-$. This follows from the facts that:

(i). $f_y(y/2)_{ij}$ is the density that given the process started from level 0 in phase i, it will be on level 0 in phase j when the total in-out fluid is y (in which case the total upward shift is $y/2$); and

(ii). integrating the density $f_y(y/2)_{ij}$ gives the expected number of visits to $(j, 0)$ given start in $(i, 0)$, by standard results in probability theory.

Consider $\mathbf{M}_{+-} = [M_{ij}]_{i\in\mathcal{S}_+, j\in\mathcal{S}_-}$, the analysis for the remaining block matrices is analogous.

Assume $i \in \mathcal{S}_+$, $j \in \mathcal{S}_-$ and $\varphi(0) = i$, $X(0) = 0$. Denote $\theta_0 = \theta(0)$, and let $\theta_n = \inf\{t > \theta_{n-1} : X(t) = 0, \varphi(t) \in \mathcal{S}_-\}$ for $n = 1, 2, \ldots$. We interpret θ_n as *the time of the n^{th} crossing of level zero from above*, for all $n = 0, 1, 2 \ldots$.

Define matrix $\mathbf{P}(n) = [P(n)_{ij}]_{i\in\mathcal{S}_+, j\in\mathcal{S}_-}$ such that, for $i \in \mathcal{S}_+$, $j \in \mathcal{S}_-$,

$$P(n)_{ij} = P(\theta_n < \infty, \varphi(\theta_n) = j \mid \varphi(0) = i, X(0) = 0) \quad (66)$$

is the probability that the n^{th} crossing of level zero from above occurs in finite time and that the process is in phase j at the time of the n^{th} crossing of level zero from above, given that the process starts with $\varphi(0) = i$, $X(0) = 0$.

By the standard theory of discrete-time Markov chains,

$$[\mathbf{M}_{+-}]_{ij} = \sum_{n=0}^{\infty} [\mathbf{P}(n)]_{ij}. \qquad (67)$$

Now, for $n = 0, 1, 2, \ldots$,

$$[\mathbf{P}(n)]_{ij} = [(\mathbf{\Psi}\mathbf{\Xi})^n \mathbf{\Psi}]_{ij}, \qquad (68)$$

and so

$$\begin{aligned} [\mathbf{M}_{+-}]_{ij} &= \sum_{n=0}^{\infty} [(\mathbf{\Psi}\mathbf{\Xi})^n \mathbf{\Psi}]_{ij} \\ &= [(\mathbf{I} - \mathbf{\Psi}\mathbf{\Xi})^{-1}\mathbf{\Psi}]_{ij}, \qquad (69) \end{aligned}$$

where the inverse $(\mathbf{I}-\mathbf{\Psi\Xi})^{-1}$ exists since the process is transient. Therefore,

$$\mathbf{M}_{+-} = (\mathbf{I}-\mathbf{\Psi\Xi})^{-1}\mathbf{\Psi}. \tag{70}$$

The other block matrices in $\mathbf{M}$ can be constructed in an analogous manner. ■

Below, we state new results for $\mathbf{\Psi}$.

Corollary 1. $\mathbf{\Psi}$ *is a solution to the Riccati equation*

$$\mathbf{M}_{+-} = \mathbf{\Psi} + \mathbf{\Psi}\mathbf{M}_{-+}\mathbf{\Psi}. \tag{71}$$

Proof: By (65),

$$\mathbf{\Xi}\mathbf{M}_{+-} \;=\; \mathbf{M}_{-+}\mathbf{\Psi}.$$

Since the process is transient, we can write

$$\begin{aligned}\mathbf{\Psi} &= (\mathbf{I}-\mathbf{\Psi\Xi})(\mathbf{I}-\mathbf{\Psi\Xi})^{-1}\mathbf{\Psi}\\ &= (\mathbf{I}-\mathbf{\Psi\Xi})\mathbf{M}_{+-}\\ &= \mathbf{M}_{+-}-\mathbf{\Psi}\mathbf{M}_{-+}\mathbf{\Psi},\end{aligned}$$

and so the result follows. ■

Corollary 2. $\mathbf{\Psi}$ *can be explicitly written as*

$$\mathbf{\Psi} = \mathbf{M}_{+-}(\mathbf{I}+\mathbf{M}_{--})^{-1}. \tag{72}$$

Proof: By (65),

$$\begin{aligned}\mathbf{M}_{+-} &= (\mathbf{I}-\mathbf{\Psi\Xi})^{-1}\mathbf{\Psi},\\ \Rightarrow\; \mathbf{M}_{+-}-\mathbf{\Psi}\mathbf{M}_{--} &= \mathbf{\Psi}, \quad \text{since } \mathbf{M}_{--}=\mathbf{\Xi}\mathbf{M}_{+-},\\ \Rightarrow\; \mathbf{\Psi} &= \mathbf{M}_{+-}(\mathbf{I}+\mathbf{M}_{--})^{-1}.\end{aligned}$$

To justify the existence of $(\mathbf{I}+\mathbf{M}_{--})^{-1}$, note that by (65),

$$\mathbf{M}_{--} = \mathbf{\Xi}(\mathbf{I}-\mathbf{\Psi\Xi})^{-1}\mathbf{\Psi} = \sum_{n=1}^{\infty}(\mathbf{\Psi\Xi})^{n},$$

and

$$\mathbf{I}+\mathbf{M}_{--} = \sum_{n=0}^{\infty}(\mathbf{\Psi\Xi})^{n} = (\mathbf{I}-\mathbf{\Psi\Xi})^{-1},$$

for all transient processes. ■

6. CONCLUSION

We have constructed a generalised reward generator $\mathbf{Z}(\underline{s})$ for the stochastic fluid model useful for tracking the accumulation of reward for different phases individually.

We have considered various projections of $\mathbf{Z}(\underline{s})$, including the fluid generators $\mathbf{Q}(s)$ [11] and $\mathbf{W}(s)$ [8].

We constructed the generator $\mathbf{Z}^{+}(s)$ which tracks the upward shift in the fluid. We applied $\mathbf{Z}^{+}(s)$ to construct the matrix $\mathbf{M}$ which records the expected number of visits to the original level before the unbounded fluid drifts off to ±∞.

We used the elements of $\mathbf{M}$ and its physical interpretation to derive a new Riccati equation and an explicit solution for the matrix $\mathbf{\Psi}$, which is a key building block to many other performance measures. Work on the algorithmic techniques resulting from this equation is in progress.

7. REFERENCES

[1] J. Abate and W. Whitt. Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing*, 7(1):36–36, 1995.

[2] S. Ahn and V. Ramaswami. Fluid flow models and queues — a connection by stochastic coupling. *Stochastic Models*, 19(3):325–348, 2003.

[3] S. Ahn and V. Ramaswami. Transient analysis of fluid flow models via stochastic coupling to a queue. *Stochastic Models*, 20(1):71–101, 2004.

[4] S. Ahn and V. Ramaswami. Efficient algorithms for transient analysis of stochastic fluid flow models. *Journal of Applied Probability*, 42(2):531–549, 2005.

[5] S. Asmussen. Stationary distributions for fluid flow models with or without Brownian noise. *Stochastic Models*, 11(1):21–49, 1995.

[6] A. Badescu, L. Breuer, A. da Silva Soares, G. Latouche, M. A. Remiche, and D. Stanford. Risk processes analyzed as fluid queues. *Scandinavian Actuarial Journal*, 2005(2):127–141, 2005.

[7] N. Bean, M. O'Reilly, and P. Taylor. Hitting probabilities and hitting times for stochastic fluid flows: The bounded model. *Probability in the Engineering and Informational Sciences*, 23(01):121–147, 2009.

[8] N. G. Bean and M. M. O'Reilly. Stochastic two-dimensional fluid model. *Stochastic Models*, 29(1):31–63, 2013.

[9] N. G. Bean, M. M. O'Reilly, and J. E. Sargison. A stochastic fluid flow model of the operation and maintenance of power generation systems. *IEEE Transactions on Power Systems*, 25(3):1361–1374, 2010.

[10] N. G. Bean, M. M. O'Reilly, and P. G. Taylor. Algorithms for return probabilities for stochastic fluid flows. *Stochastic Models*, 21(1):149–184, 2005.

[11] N. G. Bean, M. M. O'Reilly, and P. G. Taylor. Hitting probabilities and hitting times for stochastic fluid flows. *Stochastic processes and their applications*, 115(9):1530–1556, 2005.

[12] N. G. Bean, M. M. O'Reilly, and P. G. Taylor. Algorithms for the Laplace–Stieltjes transforms of first return times for stochastic fluid flows. *Methodology and Computing in Applied Probability*, 10(3):381–408, 2008.

[13] A. da Silva Soares. *Fluid Queues: Building Upon the Analogy with QBD Processes.* PhD thesis, Université Libre de Bruxelles, 2005.

[14] F. Guillemin and B. Sericola. Volume and duration of losses in finite buffer fluid queues. *Journal of Applied Probability*, 52(3):826–840, 2015.

[15] M. M. O'Reilly. Multi-stage stochastic fluid models for congestion control. *European Journal of Operations Research*, 238:514–526, 2014.

[16] M. M. O'Reilly and Z. Palmowski. Loss rates for stochastic fluid models. *Performance Evaluation*, 70(9):593–606, 2013.

[17] V. Ramaswami. Matrix analytic methods for stochastic fluid flows. In *ITC16: International Teletraffic Congress*, pages 1019–1030, 1999.

[18] V. Ramaswami. Passage times in fluid models with application to risk processes. *Methodology and Computing in Applied Probability*, 8(4):497–515, 2006.

Fluid flows with jumps at the boundary

[Extended Abstract]

Eleonora Deiana
Faculté d'Informatique
Université de Namur
eleonora.deiana@unamur.be

Guy Latouche
Université libre de Bruxelles
latouche@ulb.ac.be

Marie-Ange Remiche
Faculté d'Informatique
Université de Namur
marie-ange.remiche@unamur.be

ABSTRACT

We consider a Markov-modulated fluid flow with infinite buffer. The lower bound zero is a reactive bound: every time the process hits this boundary, it makes an instantaneous jump to a fixed level b. We use the regenerative approach to calculate the stationary distribution of this model.

CCS Concepts

•**Mathematics of computing** → **Markov processes;**

Keywords

Fluid flow; regenerative process; reactive boundaries

1. INTRODUCTION

A fluid flow model represents the evolution of the fluid content in a buffer, where the level is regulated by a continuous-time Markov process. This is a well-known model, which has been studied in the past by mean of different techniques. In the classical infinite fluid flow, the evolution of the buffer content is level-independent, and the unique bound in level 0 is generally an absorbing bound, which means that every time the buffer is empty, it remains empty until new fluid starts to enter in the buffer. We want to introduce here a reaction of the system every time the bound is reached. In a model with a reactive bound, the evolution of the buffer content does not depend anymore on the Markov process only, but also on the effective level reached by the buffer content. In particular we consider a fluid flow which makes a jump every time the buffer becomes empty.

2. THE MODEL

We define an infinite fluid flow model as a two-dimensional stochastic process $\{X(t), \phi(t)\}_{t\geq 0}$. The first dimension $X(t)$ represents the level of the fluid in the buffer at time t; as we are dealing with an infinite buffer, $X(t)$ takes values in $\mathbb{R}^+$. The second dimension is called the phase process, it is an irreducible continuous-time Markov process taking values in a finite set S, and it controls the evolution of the level. When $X(t) > 0$, the level evolves in the following way:

$$\frac{dX(t)}{dt} = c_i \quad \text{if} \quad \phi(t) = i \text{ at time } t.$$

As soon as the buffer becomes empty, we have a reaction of the system: the level $X(t)$ makes a jumps and goes instantaneously to a fixed level b. At the same time, there may be an instantaneous change of phase.

This fluid flow with a reactive bound may be used to model the buffer content of a particular type of servers. Jobs are accumulated in the buffer, while waiting to be processed by the server. Once the buffer is empty, a fixed load of b services from neighboring servers is instantaneously put into it. This policy prevents the server from being inactive.

Depending on the sign of the rates c_i, we can partition the state space S into $S^+ \cup S^-$, where $S^+ = \{i \in S : c_i > 0\}$, and $S^- = \{i \in S : c_i < 0\}$.

In order to study this fluid flow, we need to know three matrices describing its evolution. When $X(t) > 0$, the phase process $\phi(t)$ evolves following the generator T. When $X(t)$ hits the lower boundary 0, the phase instantaneously changes accordingly to the probability transition matrix W. Both matrices can be partitioned following the partition of the state space S:

$$T = \begin{bmatrix} T_{++} & T_{+-} \\ T_{-+} & T_{--} \end{bmatrix}, \text{ and } W = \begin{bmatrix} W_{-+} & W_{--} \end{bmatrix}.$$

Finally, we have the rate matrix C, which collects all the rates c_i in the diagonal. It can also be partitioned in the following way:

$$C = \begin{bmatrix} C_+ & 0 \\ 0 & C_- \end{bmatrix}.$$

We suppose that the initial level $X(0)$ is 0, and the initial phase $\phi(0)$ has stationary distribution $\boldsymbol{\alpha}$ (which means that $\boldsymbol{\alpha}T = \boldsymbol{\alpha}$ and $\boldsymbol{\alpha}\mathbf{1} = 1$). We make the assumption that the process has a negative mean drift, that is, $\boldsymbol{\alpha}C\mathbf{1} < 0$.

The aim when studying this system is to calculate the joint stationary distribution, defined as follows:

$$\Pi_j(x) = \lim_{t\to\infty} P\left[X(t) < x, \phi(t) = j\right].$$

In order to do it, we use a Markov-regenerative approach.

2.1 Markov-regenerative approach

We define a sequence of random times $\{h_n\}_{n\geq 0}$ as the epochs when the buffer becomes empty:

$$\begin{aligned} h_0 &= 0, \\ h_{n+1} &= \inf \{t > h_n | X(t) = 0\}. \end{aligned}$$

This is a sequence of regeneration points. Let us define in the following way the phases in the regeneration points and just after the jump:

$$\phi_n = \lim_{t \to h_n^-} \phi(t) \quad \text{and} \quad \phi_n^* = \lim_{t \to h_n^+} \phi(t).$$

Where $\phi_n \in S^-$, as the only way to hit the level 0 is with a negative phase, and $\phi_n^* \in S^+ \cup S^-$, as the process can leave level b either in a negative or in a positive phase.

By Çinlar [?], we know that the stationary distribution $\mathbf{\Pi}(x)$ in a regenerative process, can be expressed with the following formula:

$$\mathbf{\Pi}(x) = (\boldsymbol{\nu} \boldsymbol{m})^{-1} \boldsymbol{\nu} M(x). \tag{1}$$

The vector $\boldsymbol{\nu}$ is the stationary distribution of the phases at the regeneration times, and it is defined as:

$$\begin{aligned} &\boldsymbol{\nu} H = \boldsymbol{\nu}, \quad \text{where} \\ &H_{ij} = P[\phi(h_{n+1}) = j | \phi(h_n) = i], \quad i, j \in S^-. \end{aligned}$$

The vector $\boldsymbol{m}$ is the mean sojourn time between two regeneration points, given the phase at the beginning of the interval. Finally, $M(x)$ records the mean sojourn time in $([0, x], j)$ during the renewal interval.

2.2 Stationary distribution

Looking at the physical properties of the model, we can calculate all the quantities that we need. First of all we calculate the transition matrix of the phases between two regeneration points:

$$H = \left[\begin{array}{cc} W_{-+} & W_{--} \end{array} \right] \left[\begin{array}{c} \Psi \\ I \end{array} \right] \mathrm{e}^{Ub}.$$

Matrices Ψ and U are well known in the theory of fluid flows: the matrix Ψ gives the probability of the first return to the initial level, and the matrix U is the generator of a Markov process in the negative state space S^-, defined by

$$U = T_{--} + T_{-+} \Psi.$$

Its exponential e^{Ux} gives the probabilities, starting from a fixed level x, that the process reaches level 0 in a finite time. The details may be found in [?] or [?]. By solving the system $\boldsymbol{\nu} H = \boldsymbol{\nu}$ we can obtain the vector $\boldsymbol{\nu}$.

In order to calculate $M(x)$, the mean sojourn time in $([0, x], j)$, between two regeneration points, we can write

$$M(x) = \left[\begin{array}{cc} W_{-+} & W_{--} \end{array} \right] \widetilde{M}(x).$$

The component $\left[\widetilde{M}(x)\right]_{ij}$ is the mean sojourn time in $[0, x]$ in a phase j, given that the process starts from level b in a phase $i \in S^+ \cup S^-$, before the first return to level 0. Note that the passage from level 0 to level b is instantaneous, so there is no probability mass in 0. The matrix $\widetilde{M}(x)$ can be partitioned, depending on the initial phase, in two parts:

$$\widetilde{M}(x) = \left[\begin{array}{c} \widetilde{M}_+(x) \\ \widetilde{M}_-(x) \end{array} \right].$$

In order to calculate $\widetilde{M}(x)$, we first need to define the quantity $H_-^b(x)$. This is the mean sojourn time in $(0, x)$ starting from level b in a negative phase, before the first return either to the initial level b or to level 0. We calculate it by following the idea in [?]: we define two continuous-time processes, a lower-bounded $\{L(t), \phi(t)\}_{t \geq 0}$, with $L(t) \in \mathbb{R}^+$, and an upper-bounded $\{U(t), \phi(t)\}_{t \geq 0}$, with $U(t) \in (-\infty, b]$. Both these processes have the same transition matrix T and rate matrix C than our fluid flow, but no reaction in the boundaries. We introduce the two quantities $\Gamma(x)$ and $\widehat{\Gamma}(x)$, related to these processes: the matrix $\Gamma(x)$ is the mean sojourn time in $(0, x)$, starting from level 0, until the first return to the initial level, for the lower-bounded process; $\widehat{\Gamma}(x)$ is the mean sojourn time in $(0, x)$, starting from level b, until the first return to the initial level, for the upper-bounded process. These two quantities are calculated by taking the following integrals:

$$\begin{aligned} \Gamma(x) &= \int_0^x \mathrm{e}^{Ku} \mathrm{d}u \left[\begin{array}{cc} C_+^{-1} & \Psi |C_-^{-1}| \end{array} \right], \quad \text{and} \\ \widehat{\Gamma}(x) &= \int_{b-x}^b \mathrm{e}^{\widehat{K}u} \mathrm{d}u \left[\begin{array}{cc} \widehat{\Psi} C_+^{-1} & |C_-^{-1}| \end{array} \right]. \end{aligned}$$

The matrix $\hat{\Psi}$ is the matrix of first return probabilities to the initial level, starting from level b. The matrix K is defined as

$$K = C_+^{-1} T_{++} + \Psi |C_-|^{-1} T_{-+},$$

and its exponential e^{Kx} gives the expected number of crossing of level x, starting from level 0 before the first return to 0. Similarly, the matrix $\widehat{K}$ is defined as

$$\widehat{K} = |C_-|^{-1} T_{--} + \widehat{\Psi} C_+^{-1} T_{+-},$$

and its exponential $e^{\widehat{K}x}$ gives the expected number of crossing of level x, starting from level b before the first return to b. As the process has strictly negative mean drift, then all the eigenvalues of the matrix K have a strictly negative real part and solving the first integral is straightforward:

$$\int_0^x \mathrm{e}^{Ku} \mathrm{d}u = (-K)^{-1} \left(I - \mathrm{e}^{Kx} \right).$$

For the matrix $\widehat{K}$, things are different since $\widehat{K}$ has one eigenvalue equal to 0, so the integral becomes

$$\int_{b-x}^b \mathrm{e}^{\widehat{K}u} \mathrm{d}u = \left(-\widehat{K}^{\#} \left(\mathrm{e}^{\widehat{K}(b-x)} - \mathrm{e}^{\widehat{K}b} \right) + x \boldsymbol{v} \boldsymbol{u} \right),$$

where $\widehat{K}^{\#}$ is the group inverse of $\widehat{K}$, $\boldsymbol{v}$ and $\boldsymbol{u}$ are respectively the right and left eigenvectors of $\widehat{K}$ for the eigenvalue 0.

The quantities $\Gamma(x)$ and $\widehat{\Gamma}(x)$ are also given by the system:

$$\left[\begin{array}{c} \Gamma(x) \\ \widehat{\Gamma}(x) \end{array} \right] = \left[\begin{array}{cc} I & \mathrm{e}^{Kb} \Psi \\ \mathrm{e}^{\widehat{K}b} \widehat{\Psi} & I \end{array} \right] \left[\begin{array}{c} H_+^b(x) \\ H_-^b(x) \end{array} \right],$$

where $H_+^b(x)$, similar to $H_-^b(x)$ is the mean sojourn time in $(0, x)$ starting from level 0 in a positive phase, before the first return either to the initial level b or to level 0.

Solving the system, we obtain the quantity we are interested in:

$$H_-^b(x) = (I - e^{\widehat{K}b} \widehat{\Psi} e^{Kb} \Psi)^{-1} \left(\widehat{\Gamma}(x) - e^{\widehat{K}b} \widehat{\Psi} \Gamma(x) \right).$$

We need to separate two cases: when $0 < x < b$, and $x \geq b$. In the first case, when $0 < x < b$, we have the following system:

$$\begin{cases} \widetilde{M}_{+}(x) = \Psi \widetilde{M}_{-}(x) \\ \widetilde{M}_{-}(x) = H^{b}_{-}(x) + \widehat{\Psi}^{b} \widetilde{M}_{+}(x) \end{cases}$$

Where $\hat{\Psi}^b$ is the matrix of first return probabilities to the initial level, starting from level b, and without hitting the level 0. The first equation is straightforward, as we need the mean sojourn time in $(0, x)$, when $0 < x < b$, so the time spent above b doesn't have to be counted in. For the second equation, we first count the time spent starting from b before touching either level 0 or b, and if we touch level b before 0, then we have to consider again $\widetilde{M}_{+}(x)$. Solving the system, gives the following equations:

$$\begin{cases} \widetilde{M}_{+}(x) = \Psi(I - \widehat{\Psi}^{b}\Psi)^{-1} H^{b}_{-}(x) \\ \widetilde{M}_{-}(x) = (I - \widehat{\Psi}^{b}\Psi)^{-1} H^{b}_{-}(x). \end{cases}$$

In the second case, when $x \geq b$, $\widetilde{M}(x)$ is given by:

$$\begin{cases} \widetilde{M}_{+}(x) = \Gamma(x - b) + \Psi \widetilde{M}_{-}(x) \\ \widetilde{M}_{-}(x) = H^{b}_{-}(b) + \widehat{\Psi}^{b} \widetilde{M}_{+}(x). \end{cases}$$

For $\widetilde{M}_{-}(x)$ the equation is the same than before, for $\widetilde{M}_{+}(x)$ is similar, but we also have to count the time spent below level x, starting from level b, before coming back to level b, which is given by $\Gamma(b - x)$. The solution of the system is given by:

$$\begin{cases} \widetilde{M}_{+}(x) = (I - \Psi\widehat{\Psi}^{b})^{-1} \left(\Gamma(x - b) + \Psi H^{b}_{-}(b)\right) \\ \widetilde{M}_{-}(x) = (I - \widehat{\Psi}^{b}\Psi)^{-1} \left(\widehat{\Psi}^{b}\Gamma(x - b) + H^{b}_{-}(b)\right). \end{cases}$$

We have seen how to calculate the vector $\boldsymbol{\nu}$ and the matrix $M(x)$ for different values of x. The vector $\boldsymbol{m}$ being the mean sojourn time between two regenerative points, given the initial phase, can be calculated by taking the limit when x tends to infinite, and summing the columns:

$$\boldsymbol{m} = \left(\lim_{x \to \infty} M(x)\right) \mathbf{1}.$$

These vectors and matrix together in the formula (**??**), give us the stationary distribution we were looking for.

3. REFERENCES

[1] E. Çinlar. *Introduction to stochastic processes.* Englewood Cliffs, N.J. Prentice-Hall, 1975.

[2] A. da Silva Soares and G. Latouche. Matrix-analytic methods for fluid queues with finite buffers. *Performance Evaluation*, 63:295 – 314, 2006.

[3] G. Latouche and G. Nguyen. Feedback control: two-sided markov-modulated brownian motion with instantaneous change of phase at boundaries. 2016, to appear. arXiv:1603.01945v.

[4] V. Ramaswami. Matrix analytic methods for stochastic fluid flows. In D. Smith and P. Hey, editors, *Teletraffic Engineering in a Competitive World (Proceedings of the 16th International Teletraffic Congress)*, pages 1019–1030. Elsevier Science B.V., Edinburgh, UK, 1999.

A novel approach for the efficient waiting time and queue length analysis of Markov-modulated fluid priority queues

[Extended Abstract]

Gábor Horváth
Budapest Univ. of Technology and Economics
Department of Networked Systems and Services
ghorvath@hit.bme.hu

1. INTRODUCTION

Not many results exist for fluid queues with multi-type jobs. Among the existing results the fluid queues with priority service have been studied the most, due to their practical relevance in the performance evaluation of various telecommunication systems.

In the system considered in this paper the fluid input rates of the jobs are modulated by a Markov chain, and the service rate is constant. The same model has been studied in [10], where the partial differential equations for the joint distribution of the queue lengths are derived. However, these differential equations, especially the boundary functions are difficult to solve. [10] is able to provide the mean, the variance and the covariance of the queue lengths only in case of two-state Markov chains. In [3] the LST of the stationary joint distribution of the buffer contents is obtained in a closed form, from which the tail distributions and the queue length moments are derived. [9] follows a different approach to obtain the LST of the joint distribution based on the analysis of the idle and busy periods of the high priority queue.

The (matrix-analytic) approach presented in this paper is based on the workload process analysis, just like in [5] for the discrete case. While the main steps of the analysis are the same, adapting the method of [5] to continuous queues is not straight forward at all, since the specialties of continuous systems require different solutions in many steps of the procedure. Various performance measures are derived, including the Laplace-Stieltjes transforms (LST) and the moments of the stationary queue length distribution and the waiting time of the fluid drops. An Erlangization-based numerical method is also provided to approximate the queue length and the waiting time distributions up to an arbitrary precision.

The numerical behavior is the main focus throughout the paper. All performance measures are formulated as reward accumulation problems during busy periods of simple Markovian fluid flow models, for which matrix-analytic solutions are provided. The computation bottlenecks are the solutions of Riccati- and Sylvester-type equations, for which efficient implementations exist. As a result, the presented procedure can solve large models up to many hundreds of phases, while the past solutions mentioned above are less tractable.

2. MARKOVIAN FLUID FLOW MODELS

Both the queue length and the waiting time analysis are translated to the analysis of the amount of reward accumulated over the busy period of special Markovian fluid flow models, thus we start the paper by reviewing the related results and extend them where needed.

Formally, fluid models are two dimensional Markov processes $\{\mathcal{X}(t), \mathcal{J}(t), t > 0\}$, where $\mathcal{X}(t)$ is the fluid level and $\mathcal{J}(t)$ is the state of a CTMC (the background process) at time t. The Markov chain is assumed to be irreducible with state space $\mathcal{S}$, the number of states is $N = |\mathcal{S}|$. The generator matrix is denoted by $\mathbf{F} = [f_{ij},\ i, j \in \mathcal{S}]$. A fluid rate is associated to each state of the background process, $c_i, i \in \mathcal{S}$, that determines the rate at which the level of the fluid buffer changes. The evolution of the fluid level $\mathcal{X}(t)$ can be described as

$$\frac{d}{dt}\mathcal{X}(t) = \begin{cases} c_{\mathcal{J}(t)}, & \text{if } \mathcal{X}(t) > 0, \\ \max\{0, c_{\mathcal{J}(t)}\}, & \text{if } \mathcal{X}(t) = 0. \end{cases} \tag{1}$$

2.1 The distribution of the fluid level

The size N row vector $\pi(x)$ denotes the stationary density of the fluid level, whose ith entry is defined by $\pi_i(x) = \lim_{t\to\infty} \frac{d}{dx} P(\mathcal{X}(t) < x, \mathcal{J}(t) = i)$. At level 0 probability mass can accumulate as well. The ith entry of row vector p represents the probability mass at level 0, $p_i = P(\mathcal{X}(t) = 0, \mathcal{J}(t) = i)$.

Several numerical methods are available to obtain the stationary solution of the fluid level. For the analysis of fluid priority queues we follow the matrix-analytic solution approach ([7]). For the matrix-analytic method the state space of the Markov chain has to be partitioned into three sets, $\mathcal{S} = \mathcal{S}_+ \cup \mathcal{S}_- \cup \mathcal{S}_0$, according to the sign of the fluid rates, i.e. $\mathcal{S}_+ = \{i \in \mathcal{S}, c_i > 0\}$, $\mathcal{S}_- = \{i \in \mathcal{S}, c_i < 0\}$ and $\mathcal{S}_0 = \{i \in \mathcal{S}, c_i = 0\}$. From now on, it is assumed that matrices $\mathbf{F}$ and $\mathbf{C}$ are partitioned, thus

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{++} & \mathbf{F}_{+-} & \mathbf{F}_{+0} \\ \mathbf{F}_{-+} & \mathbf{F}_{--} & \mathbf{F}_{-0} \\ \mathbf{F}_{0+} & \mathbf{F}_{0-} & \mathbf{F}_{00} \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{C}_{+} & & \\ & \mathbf{C}_{-} & \\ & & \mathbf{0} \end{bmatrix}. \tag{2}$$

Furthermore, let us introduce matrix $\mathbf{F}^{\bullet}$ as the generator of

$\mathcal{J}(t)$ restricted to non-zero states, hence

$$\mathbf{F}^{\bullet} = \begin{bmatrix} \mathbf{F}_{++} & \mathbf{F}_{+-} \\ \mathbf{F}_{-+} & \mathbf{F}_{--} \end{bmatrix} + \begin{bmatrix} \mathbf{F}_{+\mathbf{0}} \\ \mathbf{F}_{-\mathbf{0}} \end{bmatrix} (-\mathbf{F}_{\mathbf{00}})^{-1} \begin{bmatrix} \mathbf{F}_{\mathbf{0}+} & \mathbf{F}_{\mathbf{0}-} \end{bmatrix}. \quad (3)$$

According to [7] the density function of the stationary fluid level has a matrix-exponential form if the drift of the queue is negative, hence

$$\pi(x) = \kappa\, e^{\mathbf{K}x} \underbrace{\begin{bmatrix} \mathbf{I} & \mathbf{\Psi} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{+} & \\ & |\mathbf{C}_{-}| \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{F}_{+0}(\mathbf{F}_{00})^{-1} \\ \mathbf{0} & \mathbf{I} & \mathbf{F}_{-0}(\mathbf{F}_{00})^{-1} \end{bmatrix}}_{\mathbf{A}},$$

where matrix $\mathbf{\Psi}$ is the minimal non-negative solution of the non-symmetric algebraic Riccati equation (NARE)

$$\begin{aligned} &\mathbf{\Psi}|\mathbf{C}_{-}|^{-1}\mathbf{F}^{\bullet}_{-+}\mathbf{\Psi} + \mathbf{\Psi}|\mathbf{C}_{-}|^{-1}\mathbf{F}^{\bullet}_{--} \\ &\quad + \mathbf{C}_{+}{}^{-1}\mathbf{F}^{\bullet}_{++}\mathbf{\Psi} + \mathbf{C}_{+}{}^{-1}\mathbf{F}^{\bullet}_{+-} = \mathbf{0}, \end{aligned} \quad (4)$$

and matrix $\mathbf{K}$ is obtained by $\mathbf{K} = \mathbf{C}_{+}{}^{-1}\mathbf{F}^{\bullet}_{++} + \mathbf{\Psi}|\mathbf{C}_{-}|^{-1}\mathbf{F}^{\bullet}_{-+}$. The probability mass at level zero is $p = \begin{bmatrix} 0 & p_{-} & p_{0} \end{bmatrix}$. Vectors κ, p_{-} and p_{0} are the solutions of the linear equations

$$\begin{bmatrix} \kappa & p_{-} & p_{0} \end{bmatrix} \begin{bmatrix} [& -\mathbf{A} &] \\ \mathbf{F}_{-+} & \mathbf{F}_{--} & \mathbf{F}_{-0} \\ \mathbf{F}_{0+} & \mathbf{F}_{0-} & \mathbf{F}_{00} \end{bmatrix} = 0, \quad (5)$$

and the normalization condition

$$\begin{bmatrix} \kappa & p_{-} & p_{0} \end{bmatrix} \begin{bmatrix} -\mathbf{K}^{-1}\mathbf{A}\mathbb{1} \\ \mathbb{1} \\ \mathbb{1} \end{bmatrix} = 1. \quad (6)$$

2.2 Reward accumulation in the busy period

The busy period of Markovian fluid flow models has been characterized before in the literature. The LST of the busy period duration starting from $x > 0$ amount of fluid is derived in [1]. An Erlangization based numerical method has been introduced in [8] to obtain the approximation of the distribution function of a busy period length.

In this section we consider a more general problem. To each state of the background process a reward rate is assigned, and the accumulated reward over the busy period is characterized. While this study might seem exotic, it turns out that the main performance measures of the fluid priority queue can be related to the accumulated reward over the busy period.

The introduction of reward accumulation makes the system three-dimensional, $\{\mathcal{X}(t), \mathcal{Y}(t), \mathcal{J}(t), t \geq 0\}$, where $\mathcal{Y}(t)$ represents the reward accumulated up to time t. The reward accumulation is linear, thus $\mathcal{Y}(t) = \int_0^\tau d_{\mathcal{J}(t)} dt$, where d_i denotes the reward rate associated with state i. This system is called a stochastic two-dimensional fluid model in [2].

Let us introduce the diagonal matrix of reward rates $\mathbf{D} = \mathrm{diag}\langle d_i \rangle$. If the random variable representing the busy period is denoted by $\tau = \inf(t > 0 : \mathcal{X}(t) = 0)$, then the joint distribution of the accumulated reward and the state transitions over the busy period are given by matrix $\mathbf{\Phi}(y)$ (whose LST is denoted by $\hat{\mathbf{\Psi}}_X(s)$ in [2]) defined as

$$[\mathbf{\Phi}(y)]_{ij} = P(\mathcal{Y}(\tau) < y, \mathcal{J}(\tau) = j | \mathcal{X}(0) = 0, \mathcal{Y}(0) = 0, \mathcal{J}(0) = i),$$

for $i \in \mathcal{S}_+, j \in \mathcal{S}_-$, and, if the fluid level is $x > 0$ initially, the distribution of the accumulated reward is given by matrix $\mathbf{B}(y, x)$ (LST denoted by $\hat{\mathbf{G}}^y_X(s)$ in [2]) where

$$[\mathbf{B}(y, x)]_{ij} = P(\mathcal{Y}(\tau) < y, \mathcal{J}(\tau) = j | \mathcal{X}(0) = x, \mathcal{Y}(0) = 0, \mathcal{J}(0) = i)$$

for $i \in \mathcal{S}, j \in \mathcal{S}_-$.

The following two theorems provide the LST of matrices $\mathbf{\Phi}(y)$ and $\mathbf{B}(y, x)$.

Theorem 1 ([2], Theorem 4, extended with 0 states). *For $x>0$ the blocks of matrix $\mathbf{B}^*(v, x) = \int_0^\infty e^{-vy}\, d\mathbf{B}(y, x)$ are*

$$\begin{aligned} \mathbf{B}^*_{+-}(v, x) &= \mathbf{\Phi}^*(v)\mathbf{B}^*_{--}(v, x), \\ \mathbf{B}^*_{--}(v, x) &= e^{\mathbf{H}^*_B(v)x}, \\ \mathbf{B}^*_{0-}(v, x) &= (v\mathbf{D}_0 – \mathbf{F}_{00})^{-1}\mathbf{F}_{0+}\mathbf{B}^*_{+-}(v, x) \\ &\quad + (v\mathbf{D}_0 – \mathbf{F}_{00})^{-1}\mathbf{F}_{0-}\mathbf{B}^*_{--}(v, x), \end{aligned}$$

where matrix $\mathbf{H}^*_B(v)$ is given by

$$\begin{aligned} \mathbf{H}^*_B(v) = -\mathbf{C}_{-}^{-1}\big(&\mathbf{F}_{--} - v\mathbf{D}_{-} + \mathbf{F}_{-0}(v\mathbf{D}_0 - \mathbf{F}_{00})^{-1}\mathbf{F}_{0-} \\ &+ (\mathbf{F}_{-+} + \mathbf{F}_{-0}(v\mathbf{D}_0 - \mathbf{F}_{00})^{-1}\mathbf{F}_{0+})\mathbf{\Phi}^*(v)\big). \end{aligned}$$

Theorem 2 ([2], Theorem 3). *Matrix $\mathbf{\Phi}^*(v)$ describing the LST of the accumulated reward over a busy period starting from empty buffer satisfies the NARE*

$$\begin{aligned} 0 = &\mathbf{C}_{+}^{-1} \left(\mathbf{F}_{++} - v\mathbf{D}_{+} + \mathbf{F}_{+0}(v\mathbf{D}_0 - \mathbf{F}_{00})^{-1}\mathbf{F}_{0+}\right) \mathbf{\Phi}^*(v) \\ &+ \mathbf{\Phi}^*(v)(-\mathbf{C}_{-})^{-1} \left(\mathbf{F}_{--} - v\mathbf{D}_{-} + \mathbf{F}_{-0}(v\mathbf{D}_0 - \mathbf{F}_{00})^{-1}\mathbf{F}_{0-}\right) \\ &+ \mathbf{\Phi}^*(v)(-\mathbf{C}_{-})^{-1} \left(\mathbf{F}_{-+} + \mathbf{F}_{-0}(v\mathbf{D}_0 - \mathbf{F}_{00})^{-1}\mathbf{F}_{0+}\right) \mathbf{\Phi}^*(v) \\ &+ \mathbf{C}_{+}^{-1} \left(\mathbf{F}_{+-} + \mathbf{F}_{+0}(v\mathbf{D}_0 - \mathbf{F}_{00})^{-1}\mathbf{F}_{0-}\right). \end{aligned}$$

Observe that $\mathbf{\Phi}^*(v)|_{v\to 0} = \mathbf{\Psi}$.

To obtain the moments of the performance measures of the fluid priority queue the derivatives of matrix $\mathbf{\Phi}^*(v)$ will be required at $v \to 0$, that is, $\mathbf{\Phi}^{(n)} = \frac{d^n}{dv^n}\mathbf{\Phi}^*(v)|_{v\to 0}$. It is possible to derive recursive formulas for $\mathbf{\Phi}^{(n)}$, where $\mathbf{\Phi}^{(0)} = \mathbf{\Psi}$ and for $n > 1$ matrix $\mathbf{\Phi}^{(n)}$ is obtained by the solution of a Sylvester equation involving matrices $\mathbf{\Phi}^{(k)}, k < n$. The details are omitted due to space limitations.

The rest of this section focuses on the distribution of the reward accumulated over the busy period. The Erlangization based numerical method published in [8] is extended in two aspects: first, [8] obtains the duration of the busy period, while we need the accumulated reward in this paper, second, zero rates are excluded in [8], while we need them here.

According to the Erlangization algorithm the order-n approximation of matrix $\mathbf{\Phi}(t)$ is

$$\mathbf{\Phi}_n(t) = \int_0^\infty f_{\mathcal{E}(n,n/t)}(u) \cdot \mathbf{\Phi}(u)\, du, \quad (7)$$

where $f_{\mathcal{E}(n,n/t)}(u)$ is an order-n Erlang density with rate parameter $\nu = n/t$. As n tends to infinity, $\mathbf{\Phi}_n(t)$ tends to $\mathbf{\Phi}(t)$. Matrix $\mathbf{\Phi}_n(t)$ has a probabilistic meaning as well: it is the probability that the accumulated reward in the busy period is less than an Erlang(n, ν) variable. Hence, matrix $\mathbf{\Phi}_n(t)$ is evaluated as follows. The state space of the background process is extended such that it does not only keep track of the state of $\mathcal{J}(t)$, but also the number of completed stages of the Erlang distribution. At the end of the busy period, the probabilities of those states where this counter is less that n contribute to $\mathbf{\Phi}_n(t)$. Thus, $\mathbf{\Phi}_n(t)$ can be expressed as a sum

$$\mathbf{\Phi}_n(t) = \sum_{k=0}^{n-1} \mathbf{\Psi}_k,$$

where $\mathbf{\Psi}_k$ contains the probabilities that the Erlang random variable is in stage k when the reward accumulation (and hence the busy period) ends.

It can be proven that $\mathbf{\Psi}_0 = \mathbf{\Phi}^*(\nu)$ with $\nu = n/t$, and that for $k > 0$ matrices $\mathbf{\Psi}_k$ can be obtained recursively as the solutions of Sylvester equations.

3. THE MARKOV-MODULATED FLUID PRIORITY QUEUE

3.1 The description of the system

In the system considered in the paper K different (fluid) job types are distinguished, where class 1 has the lowest, and class K the highest priority. The rates at which the fluid belonging to various job types enter the system are modulated by a continuous time Markov chain (CTMC, also referred to as the background process) with generator denoted by $\mathbf{Q}$ (assumed to be irreducible). The stationary probability vector of the CTMC is π, hence $\pi\mathbf{Q} = 0, \pi\mathbb{1} = 1$ hold ($\mathbb{1}$ denotes the column vector of ones of appropriate size). The rate at which class k fluid flows into the queue in state i is $r_i^{(k)} \geq 0$. Matrix $\mathbf{R}^{(k)}$ is a diagonal matrix composed by the class k incoming fluid rates, $\mathbf{R}^{(k)} = \text{diag}\langle r_i^{(k)}\rangle$. For simplicity, we denote the input rates of classes having priority equal to or higher than k by $r_i^{(k+)}$. The rate at which fluid can leave the queue is denoted by d.

Using the notation introduced above, the mean arrival rate of class k fluid can be obtained by $\lambda^{(k)} = \pi\mathbf{R}^{(k)}\mathbb{1}$, and the total input rate is $\lambda = \sum_{k=1}^{K} \lambda^{(k)}$. The system is assumed to be stable, hence $\lambda < d$.

3.2 Concept of the solution

The solution is based on the ”tagged customer” approach. When a class k fluid drop arrives into the queue, it finds a given amount of class $k+$ workload in the system. This amount of workload has to be served before the tagged fluid drop can leave the system. The workload of lower priority classes can be neglected. While the tagged fluid drop waits for its service in the queue, further class $(k+1)+$ fluid can arrive, which has to be served before the tagged one.

The *waiting time* of the tagged fluid drop is the time the workload found in the system at arrival, increased by the higher priority workload brought to the system while waiting, is processed by the server.

The class k *queue length* at the fluid drop departure instant is the amount of class k fluid arriving while the tagged fluid drop waits in the system.

3.3 The workload at fluid drop arrivals

As the first step of the analysis, the distribution of the total workload of classes $\geq k$ is determined at class k fluid drop arrival instants.

Let us denote the workload of classes $\geq k$ at time t by $\mathcal{V}(t)$. The joint density function of the workload and the state of the background process at time t is denoted by row vector $v(t,x)$, defined by $v_i(t,x) = \frac{d}{dx}P(\mathcal{V}(t) < x, \mathcal{J}(t) = i)$. The joint probability that the workload of the system is 0 and the state of the background process is i is stored by row vector $\alpha(t)$, thus $\alpha_i(t) = P(\mathcal{V}(t) = 0, \mathcal{J}(t) = i)$.

THEOREM 3. *Vector $v(t,x)$ is determined by the differential equation*

$$\frac{\partial}{\partial t}v(t,x) + \frac{\partial}{\partial x}v(t,x)(\mathbf{R}^{(k+)}/d - \mathbf{I}) = v(t,x)\mathbf{Q}, \qquad (8)$$

with boundary condition

$$\frac{d}{dt}\alpha_i(t) = \sum_{j:r_j<d} \alpha_j(t)q_{ji} - v_i(t,0)(r_i^{(k+)}/d - 1), \qquad (9)$$

for $i : r_i^{(k+)} \leq d$, and $\alpha_i = 0$ for $i : r_i^{(k+)} > d$.

Theorem 3 has an important corollary: the workload process behaves like an ordinary Markovian fluid flow model with parameters

$$\mathbf{F} = \mathbf{Q}, \qquad \mathbf{C} = \mathbf{R}^{(k+)}/d - \mathbf{I}, \qquad (10)$$

thus its stationary density, $v(x) = \lim_{t\to\infty} v(t,x)$, and probability mass at zero, $\alpha = \lim_{t\to\infty}\alpha(t)$, are given by $v(x) = \kappa e^{\mathbf{K}x}\mathbf{A}, \alpha = p$, where vectors κ, p and matrices $\mathbf{K}, \mathbf{A}$ are obtained as described in Section 2.1 with parameters (10). Hence, the density of the workload and the probability mass at zero embedded at class k fluid drop arrival instants are

$$v_A(x) = \frac{1}{\lambda^{(k)}}\kappa e^{\mathbf{K}x}\mathbf{A}\mathbf{R}^{(\mathbf{k})}, \qquad \alpha_A = \frac{1}{\lambda^{(k)}}p\mathbf{R}^{(\mathbf{k})}. \qquad (11)$$

3.4 The properties of the sojourn time

In order to analyze the sojourn time, a special fluid flow model and appropriate reward rates are introduced, such that the accumulated reward over a busy period of this fluid model is equal to the sojourn time of class k fluid drops.

The generator of the background process, the fluid rates, and the reward rates of this special fluid flow model are

$$\mathbf{F} = \begin{bmatrix} \mathbf{K} & \frac{\mathbf{AR}^{(k)}}{\lambda^{(k)}} \\ & \mathbf{Q} \end{bmatrix}, \mathbf{C} = \begin{bmatrix} \mathbf{I} & \\ & \frac{\mathbf{R}^{((k+1)+)}}{d} - \mathbf{I} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I} \end{bmatrix}. \qquad (12)$$

The role of the first state group of the background process is solely the accumulation of the workload a class k fluid drop finds in the system upon its arrival. The workload is increased with rate 1 and during this period no reward is accumulated, since it is not part of the sojourn time. Note that it is the matrix-exponential form of the initial workload (given by (11)) that allowed to apply this trick (see [4]).

When the appropriate workload level is reached, a transition occurs to the second group of states. In this part, the fluid level in this special model represents the workload of the system ahead of the tagged class k fluid drop. The workload brought by class $(k+1)+$ fluid flows increase the workload ahead of the tagged class k fluid drop. The rates at which the workload changes during this period are given by the diagonal elements of $\mathbf{R}^{((k+1)+)}/d - \mathbf{I}$. The reward rate is 1 in each state of the modulating CTMC, thus the reward measures the time spent in this part of the state space. The time spent till the workload ahead of the tagged drop reaches level 0 is identical to the sojourn time.

Hence, the sojourn time is given by the amount of reward accumulated during the busy period, with initial vector $\kappa' = \begin{bmatrix}\kappa & 0\end{bmatrix}$. If the random variable representing the sojourn time is denoted by $\mathcal{T}$, the LST of the distribution function $f^*_{\mathcal{T}}(s)$ and its kth moment $E(\mathcal{T}^k)$ are expressed by

$$f^*_{\mathcal{T}}(s) = \kappa'\mathbf{\Phi}^*(s)\mathbb{1}, \qquad (13)$$

$$E(\mathcal{T}^k) = (-1)^k\kappa'\mathbf{\Phi}^{(k)}\mathbb{1}, \quad k > 0, \qquad (14)$$

and the order-n approximation of the distribution function $F_{\mathcal{T}}^{(n)}(t)$ is

$$F_{\mathcal{T}}^{(n)}(t) = \alpha_A \mathbb{1} + \kappa' \mathbf{\Phi}_n(t) \mathbb{1}, \tag{15}$$

where matrices $\mathbf{\Phi}^*(s)$, $\mathbf{\Phi}^{(k)}$ and $\mathbf{\Phi}_n(t)$ are given by the reward analysis of the fluid model defined by matrices (12) (see Section 2.2). If the workload is zero when the fluid drop arrives (term $\alpha_A \mathbb{1}$), the sojourn time is zero as well.

3.5 Queue length at fluid drop departures

The approach used to characterize the amount of class k fluid in the system at class k fluid drop departures is very similar to the one presented in Section 3.4. A special Markov-modulated fluid model is constructed such that the reward accumulated during its busy period is equal to the queue length at departures. The matrices defining this special fluid model are

$$\mathbf{F} = \begin{bmatrix} \mathbf{K} & \frac{\mathbf{A}\mathbf{R}^{(k)}}{\lambda^{(k)}} \\ & \mathbf{Q} \end{bmatrix}, \mathbf{C} = \begin{bmatrix} \mathbf{I} & \\ & \frac{\mathbf{R}^{((k+1)+)}}{d} - \mathbf{I} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{0} & \\ & \mathbf{R}^{(k)} \end{bmatrix}.$$

Observe that these matrices are similar to (12), and the interpretation is similar as well. The first state group sets the initial workload seen by a class k drop arrival. The second state group follows the workload ahead of the fluid drop till it leaves the system. The given reward rate matrix $\mathbf{D}$ is such that the reward measures the amount of class k fluid arriving till the tagged fluid drop leaves, hence it provides the class k queue length at the departure of the tagged class k fluid drop.

Denoting the class k queue length at departures by $\mathcal{X}$, the LST of the distribution function by $f_{\mathcal{X}}^*(s)$, its kth moment by $E(\mathcal{X}^k)$ and the order-n approximation of its distribution function $F_{\mathcal{X}}^{(n)}(x)$, we again have that

$$f_{\mathcal{X}}^*(s) = \kappa' \mathbf{\Phi}^*(s) \mathbb{1}, \tag{16}$$

$$E(\mathcal{X}^k) = (-1)^k \kappa' \mathbf{\Phi}^{(k)} \mathbb{1}, \qquad k > 0, \tag{17}$$

$$F_{\mathcal{X}}^{(n)}(x) = \alpha_A \mathbb{1} + \kappa' \mathbf{\Phi}_n(x) \mathbb{1}. \tag{18}$$

The queue length is zero at the departure when the workload is zero at drop arrival (covered by term $\alpha_A \mathbb{1}$).

3.6 Queue length at random point in time

Let $\mathcal{Y}$ denote the class k queue length at random point in time. The distribution function, the probability mass at level 0 and the LST of the distribution function are denoted by $F_{\mathcal{Y}}(x), p_{\mathcal{Y}}$ and $f_{\mathcal{Y}}^*(s)$, respectively. Before establishing the relationship between $\mathcal{X}$ and $\mathcal{Y}$, we introduce the vector form of these quantities that include the state of the background process as well, hence characterizing $\lim_{t\to\infty}\{\mathcal{Y}(t), \mathcal{J}(t)\}$. These row vectors are denoted by $\underline{F_{\mathcal{Y}}}(x), \underline{p_{\mathcal{Y}}}$ and $\underline{f_{\mathcal{Y}}^*}(s)$.

Theorem 4. *The relation between $\underline{f_{\mathcal{X}}^*}(s)$ and $\underline{f_{\mathcal{Y}}^*}(s)$ is*

$$\underline{f_{\mathcal{Y}}^*}(s)(s\,\mathbf{R}^{(k)} - \mathbf{Q}) = \lambda^{(k)}\, s\, \underline{f_{\mathcal{X}}^*}(s). \tag{19}$$

The moments of the queue length are obtained by taking the derivatives of the LST, i.e.,

$$E(\mathcal{Y}^n) = (-1)^n \underbrace{\frac{d^n}{ds^n} \underline{f_{\mathcal{Y}}^*}(s)|_{s\to 0}}_{y^{(n)}} \mathbb{1}. \tag{20}$$

Following [6, Section 4.1] it is possible to derive a recursive algorithm to compute vectors $y^{(n)}$, based on $y^{(n-1)}$ and the queue length moments at departure instants.

Finally, it remains to obtain the order-n approximation of the class k queue length distribution at random point in time. We managed to find a recursive relation between the order-n approximations of the queue length at the departures and the queue length at arbitrary points in time.

Theorem 5. *The order-n approximation of the distribution function $F_{\mathcal{Y}}(x)$ is $F_{\mathcal{Y}}^{(n)}(x) = \underline{F_{\mathcal{Y}}^{(n)}}(x)\mathbb{1}$, where row vectors $\underline{F_{\mathcal{Y}}^{(n)}}(x)$ are defined recursively by*

$$\underline{F_{\mathcal{Y}}^{(n)}}(x) = \nu \left(\lambda^{(k)} \kappa' \mathbf{\Psi}_{n-1} + \underline{F_{\mathcal{Y}}^{(n-1)}}(x) \mathbf{R}^{(k)} \right) (\nu \mathbf{R}^{(k)} - \mathbf{Q})^{-1}$$

for $n > 1$, and

$$\underline{F_{\mathcal{Y}}^{(1)}}(x) = \nu \lambda^{(k)} (\alpha_A + \kappa' \mathbf{\Psi}_0)(\nu \mathbf{R}^{(k)} - \mathbf{Q})^{-1}$$

for $n = 1$, where $\nu = n/x$.

4. REFERENCES

[1] Soohan Ahn and Vaidyanathan Ramaswami. Efficient algorithms for transient analysis of stochastic fluid flow models. *Journal of Applied Probability*, pages 531–549, 2005.

[2] Nigel G Bean and Małgorzata M O'Reilly. A stochastic two-dimensional fluid model. *Stochastic Models*, 29(1):31–63, 2013.

[3] Bong Dae Choi and Ki Bong Choi. A Markov modulated fluid queueing system with strict priority. *Telecommunication Systems*, 9(1):79–95, 1998.

[4] Tessa Dzial, Lothar Breuer, Ana da Silva Soares, Guy Latouche, and Marie-Ange Remiche. Fluid queues to solve jump processes. *Performance Evaluation*, 62(1):132–146, 2005.

[5] Gábor Horváth. Efficient analysis of the MMAP[K]/PH[K]/1 priority queue. *European Journal of Operational Research*, 246(1):128–139, 2015.

[6] David M Lucantoni, Kathleen S Meier-Hellstern, and Marcel F Neuts. A single-server queue with server vacations and a class of non-renewal arrival processes. *Advances in Applied Probability*, pages 676–705, 1990.

[7] V. Ramaswami. Matrix analytic methods for stochastic fluid flows. In *Teletraffic Engineering in a Competitive World - Proc. of the 16th International Teletraffic Congress (ITC 16)*, pages 1019–1030. Elsevier Science B.V., 1999.

[8] V Ramaswami, Douglas G Woolford, and David A Stanford. The erlangization method for Markovian fluid flows. *Annals of Operations Research*, 160(1):215–225, 2008.

[9] Elena I Tzenova, IJBF Adan, and Vidyadhar G Kulkarni. *A two-priority fluid flow model: joint steady state distribution of the buffer content processes.* TU/e, Technische Universiteit Eindhoven, Department of Mathematics and Computing Science, 2004.

[10] Ji Zhang. Performance study of Markov modulated fluid flow models with priority traffic. In *INFOCOM'93*, pages 10–17. IEEE, 1993.

Perturbation analysis of Markov modulated fluid models

Sarah Dendievel
Ghent University,
Department of Telecommunications and
Information Processing,
SMACS Research Group,
Sint-Pietersnieuwstraat 41
B-9000 Gent, Belgium
Sarah.Dendievel@telin.ugent.be

Guy Latouche
Université libre de Bruxelles (ULB)
Faculté des Sciences
Département d'informatique, CP212
Boulevard du Triomphe 2
B-1050 Bruxelles, Belgium
latouche@ulb.ac.be

ABSTRACT

We consider regular perturbations of positive recurrent Markov modulated fluid models, that is, given an infinitesimal generator A and a perturbation matrix $\tilde{A}$, we define the infinitesimal generator $A(\varepsilon)$ as $A(\varepsilon) = A + \varepsilon\tilde{A}$, where ε denotes a small number. The perturbation performed on A is such that its effect on the performance measures of the model dissipates as ε goes to zero. For a Markov modulated fluid model, in addition to the infinitesimal generator of the phases, we may also perturb the rate matrix, and analyze the effect of those perturbations on performance measures of the model. In both cases, the key matrix to analyze is the matrix of first return probabilities to the initial level, denoted as Ψ. We show that the analysis of perturbations on the infinitesmial generator follows the usual path while the analysis for perturbations on the rate matrix is more involved. In the later case, the comparison between Ψ and its perturbed counterpart requires some re-definition. Our main contribution is the construction of a substitute for the matrix of first return probabilities, which enables us to analyze the effect of the perturbation under consideration.

Keywords

Markov modulated fluid models; Perturbation analysis; First return probabilities

1. INTRODUCTION

Most mathematical models have input parameters that are typically estimated from the real world data. Since the parameters in the modeled system represent quantities that can suffer from small errors, it is natural to analyze how the performance measures are affected by small changes in the parameters. Using the structural properties of the model, it becomes possible to assess the impact of perturbations on the key matrices of the underlying process by providing computationally feasible solutions along with probabilistic interpretation.

Markov modulated fluid models appeared in the 1960s to study the continuous-time behavior of queues and dams, an early paper being Loynes [11]. In the eighties, Markovian fluid models started to be more extensively investigated, in particular their stationary density, see for instance Rogers [14] and Asmussen [2]. The importance of the matrix of first return probabilities has been demonstrated in Ramaswami [13] and its computation has attracted much attention, see Bean *et al.* [3] and Bini *et al.* [4]. One may derive from Ψ, *the matrix of first return probabilities form above*, important performance measures of the model, such as the stationary density of the level of the fluid model.

The model $\{(X(t), \varphi(t)) : t \in \mathbb{R}^+\}$ is described as follows: $\varphi(t)$ is a Markov chain, with finite state space $\mathcal{S}$, it is called the *phase* process; $X(t)$ is a continuous function, called the *level.* The evolution of the level is continuous and may be expressed as

$$X(t) = Y(t) + \sup_{0\le s\le t} \{\max(0, -Y(s))\}$$
$$\text{where} \quad Y(t) = Y(0) + \int_0^t c_{\varphi(s)}\mathrm{d}s, \tag{1}$$

so that it varies linearly with rate c_i when $\varphi(t) = i$, $i \in \mathcal{S}$. We partition $\mathcal{S}$ into $\mathcal{S}_+ \cup \mathcal{S}_0 \cup \mathcal{S}_-$ with $\mathcal{S}_+ = \{i \in \mathcal{S} : c_i > 0\}$, $\mathcal{S}_0 = \{i \in \mathcal{S} : c_i = 0\}$ and $\mathcal{S}_- = \{i \in \mathcal{S} : c_i < 0\}$. The infinitesimal generator of the phase process is denoted by A and is written, possibly after permutation of rows and columns, as

$$A = \begin{bmatrix} A_{++} & A_{+0} & A_{+-} \\ A_{0+} & A_{00} & A_{0-} \\ A_{-+} & A_{-0} & A_{--} \end{bmatrix}, \tag{2}$$

and the *rate matrix* is denoted by

$$C = \begin{bmatrix} C_+ & & \\ & C_0 & \\ & & C_- \end{bmatrix} \tag{3}$$

with $C_+ = \operatorname{diag}(c_i : i \in \mathcal{S}_+)$, $C_- = \operatorname{diag}(c_i : i \in \mathcal{S}_-)$ and C_0 is a null matrix. Throughout the paper, we make the following assumption.

Assumption 1. *The Markov modulated fluid model is positive recurrent, that is, $\boldsymbol{\xi}C\mathbf{1} < 0$, where $\boldsymbol{\xi}$ is the stationary probability vector defined for $i, j \in \mathcal{S}$ by*

$$\xi_i = \lim_{t\to\infty} \mathbb{P}[\varphi(t) = i|\varphi(0) = j], \tag{4}$$

and is the unique solution of the equation $\boldsymbol{\xi} A = 0$ *such that* $\boldsymbol{\xi}\mathbf{1} = \mathbf{1}$.

A key matrix for Markov modulated fluid models is the matrix Ψ of *first return probabilities to the initial level from above*, with dimensions $|\mathcal{S}_+| \times |\mathcal{S}_-|$, and components

$$\Psi_{ij} = \mathbb{P}\left[\tau_- < \infty, \varphi\left(\tau_-\right) = j | Y\left(0\right) = 0, \varphi\left(0\right) = i\right], \quad (5)$$

where $\tau_- = \inf\{t > 0 : Y(t) < 0\}$, $i \in \mathcal{S}_+$ and $j \in \mathcal{S}_-$. By Rogers [14, Theorem 1], Ψ is the minimal nonnegative solution of the Riccati equation

$$C_+^{-1} Q_{+-} + C_+^{-1} Q_{++} X + X \left|C_-^{-1}\right| Q_{--} + X \left|C_-^{-1}\right| Q_{-+} X = 0, \quad (6)$$

where $|C_-^{-1}|$ denotes the entrywise absolute value of C_-^{-1} and

$$\begin{bmatrix} Q_{++} & Q_{+-} \\ Q_{-+} & Q_{--} \end{bmatrix} = \begin{bmatrix} A_{++} & A_{+-} \\ A_{-+} & A_{--} \end{bmatrix} + \begin{bmatrix} A_{+0} \\ A_{-0} \end{bmatrix} \left(-A_{00}^{-1}\right) \begin{bmatrix} A_{0+} & A_{0-} \end{bmatrix}. \quad (7)$$

Similarly, the *matrix* $\hat{\Psi}$ *of first return probabilities to the initial level from below* has components

$$\hat{\Psi}_{ij} = \mathbb{P}\left[\tau_+ < \infty, \varphi\left(\tau_+\right) = j | Y\left(0\right) = 0, \varphi\left(0\right) = i\right],$$

where $\tau_+ = \inf\{t > 0 : Y(t) > 0\}$, $i \in \mathcal{S}_-$ and $j \in \mathcal{S}_+$, it satisfies a Riccati equation similar to (6). The present article focuses on the perturbation analysis of Ψ only, as the analysis for $\hat{\Psi}$ is similar.

Two other important matrices are

$$U = |C_-^{-1}| Q_{--} + |C_-^{-1}| Q_{-+} \Psi, \quad (8)$$

$$K = C_+^{-1} Q_{++} + \Psi |C_-^{-1}| Q_{-+}. \quad (9)$$

The matrix U is the infinitesimal generator of the process of downward record and is such that for $i, j \in \mathcal{S}_-$, $(e^{Ux})_{ij}$ is the probability that, starting from (y, i), for any y, the process reaches level $y - x$ in finite time and that $(y - x, j)$ is the first state visited in level $y - x$. The matrix K defined in (9) is also an important matrix for Markov modulated fluid models and appears in the sationary density of the fluid model, see Section 4.

For a long time there has been a recurrent interest in perturbation analysis, see for instance Cao and Chen [5], Heidergott, *et al.* [7], Antunes *et al.* [1]. In this paper, we analyze the perturbation of Markov modulated fluid models. When the infinitesimal generator (2) of the phases is perturbed into $A(\varepsilon) = A + \varepsilon\tilde{A}$, the analysis follows the usual path: the perturbed first return probability matrix $\Psi(\varepsilon)$ is shown to be analytic, and computable equations are readily obtained for the derivatives of $\Psi(\varepsilon)$. We focus on the first order derivative

$$\Psi^{(1)} = \left.\frac{d\Psi(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0}$$

of a perturbed Markov modulated fluid model as it provides a good approximation of the effect of the perturbation on the system when compared to the unperturbed system. Furtermore, we are interested in the structures and going beyond the first derivative is rather computational and does not bring much more information.

We also analyze the effect on Ψ of perturbations of the rate matrix (3). When C is perturbed as $C(\varepsilon) = C + \varepsilon\tilde{C}$, phases of $\mathcal{S}_0$ may be transformed into phases of $\mathcal{S}_+$ or $\mathcal{S}_-$ in the perturbed model, with the consequence that a perturbation of the rates c_i appearing in (1) may modify the structure of $\Psi(\varepsilon)$ as the dimensions are not the same as those of Ψ. Clearly, the comparison between the matrices $\Psi(\varepsilon)$ and Ψ requires more care.

We do not consider cases where both the generator A and the rate matrix C are perturbed, as our results show that this may be done, at the cost of increased complexity in the expressions obtained.

In Section 2, we analyze perturbations of the infinitesimal generator of the phases. In Section 3, we analyze perturbations on the rate matrix C in four different cases. In Section 3.1 we assume that the phases of $\mathcal{S}_0$ are unaffected by the perturbation. In Sections 3.2–3.4 we examine what happens when the phases of $\mathcal{S}_0$ are affected by the perturbation. We propose an adapted version of Ψ which enables the analysis of the effect of the perturbation under consideration. We decompose the analysis in three subsections for the sake of clarity: firstly, we assume that all the phases in $\mathcal{S}_0$ become phases of $\mathcal{S}_+$ after perturbation, next, we assume that they all become phases of $\mathcal{S}_-$ after perturbation, finally, we assume that the phases in $\mathcal{S}_0$ are split between $\mathcal{S}_+$ and $\mathcal{S}_-$. As an application, we derive in Section 4 the first order approximation of the stationary density of a perturbed fluid model.

2. PERTURBATION OF THE INFINITESIMAL GENERATOR

In this section, the infinitesimal generator A is perturbed and becomes

$$A(\varepsilon) = A + \varepsilon\tilde{A}, \quad (10)$$

where

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{++} & \tilde{A}_{+0} & \tilde{A}_{+-} \\ \tilde{A}_{0+} & \tilde{A}_{00} & \tilde{A}_{0-} \\ \tilde{A}_{-+} & \tilde{A}_{-0} & \tilde{A}_{--} \end{bmatrix}, \quad (11)$$

$\widetilde{A}\mathbf{1} = 0$, and we assume that $A(\varepsilon)$ is an irreducible infinitesimal generator for ε sufficiently small in a neighborhood of 0.

The matrix $\Psi(\varepsilon)$ of first return probabilities for the perturbed model is the minimal nonnegative solution of the Riccati equation

$$C_+^{-1} Q_{+-}(\varepsilon) + C_+^{-1} Q_{++}(\varepsilon) X + X \left|C_-^{-1}\right| Q_{--}(\varepsilon) + X \left|C_-^{-1}\right| Q_{-+}(\varepsilon) X = 0, \quad (12)$$

where $Q(\varepsilon)$ is defined by (7), with $A(\varepsilon)$ replacing A. We write

$$\begin{bmatrix} Q_{++}(\varepsilon) & Q_{+-}(\varepsilon) \\ Q_{-+}(\varepsilon) & Q_{--}(\varepsilon) \end{bmatrix} = \begin{bmatrix} Q_{++} + \varepsilon\tilde{Q}_{++} & Q_{+-} + \varepsilon\tilde{Q}_{+-} \\ Q_{-+} + \varepsilon\tilde{Q}_{-+} & Q_{--} + \varepsilon\tilde{Q}_{--} \end{bmatrix} + O(\varepsilon^2).$$

THEOREM 1. *The matrix* $\Psi(\varepsilon)$ *of first return probabilities, minimal nonnegative solution to* (12), *for the perturbed model is analytic in a neighbourhood of zero. Furthermore,* $\Psi^{(1)}$ *is the unique solution of the Sylvester equation*

$$KX + XU = -C_+^{-1}\tilde{Q}_{+-} - C_+^{-1}\tilde{Q}_{++}\Psi - \Psi|C_-^{-1}|\tilde{Q}_{--} - \Psi|C_-^{-1}|\tilde{Q}_{-+}\Psi \quad (13)$$

where K *and* U *are defined in (9) and (8).*

PROOF. Define the continuous operator

$$\begin{aligned} F(\varepsilon, \mathcal{X}) = {} & C_+^{-1} Q_{+-}(\varepsilon) + C_+^{-1} Q_{++}(\varepsilon)\mathcal{X} \\ & + \mathcal{X}\left|C_-^{-1}\right| Q_{--}(\varepsilon) + \mathcal{X}\left|C_-^{-1}\right| Q_{-+}(\varepsilon)\mathcal{X}. \end{aligned}$$

We have $F(0, \Psi) = 0$ and $\partial_{\mathcal{X}} F(\varepsilon, \mathcal{X})$ exists in a neighborhood of $(0, \Psi)$ and is continuous at $(0, \Psi)$. For $Y, H \in \mathbb{R}^{|\mathcal{S}_+| \times |\mathcal{S}_-|}$, the equation

$$\partial_{\mathcal{X}} F(\varepsilon, \mathcal{X})\big|_{\varepsilon=0, \mathcal{X}=\Psi}(Y) = H,$$

is equivalent to the Sylvester equation

$$KY + YU = H. \qquad (14)$$

From Rogers [14] and Govorun *et al.* [6], we have $\mathrm{sp}(K) \in \{z \in \mathbb{C} : \mathrm{Re}(z) < 0\}$ and $\mathrm{sp}(-U) \in \{z \in \mathbb{C} : \mathrm{Re}(z) \geq 0\}$. Thus, K and $-U$ have no common eigenvalue and, by Lancaser and Tismenetsky [10, page 414], (14) has a unique solution, so that $\partial_{\mathcal{X}} F(\varepsilon, \mathcal{X})\big|_{\varepsilon=0, \mathcal{X}=\Psi(0)}$ is a nonsingular operator. We conclude that $\Psi(\varepsilon)$ is analytic at zero by the Implicit Function Theorem. □

Remark 1. It immediately results from Xue *et al.* [15, Theorem 2.2] that small *relative* changes to the entries of Q induce small *relative* differences between Ψ and $\Psi(\varepsilon)$. The bounding coefficient matrix in [15, Eqn. (2.12)] is the solution of a Sylvester equation with the same coefficients K and U as in (13) and a different right-hand side.

3. PERTURBATION OF THE RATE MATRIX

Define

$$C(\varepsilon) = C + \varepsilon\tilde{C} \qquad (15)$$

with

$$\tilde{C} = \begin{bmatrix} \tilde{C}_+ & & \\ & \tilde{C}_0 & \\ & & \tilde{C}_- \end{bmatrix} \qquad (16)$$

where the orders of $\tilde{C}_+$, $\tilde{C}_0$ and $\tilde{C}_-$ are equal to those of C_+, C_0 and C_-, respectively. Assume that ε is small enough so that the diagonal elements of $C_+(\varepsilon)$ are strictly positive and those of $C_-(\varepsilon)$ strictly negative.

We analyze separately the cases $\tilde{C}_0 = 0$ (in Section 3.1) and $\tilde{C}_0 \neq 0$. If $\tilde{C}_0 \neq 0$, the perturbation has the effect of changing null phases into non-null phases. To simplify the presentation, we suppose at first that all phases of $\mathcal{S}_0$ become phases of the same non-null subset $\mathcal{S}_+$ after perturbation. This is analyzed in Section 3.2. In Section 3.3, we treat the case where all the phases of $\mathcal{S}_0$ become phases of $\mathcal{S}_-$ after perturbation. Finally, we assume in Section 3.4 that the phases in $\mathcal{S}_0$ are split partially into $\mathcal{S}_+$ and into $\mathcal{S}_-$.

Clearly, Section 3.4 covers the cases analyzed in Sections 3.2 and 3.3. It is useful, nevertheless, to proceed through the special cases first, for which the results are easier to follow. In various remarks, we emphasize the unity of treatment.

The Implicit Function Theorem applies in all cases to prove the analyticity of $\Psi(\varepsilon)$, although details become more involved as we proceed from the simplest to the most general case. We show this in Theorem 3 and Theorem 4 and we omit the details for Theorem 5.

3.1 Phases in $\mathcal{S}_0$ unaffected

Assume that $\tilde{C}_0 = 0$ so that $C_0(\varepsilon) = 0$ as well. The matrix $\Psi(\varepsilon)$ of first return probabilities for the perturbed model is the minimal nonnegative solution of the Riccati equation

$$\begin{aligned} & C_+^{-1}(\varepsilon) Q_{+-} + C_+^{-1}(\varepsilon) Q_{++} X \\ & \quad + X\left|C_-^{-1}(\varepsilon)\right| Q_{--} + X\left|C_-^{-1}(\varepsilon)\right| Q_{-+} X = 0. \qquad (17) \end{aligned}$$

The next Theorem is proved by applying to (17) the same argument as in Theorem 1.

THEOREM 2. *Assume $C(\varepsilon) = C + \varepsilon\tilde{C}$, with $\tilde{C}_0 = 0$. The matrix $\Psi(\varepsilon)$ of first return probabilities for the perturbed model is analytic at zero and may be written as*

$$\Psi(\varepsilon) = \Psi + \varepsilon\Psi^{(1)} + O(\varepsilon^2),$$

where Ψ is the minimal non-negative solution to (6) *and $\Psi^{(1)}$ is the unique solution of the Sylvester equation*

$$KX + XU = -\Psi|C_-^{-1}|\tilde{C}_- U - C_+^{-1}\tilde{C}_+ \Psi U, \qquad (18)$$

where K and U are defined in (9) *and* (8) *respectively.* □

3.2 Migration of $\mathcal{S}_0$ to $\mathcal{S}_+$

Assume that $\tilde{C}_i > 0$ for all i in $\mathcal{S}_0$, this means that all phases of $\mathcal{S}_0$ become phases of fluid increase after perturbation. To make this explicit in our equations, we write $\mathcal{S}_\oplus$ instead of $\mathcal{S}_0$ and the infinitesimal generator of the phase process is written as

$$A = \left[\begin{array}{c|c|c} A_{++} & A_{+\oplus} & A_{+-} \\ \hline A_{\oplus+} & A_{\oplus\oplus} & A_{\oplus-} \\ \hline A_{-+} & A_{-\oplus} & A_{--} \end{array}\right]. \qquad (19)$$

After perturbation, it is partitioned as

$$A = \left[\begin{array}{cc|c} A_{++} & A_{+\oplus} & A_{+-} \\ A_{\oplus+} & A_{\oplus\oplus} & A_{\oplus-} \\ \hline A_{-+} & A_{-\oplus} & A_{--} \end{array}\right] \qquad (20)$$

and the set of phases with positive rates is $\mathcal{S}_+ \cup \mathcal{S}_\oplus$. The dimensions of the first return probability matrix become $(|\mathcal{S}_+| + |\mathcal{S}_\oplus|) \times |\mathcal{S}_-|$ after perturbation and Ψ may not be directly compared to $\Psi(\varepsilon)$, the matrix of first return probabilities of the perturbed model, which is partitioned as

$$\boldsymbol{\Psi}(\varepsilon) = \begin{bmatrix} \Psi_{+-}(\varepsilon) \\ \Psi_{\oplus-}(\varepsilon) \end{bmatrix}. \qquad (21)$$

The matrix $\boldsymbol{\Psi}(\varepsilon)$ is the minimal nonnegative solution of the Riccati equation

$$\begin{aligned} & \begin{bmatrix} C_+ + \varepsilon\tilde{C}_+ & \\ & \varepsilon\tilde{C}_\oplus \end{bmatrix}^{-1} \left(\begin{bmatrix} A_{+-} \\ A_{\oplus-} \end{bmatrix} + \begin{bmatrix} A_{++} & A_{+\oplus} \\ A_{\oplus+} & A_{\oplus\oplus} \end{bmatrix} X \right) \\ & \quad + X\left|C_- + \varepsilon\tilde{C}_-\right|^{-1} \left(A_{--} + \begin{bmatrix} A_{-+} & A_{-\oplus} \end{bmatrix} X\right) = 0. \qquad (22) \end{aligned}$$

As we show in the next theorem, comparisons are nevertheless possible, as Ψ is immediately recognised in the limit $\overline{\Psi} = \lim_{\varepsilon \to 0} \boldsymbol{\Psi}(\varepsilon)$.

THEOREM 3. *The matrix* (21) *of first return probabilities for the perturbed model, minimal nonnegative solution of* (22)*, is analytic near zero and may be written as*

$$\boldsymbol{\Psi}(\varepsilon) = \overline{\Psi} + \varepsilon\Psi^{(1)} + O(\varepsilon^2),$$

where

$$\overline{\Psi} = \begin{bmatrix} \Psi \\ \Psi_{\oplus -} \end{bmatrix} \quad \textit{and} \quad \Psi^{(1)} = \begin{bmatrix} \Psi^{(1)}_{+-} \\ \Psi^{(1)}_{\oplus -} \end{bmatrix}, \tag{23}$$

where Ψ *is given in* (6), $\Psi_{\oplus -} = (-A^{-1}_{\oplus\oplus})(A_{\oplus -} + A_{\oplus +}\Psi)$, $\Psi^{(1)}_{+-}$ *is the unique solution of the Sylvester equation*

$$KX + XU = -\Psi|C_-^{-1}|\tilde{C}_- U - C_+^{-1}\tilde{C}_+\Psi U + P_\oplus U, \tag{24}$$

and

$$\Psi^{(1)}_{\oplus -} = (-A^{-1}_{\oplus\oplus})\tilde{C}_\oplus \Psi_{\oplus -} U + (-A^{-1}_{\oplus\oplus})A_{\oplus +}\Psi^{(1)}_{+-}. \tag{25}$$

The matrices K *and* U *are defined in* (9) *and* (8), *and*

$$P_\oplus = K_{+\oplus}(-A^{-1}_{\oplus\oplus})\tilde{C}_\oplus(-A^{-1}_{\oplus\oplus})(A_{\oplus -} + A_{\oplus +}\Psi)$$

with $K_{+\oplus} = C_+^{-1}A_{+\oplus} + \Psi|C_-^{-1}|A_{-\oplus}$.

Proof. To remove the effect of ε^{-1} in the left-most coefficient of (22), we pre-multiply both sides by $\mathrm{diag}(I, \varepsilon I)$. For $\mathcal{X} = \begin{bmatrix} \mathcal{X}_{+-} \\ \mathcal{X}_{\oplus -} \end{bmatrix}$, we define the operator

$$F(\varepsilon, \mathcal{X}) = \begin{bmatrix} (C_+ + \varepsilon\tilde{C}_+)^{-1}(A_{+-} + A_{++}\mathcal{X}_{+-} + A_{+\oplus}\mathcal{X}_{\oplus -}) \\ \tilde{C}_\oplus^{-1}(A_{\oplus -} + A_{\oplus +}\mathcal{X}_{+-} + A_{\oplus\oplus}\mathcal{X}_{\oplus -}) \end{bmatrix}$$
$$+ \begin{bmatrix} \mathcal{X}_{+-} \\ \varepsilon\mathcal{X}_{\oplus -} \end{bmatrix} |C_- + \varepsilon\tilde{C}_-|^{-1}(A_{--} + A_{-+}\mathcal{X}_{+-} + A_{-\oplus}\mathcal{X}_{\oplus -}).$$

The equation

$$\partial_{\mathcal{X}} F(\varepsilon, \mathcal{X})|_{\varepsilon=0, \mathcal{X}=\overline{\Psi}} \begin{bmatrix} Y_{+-} \\ Y_{\oplus -} \end{bmatrix} = \begin{bmatrix} H_{+-} \\ H_{\oplus -} \end{bmatrix}$$

is equivalent to the set of equations

$$Y_{+-}U + KY_{+-} = H_{+-} + K_{+\oplus}(-A^{-1}_{\oplus\oplus})\tilde{C}_\oplus H_{\oplus -},$$
$$Y_{\oplus -} = A^{-1}_{\oplus\oplus}\tilde{C}_\oplus H_{\oplus -} + (-A^{-1}_{\oplus\oplus})A_{\oplus -}Y_{+-}.$$

This is a non-singular system, so that $\mathbf{\Psi}(\varepsilon)$ is analytic, by the Implicit Function Theorem. From (22), we obtain the two equations:

$$\Psi_{+-}(\varepsilon)|C_- + \varepsilon\tilde{C}_-|^{-1}\left(A_{--} + A_{-+}\Psi_{+-}(\varepsilon) + A_{-\oplus}\Psi_{\oplus -}(\varepsilon)\right)$$
$$+ (C_+ + \varepsilon\tilde{C}_+)^{-1}\left(A_{+-} + A_{++}\Psi_{+-}(\varepsilon) + A_{+\oplus}\Psi_{\oplus -}(\varepsilon)\right) = 0, \tag{26}$$

and

$$\varepsilon\Psi_{\oplus -}(\varepsilon)|C_- + \varepsilon\tilde{C}_-|^{-1}\left(A_{--} + A_{-+}\Psi_{+-}(\varepsilon) + A_{-\oplus}\Psi_{\oplus -}(\varepsilon)\right)$$
$$+ \tilde{C}_\oplus^{-1}\left(A_{\oplus -} + A_{\oplus +}\Psi_{+-}(\varepsilon) + A_{\oplus\oplus}\Psi_{\oplus -}(\varepsilon)\right) = 0, \tag{27}$$

in which we take the limit for $\varepsilon \to 0$. The second equation gives

$$\Psi_{\oplus -}(0) = (-A_{\oplus\oplus})^{-1}\left(A_{\oplus -} + A_{\oplus +}\Psi_{+-}(0)\right) \tag{28}$$

and the first equation gives $\Psi_{+-}(0)$ as the solution of (6), so that $\Psi_{+-}(0) = \Psi$. This proves (23).

Taking the coefficients of ε in (27) and using (28) leads directly to (25). To prove (24), we note that $\lim_{\varepsilon\to 0} U(\varepsilon) = U$ so that, taking in (26) the limit for $\varepsilon \to 0$ and using (23), we obtain

$$-\Psi U = C_+^{-1}(A_{+-} + A_{++}\Psi + A_{+\oplus}\Psi_{\oplus -}). \tag{29}$$

We take the coefficient of ε in (26) and we use (29) to obtain

$$K_{++}\Psi^{(1)}_{+-} + K_{+\oplus}\Psi^{(1)}_{\oplus -} + \Psi^{(1)}_{+-}U = -\Psi|C_-^{-1}|\tilde{C}_- U - C_+^{-1}\tilde{C}_+\Psi U$$

with $K_{++} = C_+^{-1}A_{++} + \Psi|C_-^{-1}|A_{-+}$. Using (25) and (9) gives then (24). □

Remark 2. The components of the block Ψ in $\overline{\Psi}$ are those defined in (5), for which one has a clear interpretation. The components of the second block have a probabilistic interpretation as well: the ijth entry, for $i \in \mathcal{S}_\oplus$ and $j \in \mathcal{S}_-$, is the sum of

- $[(-A^{-1}_{\oplus\oplus})A_{\oplus -}]_{ij}$, the probability that the phase process eventually goes from phase i to phase j, after some time spent in $S_\oplus$ and
- $[(-A^{-1}_{\oplus\oplus})A_{\oplus +}\Psi]_{ij}$, the probability that the phase process leaves $\mathcal{S}_\oplus$ for a phase in $\mathcal{S}_+$ and later returns to the initial level in phase j.

Remark 3. The Sylvester equations (18) and (24) for $\Psi^{(1)}_{+-}$ are nearly identical. The only difference is the last term in the right-hand side of (24), reflecting the migration of all phases of $\mathcal{S}_0$ to phases of fluid increase.

3.3 Migration of $\mathcal{S}_0$ to $\mathcal{S}_-$

Assume that $\tilde{C}_i < 0$ for all i in $\mathcal{S}_0$, so that all the phases of $\mathcal{S}_0$ become phases of $\mathcal{S}_-$ after perturbation. The set of such phases is written $\mathcal{S}_\ominus$ and the infinitesimal generator of the phases is written as

$$A = \begin{bmatrix} A_{++} & A_{+\ominus} & A_{+-} \\ A_{\ominus +} & A_{\ominus\ominus} & A_{\ominus -} \\ A_{-+} & A_{-\ominus} & A_{--} \end{bmatrix}.$$

The matrix of first return probabilities of the perturbed model is partitioned as

$$\mathbf{\Psi}(\varepsilon) = \begin{bmatrix} \mathbf{\Psi}_{+\ominus}(\varepsilon) & \mathbf{\Psi}_{+-}(\varepsilon) \end{bmatrix},$$

and it is the minimal nonnegative solution of a Riccati equation which we rewrite as the two equations

$$(C_+ + \varepsilon\tilde{C}_+)^{-1}(A_{+\ominus} + A_{++}\Psi_{+\ominus}(\varepsilon))$$
$$+ \Psi_{+\ominus}(\varepsilon)|\varepsilon\tilde{C}_\ominus|^{-1}(A_{\ominus\ominus} + A_{\ominus +}\Psi_{+\ominus}(\varepsilon))$$
$$+ \Psi_{+-}(\varepsilon)|C_- + \varepsilon\tilde{C}_-|^{-1}(A_{-\ominus} + A_{-+}\Psi_{+\ominus}(\varepsilon)) = 0, \tag{30}$$
$$(C_+ + \varepsilon\tilde{C}_+)^{-1}(A_{+-} + A_{++}\Psi_{+-}(\varepsilon))$$
$$+ \Psi_{+\ominus}(\varepsilon)|\varepsilon\tilde{C}_\ominus|^{-1}(A_{\ominus -} + A_{\ominus +}\Psi_{+-}(\varepsilon))$$
$$+ \Psi_{+-}(\varepsilon)|C_- + \varepsilon\tilde{C}_-|^{-1}(A_{--} + A_{-+}\Psi_{+-}(\varepsilon)) = 0. \tag{31}$$

Theorem 4. *The matrix* $\mathbf{\Psi}(\varepsilon)$ *of first return probabilities, minimal nonnegative solution to (30) and (31) is near zero and may be written as*

$$\mathbf{\Psi}(\varepsilon) = \overline{\Psi} + \varepsilon\Psi^{(1)} + O(\varepsilon^2), \tag{32}$$

where

$$\overline{\Psi} = \begin{bmatrix} 0 & \Psi \end{bmatrix}, \tag{33}$$
$$\Psi^{(1)} = \begin{bmatrix} \Psi^{(1)}_{+\ominus} & \Psi^{(1)}_{+-} \end{bmatrix}. \tag{34}$$

The matrix Ψ *is given in* (6), $\Psi^{(1)}_{+-}$ *is the unique solution of the Sylvester equation*

$$KX + XU = -\Psi|C_-^{-1}|\tilde{C}_- U - C_+^{-1}\tilde{C}_+\Psi U + KP_\ominus \tag{35}$$

and

$$\Psi^{(1)}_{+\ominus} = (C_+^{-1}A_{+\ominus} + \Psi|C_-^{-1}|A_{-\ominus})\left(-A^{-1}_{\ominus\ominus}\right)|\tilde{C}_\ominus|, \tag{36}$$

the matrices K and U are defined in (9) *and* (8) *and*

$$P_\ominus = \Psi^{(1)}_{+\ominus}\left(-A^{-1}_{\ominus\ominus}\right)(A_{\ominus -} + A_{\ominus +}\Psi).$$

Proof. Here, to remove the effect of ε^{-1} as a coefficient of $|\tilde{C}^{-1}_\ominus|$ in (30) and (31), we define $\Gamma(\varepsilon) = \varepsilon^{-1}\Psi_{+\ominus}(\varepsilon)$. We define the operator, for $\mathcal{X} = \left[\begin{array}{cc} \mathcal{X}_{+\ominus} & \mathcal{X}_{+-} \end{array}\right]$,

$$\begin{aligned} & F(\varepsilon, \mathcal{X}) \\ &= (C_+ + \varepsilon\tilde{C}_+)^{-1}\left[\begin{array}{cc} A_{+\ominus} + \varepsilon A_{++}\mathcal{X}_{+\ominus} & A_{+-} + A_{++}\mathcal{X}_{+-} \end{array}\right] \\ &\quad + \left[\begin{array}{cc} \mathcal{X}_{+\ominus}|\tilde{C}^{-1}_\ominus| & \mathcal{X}_{+-}|C_- + \varepsilon\tilde{C}_-|^{-1} \end{array}\right] \\ &\qquad \times \left[\begin{array}{cc} A_{\ominus\ominus} + \varepsilon A_{\ominus +}\mathcal{X}_{+\ominus} & A_{\ominus -} + A_{\ominus +}\mathcal{X}_{+-} \\ A_{-\ominus} + \varepsilon A_{-+}\mathcal{X}_{+\ominus} & A_{--} + A_{-+}\mathcal{X}_{+-} \end{array}\right] \end{aligned}$$

One shows that $\left[\begin{array}{cc} \Psi^{(1)}_{+\ominus} & \Psi \end{array}\right]$ is a solution of $F(\varepsilon, \mathcal{X}) = 0$, where $\Psi^{(1)}_{+\ominus}$ is defined in (36). Next, we take the derivative of F with respect to $\mathcal{X}$, evaluated at $\varepsilon = 0$, $\mathcal{X} = \left[\begin{array}{cc} \Psi^{(1)}_{+\ominus} & \Psi \end{array}\right]$. The system is equivalent to the set of equations

$$\begin{aligned} & Y_{+-}U + KY_{+-} = H_{+-} + H_{+\ominus}(-A^{-1}_{\ominus\ominus})(A_{\ominus -} + A_{\ominus +}\Psi), \\ &\qquad Y_{+\ominus} = Y_{+-}|C^{-1}_-|A_{-\ominus}(-A^{-1}_{\ominus\ominus})|\tilde{C}_\ominus| + H_{+\ominus}A^{-1}_{\ominus\ominus}|\tilde{C}_\ominus|, \end{aligned}$$

where, by (7), (8), (9),

$$\begin{aligned} U &= \left|C^{-1}_-\right|\left(A_{--} + A_{-\ominus}\left(-A^{-1}_{\ominus\ominus}\right)A_{\ominus -}\right) \\ &\quad + \left|C^{-1}_-\right|\left(A_{-+} + A_{-\ominus}\left(-A^{-1}_{\ominus\ominus}\right)A_{\ominus +}\right)\Psi, \qquad (37) \\ K &= C^{-1}_+\left(A_{++} + A_{+\ominus}\left(-A^{-1}_{\ominus\ominus}\right)A_{\ominus +}\right) \\ &\quad + \Psi\left|C^{-1}_-\right|\left(A_{-+} + A_{-\ominus}\left(-A^{-1}_{\ominus\ominus}\right)A_{\ominus +}\right). \qquad (38) \end{aligned}$$

The system is non-singular so that $\left[\begin{array}{cc} \Gamma(\varepsilon) & \Psi_{+-}(\varepsilon) \end{array}\right]$ is analytic.

The block components of $\overline{\Psi}$ are obtained as follows. As $\varepsilon\Gamma(\varepsilon) = \Psi_{+\ominus}(\varepsilon)$, we find that $\Psi_{+\ominus}(0) = 0$. Next, define

$$W = \lim_{\varepsilon \to 0} \Gamma(\varepsilon)|\tilde{C}_\ominus|^{-1} \qquad (39)$$

which is finite since $\Gamma(\varepsilon)$ is analytic. We rewrite (30) and find that

$$\begin{aligned} W &= C^{-1}_+ A_{+\ominus}(-A_{\ominus\ominus})^{-1} \\ &\quad + \lim_{\varepsilon \to 0} \Psi_{+-}(\varepsilon)|C_-|^{-1}A_{-\ominus}(-A_{\ominus\ominus})^{-1}. \qquad (40) \end{aligned}$$

Taking the limit as $\varepsilon \to 0$ in (31) and replacing W by (40) leads to (6). Thus, $\lim_{\varepsilon\to 0}\Psi_{+-}(\varepsilon) = \Psi$, and (33) is proved.

The block components of $\Psi^{(1)}$ are obtained as follows. Taking the coefficients of ε^0 in (30) gives directly (36). To show (35), we take the coefficients of ε^2 in (31) and get the equation

$$\begin{aligned} & \Psi^{(2)}_{+\ominus}|\tilde{C}^{-1}_\ominus|(A_{\ominus -} + A_{\ominus +}\Psi) = C^{-1}_+\tilde{C}_+C^{-1}_+(A_{+-} + A_{++}\Psi) \\ &\quad - (\Psi^{(1)}_{+-}|C^{-1}_-| + \Psi|C^{-1}_-|\tilde{C}_-|C^{-1}_-|)(A_{--} + A_{-+}\Psi) \\ &\quad - (C^{-1}_+A_{++} + \Psi|C^{-1}_-|A_{-+} + \Psi^{(1)}_{+\ominus}|\tilde{C}^{-1}_\ominus|A_{\ominus +})\Psi^{(1)}_{+-}. \quad (41) \end{aligned}$$

We equate the coefficients of ε in (30) and get

$$\begin{aligned} & \Psi^{(2)}_{+\ominus}|\tilde{C}^{-1}_\ominus| = -C^{-1}_+\tilde{C}_+C^{-1}_+A_{+\ominus}(-A^{-1}_{\ominus\ominus}) \\ &+ (C^{-1}_+A_{++} + \Psi|C^{-1}_-|A_{-+} + \Psi^{(1)}_{+\ominus}|\tilde{C}^{-1}_\ominus|A_{\ominus +})\Psi^{(1)}_{+\ominus}(-A^{-1}_{\ominus\ominus}) \\ &+ (\Psi|C^{-1}_-|\tilde{C}_- + \Psi^{(1)}_{+-})|C^{-1}_-|A_{-\ominus}(-A^{-1}_{\ominus\ominus}) \qquad (42) \end{aligned}$$

By the Riccati equation (6) and the definition (37) of U, we have

$$\begin{aligned} -\Psi U &= C^{-1}_+(A_{+-} + A_{+\ominus}\left(-A^{-1}_{\ominus\ominus}\right)A_{\ominus -}) \\ &\quad + C^{-1}_+\left(A_{++} + A_{+\ominus}\left(-A^{-1}_{\ominus\ominus}\right)A_{\ominus +}\right)\Psi. \end{aligned}$$

We replace the first coefficient $\Psi^{(1)}_{+\ominus}$ in (42) by its expression (36), then we replace $\Psi^{(2)}_{+\ominus}|\tilde{C}^{-1}_\ominus|$ in (41) by the modified right-hand side of (42). We put together the coefficients of $\Psi^{(1)}_{+-}$, use (37), (38) and eventually obtain (35). □

Remark 4. The physical justification of $\Psi_{+\ominus}(0) = 0$ goes as follows: $(\Psi_{+\ominus}(\varepsilon))_{ij}$ is the probability that the level moves to 0 in phase $j \in \mathcal{S}_\ominus$, given that the initial level is 0 and the phase is $i \in \mathcal{S}_+$, in the limit, when ε approaches 0, this probability tends to 0 because the fluid can only return to level zero in a phase of $\mathcal{S}_-$.

3.4 General case

Assume $\tilde{C}_i \neq 0$ for i in $\mathcal{S}_0$, so that all the phases of $\mathcal{S}_0$ disseminate in $\mathcal{S}_+$ and $\mathcal{S}_-$ after perturbation. The infinitesimal generator becomes

$$A = \left[\begin{array}{cccc} A_{++} & A_{+\oplus} & A_{+\ominus} & A_{+-} \\ A_{\oplus +} & A_{\oplus\oplus} & A_{\oplus\ominus} & A_{\oplus -} \\ A_{\ominus +} & A_{\ominus\oplus} & A_{\ominus\ominus} & A_{\ominus -} \\ A_{-+} & A_{-\oplus} & A_{-\ominus} & A_{--} \end{array}\right]. \qquad (43)$$

We find here a superposition of the effects observed in the two special cases examined in Sections 3.2 and 3.3. The matrix of first return probabilities from above of the perturbed system takes the form

$$\mathbf{\Psi}(\varepsilon) = \left[\begin{array}{cc} \Psi_{+\ominus}(\varepsilon) & \Psi_{+-}(\varepsilon) \\ \Psi_{\oplus\ominus}(\varepsilon) & \Psi_{\oplus -}(\varepsilon) \end{array}\right], \qquad (44)$$

it is the unique solution of the usual Riccati equation which may be rewritten as the following set of four equations:

$$\begin{aligned} & (C_+ + \varepsilon\tilde{C}_+)^{-1}(A_{+\ominus} + A_{++}\Psi_{+\ominus}(\varepsilon) + A_{+\oplus}\Psi_{\oplus\ominus}(\varepsilon)) + \Psi_{+\ominus}(\varepsilon)|\varepsilon\tilde{C}_\ominus|^{-1}(A_{\ominus\ominus} + A_{\ominus +}\Psi_{+\ominus}(\varepsilon) + A_{\ominus\oplus}\Psi_{\oplus\ominus}(\varepsilon)) \\ &\quad + \Psi_{+-}(\varepsilon)|C_- + \varepsilon\tilde{C}_-|^{-1}(A_{-\ominus} + A_{-+}\Psi_{+\ominus}(\varepsilon) + A_{-\oplus}\Psi_{\oplus\ominus}(\varepsilon)) = 0, \qquad (45) \\ & (C_+ + \varepsilon\tilde{C}_+)^{-1}(A_{+-} + A_{++}\Psi_{+-}(\varepsilon) + A_{+\oplus}\Psi_{\oplus -}(\varepsilon)) + \Psi_{+\ominus}(\varepsilon)|\varepsilon\tilde{C}_\ominus|^{-1}(A_{\ominus -} + A_{\ominus +}\Psi_{+-}(\varepsilon) + A_{\ominus\oplus}\Psi_{\oplus -}(\varepsilon)) \\ &\quad + \Psi_{+-}(\varepsilon)|C_- + \varepsilon\tilde{C}_-|^{-1}(A_{--} + A_{-+}\Psi_{+-}(\varepsilon) + A_{-\oplus}\Psi_{\oplus -}(\varepsilon)) = 0, \qquad (46) \\ & (\varepsilon\tilde{C}_\oplus)^{-1}(A_{\oplus\ominus} + A_{\oplus +}\Psi_{+\ominus}(\varepsilon) + A_{\oplus\oplus}\Psi_{\oplus\ominus}(\varepsilon)) + \Psi_{\oplus\ominus}(\varepsilon)|\varepsilon\tilde{C}_\ominus|^{-1}(A_{\ominus\ominus} + A_{\ominus +}\Psi_{+\ominus}(\varepsilon) + A_{\ominus\oplus}\Psi_{\oplus\ominus}(\varepsilon)) \\ &\quad + \Psi_{\oplus -}(\varepsilon)|C_- + \varepsilon\tilde{C}_-|^{-1}(A_{-\ominus} + A_{-+}\Psi_{+\ominus}(\varepsilon) + A_{-\oplus}\Psi_{\oplus\ominus}(\varepsilon)) = 0, \qquad (47) \\ & (\varepsilon\tilde{C}_\oplus)^{-1}(A_{\oplus -} + A_{\oplus +}\Psi_{+-}(\varepsilon) + A_{\oplus\oplus}\Psi_{\oplus -}(\varepsilon)) + \Psi_{\oplus\ominus}(\varepsilon)|\varepsilon\tilde{C}_\ominus|^{-1}(A_{\ominus -} + A_{\ominus +}\Psi_{+-}(\varepsilon) + A_{\ominus\oplus}\Psi_{\oplus -}(\varepsilon)) \\ &\quad + \Psi_{\oplus -}(\varepsilon)|C_- + \varepsilon\tilde{C}_-|^{-1}(A_{--} + A_{-+}\Psi_{+-}(\varepsilon) + A_{-\oplus}\Psi_{\oplus -}(\varepsilon)) = 0. \qquad (48) \end{aligned}$$

We show below that $\mathbf{\Psi}(\varepsilon)$ is analytic, thus we may write the matrices $U(\varepsilon)$ and $K(\varepsilon)$ as

$$U(\varepsilon) = \sum_{n=-1}^{\infty} \varepsilon^n U_n \quad \text{with} \quad U_n = \begin{bmatrix} U_{\ominus\ominus}^{(n)} & U_{\ominus-}^{(n)} \\ U_{-\ominus}^{(n)} & U_{--}^{(n)} \end{bmatrix}, \tag{49}$$

$$K(\varepsilon) = \sum_{n=-1}^{\infty} \varepsilon^n K_n \quad \text{with} \quad K_n = \begin{bmatrix} K_{++}^{(n)} & K_{+\oplus}^{(n)} \\ K_{\oplus+}^{(n)} & K_{\oplus\oplus}^{(n)} \end{bmatrix}, \tag{50}$$

in particular, the blocks

$$U_{\ominus\ominus}^{(-1)} = |\tilde{C}_{\ominus}^{-1}|A_{\ominus\ominus} + |\tilde{C}_{\ominus}^{-1}|A_{\ominus\oplus}\Psi_{\oplus\ominus}, \tag{51}$$

$$K_{\oplus\oplus}^{(-1)} = \tilde{C}_{\oplus}^{-1}A_{\oplus\oplus} + \Psi_{\oplus\ominus}|\tilde{C}_{\ominus}^{-1}|A_{\ominus\oplus}. \tag{52}$$

play an important role in what follows.

Theorem 5. *The matrix $\mathbf{\Psi}(\varepsilon)$ of first return probabilities, minimal nonnegative solution to (45-48) for the perturbed model is near zero and may be written as*

$$\mathbf{\Psi}(\varepsilon) = \overline{\Psi} + \varepsilon\Psi^{(1)} + O(\varepsilon^2),$$

where

$$\overline{\Psi} = \begin{bmatrix} 0 & \Psi \\ \Psi_{\oplus\ominus} & \Psi_{\oplus-} \end{bmatrix}. \tag{53}$$

The block Ψ is given in (6),

$$\Psi_{\oplus-} = (-K_{\oplus\oplus}^{(-1)})^{-1}\big(\tilde{C}_{\oplus}^{-1}(A_{\oplus-} + A_{\oplus+}\Psi) + \Psi_{\oplus\ominus}|\tilde{C}_{\ominus}^{-1}|(A_{\ominus-} + A_{\ominus+}\Psi)\big), \tag{54}$$

$\Psi_{\oplus\ominus}$ is the minimal nonnegative solution to the Riccati equation

$$C_{\oplus}^{-1}A_{\oplus\ominus} + C_{\oplus}^{-1}A_{\oplus\oplus}X + X\left|C_{\ominus}^{-1}\right|A_{\ominus\ominus} + X\left|C_{\ominus}^{-1}\right|A_{\ominus\oplus}X = 0. \tag{55}$$

Furthermore,

$$\Psi^{(1)} = \begin{bmatrix} \Psi_{+\ominus}^{(1)} & \Psi_{+-}^{(1)} \\ \Psi_{\oplus\ominus}^{(1)} & \Psi_{\oplus-}^{(1)} \end{bmatrix}, \tag{56}$$

with

$$\Psi_{+\ominus}^{(1)} = \big(C_{+}^{-1}(A_{+\ominus} + A_{+\oplus}\Psi_{\oplus\ominus}) + \Psi|C_{-}^{-1}| + A_{-\oplus}\Psi_{\oplus\ominus})\big)(-U_{\ominus\ominus}^{(-1)})^{-1}, \tag{57}$$

$\Psi_{\oplus\ominus}^{(1)}$ is the unique solution of the Sylvester equation

$$K_{\oplus\oplus}^{(-1)}X + XU_{\ominus\ominus}^{(-1)} = -(\tilde{C}_{\oplus}^{-1}A_{\oplus+} + \Psi_{\oplus\ominus}|C_{\ominus}^{-1}|A_{\ominus+})\Psi_{+\ominus}^{(1)} - \Psi_{\oplus-}|C_{-}^{-1}|(A_{-\ominus} + A_{-+}\Psi_{+\ominus} + A_{-\oplus}\Psi_{\oplus\ominus}), \tag{58}$$

and with

$$\Psi_{\oplus-}^{(1)} = (-K_{\oplus\oplus}^{(-1)})^{-1}\big(K_{\oplus+}^{(-1)}\Psi_{+-}^{(1)} + \Psi_{\oplus\ominus}|C_{-}^{-1}|U_{\ominus-}^{(-1)} - \Psi_{\oplus-}U_{--}^{(0)}\big), \tag{59}$$

and $\Psi_{+-}^{(1)}$ is the unique solution of the Sylvester equation

$$\begin{aligned}(C_{+}^{-1}A_{++} + \Psi|C_{-}^{-1}|A_{-+})\Psi_{+-}^{(1)} + \Psi_{+-}^{(1)}U_{--}^{(0)} \\ + (C_{+}^{-1}A_{+\oplus} + \Psi|C_{-}^{-1}|A_{-\oplus})\Psi_{+\oplus}^{(1)} + \Psi_{+\ominus}^{(2)}U_{\ominus-}^{(-1)} \\ = C_{+}^{-1}\tilde{C}_{+}C_{+}^{-1}(A_{+-} + A_{++}\Psi + A_{+\oplus}\Psi_{\oplus-}) \\ - \Psi|C_{-}^{-1}|\tilde{C}_{-}U_{--}^{(0)},\end{aligned} \tag{60}$$

where

$$\begin{aligned}\Psi_{+\ominus}^{(2)} = \Big(& - C_{+}^{-1}C_{+}C_{+}^{-1}(A_{+\ominus} + A_{+\oplus}\Psi_{\oplus\ominus}) \\ & + C_{+}^{-1}(A_{++}\Psi_{+\ominus}^{(1)} + A_{+\oplus}\Psi_{\oplus\ominus}^{(1)}) \\ & + (\Psi^{(1)} + \Psi|C_{-}^{-1}|\tilde{C}_{-})U_{-\ominus}^{(0)} \\ & + \Psi|C_{-}^{-1}|(A_{-+}\Psi_{+\ominus}^{(1)} + A_{-\oplus}\Psi_{\oplus\ominus}^{(1)})\Big)(-U_{\ominus\ominus}^{(0)})^{-1}.\end{aligned} \tag{61}$$

Proof. To remove the effect of ε^{-1} as $\varepsilon \to 0$, we need to combine the transformations of the previous two theorems. We pre-multiply the Riccati equation by $\operatorname{diag}(I, \varepsilon I)$ and we use the matrix $\Gamma(\varepsilon) = \varepsilon^{-1}\Psi_{+\ominus}(\varepsilon)$. We obtain a new fixed-point equation, from which we eventually prove, by following the same steps as in Theorem 3 and Theorem 4, that the solutions are matrices of analytic functions.

Observe the terms in ε^{-1} in the equations (45) to (48):

- we conclude from (45) that $\Psi_{+\ominus} = 0$ by a similar argument to the proof of Theorem 4;
- multiply (47) by ε and let ε tend to zero to obtain the Riccati equation (55) satisfied by $\Psi_{\oplus\ominus}$;
- multiply (48) by ε and let ε tend to zero, gives (54), taking into account that $\lim_{\varepsilon\to 0}\Psi(\varepsilon) = \Psi$, an equality that is proved below.

To determine $\Psi_{+-}(0)$ is more involved. We proceed as follows. First, from (45), we take the terms in ε^0 and we find the expression (57) for $\Psi_{+\ominus}^{(1)}$ that we replace in (46). From (46), we take the terms in ε^0 and obtain Ψ_{+-}, after a reorganization of the terms, as the minimal nonnegative solution to

$$C_{+}^{-1}T_{+-} + C_{+}^{-1}T_{++}X + X|C_{-}^{-1}|T_{--} + X|C_{-}^{-1}|T_{-+}X = 0,$$

with

$$\begin{bmatrix} T_{++} & T_{+-} \\ T_{-+} & T_{--} \end{bmatrix} = \begin{bmatrix} A_{++} & A_{+-} \\ A_{-+} & A_{--} \end{bmatrix} + \begin{bmatrix} A_{+\oplus} & A_{+\ominus} \\ A_{-\oplus} & A_{-\ominus} \end{bmatrix}\begin{bmatrix} D_{\oplus\oplus} & D_{\oplus\ominus} \\ D_{\ominus\oplus} & D_{\ominus\ominus} \end{bmatrix}\begin{bmatrix} A_{\oplus+} & A_{\oplus-} \\ A_{\ominus+} & A_{\ominus-} \end{bmatrix}.$$

where

$$\begin{aligned} D_{\oplus\oplus} &= (-K_{\oplus\oplus}^{(-1)})^{-1}\tilde{C}_{\oplus}^{-1} + \Psi_{\oplus\ominus}D_{\ominus\oplus} \\ D_{\ominus\oplus} &= (-U_{\ominus\ominus}^{(-1)})^{-1}|\tilde{C}_{\ominus}^{-1}|A_{\ominus\oplus}(-K_{\oplus\oplus}^{(-1)})^{-1}\tilde{C}_{\oplus}^{-1} \\ D_{\oplus\ominus} &= (-K_{\oplus\oplus}^{(-1)})^{-1}\Psi_{\oplus\ominus}|\tilde{C}_{\ominus}^{-1}| + \Psi_{\oplus\ominus}D_{\ominus\ominus} \\ D_{\ominus\ominus} &= (-U_{\ominus\ominus}^{(-1)})^{-1}|\tilde{C}_{\ominus}^{-1}|(I + A_{\ominus\oplus}(-K_{\oplus\oplus}^{(-1)})^{-1}\Psi_{\oplus\ominus}|\tilde{C}_{\ominus}^{-1}|) \end{aligned}$$

To prove that the matrix T is identical to the matrix Q defined in (7), we only need to show that the matrix made up of the four blocks labeled with Ds is equal to $(-A_{00}^{-1})$, partitionned according to (43), as

$$(-A_{00}^{-1}) = \begin{bmatrix} B_{\oplus\oplus} & B_{\oplus\ominus} \\ B_{\ominus\oplus} & B_{\ominus\ominus} \end{bmatrix} \tag{62}$$

where

$$\begin{aligned} B_{\oplus\oplus} &= -(A_{\oplus\oplus} + A_{\oplus\ominus}(-A_{\ominus\ominus}^{-1})A_{\ominus\oplus})^{-1} \\ B_{\ominus\oplus} &= (-A_{\ominus\ominus}^{-1})A_{\ominus\oplus}B_{\oplus\oplus} \\ B_{\oplus\ominus} &= B_{\oplus\oplus}A_{\oplus\ominus}(-A_{\ominus\ominus}^{-1}) \\ B_{\ominus\ominus} &= -(A_{\ominus\ominus} + A_{\ominus\oplus}(-A_{\oplus\oplus}^{-1})A_{\oplus\ominus})^{-1} \end{aligned}$$

By (55), we have

$$A_{\oplus\ominus} = -\tilde{C}_{\oplus} K^{(-1)}_{\oplus\oplus}\Psi_{\oplus\ominus} - \tilde{C}_{\oplus}\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}|A_{\ominus\ominus}$$

so that

$$\begin{aligned} B_{\oplus\oplus} &= -\big(A_{\oplus\oplus} - \tilde{C}_{\oplus}K^{(-1)}_{\oplus\oplus}\Psi_{\oplus\ominus}(-A^{-1}_{\ominus\ominus})A_{\ominus\oplus} \\ &\qquad + \tilde{C}_{\oplus}\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}|A_{\ominus\oplus}\big)^{-1} \\ &= (I - \Psi_{\oplus\ominus}(-A^{-1}_{\ominus\ominus})A_{\ominus\oplus})^{-1}(-K^{(-1)}_{\oplus\oplus})^{-1}\tilde{C}^{-1}_{\oplus}, \end{aligned}$$

using (52). We write

$$\begin{aligned} &(I - \Psi_{\oplus\ominus}(-A^{-1}_{\ominus\ominus})A_{\ominus\oplus})^{-1} \\ &\quad = I + \Psi_{\oplus\ominus}(I - (-A^{-1}_{\ominus\ominus})A_{\ominus\oplus}\Psi_{\oplus\ominus})^{-1}(-A^{-1}_{\ominus\ominus})A_{\ominus\oplus} \\ &\quad = I + \Psi_{\oplus\ominus}(-U^{(-1)}_{\ominus\ominus}))^{-1}|\tilde{C}^{-1}_{\ominus}|A_{\ominus\oplus}, \end{aligned}$$

so that $B_{\oplus\oplus} = D_{\oplus\oplus}$.

Next, we have

$$\begin{aligned} B_{\ominus\oplus} &= (-A^{-1}_{\ominus\ominus})A_{\ominus\oplus}(I + A_{\ominus\oplus}\Psi_{\oplus\ominus}(-U^{(-1)}_{\ominus\ominus})^{-1}|\tilde{C}^{-1}_{\ominus}|) \\ &\quad \times (-K^{(-1)}_{\oplus\oplus})^{-1}\tilde{C}^{-1}_{\oplus} \\ &= (-A^{-1}_{\ominus\ominus})(-|\tilde{C}_{\ominus}|U^{(-1)}_{\ominus\ominus}) + A_{\ominus\oplus}\Psi_{\oplus\ominus})(-U^{(-1)}_{\ominus\ominus})^{-1})|\tilde{C}^{-1}_{\ominus}| \\ &\quad \times A_{\ominus\oplus}(-K^{(-1)}_{\oplus\oplus})^{-1}\tilde{C}^{-1}_{\oplus}. \end{aligned}$$

By (51), $-|\tilde{C}_{\ominus}|U^{(-1)}_{\ominus\ominus} + A_{\ominus\oplus}\Psi_{\oplus\ominus}$ simplifies to $-A_{\ominus\ominus}$ so that $B_{\ominus\oplus} = D_{\ominus\oplus}$.

Then, we have

$$\begin{aligned} B_{\oplus\ominus} &= (-K^{(-1)}_{\oplus\oplus})^{-1}\tilde{C}^{-1}_{\oplus}A_{\oplus\ominus}(A^{-1}_{\ominus\ominus}) \\ &\quad + \Psi_{\oplus\ominus}(-U^{(-1)}_{\ominus\ominus})^{-1}|\tilde{C}^{-1}_{\ominus}|A_{\ominus\oplus}(-K^{(-1)}_{\oplus\oplus})^{-1}\tilde{C}^{-1}_{\oplus}A_{\oplus\ominus}(A^{-1}_{\ominus\ominus}), \end{aligned}$$

and we use (55) to replace $\tilde{C}^{-1}_{\oplus}A_{\oplus\ominus}$ in the first term to write

$$\begin{aligned} B_{\oplus\ominus} &= (-K^{(-1)}_{\oplus\oplus})^{-1}(-\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}|A_{\ominus\ominus} - K^{(-1)}_{\oplus\oplus}\Psi_{\oplus\ominus})(-A^{-1}_{\ominus\ominus}) \\ &\quad + \Psi_{\oplus\ominus}(-U^{(-1)}_{\ominus\ominus})^{-1}|\tilde{C}^{-1}_{\ominus}|A_{\ominus\oplus}(-K^{(-1)}_{\oplus\oplus})^{-1}\tilde{C}^{-1}_{\oplus}A_{\oplus\ominus}(-A^{-1}_{\ominus\ominus}) \\ &= (-K^{(-1)}_{\oplus\oplus})^{-1}\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}| \\ &+ \Psi_{\oplus\ominus}\big(I + (-U^{(-1)}_{\ominus\ominus})^{-1}|\tilde{C}^{-1}_{\ominus}|A_{\ominus\oplus}(-K^{(-1)}_{\oplus\oplus})^{-1}\tilde{C}^{-1}_{\oplus}A_{\oplus\ominus}\big)(-A^{-1}_{\ominus\ominus}) \end{aligned}$$

We use (51), to write the second term as

$$\begin{aligned} &\Psi_{\oplus\ominus}(-U^{(-1)}_{\ominus\ominus})^{-1}|\tilde{C}^{-1}_{\ominus}|\big(A_{\ominus\oplus}(-K^{(-1)}_{\oplus\oplus})^{-1}\tilde{C}^{-1}_{\oplus}A_{\oplus\ominus} \\ &\qquad\qquad + (-A_{\ominus\ominus} - A_{\ominus\oplus}\Psi_{\oplus\ominus})\big)(-A^{-1}_{\ominus\ominus}) \\ &\quad = \Psi_{\oplus\ominus}(-U^{(-1)}_{\ominus\ominus})^{-1}|\tilde{C}^{-1}_{\ominus}| \\ &\qquad \times \big(I - A_{\ominus\oplus}(-K^{(-1)}_{\oplus\oplus})^{-1}(\tilde{C}^{-1}_{\oplus}A_{\oplus\ominus} - K^{(-1)}_{\oplus\oplus}\Psi_{\oplus\ominus})(-A^{-1}_{\ominus\ominus})\big) \\ &\quad = \Psi_{\oplus\ominus}(-U^{(-1)}_{\ominus\ominus})^{-1}|\tilde{C}^{-1}_{\ominus}|\big(I - A_{\ominus\oplus}(-K^{(-1)}_{\oplus\oplus})^{-1}\Psi_{\oplus\ominus}|\tilde{C}_{\ominus}|\big) \end{aligned}$$

were we used (55) to replace $K^{(-1)}_{\oplus\oplus}\Psi_{\oplus\ominus}$. We find thus $B_{\oplus\ominus} = D_{\oplus\ominus}$.

Finally, we use the definition of $U^{(-1)}_{\ominus\ominus}$ to write

$$\begin{aligned} B_{\ominus\ominus} &= -(A_{\ominus\ominus} + A_{\ominus\oplus}\Psi_{\oplus\ominus} - A_{\ominus\oplus}(-A^{-1}_{\oplus\oplus})\tilde{C}_{\oplus}\Psi_{\oplus\ominus}U^{(-1)}_{\ominus\ominus})^{-1} \\ &= (-U^{(-1)}_{\ominus\ominus})^{-1}|\tilde{C}^{-1}_{\ominus}|(I - A_{\ominus\oplus}(-A^{-1}_{\oplus\oplus})\tilde{C}_{\oplus}\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}|)^{-1} \end{aligned}$$

We write

$$\begin{aligned} &(I - A_{\ominus\oplus}(-A^{-1}_{\oplus\oplus})\tilde{C}_{\oplus}\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}|)^{-1} \\ &\qquad = I + A_{\ominus\oplus}\big(I - (-A^{-1}_{\oplus\oplus})\tilde{C}_{\oplus}\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}|A_{\ominus\oplus}\big)^{-1} \\ &\qquad\qquad \times (-A^{-1}_{\oplus\oplus})\tilde{C}_{\oplus}\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}| \\ &\qquad = I + A_{\ominus\oplus}\big(-A_{\oplus\oplus} - \tilde{C}_{\oplus}\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}|A_{\ominus\oplus}\big)^{-1} \\ &\qquad\qquad \times \tilde{C}_{\oplus}\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}| \\ &\qquad = I + A_{\ominus\oplus}(-K^{(-1)}_{\oplus\oplus})^{-1}\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}| \end{aligned}$$

by (52), so that $B_{\ominus\ominus} = D_{\ominus\ominus}$.

We find the block $\Psi^{(1)}_{\oplus-}$ of $\Psi^{(1)}$ given in (59) by observing the terms in ε^0 in (48). From (47), we obtain the Sylvester equation (59) for $\Psi^{(1)}_{\oplus\ominus}$. Taking the terms in ε in (45) and (47) leads respectively to (60) and (61). □

Remark 5. Not surprisingly, $\Psi_{+\ominus} = 0$, as we found in (33).

As in Section 3.2, (54) is a function of Ψ but also of the supplementary component $\Psi_{\oplus\ominus}$. This generalizes $\Psi_{\oplus-}$ given in (23). There is a probabilistic interpretation similar to the one given in (23), with, here, a correction term due to the introduction of $\mathcal{S}_{\ominus}$: $[\Psi_{\oplus-}]_{ij}$ is the sum of

- $[(-K^{(-1)}_{\oplus\oplus})^{-1}\tilde{C}^{-1}_{\oplus}A_{\oplus-}]_{ij}$, the probability that the phase process goes from i to j, after some time spent in phases of $\mathcal{S}_{\oplus}$ or $\mathcal{S}_{\ominus}$,
- $[(-K^{(-1)}_{\oplus\oplus})^{-1}\tilde{C}^{-1}_{\oplus}A_{\oplus+}\Psi]_{ij}$, the probability that the process leaves i for a phase in $\mathcal{S}_{+}$ and later returns to the initial level in j,
- $[(-K^{(-1)}_{\oplus\oplus})^{-1}\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}|A_{\ominus-}]_{ij}$ the probability that the process comes back to the initial level in a phase of $\mathcal{S}_{\ominus}$ and goes to j,
- $[(-K^{(-1)}_{\oplus\oplus})^{-1}\Psi_{\oplus\ominus}|\tilde{C}^{-1}_{\ominus}|A_{\ominus+}\Psi]_{ij}$ the process comes back to the initial level in a phase of $\mathcal{S}_{\ominus}$, goes to a phase of $\mathcal{S}_{+}$ and later returns to the initial level in j,

for $i \in \mathcal{S}_{\oplus}$, $j \in \mathcal{S}_{-}$.

Remark 6. Higher order terms (in particular, the coefficients of ε^2) may be of interest in some cases. It is clear that the principal difficulty lies in the necessity to deal with calculations that are steadily more cumbersome, but no more. We expect that coefficients of $\Psi_{+\ominus}$ or $\Psi_{\oplus-}$ will be given explicitly and that each successive coefficients of Ψ_{+-} and $\Psi_{\oplus\ominus}$ will be solutions of Sylvester equations.

4. IMPACT ON THE STATIONARY PROBABILITY

For $j \in \mathcal{S}$ and $x \in \mathbb{R}^+$, we define the joint distribution function of the level and the phase at time t, $F_j(x,t) = \mathbb{P}\left[X\left(t\right) \le x, \varphi\left(t\right) = j\right]$, and its density by

$$f_j(x,t) = \frac{\partial}{\partial x}F_j(x,t), \quad \text{with} \quad f_j(0,t) = \lim_{x\to 0} f_j(x,t).$$

The stationary density vector $\boldsymbol{\pi}(x) = (\pi_j(x) : j \in \mathcal{S})$ of the fluid model, where, for $j \in \mathcal{S}$, $\pi_j(x) = \lim_{t\to\infty} f_j(x,t)$, exists if and only if the mean stationary drift is negative, that is, if and only if $\sum_{i\in\mathcal{S}} \xi_i c_i < 0$, where ξ_i is defined in (4) for all i. When the mean stationary drift of the fluid model is negative, from Govorun *et al.* [6], we have, for $x > 0$,

$$\boldsymbol{\pi}\,(x) = \boldsymbol{q}e^{Kx}\left[\ C_+^{-1}\,;\Psi\,|C_-|^{-1}\ ;\Theta\ \right], \tag{63}$$

and the mass at zero is $[0\,;\boldsymbol{p}_-\ ;\boldsymbol{p}_0]$ where

$$K = C_+^{-1}Q_{++} + \Psi\,|C_-|^{-1}\,Q_{-+}, \tag{64}$$

$$\Theta = \left(C_+^{-1}A_{+0} + \Psi\,|C_-|^{-1}\,A_{-0}\right)(A_{00})^{-1}, \tag{65}$$

$$\boldsymbol{q} = \boldsymbol{p}_-A_{-+} + \boldsymbol{p}_0A_{0+} \tag{66}$$

and $[\boldsymbol{p}_-\ ;\boldsymbol{p}_0]$ is the unique solution of the system

$$[\boldsymbol{p}_-\ ;\boldsymbol{p}_0]\begin{bmatrix}A_{--}+A_{-+}\Psi & A_{-0}\\ A_{0-}+A_{0+}\Psi & A_{00}\end{bmatrix} = \mathbf{0} \tag{67}$$

$$[\boldsymbol{p}_-\ ;\boldsymbol{p}_0]\mathbf{1} + \boldsymbol{q}_-(-K)^{-1}(C_+^{-1} + \Psi\,|C_-|^{-1} + \Theta)\mathbf{1} = 1. \tag{68}$$

Expression (63) is numerically stable and has a physical interpretation (da Silva Soares [38, Chapter 1, Section 1.3]). Furthermore, it appears clearly that all the quantities appearing in the expression of the stationary density are functions of Ψ.

The stationary density of (10) may be formulated as

$$\boldsymbol{\pi}\,(x,\varepsilon) = \boldsymbol{q}(\varepsilon)e^{K(\varepsilon)x}\left[\ C_+^{-1}\,;\Psi(\varepsilon)\,|C_-|^{-1}\ ;\Theta(\varepsilon)\ \right], \tag{69}$$

where $K(\varepsilon)$, $\Theta(\varepsilon)$ and $\boldsymbol{q}(\varepsilon)$ are defined similary to (64),(65) and (66) respectively. It is well known that the stationary density vector $\boldsymbol{\pi}(x,\varepsilon)$ is differentiable (see Kato [9, Section 2]) and such that $\boldsymbol{\pi}(x,\varepsilon)$ may be written as

$$\boldsymbol{\pi}(x,\varepsilon) = \boldsymbol{\pi}(x) + \varepsilon\boldsymbol{\pi}^{(1)}(x,0) + O(\varepsilon^2),$$

where

$$\boldsymbol{\pi}^{(1)}(x,0) = \lim_{\varepsilon\to 0}\frac{\boldsymbol{\pi}(x,\varepsilon) - \boldsymbol{\pi}(x,0)}{\varepsilon}, \tag{70}$$

for all $x \in \mathbb{R}^+$. We find

$$\begin{aligned}\boldsymbol{\pi}^{(1)}(x,0) &= \boldsymbol{q}e^{Kx}\left[\ 0\,;\ \Psi^{(1)}\left|C_-^{-1}\right|\ ;\ \Theta^{(1)}\ \right]\\ &\quad + (\boldsymbol{q}^{(1)}e^{Kx} + \boldsymbol{q}L^{(1)}(x))\left[\ C_+^{-1}\,;\Psi\,|C_-|^{-1}\ ;\Theta\ \right],\end{aligned}$$

where $\Psi^{(1)}$ is given in Theorem 1 and

$$\begin{aligned}\Theta^{(1)} &= (C_+^{-1}A_{+0} + \Psi|C_-^{-1}|A_{-0})(-A_{00}^{-1})\tilde{A_{00}}A_{00}^{-1}\\ &\quad + C_+^{-1}\tilde{A}_{+0} + \Psi^{(1)}|C_-^{-1}|A_{-0} + \Psi|C_-^{-1}|\tilde{A}_{-0}.\end{aligned}$$

The vector $\boldsymbol{q}(\varepsilon)$ is differentiable by Kato [9, Section 2] and

$$\boldsymbol{q}^{(1)} = \boldsymbol{p}_-^{(1)}A_{-+} + \boldsymbol{p}_0^{(1)}A_{0+} + \boldsymbol{p}_-\tilde{A}_{-+} + \boldsymbol{p}_0^{(1)}\tilde{A}_{0+}$$

with

$$\begin{aligned}\left[\boldsymbol{p}_-^{(1)}\ ;\boldsymbol{p}_0^{(1)}\right] = -\left[\boldsymbol{p}_-\ ;\boldsymbol{p}_0\right]&\begin{bmatrix}\tilde{A}_{--}+\tilde{A}_{-+}\Psi+A_{-+}\Psi^{(1)} & \tilde{A}_{-0}\\ \tilde{A}_{0-}+\tilde{A}_{0+}\Psi+A_{0+}\Psi^{(1)} & \tilde{A}_{00}\end{bmatrix}\\ &\begin{bmatrix}A_{--}+A_{-+}\Psi & A_{-0}\\ A_{0-}+A_{0+}\Psi & A_{00}\end{bmatrix}^{\#} + c\boldsymbol{\pi}(x),\end{aligned} \tag{71}$$

where $M^{\#}$ denotes the group inverse of the matrix M. We find (71) by solving the Poisson equation (see Meyer [12]) satsified by $[\boldsymbol{p}_-^{(1)}\ ;\boldsymbol{p}_0^{(1)}]$, deduced from (67), where c is a normalisation found through (68). Finally,

$$L^{(1)}(x) = \int_0^x e^{K(x-s)}K^{(1)}e^{Ks}\mathrm{d}s,$$

where $K^{(1)} = C_+^{-1}\tilde{Q}_{++} + \Psi^{(1)}|C_-^{-1}|Q_{-+} + \Psi|C_-^{-1}|\tilde{Q}_{-+}$. In order to actually compute $L^{(1)}(x)$, we refer the reader to Higham [8, Theorem 10.13, Equation (10.17a)].

5. REFERENCES

[1] N. Antunes, C. Fricker, F. Guillemin, and P. Robert. Perturbation analysis of a variable M/M/1 queue: A probabilistic approach. *Advances in Applied Probability*, 38(1):263–283, 2006.

[2] S. Asmussen. Stationary distributions for fluid flow models with or without Brownian noise. *Communications in Statistics. Stochastic Models*, 11(1):21–49, 1995.

[3] N. G. Bean, M. M. O'Reilly, and P. G. Taylor. Algorithms for return probabilities for stochastic fluid flows. *Stochastic Models*, 21(1):149–184, 2005.

[4] D. A. Bini, B. Iannazzo, G. Latouche, and B. Meini. On the solution of algebraic Riccati equations arising in fluid queues. *Linear Algebra and its Applications*, 413(2):474–494, 2006.

[5] X.-R. Cao and H.-F. Chen. Perturbation realization, potentials, and sensitivity analysis of Markov processes. *IEEE Transactions on Automatic Control*, 42(10):1382–1393, 1997.

[6] M. Govorun, G. Latouche, and M.-A. Remiche. Stability for fluid queues: Characteristic inequalities. *Stochastic Models*, 29(1):64–88, 2013.

[7] B. Heidergott, A. Hordijk, and N. Leder. Series expansions for continuous-time Markov processes. *Operations Research*, 58(3):756–767, 2010.

[8] N. J. Higham. *Functions of Matrices: Theory and Computation.* SIAM, 2008.

[9] T. Kato. *Perturbation Theory for Linear Operators*, volume 132. Springer Science & Business Media, 2013.

[10] P. Lancaster and M. Tismenetsky. *The Theory of Matrices.* Computer Science and Applied Mathematics. Academic Press, 1985.

[11] R. Loynes. A continuous-time treatment of certain queues and infinite dams. *Journal of the Australian Mathematical Society*, 2(04):484–498, 1962.

[12] C. D. Meyer, Jr. The role of the group generalized inverse in the theory of finite Markov chains. *SIAM Review*, 17(3):443–464, 1975.

[13] V. Ramaswami. Matrix analytic methods for stochastic fluid flows. In D. Smith and P. Hey, editors, *Teletraffic Engineering in a Competitive World (Proceedings of the 16th International Teletraffic Congress)*, pages 1019–1030. Elsevier Science B.V., Edinburgh, UK, 1999.

[14] L. Rogers. Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. *The Annals of Applied Probability*, pages 390–413, 1994.

[15] J. Xue, S. Xu, and R.-C. Li. Accurate solutions of M-matrix algebraic Riccati equations. *Numerische Mathematik*, 120(4):671–700, 2012.

Exact Stationary Tail Asymptotics for a Markov Modulated Two-Demand Model — In Terms of a Kernel Method

Yuanyuan Liu
School of Mathematics and Statistics
New Campus
Central South University
Changsha, Hunan
China, 410083
liuyy@csu.edu.cn

Pengfei Wang
School of Mathematics and Statistics
New Campus
Central South University
Changsha, Hunan
China, 410083
wangpengfei_csu@163.com

Yiqiang Q. Zhao
School of Mathematics and Statistics
Carleton University
Ottawa, Ontario
Canada, K1S 5B6
zhao@math.carleton.ca

ABSTRACT

In this paper, we extend the kernel method (for scalar random walks in the quarter plane) to study exact tail asymptotic properties in the stationary distribution for a Markov modulated random walk in the quarter plane. Specifically, we demonstrate how the extended kernel method works by a Markov modulated two-demand queueing model, or a two-demand model with a Markovian arrival process. The key ideal in this extension is to decompose the block-form fundamental form to a scalar one, for which the asymptotic property is the same as that for the modulated walk that can be obtained through recently developed kernel method techniques. We should point it out that research in this paper is not complete and only contains preliminary results. It is our plan to continue our study on this extension of the kernel method and to make the study more comprehensive.

Keywords

random walks in the quarter plane; Markov modulated random walks in the quarter plane; stationary distribution; generating function; kernel method; singularity analysis; exact tail asymptotics

1. INTRODUCTION

The standard kernel method was originated from analytic combinatorics [10, 2], which is a very efficient approach to study tail asymptotic properties in a sequence of nonnegative numbers, including stationary probability sequences in one-dimensional models. This method deals with a functional equation for the unknown generating functions through analysis of the so-called kernel equation using one branch of an algebraic function determined by the kernel equation. Analytic continuation and singularity analysis are two key components in the method before one can finally obtain exact tail asymptotic properties (through a Tauberian-like theorem, for example Theorem 4.1 or Theorem 4.2 in [14], a consequence of Corollary VI.1 or Theorem VI.5, respectively, in [5]). For one-dimensional models, this is relatively simple since the resulting functional equation defined by the kernel equation contains only one unknown generating function.

However, when the key idea in the kernel method applies to two-dimensional queueing models, the functional equation defined by the kernel equation contains two unknown generating functions. The tail asymptotic analysis becomes much more challenging. A few successful methods, including the kernel method, are now available for so-called exact stationary tail asymptotics (for example, see the review paper [16]). As indicated above, in the kernel method, before we can apply the Tauberian-like theorem, analytic continuation of the unknown generating functions has to be established, and singularity analysis of the unknown functions at their dominant singulatiries has to be carried out. For two-dimensional models, readers may refer to [13, 14, 11] for more details.

In this paper, we further extend the kernel method to study exact tail asymptotic properties for Markov modulated random walks, i.e. the random walks in the quarter pane modulated by a finite-state Markov chain. Recall that for the (scalar) random walk in the quarter plane, the transitions from a state in the interior region, on the horizontal boundary, on the vertical boundary, and at the origin are characterized by the distributions $p_{i,j}$ and $p_{i,j}^{(k)}$ for $k = 1, 2, 0$, respectively, to a state with increments $i, j = 0, \pm 1$ in horizontal and vertical directions subject to the reflective boundary constrain. For the Markov modulated random walk, we can formulate the model as a Markov chain in such a way that $p_{i,j}$ and $p_{i,j}^{(k)}$ are now generalized to square matrices $A_{i,j}$ and $A_{i,j}^{(k)}$, respectively. Under a stability condition, let $\pi_{m,n;k}$ be the unique stationary probability vector. Our purpose is to study exact tail asymptotic properties in $\pi_{m,n,k}$ in terms of the kernel method. The same problem (with the focus on logarithmic asymptotics) has been studied by Ozawa [20] and Miyazawa [17] using a geometric method developed by Miyazawa [15] (also refer to Kobayashi and Miyazawa [8]). Tail asymptotic properties for specific cases of a Markov modulated random walks have also been considered using other methods, for example, Li and Zhao [12], Sakuma, Miyazawa and Zhao [21], Avrachenkov, Nain and Yechiali [1], and Song, Liu and Zhao [22].

The main contributions made in this paper include: (1) a set up of the key components (in block-form or vector-form) in the kernel method for the Markov modulated random walk; (2) in terms of matrix-analytic methods, convert the vector-form fundamental form to a scalar one (that has the same tail asymptotic properties as that for the vector-form model) such that results for (scalar) random walks can be used here; and (4) demonstrate how the extended kernel method works by a Markov modulated two-demand queueing model.

The rest of the paper is organized as follows: In Section 2, the Markov modulated random walk is introduced as discrete-time Markov chain; Section 3 establishes the block-form fundamental form for the modulated random walk, and obtain a decomposition result by converting the block-form fundamental form to a scalar form. We claim that the exact tail asymptotic properties for the probability sequences involved in the scalar fundamental form and in the block-form fundamental form are the same. Section 4 demonstrates how this extended kernel method works in terms of a Markov modulated two-demand model. The final section contains concluding remarks.

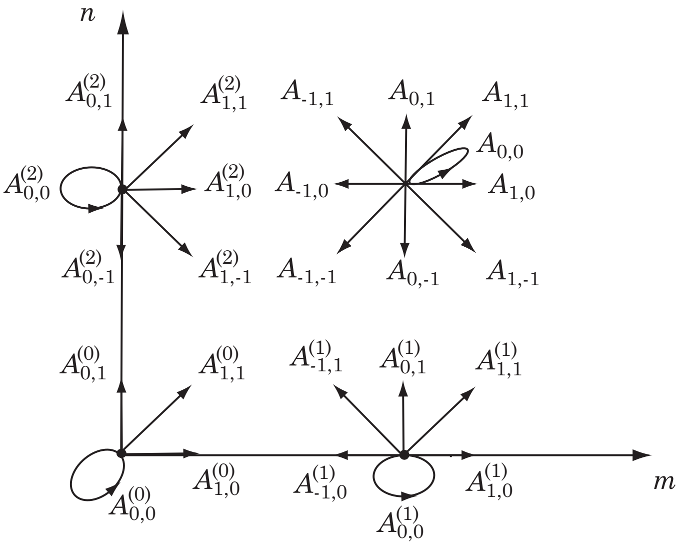


Figure 1: Transition diagram for Markov modulated random walks

2. MARKOV MODULATED RANDOM WALKS

The model studied in this paper is a Markov modulated random walk, which is a discrete-time Markov chain with state space $S = \{(m, n; l) : m, n = 0, 1, \ldots; l = 1, 2, \ldots, M\}$. The Markov modulated random walk can be considered as a generalization of the (scalar) random walk in the quarter plane, or random walks with two reflective boundaries. Specifically, the one-step transition probability distributions $p_{i,j}$ and $p^{(k)}_{i,j}$ $(k = 0, 1, 2)$ (for the scalar random walk) are now generalized to blocks $A_{i,j}$ and $A^{(k)}_{i,j}$ of transition probabilities (for the modulated one), where $A_{i,j}$ and $A^{(k)}_{i,j}$ are all nonnegative matrices of size $M \times M$, and $\sum_{i,j} A_{i,j}$ and $\sum_{i,j} A^{(k)}_{i,j}$ are stochastic. The transition diagram for the modulated random walk is depicted in Figure 1.

If we use the state variable m as level and n as background or phase, then the transition matrix P is given by:

$$P = \begin{pmatrix} B_0 & B_1 & & & \\ A_{-1} & A_0 & A_1 & & \\ & A_{-1} & A_0 & A_1 & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

where

$$B_i = \begin{pmatrix} A^{(0)}_{i,0} & A^{(0)}_{i,1} & & & \\ A^{(2)}_{i,-1} & A^{(2)}_{i,0} & A^{(2)}_{i,1} & & \\ & A^{(2)}_{i,-1} & A^{(2)}_{i,0} & A^{(2)}_{i,1} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

and

$$A_i = \begin{pmatrix} A^{(1)}_{i,0} & A^{(1)}_{i,1} & & & \\ A_{i,-1} & A_{i,0} & A_{i,1} & & \\ & A_{i,-1} & A_{i,0} & A_{i,1} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Similarly, if we use n as level and m as background, the transition matrix $\tilde{P}$ is given by:

$$\tilde{P} = \begin{pmatrix} \tilde{B}_0 & \tilde{B}_1 & & & \\ \tilde{A}_{-1} & \tilde{A}_0 & \tilde{A}_1 & & \\ & \tilde{A}_{-1} & \tilde{A}_0 & \tilde{A}_1 & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

where

$$\tilde{B}_i = \begin{pmatrix} A^{(0)}_{0,i} & A^{(0)}_{1,i} & & & \\ A^{(1)}_{-1,i} & A^{(1)}_{0,i} & A^{(1)}_{1,i} & & \\ & A^{(1)}_{-1,i} & A^{(1)}_{0,i} & A^{(1)}_{1,i} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

and

$$\tilde{A}_i = \begin{pmatrix} A^{(2)}_{0,i} & A^{(2)}_{1,i} & & & \\ A_{-1,i} & A_{0,i} & A_{1,i} & & \\ & A_{-1,i} & A_{0,i} & A_{1,i} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Our purpose is to study this model with the main focus on the exact tail asymptotic property in the stationary probability distribution $\pi_{m,n;k}$ under the stability condition of the system.

3. FUNDAMENTAL FORM AND KERNEL METHOD

It is well-known that the fundamental form (for example, (1.3) on page 151 in Cohen [3], or (1.3.6) in Fayolle, Iasnogorodski and Makysheve [4]) plays a key role in the analysis of a (scalar) random walk. Similarly, the following vector-form fundamental form plays a fundamental role in the analysis of a Markov modulated random walk.

For the Markov modulated random walk, we have the following (vector-form) fundamental form: for $|x| \leq 1$ and

$|y| \le 1$,

$$-\Pi(x,y)H(x,y) = \Pi_1(x)H_1(x,y)+\Pi_2(y)H_2(x,y)+\Pi_0 H_0(x,y), \tag{3.1}$$

where

$$H(x,y) = xy\left(I - \sum_{i=-1}^{1}\sum_{j=-1}^{1} x^i y^j A_{ij}\right),$$

$$H_1(x,y) = x\left(I - \sum_{i=-1}^{1}\sum_{j=0}^{1} x^i y^j A_{ij}^{(1)}\right),$$

$$H_2(x,y) = y\left(I - \sum_{i=0}^{1}\sum_{j=-1}^{1} x^i y^j A_{ij}^{(2)}\right),$$

$$H_0(x,y) = \left(I - \sum_{i=0}^{1}\sum_{j=0}^{1} x^i y^j A_{ij}^{(0)}\right),$$

$$\Pi(x,y) = \left(\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\pi_{i,j;1}x^{i-1}y^{j-1}, \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\pi_{i,j;2}x^{i-1}y^{j-1}, \dots, \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\pi_{i,j;M}x^{i-1}y^{j-1}\right)_{1\times M},$$

$$\Pi_1(x) = \left(\sum_{i=1}^{\infty}\pi_{i,0;1}x^{i-1}, \sum_{i=1}^{\infty}\pi_{i,0;2}x^{i-1}, \dots, \sum_{i=1}^{\infty}\pi_{i,0;M}x^{i-1}\right)_{1\times M},$$

$$\Pi_2(y) = \left(\sum_{j=1}^{\infty}\pi_{0,j;1}y^{j-1}, \sum_{j=1}^{\infty}\pi_{0,j;2}y^{j-1}, \dots, \sum_{j=1}^{\infty}\pi_{0,j;M}y^{j-1}\right)_{1\times M},$$

$$\Pi_0 = (\pi_{0,0;1}, \pi_{0,0;2}, \dots, \pi_{0,0;M})_{1\times M}.$$

Remark 3.1. *When $M = 1$, (3.1) is reduced to the (scalar) fundamental form (for example, (1.3) on page 151 in [3], which is the same as that in (1.3.6). In this case, $\Pi H = 0$ for Π finite is equivalent to the kernel equation $H = 0$. In the following, we establish the kernel equation for the vector case.*

We also mention here an analogous result to Theorem 1.3.1 of [4], which is used in Remark 3.2.

Theorem 3.1. *For the irreducible aperiodic Markov modulated random walk to be ergodic, if and only if there exist $\Pi(x,y)$, $\Pi_1(x)$ and $\Pi_2(y)$ holomorphic in $|x|$, $|y| < 1$, and a constant vector Π_0, satisfying the fundamental form (3.1) together with the l_1-condition $\sum_{m,n,j}|\pi_{m,n;k}| < \infty$. In this case, these functions are unique.*

In the standard kernel method, we study the kernel equation: $h = 0$. For the modulated model, we study the equation defined by $\det H(x,y) = 0$. For this purpose, we define

$$C(x,y) = \sum_{i=-1}^{1}\sum_{j=-1}^{1} x^i y^j A_{i,j}.$$

It is clear that if $\chi_k(x,y)$ is an eigenvalue of $C(x,y)$ then $xy(1-\chi_k(x,y)$ is an eigenvalue of $H(x,y)$. For our purpose, let us consider the Perron-Frobenius eigenvalue, denoted by $\chi(x,y)$. Define $\Gamma = \{(s_1,s_2) \in \mathbb{R}^2 : \chi(e^{s_1}, e^{s_2}) = 1\}$. We provide the following decomposition lemma.

Lemma 3.1. *For many enough (x,y) (containing a region in $|x| < 1$ and $|y| < 1$), $\det H(x,y) = 0$ can be factored as*

$$\begin{aligned}\det H(x,y) =& [a(x)y^2 + b(x)y + c(x)]q(x,y)\\ =& [\tilde{a}(y)x^2 + \tilde{b}(y)x + \tilde{c}(y)]q(x,y) = 0,\end{aligned}$$

where $a(x)$ ($\tilde{a}(y)$), $b(x)$ ($\tilde{b}(y)$) and $c(x)$ ($\tilde{c}(y)$) are polynomials of degree at most two, and $q(x,y)$ is a polynomial of x and y.

Furthermore, let $h(x,y) = a(x)y^2+b(x)y+c(x) = \tilde{a}(y)x^2+\tilde{b}(y)x+\tilde{c}(y)$, which corresponds to the kernel function of a (scalar) random walk, and let $x_2 = \min_{(x,y)\in\Gamma} x$ and $x_3 = \max_{(x,y)\in\Gamma} x$, then x_2 and x_3 are two branch points (or zeros of $D_1(x) = b(x)^2 - 4a(x)c(x)$) satisfying $0 < x_2 < 1 < x_3$ if the scalar random walk is non-singular (or $h(x,y)$ is irreducible and quadratic in both x and y). Similarly, let $y_2 = \min_{(x,y)\in\Gamma} y$ and $y_3 = \max_{(x,y)\in\Gamma} y$, then y_2 and y_3 are two branch points (or zeros of $D_2(y) = \tilde{b}(y)^2 - 4\tilde{a}(y)\tilde{c}(y)$) satisfying $0 < y_2 < 1 < y_3$ if the scalar random walk is non-singular.

Proof. First based on the results proved in [20], $\bar{\Gamma} = \{(s_1,s_2) \in \mathbb{R}^2 : \chi(e^{s_1}, e^{s_2}) \le 1\}$ is convex, and notice that $\chi(x,y)$ is continuous, we can conclude that Γ defines a simple closed curve. Furthermore, we conclude that for each $x_2 < x < x_3$, there are exactly two points on the curve Γ, while for $x = x_2$ and $x = x_3$ respectively, there is only one point on the curve. It means that

$$\det H(x,y) = xy(1 - \chi(x,y)\tilde{q}(x,y) = 0,$$

or for a fixed x, $\det H(x,y) = 0$, as a polynomial of y, has exactly two solutions since $\chi(x,y)$ is the Perron-Frobenius eigenvalue of $C(x,y)$. Therefore, we can write $\det H(x,y) = h(x,y)q(x,y) = 0$, where both $h(x,y)$ and $q(x,y)$ are polynomial. Furthermore, for each $x_2 < x < x_3$, $h(x,y) = 0$ has exactly two roots, denoted by $(x, Y_0(x))$ and $(x, Y_1(x)$ where Y_0 has the smaller modulus and Y_1 has the bigger modulus, and for $x = x_i$ ($i = 2,3$), $h(x,y) = 0$ has only one root, or $Y_0(x_i) = Y_1(x_i)$ for $i = 2,3$. It concludes that $h(x,y)$ is a polynomial of degree two and x_2 and x_3 are two branch points, or zeros of the discriminant $D_1(x)$. If the scalar random variable is non-singular, then both $x_2 \ne 1$ and $x_3 \ne 1$. We know $(1,1)$ is on Γ, which leads to $0 < x_2 < 1 < x_3$. □

Remark 3.2. *Consider pairs of $(x, Y_0(x))$ defined by $h(x,y) = 0$ given in the above lemma. We can argue that there are plenty of pairs (x,y) such that $\Pi(x,y)H(x,y) = 0$. To see it, we start with $\Pi_2(y)$ which is analytic in $|y| < 1$. Consider the function $\tilde{\Pi}_1(x)$ defined by $\Pi_1(x)H_1(x,Y_0(x)) + \Pi_2(Y_0(x))H_2(x,Y_0(x)) + \Pi_0 H_0(x,Y_0(x))$, which can be proved analytic in some region. We can then consider the function $\tilde{\Pi}(x,y)$ defined through the fundamental form and prove, through using Theorem 3.1 that $\tilde{\Pi}(x,y) = \Pi(x,y)$ and $\tilde{\Pi}_1(x) = \Pi_1(x)$. It implies that there are plenty of pairs (x,y) such that $\Pi(x,y)H(x,y) = 0$.*

Based on the above discussion, let us consider $\Pi(x,y)H(x,y) = 0$, for (x,y) such that $\Pi(x,y) \ne 0$. By linear algebra, we have $\det H(x,y) = 0$ for enough many (x,y). In this case, if $|\Pi(x,y)| < \infty$, we then obtain a relationship between two unknown vector-form functions Π_1 and Π_2. Therefore,

$\det H = 0$ plays a similar role to that played by $h = 0$ in the kernel method for scalar random walks. It leads to

$$\Pi_1(x)H_1(x,y) = -[\Pi_2(y)H_2(x,y) + \Pi_0 H_0(x,y)]. \quad (3.2)$$

Similarly,

$$\Pi_2(y)H_2(x,y) = -[\Pi_1(x)H_1(x,y) + \Pi_0 H_0(x,y)].$$

For the first equation, based on the fact that $Y_0(x)$ is analytic in the cut plane (for example, see [4]), $\Pi_1(x)$ can be analytically continued to a region with its dominant singularity equal to the minimum of x_3, the dominant singularity of $\Pi_2(Y_0(x))$ and the smallest zero in $(1, x_3]$ of $\det H_1(x, Y_0(x))$. For $\det H_1(x,y) = 0$, we establish the following decomposition property (similarly for $\det H_2 = 0$).

Lemma 3.2. $\det H_1(x,y) = 0$ *can be factored as*

$$\begin{aligned}\det H_1(x,y) =& [b_1(x)y + c_1(x)]q_1(x,y)\\ =& [\tilde{a}_1(y)x^2 + \tilde{b}_1(y)x + \tilde{c}_1(y)]q_1(x,y),\end{aligned}$$

or $h_1(x,y) = b_1(x)y + c_1(x) = \tilde{a}_1(y)x^2 + \tilde{b}_1(y)x + \tilde{c}_1(y)$ *is a polynomial of degree one in* y *and degree two in* x*. Similarly,*

$$\begin{aligned}\det H_2(x,y) =& [a_2(x)y^2 + b_2(x)y + c_2(x)]q_2(x,y)\\ =& [\tilde{b}_2(y)x + \tilde{c}_2(y)]q_2(x,y),\end{aligned}$$

or $h_2(x,y) = a_2(x)y^2 + b_2(x)y + c_2(x) = \tilde{b}_2(y)x + \tilde{c}_2(y)$ *is a polynomial of degree one in* x *and degree two in* y*.*

Proof. The proof follows from a similar argument to Lemma 3.1. Specifically, define

$$C_1(x,y) = \sum_{i=-1}^{1}\sum_{j=0}^{1} x^i y^j A_{i,j}^{(1)},$$

and denote by $\chi_1(x,y)$ the Perron-Frobenius eigenvalue of $C_1(x,y)$. Then, it follows from a similar argument to that in [20] that the region $\bar{\Gamma}_1 = \{(s_1,s_2) \in \mathbb{R}^2 : \chi_1(e^{s_1}, e^{s_2}) \leq 1\}$ is convex. Furthermore, we can show that for each x there is only one point on the curve $\Gamma_1 = \{(s_1,s_2) \in \mathbb{R}^2 : \chi_1(e^{s_1}, e^{s_2}) = 1\}$ and for each y there are exactly two points on the curve, which implies that $\det H_1(x,y) = h_1(x,y)q_1(x,y) = 0$. The other result in the lemma can be proved in parallel. □

Now, consider the asymptotic property of the kth component function in $\Pi_1(x)$ defined by equation (3.2). Based on the above discussion, its asymptotic property as x goes to its dominant singularity is either the branch point x_3, or a pole, which is the smallest zero of $\det H_1(x, Y_0(x))$, or the smallest one of the dominant singularities of all component functions in $\Pi_2(Y_0(x))$. A similar argument applies to $\Pi_2(y)$ to conclude that the dominant singularity of the kth component function in $\Pi_1(x)$ is independent of k and the same as that for the function $\pi_1(x)$, determined by:

$$h_1(x,y)\pi_1(x) = -[\Pi_2(y)H_2(x,y) + \Pi_0 H_0(x,y)]\big|_i, \quad (3.3)$$

where the right hand side is the ith component of the right hand side vector in (3.2) for any i. For this component function, its asymptotic property at its dominant singularity is the same as as that for the function $\pi_2(y)$, determined by:

$$h_2(x,y)\pi_2(y) = -[h_1(x,y)\pi_2(y) + h_0(x,y)\pi_{0,0}],$$

where $h_0(x,y)\pi_{0,0}$ is the ith component of $\Pi_0 H_0(x,y)$ that will not impact on the asymptotic property of $\pi_2(y)$. As a summary, the above discussion convert the tail asymptotic problem for $\Pi_1(x)$ and $\Pi_2(y)$ (and therefore for $\Pi(x,y)$) to the tail asymptotic problem for $\pi_1(x)$ and $\pi_2(y)$ (and therefore for $\pi_{(}x,y)$), respectively, defined by the following scalar fundamental form for a random walk in the quarter plane:

$$h(x,y)\pi(x,y) = h_1(x,y)\pi_1(x) + h_2(x,y)\pi_2(y) + h_0(x,y)\pi_{0,0}. \quad (3.4)$$

As a consequence, the modulated walk and the above scalar walk have the same stability condition.

4. A MARKOV MODULATED TWO-DEMAND QUEUEING MODEL

In this section, we consider a Markov modulated two-demand model to demonstrate how the extended kernel method works.

4.1 Model description

The Markov modulated two-demand model considered here is a generalization of the (standard) two-demand model studied in Flatto and Hahn [6]. This model differs from the standard one in its arrival process. Instead of the Poisson process in the standard model, we assume that the arrival rate is λ_k when the modulating Markov chain is in state k. For convenience, we only consider a two-state Markov chain (state 0 and state 1) with the transition matrix given by

$$J = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} p & \bar{p} \\ \bar{q} & q \end{bmatrix} \end{array},$$

where $\bar{a} = 1 - a$, and $0 < p, q < 1$ to avoid triviality. Equivalently, this is a model with a Markovian arrival process, each arrival creates simultaneously two jobs to two parallel queues served by two exponential servers with rates μ_1 and μ_2, respectively. The Markovian arrival process is characterized by

$$D_1 = \begin{bmatrix} \lambda_0 p & \lambda_0 \bar{p} \\ \lambda_1 \bar{q} & \lambda_1 q \end{bmatrix}, \quad D_0 = \begin{bmatrix} -\lambda_0 & 0 \\ 0 & -\lambda_1 \end{bmatrix}.$$

Let $Q_i(t)$ be the number of jobs in queue i for $i = 1, 2$, including the job in service if there is any, and let $J(t)$ be the phase of the Markovian arrival process. Then, $\{(Q_1(t), Q_2(t); J(t)) : t \geq 0\}$ is a continuous-time Markov chain. Upon uniformization (assuming that $\lambda_0 + \lambda_1 + \mu_1 + \mu_2 = 1$), the discrete-time Markov chain is an example of the Markov modulated random walk considered in this paper. Specifically, we have

$$A_{1,1} = \begin{bmatrix} \lambda_0 p & \lambda_0 \bar{p} \\ \lambda_1 \bar{q} & \lambda_1 q \end{bmatrix}, \quad A_{0,-1} = \mu_2 \begin{bmatrix} p & \bar{p} \\ \bar{q} & q \end{bmatrix},$$

$$A_{-1,0} = \mu_1 \begin{bmatrix} p & \bar{p} \\ \bar{q} & q \end{bmatrix}, \quad A_{0,0} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_0 \end{bmatrix},$$

and

$$A_{0,0}^{(0)} = \begin{bmatrix} 1 - \lambda_0 & 0 \\ 0 & 1 - \lambda_1 \end{bmatrix}, \quad A_{0,0}^{(1)} = \begin{bmatrix} \lambda_1 + \mu_2 & 0 \\ 0 & \lambda_0 + \mu_2 \end{bmatrix},$$

$$A_{0,0}^{(2)} = \begin{bmatrix} \lambda_1 + \mu_1 & 0 \\ 0 & \lambda_0 + \mu_1 \end{bmatrix},$$

$A_{1,1}^{(0)} = A_{1,1}^{(1)} = A_{1,1}^{(2)} = A_{1,1}$, $A_{-1,0}^{(1)} = A_{-1,0}$ and $A_{0,-1}^{(2)} = A_{0,-1}$.

4.2 Stability condition

For a stability condition, solve

$$\pi D = 0, \quad \pi_0 + \pi_1 = 1,$$

to have

$$\pi_0 = \frac{\lambda_1 \bar{q}}{\lambda_1 \bar{q} + \lambda_0 \bar{p}}, \quad \pi_1 = \frac{\lambda_0 \bar{p}}{\lambda_1 \bar{q} + \lambda_0 \bar{p}}.$$

The fundamental arrival rate for the arrival process is given by $\lambda^* = \pi D_1 e$. When $p = q = 1/2$,

$$\lambda^* = \pi D_1 e = \frac{2\lambda_0 \lambda_1}{\lambda_0 + \lambda_1}.$$

By standard stability analysis for queues, we obtain the following assertion.

Theorem 4.1. *The Markov modulated two-demand queueing model is stable if and only if $\lambda^* < \min\{\mu_1, \mu_2\}$. When $p = q = 1/2$, the condition is simplified to*

$$\frac{2\lambda_0 \lambda_1}{\lambda_0 + \lambda_1} < \min\{\mu_1, \mu_2\}.$$

4.3 Reduced to a scalar random walk

We first calculate $\det H(x, y)$.

$$H(x,y) = \left[xy\big(I - \sum_i \sum_j x^i y^j A_{i,j}\big) \right]$$

$$= \begin{bmatrix} xy(1-\lambda_1) - p g_0(x,y) & -\bar{p} g_0(x,y) \\ -\bar{q} g_1(x,y) & xy(1-\lambda_0) - q g_1(x,y) \end{bmatrix},$$

where

$$g_0(x,y) = x^2 y^2 \lambda_0 + x\mu_2 + y\mu_1$$

and

$$g_1(x,y) = x^2 y^2 \lambda_1 + x\mu_2 + y\mu_1.$$

For simplicity, assume $p = q = 1/2$, which leads to

$$\det H(x,y) = \left(\frac{1}{2}\right)^2$$

$$\det \begin{bmatrix} 2xy(1-\lambda_1) - g_0(x,y) & -g_0(x,y) \\ -g_1(x,y) & 2xy(1-\lambda_0) - g_1(x,y) \end{bmatrix}$$

$$= -\frac{x^2 y^2}{2} h(x,y),$$

where

$$\begin{aligned} h(x,y) = & [\lambda_0(1-\lambda_0) + \lambda_1(1-\lambda_1)]x^2y^2 - 2(1-\lambda_0)(1-\lambda_1)xy \\ & + [(1-\lambda_0) + (1-\lambda_1)](\mu_2 x + \mu_1 y). \end{aligned}$$

Remark 4.1. *If $\lambda_0 = \lambda_1$, then $\det H(x, y) = 0$ becomes $\lambda_0 x^2 y^2 - (1-\lambda_0)xy + (\mu_2 x + \mu_1 y) = 0$, which degenerates to the kernel equation for the standard two-demand model (by noticing the difference of the normalizing constant: here $\lambda_0 + \lambda_1 + \mu_1 + \mu_2 = 1$ is assumed instead of $\lambda + \mu_1 + \mu_2 = 1$ for the standard two-demand model, where λ is the arrival rate for the standard two-demand model).*

Remark 4.2. *To see how $xy(1-\chi(x,y)) = 0$ and $h(x,y) = 0$ are related, we calculated the two eigenvalues for $C(x,y)$, which are given by*

$$\chi(x,y), \phi(x,y) = \frac{(a+d) \pm \sqrt{\Delta(x,y)}}{2},$$

where $a = xy\lambda_1 + g_0(x,y)/2$, $b = -g_0(x,/2)$, $c = -g_1(x,y)/2$, $d = xy\lambda_0 + g_1(x,y)/2$, and $\Delta(x,y) = (a+d)^2 - 4(ad-bc)$. We can observe that $xy(1-\chi(x,y)$ is not a polynomial and $h(x,y) = 2xy(1-\chi(x,y))(1-\phi(x,y))$.

We then calculate $\det H_1(x, y)$.

$$H_1(x,y) =$$

$$\begin{bmatrix} x(\lambda_0+\mu_1) - (\lambda_0 x^2 y + \mu_1)p & -(\lambda_0 x^2 y + \mu_1)\bar{p} \\ -(\lambda_1 x^2 y + \mu_1)\bar{q} & x(\lambda_1+\mu_1) - (\lambda_1 x^2 y + \mu_1)q \end{bmatrix}$$

with

$$\det H_1(x,y) =$$

$$\begin{aligned} & (pq - \bar{p}\bar{q})\lambda_0\lambda_1 y^2 x^4 - \big[q(\lambda_0+\mu_1)\lambda_1 + p(\lambda_1+\mu_1)\lambda_0\big] y x^3 \\ & + \big[(\lambda_0+\mu_1)(\lambda_1+\mu_1) + (pq - \bar{p}\bar{q})\mu_1(\lambda_+\lambda_1)y\big] x^2 \\ & - \big[q(\lambda_0+\mu_1) + p(\lambda_1+\mu_1)\big]\mu_1 x + \mu_1^2(pq - \bar{p}\bar{q}). \end{aligned}$$

When $p = q = 1/2$, we have

$$\det H_1(x,y) = \left(-\frac{x}{2}\right) h_1(x,y),$$

where

$$\begin{aligned} h_1(x,y) = & \big[(\lambda_0+\mu_1)\lambda_1 + (\lambda_1+\mu_1)\lambda_0\big] y x^2 - 2(\lambda_0+\mu_1)(\lambda_1+\mu_1)x \\ & + \big[(\lambda_0+\mu_1) + (\lambda_1+\mu_1)\big]\mu_1. \end{aligned}$$

Finally, we calculate $\det H_2(x, y)$.

$$H_2(x,y) =$$

$$\begin{bmatrix} y(\lambda_0+\mu_2) - (\lambda_0 x y^2 + \mu_2)p & -(\lambda_0 x y^2 + \mu_2)\bar{p} \\ -(\lambda_1 x y^2 + \mu_2)\bar{q} & y(\lambda_1+\mu_2) - (\lambda_1 x y^2 + \mu_2)q \end{bmatrix}$$

with

$$\det H_2(x,y) =$$

$$\begin{aligned} & (pq - \bar{p}\bar{q})\lambda_0\lambda_1 x^2 y^4 - \big[q(\lambda_0+\mu_2)\lambda_1 + p(\lambda_1+\mu_2)\lambda_0\big] x y^3 \\ & + \big[(\lambda_0+\mu_2)(\lambda_1+\mu_2) + (pq - \bar{p}\bar{q})\mu_2(\lambda_+\lambda_1)x\big] y^2 \\ & - \big[q(\lambda_0+\mu_2) + p(\lambda_1+\mu_2)\big]\mu_2 x + \mu_2^2(pq - \bar{p}\bar{q}). \end{aligned}$$

When $p = q = 1/2$, we have

$$\det H_2(x,y) = \left(-\frac{y}{2}\right) h_2(x,y),$$

where

$$\begin{aligned} h_2(x,y) = & \big[(\lambda_0+\mu_2)\lambda_1 + (\lambda_1+\mu_2)\lambda_0\big] x y^2 - 2(\lambda_0+\mu_2)(\lambda_1+\mu_2)y \\ & + \big[(\lambda_0+\mu_2) + (\lambda_1+\mu_2)\big]\mu_2. \end{aligned}$$

The exact tail asymptotic property for the Markov modulated walk is equivalent to the property of the following (scalar) walk given in (3.4) defined by

$$
\begin{aligned}
p_{1,1} =&\lambda_0(1-\lambda_0)+\lambda_1(1-\lambda_1),\\
p_{0,-1} =&[(1-\lambda_0)+(1-\lambda_1)]\mu_2,\\
p_{-1,0} =&[(1-\lambda_0)+(1-\lambda_1)]\mu_1,\\
p_{0,0} =&1-2(1-\lambda_0)(1-\lambda_1),\\
p^{(1)}_{1,1} =&\lambda_0(\lambda_1+\mu_1)+\lambda_1(\lambda_0+\mu_1),\\
p^{(1)}_{-1,0} =&[(\lambda_1+\mu_1)+(\lambda_0+\mu_1)]\mu_1,\\
p^{(1)}_{0,0} =&1-2(\lambda_0+\mu_1)(\lambda_1+\mu_1)\\
p^{(2)}_{1,1} =&\lambda_0(\lambda_1+\mu_2)+\lambda_1(\lambda_0+\mu_2),\\
p^{(2)}_{0,-1} =&[(\lambda_0+\mu_2)+(\lambda_1+\mu_2)]\mu_2,\\
p^{(2)}_{0,0} =&1-2(\lambda_0+\mu_2)(\lambda_1+\mu_2).
\end{aligned}
$$

4.4 Exact tail asymptotic property

Based on the decomposition results, for fixed k, $\pi_{m,0;k}$ and $\pi_{0,n;k}$ (for the modulated model|) have the same exact tail asymptotic property for $\pi_{m,0}$ and $\pi_{0,n}$ (for the scalar random walk obtained based on the decomposition result), respectively. Let us consider the exact tail asymptotic property for $\pi_{m,0}$ (the tail asymptotic property for $\pi_{0,n}$ can be considered in parallel), which is determined based on the asymptotic property of the generating function $\pi_1(x)$ at it dominant singularity (through the Tauberian-like theorem). For a general scalar random walk, this dominant singularity x_{dom} can be a pole, either x^* or $\tilde{x}_1$, or a branch point x_3, where x^* is a zero of $h_1(x, Y_0(x)$, $\tilde{x}_1$ is such that $h_2(X_0(y), y) = 0$ with $y = Y_0(\tilde{x}_1)$. There are two important steps involved in the analysis: the first one to determine the existence of a finite solution $x^* > 1$ ($\tilde{x}_1 > 1$), and the next step is determine which candidate is the smallest one (or which ones are smallest).

Towards this end, for the modulated two-demand model, recall

$$a(x) = [\lambda_0(1-\lambda_0)+\lambda_1(1-\lambda_1)]\,x^2, \qquad (4.1)$$

$$b(x) = \mu_1(2-\lambda_0-\lambda_1) - 2(1-\lambda_0)(1-\lambda_1)x, \quad (4.2)$$

$$c(x) = \mu_2(2-\lambda_0-\lambda_1)x, \qquad (4.3)$$

and the discriminant $D_1(x) = b^2(x) - 4a(x)c(x)$, which is a cubic polynomial. We an first show:

Lemma 4.1. *$D_1(x)$ has three zeros: $0 < x_1 < x^* < x_2 < 1 < x_3 < +\infty$, i.e.*

$$D_1(x_i) = 0, \quad i = 1, 2, 3. \qquad (4.4)$$

Moreover, $D_1(x) > 0$ in $(-\infty, x_1)\bigcup(x_2, x_3)$ and $D_1(x) < 0$ in $(x_1, x_2)\bigcup(x_3, +\infty)$. Here,

$$x^* = \frac{\mu_1(2-\lambda_0-\lambda_1)}{2(1-\lambda_0)(1-\lambda_1)}$$

is the unique solution to $b(x) = 0$.

We are then expected to show (following a similar process to that in [13, 14]):

1. $h_1(x, Y_0(x))$ has a unique zero x^* that is greater than one;

2. $h_2(X_0(y), y)$ does not have any zero y such that $y = X_0(\tilde{x}_1)$ for some $\tilde{x}_1 > 1$.

Finally, based on which one is the dominant singularity, there are three types of tail asymptotic properties for $\pi_{m,0}$:

Type one: If $x^* < x_3$, then $\pi_{m,0} \sim c(1/x^*)^m$;

Type two: If $x_3 < x^*$, then $\pi_{m,0} \sim cm^{-3/2}(1/x_3)^m$;

Type three; If $x^* = x_3$, then $\pi_{m,0} \sim cm^{-1/2}(1/x^*)^m = cm^{-1/2}(1/x_3)^m$.

Unfortunately, there is no simple characterization for the three regions, on which type 1, 2 or 3 tail asymptotic property holds. However, for a specific set of system parameters, a simple numerical calculation will always tell the type of the tail property.

To end this section, we demonstrate how to show $h_1(x, Y_0(x))$ has a unique zero x^* that is greater than one. To convert $h_1(x, Y_0(x))$ to a polynomial, calculate:

$$\bar{\lambda}h_1(x, Y_0(x))h_1(x, Y_1(x)) = \mu_2(x{-}1)\left[Ax^2 + (A+B)x + C\right],$$

where

$$
\begin{aligned}
\bar{\lambda} =&p_{1,1} = \lambda_0(1-\lambda_0)+\lambda_1(1-\lambda_1)\\
A =&\big[\lambda_0(\lambda_1+\mu_1)+\lambda_1(\lambda_0+\mu_1)\big]^2(2-\lambda_0-\lambda_1),\\
B =&-4\big[\lambda_0(1-\lambda_0)(\lambda_1+\mu_1)+\lambda_1(1-\lambda_1)(\lambda_0+\mu_1)\big](\lambda_0+\mu_1)(\lambda_1+\mu\\
C =&-\ \big[(\lambda_0+\mu_1)+(\lambda_1+\mu_1)\big](\lambda_0-\lambda_1)^2\mu_1^2.
\end{aligned}
$$

The rest is now routine and tedious.

5. CONCLUDING REMARKS

In this paper, we demonstrated how to extend the kernel method to study tail asymptotic properties of a Markov modulated random walk, in terms of a modulated two-demand queueing model. It is interesting to compare this method to other available methods. Also, we should remark here that this paper contains results from our preliminary studies. It is our plan to continue this research to provide a comprehensive study of the extended kernel method with all details of the proof.

Acknowledgements: The authors thank all reviewers of an earlier version of this paper for their valuable comments and critics. This work was supported in part by National Natural Science Foundation of China (Grant No. 11571372) and the Innovation Program of Central South University, and a Discovery Grant from NSERC of Canada. The authors acknowledge the private communications [7] of Yanfei Gao about her preliminary studies on Markov modulated random walks.

6. REFERENCES

[1] Avrachenkov, K., Nain, P. and Yechiali, U. (2014) A retrial system with two input streams and two orbit queues, *Queueing Systems*, **77(1)**, 1-âĂŞ31.

[2] Banderier, C., Bousquet-Mélou, M, Denise, A, Flajolet, P., Gardy, D. and Gouyou-Beauchamps, D. (2002) Generating functions of generating trees, *Discrete Math.*, **246**, 29–55.

[3] Cohen, J.W. (1992) *Analysis of Random Walks*, IOS Press, Amsterdam.

[4] Fayolle, G., Iasnogorodski, R. and Malyshev, V. (1999) *Random Walks in the Quarter-Plane*, Springer, New York.

[5] Flajolet, F. and Sedgewick, R. (2009) *Analytic Combinatorics*, Cambridge University Press.

[6] Flatto, L. and Hahn, S. (1984) Two parallel queues created by arrivals with two demands I, *SIAM J. Appl. Math.*, **44**, 1041–1053.

[7] Gao, Y. (2012) Private communications.

[8] Kobayashi, M. and Miyazawa, M. (2013) Revisit to the tail asymptotics of the double QBD process: Refinement and complete solutions for the coordinate and diagonal directions, Chapter 8 in *Matrix-Analytic Methods in Stochastic Models*, 145–185, Springer.

[9] Kobayashi, M., Miyazawa, M. (2014) Tail asymptotics of the stationary distribution of a two dimensional reflecting random walk with unbounded upward jumps. *Advances in Applied Probability*, **46(2)**, 365–399.

[10] Knuth, D.E. (1969) *The Art of Computer Programming, Fundamental Algorithms*, vol. 1, second ed., Addison-Wesley.

[11] Li, H., Tavakoli, J. and Zhao, Y.Q. (2013) Analysis of exact tail asymptotics for singular random walks in the quarter plane. *Queueing Systems*, **74**, 151–179.

[12] Li, H. and Zhao, Y.Q. (2005) A retrial queue with a constant retrial rate, server break downs and impatient customers, *Stochastic Models*, **21**, 531–550.

[13] Li, H. and Zhao, Y.Q. (2011) Tail asymptotics for a generalized two-demand queueing model — A kernel method, *Queueing Systems*, **69**, 77–100.

[14] Li, H. and Zhao, Y.Q. (2011) A kernel method for exact tail asymptotics—Radom walks in the quarter plane, submitted. (arXiv:1505.04425)

[15] Miyazawa, M. (2009) Tail decay rates in double QBD processes and related reflected random walks, *Mathematics of Operations Research*, **34**, 547–575.

[16] Miyazawa, M. (2011) Light tail asymptotics in multidimensional reflecting processes for queueing networks, *TOP*, **19**, 233–299.

[17] Miyazawa, M. (2015) A superharmonic vector for a nonnegative matrix with QBD block structure and its application to a Markov modulated two dimensional reflecting process, *Queueing Systems*, **81**, 1–48.

[18] Ozawa, T. (2012) Ozawa, T. (2012). Positive recurrence and transience of multidimensional skip-free reflecting random walks with a background process, submitted. (arXiv: 1208.3043)

[19] Ozawa, T. (2013) Positive recurrence and transience of a two-station network with server states, submitted. (arXiv: 1308.6104)

[20] Ozawa, T. (2013) Asymptotics for the stationary distribution in a discrete-time two-dimensional quasi-birth-and-death process. *Queueing Systems*, **74**, 109-âĂŞ149.

[21] Sakuma, Y., Miyazawa, M. and Zhao, Y.Q. (2006) Decay rate for a PH/M/2 queue with shortest queue discipline, *Queueing Systems*, **53**, 189–201.

[22] Song, Y., Liu, Z. and Zhao, Y.Q. (2015) Exact tail asymptotics — Revisit of a retrial queue with two input streams and two orbits, accepted by *Ann. of Operations Research.*

Martingale decomposition for large queue asymptotics

[Extended Abstract]

Masakiyo Miyazawa
Tokyo University of Science
2641 Yamazaki, Noda
Chiba 276-8510, Japan
miyazawa@rs.tus.ac.jp

ABSTRACT

We consider asymptotic problems for large queues in the steady state. There are two types of asymptotics. One is large deviations for a fixed model. Another is weak convergence of a sequence of stationary distributions in heavy traffic, called heavy traffic approximation. In this paper, we focus on large deviations, and study tail asymptotic behaviors of the stationary joint queue length distribution of a generalized Jackson network. For this, we use an approach using martingale in [6].

1. INTRODUCTION

Asymptotic analyses have been actively studied in the recent queueing theory. This is because queueing models, particularly, queueing networks, become very complicated and their exact analyses are getting harder. We focus on asymptotic analyses for large queues, and aim to understand their limiting behaviors through their modeling primitives.

There are two different types of asymptotic analyses for large queues. One is large deviations, which is typically studied for a fixed model. Another is the limit of a sequence of queueing models in heavy traffic under appropriate scaling of time, space and/or modeling primitives. This gives a theoretical support for the limiting model, which is called heavy traffic approximation. In this paper, we focus on the large deviations. Among them, we are particularly interested in the tail asymptotic behaviors for the stationary distribution of a generalized Jackson network, GJN for short. Its heavy traffic approximation has been studied in [1].

Those two asymptotic problems have been studied separately in the literature. Recently, the author [6] studied a unified approach which is applicable to both asymptotic analyses. We use this unified approach, which uses a piecewise deterministic Markov process, PDMP for short, is used to describe queueing models.

A sample path of the PDMP is composed of two parts, deterministic continuous part and discontinuous part, called jumps, by which randomness is created. This PDMP is widely applicable, but known to be hard to analysis. Because of this, PDMP is often used for describing models, but rarely used for analytical study. So, other methods have been employed for analysis. Using phase type distributions, the state space can be discretized, and Markov chains are applicable for asymptotic analysis, which may be most popular in queueing theory.

Contrary to the analytical difficulty, the PDMP has a simple sample path. Its time evolution is easily presented by a stochastic integral equation using a test function, which maps the states of the PDMP to real values (see (3)). In this stochastic equation, randomness is created at its jump instants due to arrivals and service completions. This causes difficulty for analysis. Davis [2] who introduced PDMP replaces them by a martingale, imposing the so called boundary condition on the test function. However, it is not easy to find a class of the test functions which characterize a distribution on the state space of the PDMP.

A basic idea in [6] is to choose a smaller class of test functions to overcome those difficulties. We then have a martingale, which will be used for change of measure. They can not characterize a distribution on the state space, but still retains full information to study large queues under the stationary distribution.

In this paper, we focus on the tail decay rates of the stationary distribution of the joint queue length in the GJN. This problem has been solved for a two node GJN assuming phase-type distributions in [5]. We derive upper and lower bounds for them for the GJN with any $d \geq 2$ and general inter-arrival and service time distributions, which may have heavy tails, but those bounds may not be sharp for $d \geq 3$.

2. PIECEWISE DETERMINISTIC MARKOV PROCESS, PDMP

We introduce a piecewise deterministic Markov process $\{X(t); t \geq 0\}$. For each $t \geq 0$, $X(t)$ has two components $Z(t)$ and $C(t)$, namely,

$$X(t) = (Z(t), C(t)), \qquad t \geq 0,$$

where $Z(t)$ is a state to be interested, and $C(t)$ is a time counter for the next jump. As always, $X(t)$ is assumed to be right-continuous and has the left-hand limit at each time t. Its state space has the following structure.

(a) $Z(t)$ takes values in a complete and separable topological space S_1.

(b) For each $Z(t)$, there is a finite set $\mathcal{K}(Z(t))$, and $C(t)$ takes values in $\mathbb{R}_+^{\mathcal{K}(Z(t))}$ which is the set of all vectors

$\boldsymbol{y}$ whose entry y_i takes values in $\mathbb{R}_+$ for $i \in \mathcal{K}(Z(t))$, where $\mathbb{R}_+$ is the set of all nonnegative real numbers. The entry of $C(t)$ with index i is denoted by $C_i(t)$.

The state space S_1 which we use is the d-dimensional nonnegative integer orthant $\mathbb{Z}_+^d$ with discrete topology, where $\mathbb{Z}_+$ is the set of all nonnegative integers.

Let $S = \{(z, \boldsymbol{y}); z \in S_1, \boldsymbol{y} \in \mathbb{R}_+^{\mathcal{K}(z)}\}$, which is the state space of $X(t)$. We assume the following dynamics.

(c) $X(t)$ is a continuously differentiable deterministic function of t except for jump instants, which are denoted by an increasing sequence $\{t_n; n = 1, 2, \ldots\}$, with $t_0 = 0$.

(d) There is a set $\mathcal{M}(S)$ of continuous functions from S to $\mathbb{R}$ such that

(d1) The distribution of $X(t)$ is determined by $\mathbb{E}(f(X(t))) < \infty$ for $f \in \mathcal{M}(S)$.

(d2) For $t \in (t_{n-1}, t_n)$, $f(X(t))$ has a continuous derivative $\mathcal{A}f(X(t))$ for $f \in \mathcal{M}(S)$, where $\mathcal{A}$ is an operator on $\mathcal{M}(S)$, that is, $\mathcal{A}f \in \mathcal{M}(S)$ for $f \in \mathcal{M}(S)$.

(d3) For $t \in (t_{n-1}, t_n)$, $\mathcal{K}(t)$ is unchanged, and $C_i(t)$ is non-increasing in t for each $i \in \mathcal{K}(t)$.

(e) For $s > 0$ and $n \geq 1$, $t_n > t_{n-1} + s$ if and only if $C_i((t_{n-1} + u)-) > 0$ for all $u \in [0, s]$ and all $i \in \mathcal{K}(t_{n-1})$, where $C_i(t-) = \lim_{\epsilon \downarrow 0} C_i(t - \epsilon)$.

(f) The conditional distribution of $X(t_n)$ given $\{X(s); s < t_n\}$ is a function of $X(t_n-)$ for $n \geq 1$, which is characterized by the transition kernel Q given below.

$$Qf\big(X(t-)\big) = \mathbb{E}\big(f(X(t))|X(t-)\big), \quad X(t-) \in \Gamma, \quad (1)$$

for $f \in \mathcal{M}(S)$, where Γ is the subset of S such that some entries of $\boldsymbol{y}$ vanish for $(z, \boldsymbol{y}) \in S$. This Γ is referred to as a terminal set, while Q is referred to as a jump kernel.

Obviously, $\{X(t); t \geq 0\}$ satisfying the conditions (a)–(f) is a Markov process, whose dynamics is specified by $\mathcal{A}$ and Q. This process is essentially the same as a piecewise deterministic Markov process, PDMP for short, introduced by Davis [2]. We refer to it as the same name. It is noticed that we exclude jumps generated by the main component $Z(t)$, but they may be included in $C(t)$.

2.1 Martingale decomposition of the PDMP

Let $X(\cdot)$ be a PDMP satisfying the conditions (a)–(f). We consider its evolution in time by a stochastic integral equation. Let $\mathcal{F}_t = \sigma(X(s); s \leq t)$, where $\sigma(\cdot)$ stands for the minimal σ-field. $\{\mathcal{F}_t; t \geq 0\}$ is called a filtration. Then, $X(\cdot)$ is a strong Markov process with respect to $\{\mathcal{F}_t; t \geq 0\}$. Define the counting process $N(\cdot) \equiv \{N(t); t \geq 0\}$ for the jump instants of this PDMP as

$$N(t) = \sum_{n=1}^{\infty} 1(t_i \leq t), \qquad t \geq 0. \qquad (2)$$

Then, we obviously have a stochastic integral equation.

$$\begin{aligned} f(X(t)) =& f(X(0)) + \int_0^t \mathcal{A}f(X(s))ds \\ &+ \int_0^t \Delta f(X(s))dN(s), \quad f \in \mathcal{M}(S), \quad (3) \end{aligned}$$

where $\Delta f(X(s)) = f(X(s)) - f(X(s-))$. Note that $\Delta N(s) > 0$ if and only if $X(s-) \in \Gamma$, which causes a jump.

Our arguments will be based on the following fact due to Davis [2].

Lemma 2.1 (A special case of Theorem 5.5 of [2]). *For $f \in \mathcal{M}(S)$, if*

$$\begin{aligned} M(t) \equiv& f(X(t)) - f(X(0)) - \Big(\int_0^t \mathcal{A}f(X(s))ds \\ &+ \int_0^t (Qf(X(s-)) - f(X(s-)))dN(s)\Big) \quad (4) \end{aligned}$$

satisfies that $\mathbb{E}(|M(t)|) < \infty$, then $M(\cdot) \equiv \{M(t); t \geq 0\}$ is an $\mathcal{F}_t$-martingale. In particular, if f satisfies that

$$Qf(\boldsymbol{x}) = f(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \Gamma, \qquad (5)$$

then the $\mathcal{F}_t$-martingale $M(t)$ is simplified to

$$M(t) = f(X(t)) - f(X(0)) - \int_0^t \mathcal{A}f(X(s))ds. \qquad (6)$$

Although Davis [2] refers to (5) as a boundary condition, we refer to (5) as a terminal condition following the terminology of [6]. Note that (6) can be written as

$$f(X(t)) = f(X(0)) + \int_0^t \mathcal{A}f(X(s))ds + M(t). \qquad (7)$$

Apart from the terminal condition (5), this representation is standard for a Markov process which has $\mathcal{A}$ as an extended generator.

3. GENERALIZED JACKSON NETWORK

We first introduce this network model, then describe it by a PDMP. Let d be a positive integer, and consider a d-node queueing network with single servers at nodes in which exogenous customers arrive at each node subject to a renewal process and service times at each node are independent and identically distributed which are independent of everything else. Each node has an infinite buffer, customers are served in the FCFS manner, and they are routed to the next nodes or leave the network according to a given probability which only depends on the current node when their service completed. We refer this queueing network as a generalized Jackson network, GJN for short.

3.1 Notations and assumptions

Let $\mathcal{J} = \{1, 2, \ldots, d\}$, and let $\mathcal{E}$ be the set of nodes which have exogenous arrivals. For time t and node $i \in \mathcal{J}$, let $L_i(t)$ be the number of customers, and let $R_{s,i}(t)$ be the residual service times, respectively, where we set $R_{s,i}(t) = 0$ when $L_i(t) = 0$. For $i \in \mathcal{E}$, let $R_{e,i}(t)$ be the residual time to the next exogenous arrival at node i. Let p_{ij} be the probability that a customer completing service at node i is routed to node j for $i, j \in \mathcal{J}$, where those customer leave the outside of the network with probability:

$$p_{i0} \equiv 1 - \sum_{i \in \mathcal{J}} p_{ij}.$$

For each node i, let $F_{e,i}$ be the interarrival time distribution of exogenous customers, and let $F_{s,i}$ be the service time distribution. Denote the vectors whose i-th entries are

$L_i(t), R_{e,i}(t)$ for $i \in \mathcal{E}$, $R_{s,i}(t)$ by $\boldsymbol{L}(t), \boldsymbol{R}_e(t), \boldsymbol{R}_s(t)$, respectively, and define $X(t)$ as

$$X(t) = (\boldsymbol{L}(t), \boldsymbol{R}_e(t), \boldsymbol{R}_s(t)), \qquad t \geq 0.$$

Then, $\{p_{ij}; i, j \in \mathcal{J}\}$, $\{F_{e,i}; i \in \mathcal{E}\}$ and $\{F_{s,i}; i \in \mathcal{J}\}$ are the modeling primitives, and it is not hard to see that $X(t)$ is a PDMP. The state space S is given by

$$S = \{(\boldsymbol{z}, \boldsymbol{y}_e, \boldsymbol{y}_s); z \in \mathbb{Z}_+^d, \boldsymbol{y}_e \in \mathbb{R}_+^{\mathcal{E}}, \boldsymbol{y}_s \in \mathbb{R}_+^{\mathcal{K}(z)}\}.$$

Let $\widehat{F}_{e,i}$ and $\widehat{F}_{s,i}$ be the moment generating functions, MGF for short, of the distributions $F_{e,i}$ and $F_{s,i}$, respectively. We define $\beta_{w,i}$ and $\theta_{w,i}$ for $w = e, s$ as

$$\beta_{w,i} = \sup\{\theta \in \mathbb{R}; \widehat{F}_{w,i}(\theta) < \infty\}. \tag{8}$$

Note that $\beta_{w,i}$ may be infinite. If $\beta_{w,i} = 0$, then the distribution $F_{w,i}$ is said to have a heavy tail. Otherwise, it is said to have a light tail.

3.2 Terminal condition for the GJN

We consider a test function for the martingale decomposition in Lemma 2.1 to be available. For this, we first truncate distribution $F_{w,i}$ for $w = e, s$. It will turn out that this is not only for handling heavy tailed distributions but also for well controlling $\boldsymbol{R}_e(t)$ and $\boldsymbol{R}_s(t)$.

Let $T_{e,i}$ and $T_{s,i}$ be the random variables subject to the distributions $F_{e,i}$ and $F_{s,i}$. For $v > 0$, we denote the distributions of $T_{e,i} \wedge v$ and $T_{s,i} \wedge v$ by $F_{e,i}^{(v)}$ and $F_{s,i}^{(v)}$, where $a \wedge b = \min(a, b)$ for $a, b \in \mathbb{R}$. Let

$$t_i(\boldsymbol{\theta}) = e^{-\theta_i}\Big(\sum_{j \in \mathcal{J}} p_{ij} e^{\theta_j} + p_{i0}\Big), \qquad \boldsymbol{\theta} \in \mathbb{R}^d, i \in \mathcal{J}.$$

Using parameters $\boldsymbol{\theta}, \boldsymbol{\zeta} \in \mathbb{R}^d$ and $\boldsymbol{\eta} \in \mathbb{R}^{\mathcal{E}}$, we choose the following test function for $v > 0$,

$$f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{y}_e, \boldsymbol{y}_s) = e^{\langle \boldsymbol{\theta}, \boldsymbol{z} \vee \boldsymbol{1}\rangle + \langle \boldsymbol{\eta}(\boldsymbol{u},\boldsymbol{\theta}), \boldsymbol{y}_e \wedge \boldsymbol{u}\rangle + \langle \boldsymbol{\zeta}(\boldsymbol{v},\boldsymbol{\theta}), \boldsymbol{y}_s \wedge \boldsymbol{v}\rangle},$$

where $\boldsymbol{z} \vee \boldsymbol{1}$ is the d-dimensional vector whose i-th entry is $\max(z_i, 1)$, $\boldsymbol{y} \wedge \boldsymbol{u}$ is the d-dimensional vector whose i-th entry is $\min(y_i, u_i)$, and $\boldsymbol{\eta}(\boldsymbol{u}, \boldsymbol{\theta})$ and $\boldsymbol{\zeta}(\boldsymbol{v}, \boldsymbol{\theta})$ are the solutions of $\boldsymbol{\eta} \equiv \{\eta_i; i \in \mathcal{E}\}$ and $\boldsymbol{\zeta} \equiv \{\zeta_i; i \in \mathcal{J}\}$, respectively, of the following equations.

$$e^{\theta_i} \widehat{F}_{e,i}^{(u_i)}(\eta_i) = 1, \quad i \in \mathcal{E}, \quad t_i(\boldsymbol{\theta}) \widehat{F}_{s,i}^{(v_i)}(\zeta_i) = 1, \quad i \in \mathcal{J},$$

where $\widehat{F}_{e,i}^{(u_i)}$ and $\widehat{F}_{s,i}^{(v_i)}$ are the MGF's of $F_{e,i}^{(u_i)}$ and $F_{s,i}^{(v_i)}$, respectively. Because of the truncations, $\boldsymbol{\eta}(\boldsymbol{u}, \boldsymbol{\theta})$ and $\boldsymbol{\zeta}(\boldsymbol{v}, \boldsymbol{\theta})$ are well defined, and the terminal condition (5) is satisfied for all $\boldsymbol{\theta} \in \mathbb{R}^d$ (see Lemma 2.3 of [6]). Let

$$\begin{aligned}
\eta^R_{\boldsymbol{u},\boldsymbol{\theta}}(s) &= \sum_{i \in \mathcal{E}} \eta_i(u_i, \theta_i) 1(R_{e,i}(s) > u_i), \\
\zeta^R_{\boldsymbol{v},\boldsymbol{\theta}}(s) &= \sum_{i \in \mathcal{E}} \zeta_i(v_i, \boldsymbol{\theta}) 1(R_{s,i}(s) > v_i), \\
\gamma_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{\theta}) &= -\sum_{i \in \mathcal{E}} \eta_i(u_i, \theta_i) - \sum_{i \in \mathcal{J}} \zeta_i(v_i, \boldsymbol{\theta}).
\end{aligned}$$

Then, we have the following martingale under the assumption $\mathbb{E}(|X(t)|) < \infty$ for all $t \geq 0$.

$$\begin{aligned}
M_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(t) &\equiv f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(X(t)) - f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(X(0)) \\
&\quad - \int_0^t \Big(\gamma_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{\theta}) + \eta^R_{\boldsymbol{u},\boldsymbol{\theta}}(s) + \zeta^R_{\boldsymbol{v},\boldsymbol{\theta}}(s) \\
&\qquad + \sum_{i \in \mathcal{J}} \zeta_i(v_i, \boldsymbol{\theta}) 1(L_i(s) = 0)\Big) f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(X(s)) ds.
\end{aligned} \tag{9}$$

We now consider $\eta_i(u_i, \theta_i)$ and $\zeta_i(v_i, \boldsymbol{\theta})$. For this, we first consider a nonnegative random variable T with distribution F. Denote the MGF of its truncation by $\widehat{F}^{(u)}$.

Lemma 3.1. *Let $h(\boldsymbol{\theta})$ be a positive valued function such that $\log h(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta} \in \mathbb{R}^d$. Define $\xi(u, \theta)$ as the solution of ξ of the following equation.*

$$h(\boldsymbol{\theta}) \widehat{F}^{(u)}(\xi) = 1,$$

then $\xi(u, \theta)$ is concave in $\boldsymbol{\theta}$ for each fixed $u > 0$, and is negative and increasing (positive and decreasing) in $u > 0$ for each fixed $\boldsymbol{\theta}$ satisfying $h(\boldsymbol{\theta}) > 1$ (< 1, respectively).

Using the fact that $\log \widehat{F}^{(u)}(s)$ is convex in $s \in \mathbb{R}$, this lemma can be proved in the same way as Lemma 2.4 of [6].

Lemma 3.1 enables us to define

$$\xi(\triangle, \boldsymbol{\theta}) = \lim_{u \uparrow \infty} \xi(u, \boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \mathbb{R}^d. \tag{10}$$

We also define $\xi(\boldsymbol{\theta}) = \xi(\infty, \boldsymbol{\theta})$ as long as it is well defined and finite for $\boldsymbol{\theta}$. It is notable that $\xi(\boldsymbol{\theta})$ may not equal $\xi(\triangle, \boldsymbol{\theta})$.

Since $\log t_i(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$, we can take it or e^{θ_i} for $h(\boldsymbol{\theta})$ in Lemma 3.1, and therefore $\eta_i(\triangle, \theta_i)$ and $\zeta_i(\triangle, \boldsymbol{\theta})$ can be defined in the same way as in (10). $\eta_i(\theta_i)$ and $\zeta_i(\boldsymbol{\theta})$ are similarly defined. Similar to Lemma 2.5 of [6], we have the following lemma.

Lemma 3.2. *(a) $\eta_i(\triangle, \theta_i)$ and $\zeta_i(\triangle, \boldsymbol{\theta})$ are finite and concave for all $\theta_i \in \mathbb{R}$ and $\boldsymbol{\theta} \in \mathbb{R}^d$. (b) $\eta_i(\triangle, \theta_i) \leq \beta_{e,i}$ for all $\theta_i \in \mathbb{R}$ with equality for $\theta_i \leq \theta^*_{e,i}$, and $\eta_i(\triangle, \theta_i) = \eta_i(\theta_i)$ for $\theta_i > \theta^*_{e,i}$, where $\theta^*_{e,i} = -\log \widehat{F}_{e,i}(\beta_{e,i})$. (c) $\zeta_i(\triangle, \boldsymbol{\theta}) \leq \beta_{s,i}$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$ with equality for $\boldsymbol{\theta}$ satisfying $t_i(\boldsymbol{\theta})\widehat{F}_{s,i}(\beta_{s,i}) \leq 1$, and $\zeta_i(\triangle, \boldsymbol{\theta}) = \zeta_i(\boldsymbol{\theta})$ for for $\boldsymbol{\theta}$ satisfying $t_i(\boldsymbol{\theta})\widehat{F}_{s,i}(\beta_{s,i}) > 1$. (d) $\widehat{F}(\eta_i(\triangle, \theta_i)) < \infty$ for all $\theta_i \in \mathbb{R}$, and $\widehat{F}(\zeta_i(\triangle, \boldsymbol{\theta})) < \infty$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$.*

For $J \subset \mathcal{E}$, $\boldsymbol{u}_J$ is the vector in $\mathbb{R}_+^{|\mathcal{E}|}$ whose i-th entry is u_i if $i \in J$ and otherwise is $\triangle$. Similarly, $\boldsymbol{v}_K$ is defined for $K \subset \mathcal{J}$. For convenience, $\boldsymbol{u}_\emptyset$ and $\boldsymbol{v}_\emptyset$ are denoted by $\blacktriangle$.

$$\gamma_{\boldsymbol{u}_J, \boldsymbol{v}_K}(\boldsymbol{\theta}) = \lim_{u_j \uparrow \infty, j \not\in J, v_k \uparrow \infty, k \not\in K} \gamma_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{\theta}).$$

In particular, $\gamma_\triangle(\boldsymbol{\theta}) = \gamma_{\blacktriangle,\blacktriangle}(\boldsymbol{\theta})$. By Lemma 3.2, $\gamma_\triangle(\boldsymbol{\theta})$, $\gamma_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{\theta})$ and $\gamma_{\boldsymbol{u}_J,\boldsymbol{v}_K}(\boldsymbol{\theta})$ are all convex in $\boldsymbol{\theta} \in \mathbb{R}^d$.

3.3 Bounds for the tail decay rates

We consider the tail asymptotic problem for the GJN. Denote the means of $T_{e,i}, T_{s,i}$ by $m_{e,i}$ and $m_{s,i}$, respectively. Let $\lambda_{e,i} = 1/m_{e,i}$ for $i \in \mathcal{E}$. Let α_i for $i \in \mathcal{J}$ be the solutions of the following traffic equation.

$$\alpha_i = \lambda_i 1(i \in \mathcal{E}) + \sum_{j \in \mathcal{J}} \alpha_j p_{ji}, \qquad i \in \mathcal{J}.$$

It is easy to see that the solutions uniquely exist if the $d \times d$ matrix $P \equiv \{p_{ij}; i, j \in \mathcal{J}\}$ is strictly substochastic and if $\overline{P} \equiv \{p_{ij}; i,j \in \{0\} \cup \mathcal{J}\}$ is irreducible, where $p_{00} = 0$, and $p_{0i} = \lambda_i 1(i \in \mathcal{E}) / \sum_{j \in \mathcal{E}} \lambda_j$ for $i \in \mathcal{J}$. We assume these conditions. Let $\rho_i = \alpha_i m_i$, and assume the stability condition that $\rho_i < 1$ for all $i \in \mathcal{J}$.

We present a main result of this paper. For $K \subset \mathcal{J}$, let

$$\Gamma_K^+ = \{\boldsymbol{\theta} \in \mathbb{R}^d; \gamma_\triangle(\boldsymbol{\theta}) < 0, \zeta_j(\triangle, \boldsymbol{\theta}) < 0, \forall j \in K\},$$
$$\Gamma_K^- = \{\boldsymbol{\theta} \in \mathbb{R}^d; \gamma_\triangle(\boldsymbol{\theta}) \geq 0, \zeta_j(\triangle, \boldsymbol{\theta}) \geq 0, \forall j \in K\}.$$

Lemma 3.3. *Assume that the GJN is stable, and let $\varphi_i(\boldsymbol{\theta}) = \mathbb{E}(e^{\langle \boldsymbol{\theta}, \boldsymbol{L} \rangle} 1(L_j = 0))$ for $\boldsymbol{\theta} \in \mathbb{R}^d, i \in \mathcal{J}$. Then, we have for $K \subset \mathcal{J}$ satisfying $|K| \geq d-1$, where $|K|$ is the number of elements of K,*

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}(\boldsymbol{L} \geq n\boldsymbol{c}) \leq -\sup\{\langle \boldsymbol{c}, \boldsymbol{\theta} \rangle; \boldsymbol{\theta} \in \Gamma_K^+, \varphi_i(\boldsymbol{\theta}) < \infty, \forall i \notin K\}, \qquad (11)$$

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(\boldsymbol{L} \geq n\boldsymbol{c}) \geq -\inf\{\langle \boldsymbol{c}, \boldsymbol{\theta} \rangle; \boldsymbol{\theta} \in \Gamma_K^-, \varphi_i(\boldsymbol{\theta}) < \infty, \forall i \notin K\}. \qquad (12)$$

Outline of proof. We first construct an exponential martingale for change of measure. Let ν be the distribution of $X(0)$, and denote the probability measure under this initial distribution by $\mathbb{P}_\nu$. Assume that $\mathbb{E}_\nu(e^{\langle \boldsymbol{\theta}, \boldsymbol{L}(0) \rangle}) < \infty$, from which it can be proved that $\mathbb{E}_\nu(f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(X(t))) < \infty$ for $t \geq 0$. Let

$$Y(t) = \frac{1}{f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(X(0))} \exp\Big(-\int_0^t \frac{\mathcal{A} f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(X(s))}{f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(X(s))} ds \Big),$$

which is obviously continuous in t, so $\mathcal{F}_{t-}$-measurable. Then,

$$Y \cdot M_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(t) \equiv 1 + \int_0^t Y(s) M_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(ds)$$

is a $\mathcal{F}_t$-martingale under $\mathbb{P}_\nu$ (see Section 4d of Chapter I of [3] and [7] for PDMP). Denote $Y \cdot M_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(t)$ by $E^{f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}}$, then

$$E^{f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}}(t) = \frac{f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(X(t))}{f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(X(0))} \exp\Big(-\int_0^t \frac{\mathcal{A} f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(X(s))}{f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(X(s))} ds \Big).$$

Since $E^{f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}}(t)$ is positive and $E^{f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}}(0) = 1$, we can define a new probability measure $\widetilde{\mathbb{P}}_\nu^{(\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta})}$ by

$$\frac{d\widetilde{\mathbb{P}}_\nu^{(\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta})}}{d\mathbb{P}_\nu}\Big|_{\mathcal{F}_t} = E^{f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}}(t), \qquad t \geq 0.$$

This implies that

$$d\mathbb{P}_\nu = (E^{f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}}(t))^{-1} d\widetilde{\mathbb{P}}_\nu^{(\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta})} \quad \text{on } \mathcal{F}_t. \qquad (13)$$

To choose the initial distribution ν, let τ_A^+, τ_A^- be the first exit and return times of $L(t)$ to $A \subset S_1$ such that $\tau_A^+ < \tau_A^-$. Let ν_A^+ the distribution of $X(t)$ at time τ_A^+ given that $X(0)$ is subject to the normalized stationary distribution ν_A limited on $\{(z, \boldsymbol{y}) \in S; z \in A\}$. Denote a random vector subject to the stationary distribution of $X(t)$ by $X \equiv (\boldsymbol{L}, \boldsymbol{R}_e, \boldsymbol{R}_s)$. Then, the cycle formula yields, for measurable $B \subset S_1 \setminus A$,

$$\mathbb{P}(\boldsymbol{L} \in B) = c(A) \mathbb{E}_{\nu_A^+}\Big(\int_0^{\tau_A^-} 1(\boldsymbol{L}(s) \in B) ds \Big), \qquad (14)$$

where $c(A) = \mathbb{P}(\boldsymbol{L} \in A) / \mathbb{E}_{\nu_A^+}(\tau_A^- - \tau_A^+)$. For $K \in 2^{\mathcal{J}}$, let

$$F_K = \cup_{i \in \mathcal{J} \setminus K} \{\boldsymbol{z} \in S_1; z_i = 0\}.$$

For $\boldsymbol{c} \in S_1 \setminus F_K$ and $n \geq 1$, let $A = F_K$ and let

$$\tau_{\boldsymbol{c},n}^+ = \inf\{t > 0; \boldsymbol{L}(t) \geq n\boldsymbol{c}\},$$

$$Y_{\tau_{\boldsymbol{c},n}^+ -} = \mathbb{E}_{\nu_A^+}\Big(\int_{\tau_{\boldsymbol{c},n}^+}^{\tau_{F_K}^-} 1(\boldsymbol{L}(s) \geq n\boldsymbol{c}) ds \Big| \mathcal{F}_{\tau_{\boldsymbol{c},n}^+ -} \Big) 1(\tau_{\boldsymbol{c},n}^+ < \tau_{F_K}^-),$$

and applying (13) with $\nu = \nu_{F_K}^+$, we have

$$\mathbb{P}(\boldsymbol{L} \geq n\boldsymbol{c}) = c(F_K) \widetilde{\mathbb{E}}_{\nu_{F_K}^+}^{(\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta})} \Big(f_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta}}(X(0)) Y_{\tau_{\boldsymbol{c},n}^+ -} e^{-\langle \boldsymbol{\theta}, \boldsymbol{L}(\tau_{\boldsymbol{c},n}^+ -) \rangle}$$
$$\times e^{\gamma_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{\theta}) \tau_{\boldsymbol{c},n}^+ - \langle \boldsymbol{\eta}(\boldsymbol{u},\boldsymbol{\theta}), \boldsymbol{R}_e(\tau_{\boldsymbol{c},n}^+ -) \wedge \boldsymbol{u} \rangle - \langle \boldsymbol{\zeta}(\boldsymbol{v},\boldsymbol{\theta}), \boldsymbol{R}_s(\tau_{\boldsymbol{c},n}^+ -) \wedge \boldsymbol{v} \rangle}$$
$$\times e^{\int_0^{\tau_{\boldsymbol{c},n}^+} [\eta_{\boldsymbol{u},\boldsymbol{\theta}}^R(s) + \zeta_{\boldsymbol{v},\boldsymbol{\theta}}^R(s)) + \sum_{i \in K} \zeta_i(v_i, \boldsymbol{\theta}) 1(L_i(s) = 0)] ds} \Big). \qquad (15)$$

To prove (11) for $K \in 2^{\mathcal{J}}$, we take the following steps.

(a) Choose $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $\boldsymbol{\theta} \in \Gamma_K^+, \varphi_i(\boldsymbol{\theta}) < \infty, \forall i \notin K$, and sufficiently large v_i for all $i \in K$ such that $\zeta_i(v_i, \boldsymbol{\theta})$ and $\zeta_i(\triangle, \boldsymbol{\theta})$ have the same sign.

(b) For all $j \in \mathcal{J} \setminus K$ such that $\zeta_i(\triangle, \boldsymbol{\theta}) > 0$, let $v_j \to \infty$ in (15). Then, $\langle \boldsymbol{\theta}, \boldsymbol{c} \rangle n + \log \mathbb{P}(\boldsymbol{L} \geq n\boldsymbol{c})$ is bounded by a function of $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\theta}$ not depending on n.

(c) Divide both sides of the above inequality by n and let $n \to \infty$, then take the supremum on $\boldsymbol{\theta}$.

Similarly, (12) is obtained, and the proof is completed. ∎

The condition that $|K| \geq d-1$ may be too strong for $d \geq 3$, but is needed to verify (b). We conjecture that it can be removed. Note that the upper bound (11) is similar to those for the d-dimensional reflecting random walk (see Theorem 6.1 and (6.9) of [4]). We expect that the decay rate in an arbitrary direction can be obtained from (11) and (12) in a similar way as Theorem 6.1 of [4] if the condition that $|K| \geq d-1$ is unnecessary in Lemma 3.3.

References

[1] Braverman, A., Dai, J. and Miyazawa, M. (2015). Heavy traffic approximation for the stationary distribution of a generalized jackson network: the BAR approach. Submitted for publication.

[2] Davis, M. H. A. (1984). Piecewise deterministic Markov processes: a general class of non-diffusion stochastic models. *Journal of Royal Statist. Soc. series B*, **46** 353–388.

[3] Jacod, J. and Shiryaev, A. N. (2003). *Limit Theorems for stochastic processes.* 2nd ed. Springer, Berlin.

[4] Miyazawa, M. (2011). Light tail asymptotics in multidimensional reflecting processes for queueing networks. *TOP*, **19** 233–299.

[5] Miyazawa, M. (2015). A superharmonic vector for a nonnegative matrix with QBD block structure and its application to a Markov modulated two dimensional reflecting process. *Queueing Systems*, **81** 1–48.

[6] Miyazawa, M. (2017). A unified approach for large queue asymptotics in a heterogeneous multiserver queue. *Advances in Applied Probability (to appear).*

[7] Palmowski, Z. and Rolski, T. (2002). A technique of the exponential change of measure for Markov processes. *Bernoulli*, **8** 767–785.

Stability criterion of a MAP/PH-multiserver model with simultaneous service

A. Rumyantsev
Institute of Applied Mathematical Research,
Karelian Research Centre of RAS
11 Pushkinskaya Str.
Petrozavodsk, Russian Federation
ar0@krc.karelia.ru

E. Morozov
Institute of Applied Mathematical Research,
Karelian Research Centre of RAS
11 Pushkinskaya Str.
Petrozavodsk, Russian Federation
emorozov@karelia.ru

ABSTRACT

In this paper, we study the stability conditions of the multiserver system in which each customer requires a random number of servers simultaneously and a random (but identical) service time at all occupied servers. We call it cluster model, because this model describes the dynamics of multicore high performance clusters. Stability criteria of an $M/M/s$ cluster model has been found earlier. In this work we, again using the matrix-analytic approach, show that the stability criteria of an $MAP/M/s$ cluster model has the same form as for $M/M/s$ system. We believe, that the result holds true for a more general $MAP/PH/s$ cluster model. However, the proof is still in progress, and we illustrate the result with simulations.

CCS Concepts

•**Mathematics of computing** → **Queueing theory; Markov processes;** •**Computer systems organization** → *Parallel architectures;*

Keywords

Stability condition; high performance cluster; MAP arrivals; PH service time; simultaneous service multiserver system

1. INTRODUCTION

A major pivot from frequency scaling of a Central Processing Unit (CPU) to the massive use of multicore and multi-CPU architectures [15] caused a rebirth of interest in studying of stochastic models of modern multiserver systems. A separate class of multiserver models that allow a single customer to be served simultaneously by a number of servers, is practically motivated by computing systems such as high performance clusters (as well as cloud/distributed computing) containing a huge number of servers working in parallel. According to [16], the class of systems with simultaneous service has two major subclasses: i) systems with independent service (service times of a given customer are independent) and ii) systems with concurrent service (service times of a customer are identical through all the occupied servers). The key feature of the systems ii) is that there is a possibility to have idle servers and a non-empty queue simultaneously, which significantly complicates the stability analysis. While for the subclass i), the stability conditions in an explicit form have been obtained in a number of papers, see e.g. [3, 5], the subclass ii) requires a more delicate analysis. In the work [1] the stationary distribution of class-dependent delay has been obtained by the system point approach, however the stability problem for the general multiserver system has not been addressed there. The stability condition obtained in [8] requires a numerical solution of a large dimension matrix equation, while the corresponding matrices are not explicitly defined. In recent works [4, 2], a *two-server system* is investigated by means of the matrix-analytic method (more on this method see [10, 9, 7, 6]), and the stability condition (earlier stated with no proof in the paper [1]) has been strictly proven. The work [4] deals with exponential distributions, whereas the work [2] extends the stability condition from [4] to the MAP input. The stability criterion of the exponential model with *arbitrary number of servers* and arbitrary distribution of the required number of servers has been obtained in [14] by means of matrix-analytic approach. Moreover, a computationally effective verification of the stability criterion has been proposed in [13].

The main contribution of this paper is the extention of the stability criterion to the cluster model with MAP input.

The paper is organised as follows. In Section 2, we describe general $MAP/PH/s$ cluster model. Then, in Section 3.2, we sketch the proof of the stability condition of a $MAP/M/s$ cluster model. We show that the earlier found stability criteria of $M/M/s$ model (described briefly in Section 3.1) holds true for the MAP arrivals as well. In Section 3.3 we demonstrate by simulation that (an appropriately modified) condition obtained earlier allows to delimit the stability/instability zones of the $MAP/PH/s$ model.

2. DESCRIPTION OF THE MODEL

In this section we describe the $MAP/PH/s$ cluster model. For more details on MAP input and PH distributions see [6, 9].

We consider a FCFS s-server simultaneous service queueing system with input flow driven by a MAP (D_0, D_1) with k states. Customer i occupies N_i servers simultaneously for the same service time S_i having a PH distribution (τ, T)

with l states (and $l+1$ being the absorbtion state). It is assumed that $\tau\mathbf{1} = 1$ (where $\mathbf{1}$ is a vector of ones), implying $S_i > 0$. Denote $\lambda = \theta D_1 \mathbf{1}$ the fundamental rate of the MAP, where vector θ satisfies the following equations

$$\begin{cases} \theta D &= 0, \\ \theta \mathbf{1} &= 1, \end{cases} \tag{1}$$

with matrix $D := D_0 + D_1$. Let $\mu := \left(-\tau T^{-1}\mathbf{1}\right)^{-1}$ be the service rate of PH distribution, and define the l-dimensional vector $\pi = \mu(\tau T^{-1})$. It is known that the vector π is the stationary distribution of the PH (renewal) process [9].

We call customer i *class-j* one if $N_i = j$. The sequence $\{N_i\}$ is assumed to be i.i.d. with a given distribution

$$p_j := \mathbb{P}(N = j), \quad j = 1, \dots, s \quad (\sum_{j=1}^{s} p_j = 1). \tag{2}$$

(We omit the serial index to denote a generic element of an i.i.d. sequence.)

Let $\nu(t)$ be the number of customers in the system at instant t, $t \geqslant 0$. Following [17], we call the vector $m(t) = (m_1(t), \dots, m_s(t))$ a *macrostate*, where $m_i(t)$ is the class of the ith oldest customer in the system (if $\nu(t) < s$, then $m_i(t) := 0$ for $i > \nu(t)$). Let $\varphi(t) \in \{1, \dots, k\}$ be the phase of the MAP at instant t. Denote $\psi(t) = (\psi_1(t), \dots, \psi_s(t))$, where $\psi_i(t)$ is the phase of the ith oldest customer being served (we put $\psi_i(t) := 0$ if the i-th oldest customer is waiting in the queue). Note that $\psi(t) \in \{0, \dots, l\}^s$.

Denote the set of macrostates $\mathcal{M} = \{1, \dots, s\}^s$, and let $\mathbb{Z}_+ := \{1, 2, \dots\}$. For a fixed $m \in \mathcal{M}$, define

$$\sigma(m) := \max\left\{ i : \sum_{j=1}^{i} m_j \leqslant s \right\} \leqslant s \tag{3}$$

the number of customers *being served* in the macrostate m, and let

$$\Psi(m) := \{\psi : \psi_k > 0, k \leqslant \sigma(m), \; \psi_k = 0, k > \sigma(m)\}. \tag{4}$$

Then the set$\Psi := \bigcup_{m \in \mathcal{M}} \Psi(m)$ contains $l(l^s - 1)/(l-1)$ combinations of PH phases of customers being served. Finally, let $\Omega := \mathbb{Z}_+ \times \mathcal{M} \times \Psi \times \{1, \dots, k\}$. For convenience, we use the lexicographical order to enumerate the phases $(m, \psi, \varphi) \in \mathcal{M} \times \Psi \times \{1, \dots, k\}$, and use this multi-dimensional index to refer to the components of matrices.

As we show below, the process

$$\Big\{\Theta(t) := \big(\nu(t), m(t), \psi(t), \varphi(t)\big) \in \Omega; \; t \geq 0\Big\}, \tag{5}$$

is a QBD process living in Ω, where $\nu(t)$ is called the *level* of the process.

Consider a fixed state (n, m, x, y) of the process and one-step transitions $(n, m, y, x) \to (n', m', y', x')$. The following events (transitions) are possible for levels $n > s$:

1. A *change of MAP phase* with no arrivals: $n' = n, m' = m, y' = y$ and $x' \neq x$.

2. The phase of service time of the ith oldest customer being served *is changed with no completion of service*: $n' = n$, $m' = m$, $x' = x$, $y_i \neq y_i' < l+1$, and $y_j' = y_j, j \neq i$.

3. *An arrival* to the system: $n' = n+1, m' = m, y' = y$.

4. *Departure of the ith oldest customer*: $n' = n-1, x' = x$,

   $$m_j' = m_j, j < i, m_{j-1}' = m_j, j > i,$$

   and m_s' is chosen from (2). Moreover,

   $$y_j' = y_j, j < i, \; y_{j-1}' = y_j, i < j \leqslant \sigma(m),$$

   whereas, for $\sigma(m) < j \leqslant \sigma(m')$ (if any), y_j' is chosen from the initial distribution τ. We also set $y_j' = 0$ for $j > \sigma(m')$, if any.

The infinitesimal generator of the QBD process $\{\Theta(t)\}$ with a finite number of phases $d := s^s kl(l^s - 1)/(l-1)$ has the following block-tridiagonal form [9]:

$$\begin{pmatrix} B_1 & B_0 & 0 & 0 & \dots \\ B_2 & A_1 & A_0 & 0 & \dots \\ 0 & A_2 & A_1 & A_0 & \dots \\ 0 & 0 & A_2 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{6}$$

where $A_i, i = 0, 1, 2$ are the square matrices of order d, which contain the intensity of transitions of $\{\Theta(t)\}$ caused by event 3, events 1 and 2, and event 4 defined above, respectively. The matrices $B_i, i = 0, 1, 2$ are related to the initial conditions (for levels $n \leqslant s$) and are not used in the stability analysis (for more details see [9]).

Recall also the basic property of the infinitesimal generator of the QBD process

$$A\mathbf{1}_d = \mathbf{0}_d, \tag{7}$$

where the matrix $A := A_0 + A_1 + A_2$.

3. STABILITY ANALYSIS

Below we use the Kronecker product $\otimes$ and Kronecker sum $\oplus$, which is defined as $A \oplus B := A \otimes I + I \otimes B$ for two matrices A, B and the identity matrix I of the appropriate size. We also use the following property of the Kronecker product. Let A, B, C, D be the matrices of such a size, that AC and BD are possible. Then

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD). \tag{8}$$

The basic result for stability analysis is the following *Neuts condition for ergodicity* of a QBD process with finite number of states (see [9], theorem 7.2.4, also [6, 10]). The QBD is positive recurrent if and only if

$$\alpha A_2 \mathbf{1} > \alpha A_0 \mathbf{1}, \tag{9}$$

where α is the unique solution of the system

$$\begin{aligned} \alpha A &= \mathbf{0}, \\ \alpha \mathbf{1} &= 1. \end{aligned} \tag{10}$$

3.1 $M/M/s$ cluster model

In this case the QBD process (5) becomes two-dimensional, $\Theta(t) := \{\nu(t), m(t)\}, t \geqslant 0$. with $d = s^s$ phases. It has been shown in [14], that the QBD process $\{\Theta(t)\}$ has the infinitesimal generator of block-tridiagonal form. Moreover, the intensity of transitions caused by an arrival/departure at levels $\nu(t) > s$ is governed by the square matrices of size $s^s \times s^s$. Let Q_0 be the matrix of transitions caused by an arrival, that is, the component $Q_0(m, m')$ is the intensity of transition from a macrostate m to a macrostate m' (at some instant t). Note that $Q_0(m,m') = \lambda$ for $m = m'$, and 0 otherwise (since an arrival does not change the macrostate). Let matrix Q_2 contain the intensities of transitions caused by a departure. It was shown in [14], that the knowledge of the detailed structure of Q_2 is not required for the stability analysis. Finally, a diagonal matrix Q_1 contains the rates of holding times of the process. It was proved in [14], that the solution of the system

$$\begin{array}{rcl} \gamma(Q_0+Q_1+Q_2) &=& \mathbf{0}, \\ \gamma\mathbf{1} &=& 1, \end{array}$$

has the following form:

$$\gamma_m = \frac{1}{C}\frac{\prod_{i=1}^s p_{m_i}}{\sigma(m)}, \quad m \in \mathcal{M}. \tag{11}$$

Note that the vector $\gamma = (\gamma_m, m \in \mathcal{M})$ can be interpreted as an approximation for distribution of the macrostates for high levels of $\nu(t)$, see [6]. The following result has been proved in [14] for $M/M/s$ cluster model (that is, for $k = l = 1$).

Theorem 1. *The irreducible continuous-time QBD process* (5) *is positive recurrent if and only if*

$$\rho := \frac{\lambda}{\mu} C < 1, \tag{12}$$

where

$$C = \sum_{m\in\mathcal{M}} \frac{\prod_{i=1}^s p_{m_i}}{\sigma(m)}. \tag{13}$$

It is null recurrent if $\rho = 1$ and transient if $\rho > 1$.

3.2 $MAP/M/s$ cluster model

In this section we assume that PH distribution has a single state, that is, the service times are exponential with rate μ. Consider the process $\Theta(t) := \{\nu(t), m(t), \varphi(t)\}$, $t \geqslant 0$, with $d = s^s k$ phases.

Now we define the matrices A_i explicitly. Indeed, the matrix A_0 corresponds to the arrivals into the system at high levels $\nu(t) > s$. Then the arrivals do not change the macrostate (since the macrostate is defined only by the s oldest customers). However, an arrival may change the MAP-phase $\varphi(t)$ according to the intensity matrix D_1. As a result, we obtain

$$A_0 := \mathbb{O}_{s^s} \oplus D_1 = I_{s^s} \otimes D_1, \tag{14}$$

where $\mathbb{O}_i$ (I_i) is the square zero (identity) matrix of size i.

Consider the matrix A_2, which corresponds to a departure. Since the MAP-phase is not changed at the departure epoch, and the change of a macrostate is still described by the matrix Q_2, then we obtain

$$A_2 := Q_2 \oplus \mathbb{O}_k = Q_2 \otimes I_k. \tag{15}$$

Denote by $\mathbf{1}_{ks^s}$ the ks^s-dimensional vector of ones. It follows from the properties of matrix Q_2, definition (15) and property (8), that

$$A_2\mathbf{1}_{ks^s} = (Q_2 \otimes I_k)(\mathbf{1}_{s^s} \otimes \mathbf{1}_k) = \mu\sigma \otimes \mathbf{1}_k, \tag{16}$$

where the m-th component of column vector σ is defined as $\sigma(m)$, and the equality $Q_2\mathbf{1}_{s^s} = \mu\sigma$ is proved in [14] (see equality (11) there). To explain the equality, we recall that, for each fixed macrostate m, there are exactly $\sigma(m)$ customers being served, with exponential service times (with intensity μ).

The matrix A_1 corresponds to the transitions of QBD process, which do not change the level. There is the only possibility for this transition, namely, the MAP-phase $\varphi(t)$ may change with no changes of the macrostate $m(t)$. The corresponding transition of the MAP-phase is governed by matrix D_0. Define the square matrix $J := \operatorname{diag}(\sigma)$ of order s^s. Then it follows from the balance condition (7) and equality (16), that the matrix A_1 has the following form:

$$A_1 := -\mu J \oplus D_0. \tag{17}$$

Then it follows from (14) and (17), that

$$A_0 + A_1 = -\mu J \oplus D. \tag{18}$$

Now we are ready to find a solution of the system of equations (10). By (15) and (18), the first equation of the system (10) is equivalent to

$$\alpha(Q_2 \otimes I_k) = \mu\alpha(J \oplus D). \tag{19}$$

Recall (1), (11) and define vector

$$\alpha := \gamma \otimes \theta. \tag{20}$$

It is easily seen (e.g. by property (8)), that α satisfies the second equation of the system (10).

Using (1) and (8), we obtain

$$\mu\alpha(I_{s^s} \otimes D) = \mu(\gamma\otimes\theta)(I_{s^s}\otimes D) = \mu\gamma \otimes \mathbb{O}_k = \mathbb{O}_{ks^s}. \tag{21}$$

Then, using (20), (21), we can rewrite (19) as

$$(\gamma\otimes\theta)(Q_2 \otimes I_k) = \mu(\gamma\otimes\theta)(J\otimes I_k + I_{s^s}\otimes D).$$

By the property (8), it now follows that

$$(\gamma Q_2 \otimes \theta I_k) = \mu(\gamma J \otimes \theta I_k + \gamma I_{s^s} \otimes \theta D).$$

Finally, a simplification of both parts of the last equality yields

$$\gamma Q_2 \otimes \theta = \mu\gamma J \otimes \theta. \tag{22}$$

It was proved in [14], that $\gamma Q_2 = \mu\gamma J$. Then (22) is also true, that is, the vector α defined in (20) is the solution of system (10).

It follows from (14) and (20), that

$$\begin{aligned} \alpha A_0 \mathbf{1} =& (\gamma\otimes\theta)(I_{s^s}\otimes D_1)\mathbf{1}_{ks^s} \\ =& (\gamma I_{s^s}\mathbf{1}_{s^s}) \otimes (\theta D_1 \mathbf{1}_k) = \theta D_1 \mathbf{1}_k =: \lambda. \end{aligned} \tag{23}$$

Note that the equations (16), (23), (11) together with (1) give

$$\alpha A_2 \mathbf{1} = (\gamma\otimes\theta)(\mu\sigma\otimes\mathbf{1}_k) = \mu\gamma\sigma = \frac{\mu}{C}, \tag{24}$$

where the last equality follows since

$$\gamma\sigma = \frac{1}{C}\sum_{m\in\mathcal{M}} \frac{\prod_{i=1}^s p_{m_i}}{\sigma(m)}\sigma(m) = \frac{1}{C}.$$

Thus, theorem 1 holds for $MAP/M/s$ cluster model.

3.3 Simulation results of $MAP/PH/s$ model

To illustrate the applicability of result (12) to the $MAP/PH/s$ model, we perform a simple numerical experiment. We set $s = 50$ and observe 5000 customers. First, we consider an underloaded system with $\rho = 0.9$. We generate the interarrival times $\{T_i\}$ by means of a two-phase MAP process with matrices

$$D_0 = \begin{pmatrix} -4.44 & 2 \\ 1 & -1.49 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 2.44 & 0 \\ 0 & 0.49 \end{pmatrix},$$

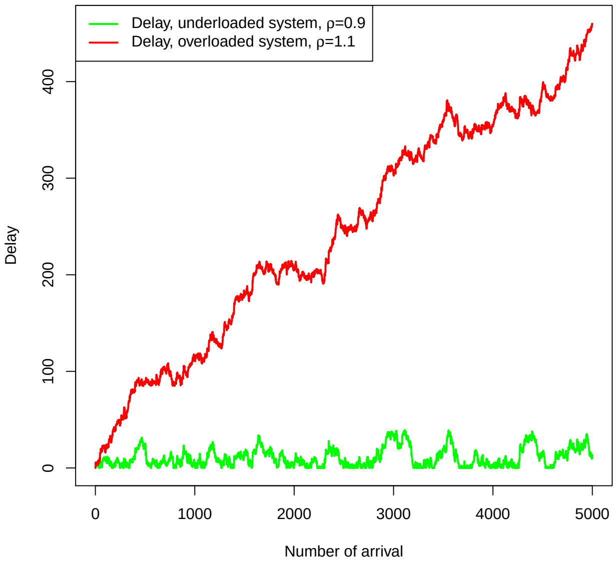


Figure 1: Comparison of the delays for 5000 customers of a $MAP/PH/s$ cluster model, when the stability condition holds (green), and is violated (red).

and generate the service times $\{S_i\}$ by a three-phase PH distribution with $\tau = (0.2, 0.4, 0.4)$ and

$$T = \begin{pmatrix} -4 & 2 & 0 \\ 2 & -5 & 1 \\ 1 & 0 & -1 \end{pmatrix}.$$

It then follows, that $\lambda = 1.14$ and $\mu = 0.909$. Next, we generate the values p_k in such a way to obtain $C = 0.717$. It implies $\rho = \lambda C/\mu = 0.9$. To obtain the delay of each customer, we apply the `hpcwld` package [12].

Next, we consider the overloaded system, with $\rho = 1.1$. The interarrival times $\{T_i\}$ follow the two-phase MAP with matrices

$$D_0 = \begin{pmatrix} -4.99 & 2 \\ 1 & -1.6 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 2.99 & 0 \\ 0 & 0.6 \end{pmatrix}.$$

We generate service times $\{S_i\}$ from the same three-phase PH distribution defined above. We also use the same values $\{p_k\}$. Then $\lambda = 1.39$, which implies $\rho = 1.1$.

The resulting delays shown on Fig. 1 demonstrate the (approximate) linear growth of the overloaded system, and a stable behavior of the underloaded system. It indicates that condition (12) allows to delimit the stability/instability zones of the $MAP/PH/s$ cluster model.

4. ACKNOWLEDGMENTS

This research is partially supported by Russian Foundation for Basic Research, grants 15-07-02341, 15-07-02354, 15-07-02360, 15-29-07974, 16-07-00622 and the Program of Strategic Development of Petrozavodsk State University. The authors thank Udo Krieger for a few useful comments.

5. REFERENCES

[1] P. Brill and L. Green. Queues in which customers receive simultaneous service from a random number of servers: A system point approach. *Management Science*, 30(1):51–68, 1984.

[2] S. Chakravarthy and H. Karatza. Two-server parallel system with pure space sharing and markovian arrivals. *Computers & Operations Research*, 40(1):510 – 519, 2013.

[3] A. Federgruen and L. Green. An M/G/c queue in which the number of servers required is random. *Journal of Applied Probability*, 21(3):583, 1984.

[4] D. Filippopoulos and H. Karatza. An M/M/2 parallel system model with pure space sharing among rigid jobs. *Mathematical and Computer Modelling*, 45(5–6):491–530, 2007.

[5] F. Gillent and G. Latouche. Semi-explicit solutions for M/PH/1-like queuing systems. *European Journal of Operational Research*, 13(2):151–160, 1983.

[6] Q.-M. He. *Fundamentals of Matrix-Analytic Methods.* Springer New York, 2014.

[7] O. C. Ibe. *Markov processes for stochastic modeling.* Academic Press, Amsterdam; Boston, 2009.

[8] S. Kim. *M/M/s Queueing System Where Customers Demand Multiple Server Use.* PhD thesis, Southern Methodist University, 1979.

[9] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* ASA–SIAM, Philadelphia, 1999.

[10] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models.* Johns Hopkins University Press, Baltimore, 1981.

[11] *R Foundation for Statistical Computing.* Vienna, Austria. ISBN 3-900051-07-0. [Online]. Available: http://www.r-project.org/

[12] A. Rumyantsev *hpcwld: High Performance Cluster Models Based on Kiefer-Wolfowitz Recursion. [Online].* Available: http://cran.r-project.org/web/packages/hpcwld/index.html

[13] A. Rumyantsev and E. Morozov. Accelerated Verification of Stability of Simultaneous Service Multiserver Systems. In *Proceedings of 2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Brno, Czech Republic, 6-8 October 2015.*, pages 239–242. IEEE, 2015.

[14] A. Rumyantsev and E. Morozov. Stability criterion of a multiserver model with simultaneous service. *Annals of Operations Research*, pages 1–11, June 2015.

[15] H. Sutter and J. Larus. Software and the concurrency revolution. *Queue*, 3(7):54–62, Sept. 2005.

[16] N. M. Van Dijk. Blocking of finite source inputs which require simultaneous servers with general think and holding times. *Oper. Res. Lett.*, 8(1):45–52, 1989.

[17] D. Wagner, V. Naumov, and U. R. Krieger. *Analysis of a multi-server delay-loss system with a general Markovian arrival process.* Techn. Hochschule, FB 20, Inst. für Theoretische Informatik, Darmstadt, Jan. 1994.

The computational complexity of analyzing infinite-state structured Markov chains and structured Markov decision processes

[Invited talk]

Kousha Etessami
School of Informatics
University of Edinburgh
Scotland, UK
kousha@staffmail.ed.ac.uk

ABSTRACT

I will survey a series of results over recent years on the computational complexity of key analysis problems for several families of infinite-state structured Markov chains (MCs) and structured Markov decision processes (MDPs).

A key aspect of our results is algorithmic bounds for computing the least fixed point (the least non-negative solution) for monotone systems of nonlinear (min/max)-polynomial equations which arise for these stochastic models and MDPs.

In particular, I will describe algorithms based on Newton's method, combined with P-time preprocessing steps, which yield polynomial time algorithms (in the standard Turing model of computation) for computing, to arbitrary desired accuracy, the G matrix of quasi-birth death processes (QBDs), and the vector of extinction probabilities of multi-type branching processes (or Markovian trees).

We then consider extensions of these purely stochastic models to MDPs. We describe a Generalized Newton's Method (GNM), which employs linear programming in each iteration, and we use GNM to obtain a P-time algorithm for computing, to desired accuracy, the vector of optimal (maximum or minimum) extinction probabilities for a given Branching MDP.

We also study one-counter MDPs, which generalize discrete-time QBDs with control, and we give algorithms and complexity bounds for both qualitative and quantitative analysis of optimal termination probabilities for one-counter MDPs.

Finally, we consider a model more general than the above stochastic models and MDPs, called recursive Markov chains (RMCs) and Recursive MDPs (RMDPs). RMCs are closely related to probabilistic pushdown systems and to tree-like-QBDs. We show that, in the worst-case, any non-trivial approximation of the termination probabilities of a RMC (or tree-like-QBD) is PosSLP-hard, and thus approximation in NP would already yield a breakthrough in exact numerical computation. For RMDPs (or tree-like-QBD-MDPs), we show that computing any non-trivial approximation of their optimal termination probability is not even computable at all (with any complexity).

(This talk describes a series of results together with Alistair Stewart and Mihalis Yannakakis, as well as some older results with other co-authors.)

Keywords

quasi-birth death processes (QBDs), Markovian trees, Recursive Markov decision processes (RMDPs).

An algorithmic approach to the extinction of branching processes with countably infinitely many types

[Extended Abstract]

Peter Braunsteins
The University of Melbourne
Department of Mathematics and Statistics
Melbourne, 3010, Australia
petertb@student.unimelb.edu.au

Sophie Hautphenne
The University of Melbourne
Department of Mathematics and Statistics
Melbourne, 3010, Australia
Ecole Polytechnique Fédérale de Lausanne
Institute of Mathematics
Lausanne 1015, Switzerland
sophiemh@unimelb.edu.au
sophie.hautphenne@epfl.ch

ABSTRACT

We consider the extinction events of Galton-Watson processes with countably infinitely many types. In particular, we construct truncated and augmented Galton-Watson processes with finite but increasing sets of types. A pathwise approach is then used to show that, under some sufficient conditions, the corresponding sequence of extinction probability vectors converge to the global extinction probability vector of the Galton-Watson processes with countably infinitely many types. This gives rise to a number of iterative methods for the computation of the global extinction probability vector.

Keywords

Multi-type branching process; extinction probability; iterative methods

1. INTRODUCTION

Consider a multi-type Galton-Watson process with type set $\mathcal{S} \subseteq \mathbb{N}$. Let $\{\boldsymbol{Z}_n : n \in \mathbb{N}\}$ be such that $\boldsymbol{Z}_n$ is a vector whose ith entry, $Z_{n,i}$, contains the number of type i individuals in generation n. We assume the branching process initially contains a single individual, whose type will be denoted by φ_0. The process then evolves according to the following rules:

(i) each individual lives for a single generation, and

(ii) at death it gives birth to $\boldsymbol{r} = (r_1, r_2, ...)$ offspring, that is, r_1 individuals of type 1, r_2 individuals of type 2, etc., where the vector $\boldsymbol{r}$ is chosen independently of all other individuals according to a probability distribution, $p_i(\cdot)$, specific to the parental type $i \in \mathcal{S}$.

From the set of probability distributions $\{p_i(\cdot)\}_{i\in\mathcal{S}}$ we define the *progeny generating function* $\boldsymbol{G} : [0,1]^{\mathcal{S}} \to [0,1]^{\mathcal{S}}$, which has entries,

$$G_i(\boldsymbol{s}) = \sum_{\boldsymbol{r}\in(\mathbb{N}_0)^{\mathcal{S}}} p_i(\boldsymbol{r})\boldsymbol{s}^{\boldsymbol{r}} = \sum_{\boldsymbol{r}\in(\mathbb{N}_0)^{\mathcal{S}}} p_i(\boldsymbol{r}) \prod_{k=1}^{\infty} s_k^{r_k}. \qquad (1)$$

The mean progeny matrix M is an infinite matrix whose entries are given by

$$M_{ij} = \left.\frac{\partial G_i(\boldsymbol{s})}{\partial s_j}\right|_{\boldsymbol{s}=\boldsymbol{1}}, \qquad \text{for } i, j \in \mathcal{S},$$

where M_{ij} can be interpreted as the expected number of type j children born to a parent of type i. We say there is a path from type i to j if there exists ℓ such that $(M^{\ell})_{ij} > 0$. We assume that the row sums of M are finite, that is, the expected total number of direct offspring of an individual of any type is finite.

The branching process, $\{\boldsymbol{Z}_n\}$, is then an infinite dimensional Markov-chain in which the state $\boldsymbol{0}$ is absorbing. We say that $\{\boldsymbol{Z}_n\}$ becomes *globally extinct* once it reaches the absorbing state and we define the global extinction probability vector $\boldsymbol{q}$, with entries

$$q_i = \mathbb{P}(\lim_{n\to\infty} \boldsymbol{Z}_n = \boldsymbol{0} | \varphi_0 = i).$$

The global extinction probability vector is the minimal nonnegative solution to the fixed point equation

$$\boldsymbol{s} = \boldsymbol{G}(\boldsymbol{s}). \qquad (2)$$

In all but a few special cases this fixed point equation cannot be solved analytically. When the set of types is finite, (2) describes a finite system of equations, which can be used to compute $\boldsymbol{q}$ algorithmically, for instance through functional iteration. It is then natural to question whether a similar computational technique exists when the set of types becomes countably infinite, $\mathcal{S} = \{1, 2, 3, ...\}$. However, little work had been done in this area until [1], of which this paper is a continuation.

Importantly, to allow the set of types to become infinite gives rise to a second kind of extinction event, which we refer to as *partial extinction*. This corresponds to the event that every type becomes extinct. We denote by $\tilde{\boldsymbol{q}}$ the partial

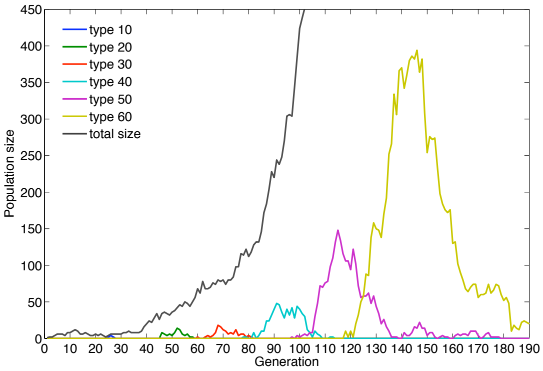


Figure 1: Simulation of the evolution of the population size of different types and the total population size of a branching process with $\boldsymbol{q} < \tilde{\boldsymbol{q}} = \boldsymbol{1}$. Sourced from [1].

extinction probability vector, with entries,

$$\tilde{q}_i = \mathbb{P}(\forall l \in \mathcal{S}, \lim_{n\to\infty} Z_{n,l} = 0 \big| \varphi_0 = i).$$

Although it is clear that global extinction implies partial extinction and thus $\boldsymbol{q} \leq \tilde{\boldsymbol{q}}$, it is indeed possible for every type to become to eventually disappear while the total number of individuals approaches infinity so we may have $\boldsymbol{q} < \tilde{\boldsymbol{q}}$ (see Figure 1).

To calculate $\tilde{\boldsymbol{q}}$ and $\boldsymbol{q}$ the authors of [1] considered two sequences of branching processes which were constructed by modifying the offspring distributions of $\{\boldsymbol{Z}_n\}$. The first sequence of branching processes, $\{\tilde{\boldsymbol{Z}}_n^{(k)}\}$, was modified by making all types strictly greater than k produce no offspring. We refer to these individuals as *sterile*. The second sequence of branching processes, $\{\boldsymbol{Z}_n^{(k)}\}$, was modified by instantaneously replacing each individual with type greater than k by a type-Δ individual, which at death has a single type-Δ offspring with probability 1. These type-Δ individuals can be thought of as *immortal*. The extinction probability vectors of these sequences of branching processes are denoted by $\tilde{\boldsymbol{q}}^{(k)}$ and $\boldsymbol{q}^{(k)}$, respectively. Through application of the monotone convergence theorem it was shown that

$$\tilde{\boldsymbol{q}}^{(k)} \searrow \tilde{\boldsymbol{q}} \quad \text{and} \quad \boldsymbol{q}^{(k)} \nearrow \boldsymbol{q},$$

pointwise as $k \to \infty$. It was observed that the vectors $\tilde{\boldsymbol{q}}^{(k)}$ and $\boldsymbol{q}^{(k)}$ can be calculated using established methods for Galton-Watson processes with finitely many types and consequently, by taking $k \to \infty$ one can then compute $\tilde{\boldsymbol{q}}$ and $\boldsymbol{q}$.

The objective of the present paper is to investigate the intermediate case where, rather than replacing by a sterile or immortal type, we replace all types larger than k with a type randomly selected from the set $\{1, \ldots, k\}$ according to some probability distribution $\boldsymbol{\alpha}^{(k)}$; this defines a new sequence $\{\bar{\boldsymbol{q}}^{(k)}\}_{k\geq 1}$ of extinction probability vectors that can again be computed using established methods for branching processes with finitely many types. In particular, through Theorem 1 we provide sufficient conditions on the branching process $\{\boldsymbol{Z}_n\}$ and on the sequence of replacement distributions $\{\boldsymbol{\alpha}^{(k)}\}$ for

$$\bar{\boldsymbol{q}}^{(k)} \to \boldsymbol{q} \tag{3}$$

pointwise as $k \to \infty$. Observe that when this convergence occurs it is not necessarily monotone. To establish (3) we therefore employ a different method of proof to that of [1]. The statement of Theorem 1 and an outline of the main steps of the proof is given in Section 2. In Section 3 a numerical example is provided. In particular, we give an example where $\bar{\boldsymbol{q}}^{(k)} \nrightarrow \boldsymbol{q}$ when $\{\boldsymbol{\alpha}^{(k)}\}$ does not satisfy the sufficient condition of Theorem 1. The example is also used to compare the rate at which $\boldsymbol{q}^{(k)}$ and $\bar{\boldsymbol{q}}^{(k)}$ converge to $\boldsymbol{q}$. For a more detailed account of the material covered in the remainder of this paper we refer the reader to [2].

2. RANDOM REPLACEMENT

In Theorem 1 we impose two primary conditions:

Assumption 1. *There exists $\beta > 0$ such that*

$$\inf_i q_i \geq \beta.$$

Assumption 2. *There exist constants $N_1, N_2 \geq 1$ and $a > 0$, all independent of k, such that*

$$\sum_{i=1}^{\min\{N_1,k\}} \alpha_i^{(k)} \geq a \qquad \textit{for all } k \geq N_2.$$

Assumption 1 is a commonly applied sufficient condition for $|\boldsymbol{Z}_n| := \sum_{i=1}^{\infty} Z_{n,i}$ to satisfy the dichotomy property. That is, with probability 1, either $|\boldsymbol{Z}_n| \to 0$ or $|\boldsymbol{Z}_n| \to \infty$ as $n \to \infty$. When Assumption 1 is not satisfied (but Assumption 2 is satisfied) one can construct examples in which $\boldsymbol{q} < \lim_{n\to\infty} \bar{\boldsymbol{q}}^{(k)} < \tilde{\boldsymbol{q}}$, however we do not go into detail here. Assumption 2 can be viewed as a more general version of tightness on the sequence of probability distributions $\{\boldsymbol{\alpha}^{(k)}\}$. It ensures that for $k \geq N_2$ with at least probability a an individual is replaced by a type in the set $\{1, \ldots, N_1\}$. Note that Assumption 2 is satisfied if we replace by type 1, $\boldsymbol{\alpha}^{(k)} = \boldsymbol{e}_1$, for example. We consider replacement distributions that do not satisfy Assumption 2 in Section 3.

Theorem 1. *Suppose Assumptions 1 and 2 hold. In addition, assume that there exists N_1 such that either*

(i) *$\tilde{q}_j < 1$ for all $j \in \{1, \ldots, N_1\}$, or*

(ii) *$\tilde{q}_j = 1$ for all $j \in \{1, \ldots, N_1\}$, and there is a path from any $j \in \{1, \ldots, N_1\}$ to the initial type i.*

Then

$$\lim_{k\to\infty} \bar{q}_i^{(k)} \to q_i,$$

for any initial type i.

Observe that either ***(i)*** or ***(ii)*** is satisfied if M is irreducible. However, the more general conditions also apply to many reducible cases.

We now give a sketch of the main steps in the proof of Theorem 1.

Step 1. For each $k \geq 1$, we place $\{\tilde{\boldsymbol{Z}}_n^{(k)}\}$, $\{\boldsymbol{Z}_n^{(k)}\}$ and $\{\bar{\boldsymbol{Z}}_n^{(k)}\}$ on a common probability space by constructing each process from the same outcome of $\{\boldsymbol{Z}_n\}$. More specifically, we construct $\{\tilde{\boldsymbol{Z}}_n^{(k)}\}$ from $\{\boldsymbol{Z}_n\}$ by removing all descendants

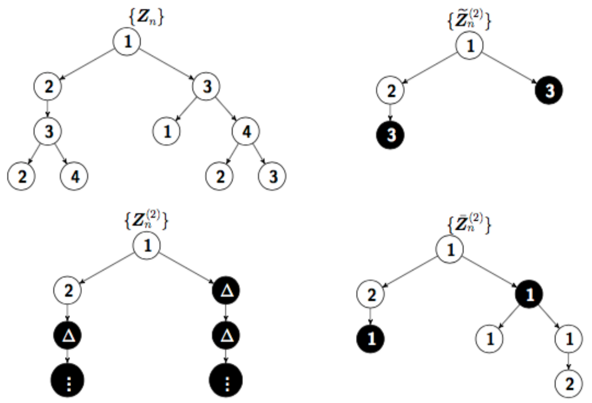

$\{\boldsymbol{Z}_n\}$
1
2
3
3
1
4
2
4
2
3
$\{\tilde{\boldsymbol{Z}}_n^{(2)}\}$
1
2
3
3
$\{\boldsymbol{Z}_n^{(2)}\}$
1
2
Δ
Δ
Δ
⋮
⋮
$\{\bar{\boldsymbol{Z}}_n^{(2)}\}$
1
2
1
1
1
1
2

Figure 2: A realisation of $\{\boldsymbol{Z}_n\}$ and the corresponding outcomes of $\{\tilde{\boldsymbol{Z}}_n^{(k)}\}$, $\{\boldsymbol{Z}_n^{(k)}\}$ and $\{\bar{\boldsymbol{Z}}_n^{(k)}\}$ when $k = 2$.

from individuals with type greater than k (which therefore become sterile). The process $\{\boldsymbol{Z}_n^{(k)}\}$ is then constructed from $\{\tilde{\boldsymbol{Z}}_n^{(k)}\}$ by replacing each sterile individual by an infinite line of descent of type-Δ individuals. Similarly, $\{\bar{\boldsymbol{Z}}_n^{(k)}\}$ is constructed from $\{\tilde{\boldsymbol{Z}}_n^{(k)}\}$ by replacing each sterile individual by an independent copy of $\{\bar{\boldsymbol{Z}}_n^{(k)}\}$ whose root is randomly selected using the probability distribution $\boldsymbol{\alpha}^{(k)}$. A visualisation of each branching process for an outcome on this probability space is given in Figure 2.

Step 2. We define $\{S_k : k \in \mathbb{N}\}$ to count the number of sterile individuals produced over the lifetime of $\{\tilde{\boldsymbol{Z}}_n^{(k)}\}$ (these are illustrated in black in the top right of Figure 2) for different truncation points k. To ensure $S_k < \infty$ with probability 1, we let $S_k = 0$ if $\{\tilde{\boldsymbol{Z}}_n^{(k)}\}$ does not become extinct. That is, we let

$$S_k = \begin{cases} \sum_{n=1}^{\infty} \sum_{i=k+1}^{\infty} (\tilde{\boldsymbol{Z}}_n)_i, & \text{if } \{\tilde{\boldsymbol{Z}}_n^{(k)}\} \text{ becomes extinct} \\ 0, & \text{otherwise.} \end{cases}$$

Each member of S_k can be thought of as giving $\{\bar{\boldsymbol{Z}}_n^{(k)}\}$ an i.i.d chance to survive. We therefore have,

$$\bar{q}_i^{(k)} - q_i^{(k)} = \mathbb{E}_i \left[\left(\sum_{\ell=1}^{k} \alpha_\ell^{(k)} \bar{q}_\ell^{(k)} \right)^{S_k} - \mathbb{1}\{S_k = 0\} \right], \qquad (4)$$

where $\mathbb{E}_i(\cdot) \equiv \mathbb{E}(\cdot|\varphi_0 = i)$. Under Assumption 1 we then show that S_k also satisfies the dichotomy property. That is,

$$\mathbb{P}_i(S_k \to 0 \text{ or } \infty) = 1. \qquad (5)$$

Step 3. We prove that under Assumption 2 and the additional conditions of Theorem 1 that,

$$\mathbb{P}_i(S_k \to \infty) > 0$$

implies that there exists $\varepsilon > 0$ and $N \in \mathbb{N}$ such that

$$\sum_{\ell=1}^{k} \alpha_\ell^{(k)} \bar{q}_\ell^{(k)} < 1 - \varepsilon \qquad (6)$$

for all $k > N$.

Step 4. From Equation (4) for any $K \in \mathbb{N}$,

$$\begin{aligned} \bar{q}_i^{(k)} - q_i^{(k)} \le \\ \mathbb{E}_i \left[\left(\sum_{\ell=1}^{k} \alpha_\ell^{(k)} \bar{q}_\ell^{(k)} \right)^{S_k} \middle| \, 0 < S_k < K \right] \mathbb{P}_i(0 < S_k < K) \\ (7) \\ + \left(\sum_{\ell=1}^{k} \alpha_\ell^{(k)} \bar{q}_\ell^{(k)} \right)^{K} \mathbb{P}_i(S_k \ge K). \qquad (8) \end{aligned}$$

Using Equations (5) and (6), we then show that for any $\varepsilon_2 < 0$ there exists K and $N(K)$ such that for all $k > N(K)$ we have $\bar{q}_i^{(k)} - q_i^{(k)} < \varepsilon_2$. This concludes the proof.

3. NUMERICAL EXAMPLE

Theorem 1 proves that $\bar{\boldsymbol{q}}^{(k)} \to \boldsymbol{q}$ for a large class of replacement distributions $\{\boldsymbol{\alpha}^{(k)}\}$. For instance it implies convergence when replacement is made by type 1, $\boldsymbol{\alpha}^{(k)} = \boldsymbol{e}_1$. In the following example we also consider replacement distributions that do not satisfy Assumption 2. These are replacement by type k, $\boldsymbol{\alpha}^{(k)} = \boldsymbol{e}_k$, and replacement by a type uniformly distributed on $\{1, \ldots, k\}$, $\boldsymbol{\alpha}^{(k)} = \mathbf{1}/k$.

We consider a modified version of the example of [1, Section 5.1]. That is, we assume $a, c > 0$, $d > 1$ and define

$$G_1(\boldsymbol{s}) = \frac{cd}{t}\, s_2^t + 1 - \frac{cd}{t},$$

and for $i \ge 2$,

$$G_i(\boldsymbol{s}) = \begin{cases} \dfrac{cd}{u}\, s_{i+1}^u + \dfrac{ad}{u}\, s_{i-1}^u + 1 - \dfrac{d(a+c)}{u} & \text{when } i \text{ is odd,} \\ \dfrac{c}{dv}\, s_{i+1}^v + \dfrac{a}{dv}\, s_{i-1}^v + 1 - \dfrac{(a+c)}{dv} & \text{when } i \text{ is even,} \end{cases}$$

where $t = \lceil dc \rceil + 1$, $u = \lceil d(c+a) \rceil + 1$ and $v = \lceil (c+a)/d \rceil + 1$. When $i \ge 2$ the mean progeny matrix M has entries,

$$M_{i,i-1} = ad \quad \text{and} \quad M_{i,i+1} = cd$$

for i odd and

$$M_{i,i-1} = a/d \quad \text{and} \quad M_{i,i+1} = c/d$$

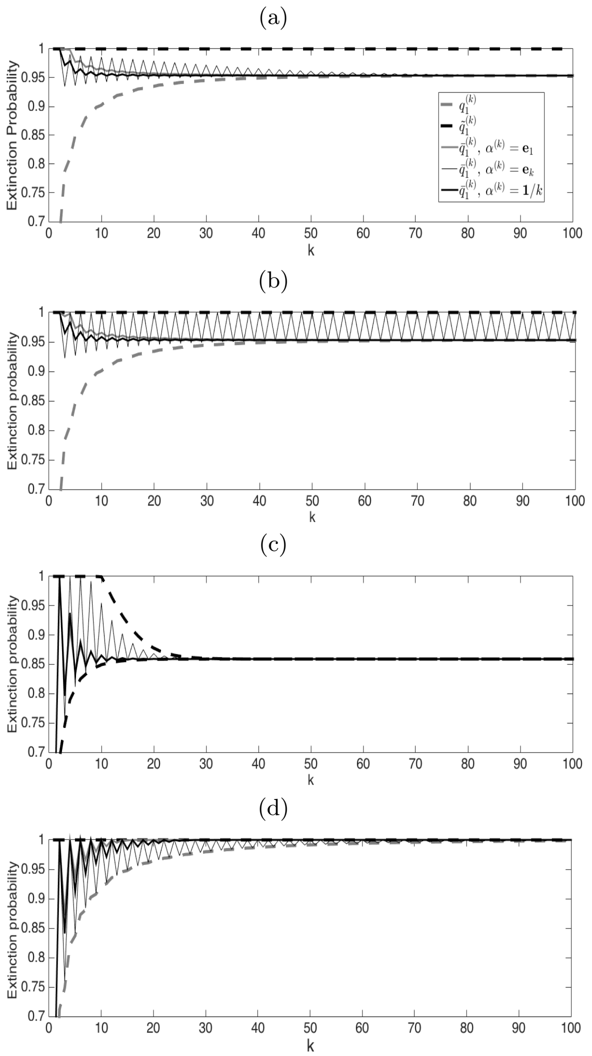
for i even. In this sense one can think of the odd types as stronger than the even types.

In Figure 3 we plot $\tilde{q}_1^{(k)}$ (black dashed), $q_1^{(k)}$ (grey dashed) and $\bar{q}_1^{(k)}$ for $\boldsymbol{\alpha}^{(k)} = \boldsymbol{e}_1$ (solid grey bold), $\boldsymbol{\alpha}^{(k)} = \mathbf{1}/k$ (solid black bold) and $\boldsymbol{\alpha}^{(k)} = \boldsymbol{e}_k$ (solid fine). In the top two plots we let $a = 1/6$ and $c = 7/8$, and choose $d^{-1} = 0.95$ (panel (a)) and $d^{-1} = 0.93$ (panel (b)). We observe that for these values $q_1 < \tilde{q}_1 = 1$. In panel (a) we see that for each sequence $\{\boldsymbol{\alpha}^{(k)}\}$, $\bar{q}_1^{(k)} \to q_1$, with $\boldsymbol{\alpha}^{(k)} = \mathbf{1}/k$ having the fastest rate of convergence. Similar behaviour is displayed in panel (b) except that in this case, when $\boldsymbol{\alpha}^{(k)} = \boldsymbol{e}_k$, we have $\bar{q}_1^{(2k+1)} \to q_1 < 1$ and $\bar{q}_1^{(2k)} \to \tilde{q}_1 = 1$. In panel (c) we let $a = 1/3$ and $c = 13/16$, in which case $q_1 = \tilde{q}_1 < 1$ and in panel (d) we let $a = 1/6$, $c = 13/16$ and $d = 2$ in which case $q_1 = \tilde{q}_1 = 1$. In both panels (c) and (d) $\bar{q}_1^{(k)}$ with $\boldsymbol{\alpha}^{(k)} = \boldsymbol{e}_1$ and $\boldsymbol{\alpha}^{(k)} = \mathbf{1}/k$ perform similarly and again converge to q_1 at a faster rate than $q_1^{(k)}$.

This example demonstrates that when $\boldsymbol{\alpha}^{(k)} = \boldsymbol{e}_k$, the limit of the sequence $\bar{\boldsymbol{q}}^{(k)}$ does not necessarily exist. However, one can prove that, in this particular example, for any values $a, c > 0$ and $d > 1$,

$$\liminf_{k\to\infty} \bar{\boldsymbol{q}}^{(k)} = \boldsymbol{q}.$$

Under Assumption 1 we believe this to be true in general but we are yet to formally prove this claim. Similarly, when $\boldsymbol{\alpha}^{(k)} = \mathbf{1}/k$ one may question whether $\lim_{k\to\infty} \bar{\boldsymbol{q}}^{(k)} = \boldsymbol{q}$, however in this case we can construct an example satisfying Assumption 1 in which $\boldsymbol{q} < \lim_{k\to\infty} \bar{\boldsymbol{q}}^{(k)} = \tilde{\boldsymbol{q}}$.


(a)

(b)

(c)

(d)

Figure 3: Sequences of extinction probabilities $\tilde{q}_1^{(k)}$, $q_1^{(k)}$ and $\bar{q}_1^{(k)}$ for different replacement distributions and different parameters values, corresponding to the example given in Section 3. Details are given the text.

4. REFERENCES

[1] S. Hautphenne, G. Latouche and G. Nguyen. Extinction probabilities of branching processes with countably infinitely many types. *Advances in Applied Probability*, 45(4):1068–1082, 2013

[2] P. Braunsteins, G. Decrouez and S. Hautphenne. A pathwise iterative approach to the extinction of branching processes with countably many types. *arXiv preprint arXiv:1605.03069*, 2016.

Optimal Control of Service Rates in a MAP/M/1 Queue [*]

[Abstract] [†]

Li Xia
CFINS, Department of Automation, TNList
Tsinghua University
Beijing 100084, China
xial@tsinghua.edu.cn

Qi-Ming He
Department of Management Sciences
University of Waterloo
Waterloo, Ontario Canada
q7he@uwaterloo.ca

Attahiru Sule Alfa
Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, Manitoba Canada
attahiru.alfa@umanitoba.ca

ABSTRACT

Service rate control is a classical optimization problem of queueing theory. There are considerable research efforts on this topic, from the simple M/M/1 queue, to tandem queue, cyclic queue, and closed Jackson networks. The objective of the service rate control is to identify the optimal values of service rates of every server at every state such that the system average performance (cost) is maximized (minimized).

In this paper, we study the service rate control problem in a MAP/M/1 queue. The arrival process is a Markovian arrival process (MAP). The service rate is allowed to be state-dependent, i.e. we can adjust the service rate according to the queue length and the phase of the MAP. The cost function consists of holding cost and operating cost. We use the matrix-analytic methods (MAM) together with the sensitivity-based optimization (SBO) theory to study this problem. A performance difference formula is derived, which can quantify the difference of the long-run average total cost under any two different settings of service rates. Based on the difference formula, we show that the long-run average total cost is monotone in the service rate and the optimal control is a bang-bang control. We also show that, under some mild conditions, the optimal control policy of service rates is of a quasi threshold-type. By utilizing the MAM theory, we propose a recursive algorithm to compute the value function related quantities. An iterative algorithm to efficiently find the optimal policy, which is similar to policy iteration, is proposed based on the SBO theory. Numerical examples demonstrate the main results and explore the impact of the MAP phase on the optimization in the MAP/M/1 queue.

This paper shows the potentials of combining MAM and SBO theories to study the performance optimization of queueing systems. For future work, we may further consider the server assignment problem, optimization with constraints such as finite buffer or resource constraint, and load-dependent service rate control, etc.

Keywords

Queueing system, Markov decision process, matrix-analytic methods, sensitivity-based optimization, service rate control

[*] This work was supported in part by the National Natural Science Foundation of China (61573206, 61203039, U1301254), the National 111 International Collaboration Project (B06002), the Specialized Research Fund for the Doctoral Program of Higher Education (20120002120009).

[†] A full version of this paper is under review of the *IEEE Transactions on Automatic Control* as a regular paper.

Stationary analysis of finite MAP/PH/1 queue with bi-level hysteretic control of arrivals

[Extended Abstract]

Rostislav Razumchik
Institute of Informatics Problems
of the Federal Research Center “Computer Science and Control”
of the Russian Academy of Sciences
Vavilova, 44-2
Moscow, Russia
rrazumchik@ipiran.ru

Introduction

This short note provides an overview of the motivation and results of the ongoing research devoted to the development of analytic methods for the steady-state analysis of one class of queueing systems: queueing systems with hysteretic control of arrivals. In fact queueing systems with different types of hysteretic control have been a subject of extensive research for many years and a plenty of models and results are available in the literature. But, to our knowledge, among them the least concern queueing systems with hysteretic control of arrivals. As it will be mentioned further such systems can be used for performance evaluation of network signalling nodes (like SIP-servers). Here consideration is given to one of representatives of this class – finite-capacity $MAP/PH/1/r$ queueing system with bi-level hysteretic control of arrivals. The system consists of a single server and a queue of finite capacity r. The arrival process is a MAP with representation $(\mathbf{D_0}, \mathbf{D_1})$ of order N. We assume that an arrival, whenever it occurs, can be of one of the two types: either a priority arrival or a non-priority. Thus the matrix $\mathbf{D_1}$ is assumed to have the form $\mathbf{D_1} = \mathbf{D_{1,1}} + \mathbf{D_{1,2}}$, where $\mathbf{D_{1,1}}$ $(\mathbf{D_{1,2}})$ describes state transitions with an arrival of priority (non-priority) customer. Bi-level hysteretic control of arrivals is assumed to be implemented in the system. It operates as follows (see Fig.1). There are three operation modes for the system: “normal”, “overload”, “blocking”. Let L and H be arbitrary whole numbers such that $0 < L < H < r + 1$. The system starts empty and as long as the total number of customers in the system remains below H, the system is considered to be in “normal“ mode and accepts all arrivals (both priority and non-priority). When the total number of customers reaches H for the first time, the system changes its mode to “overload“ and stays in it as long as the total number of customers remains between L and r. When overloaded, the system accepts only priority customers (non-priority customers are lost) till the total number of customers either drops down below L after which it changes its mode back to “normal“, or exceeds r after which it changes its state to “blocking“. In the “blocking“ mode the system does not accept newly arriving customers until the total number of customers drops down below $(H+1)$, after which the system changes mode back to “overload“ and the process goes on. The service time of both priority and non-priority customers is PH distributed with representation $(\vec{f}, G)$ of order M and $\vec{g} = -G\vec{1}$. We are interested in the development of the analytic method, which allows the efficient calculation of the steady-state characteristics of the system such as blocking probability, moments of the sojourn times in different modes etc.

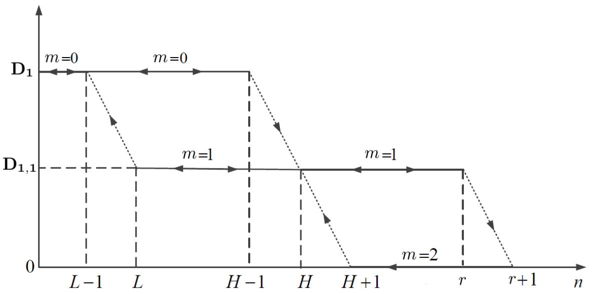


Figure 1: Sketch of the bi-level hysteretic control of arrivals in the $MAP/PH/1/r$ system.

The impulse for this research was initially given by two observations. The first one is the application of such models for the performance evaluation of network signalling nodes, which suffer from overloads. In a number of recent papers it is being reported that overload protection of the essential components of telecommunication networks (like SIP-servers handling signalling messages) is crucial in next and future generation networks. For the detailed description of the SIP-server overload problem and methods of its solution one can refer to IETF RFC’s and several research papers, for example, [1–9]. One of the approaches towards the solution of the overload problem in SIP-servers is the use of the thresholds and specifically hysteretic control of arrivals. Our contribution into this field is that we consider an analytic model of SIP-server based on the same (technical) assumption as in a series of papers (see, for example, [13, 14]), but under more general assumptions about the incoming flow and service times, which are justified by the results of our recent experimental research (see [15]).

The second observation concerns the analytical methods used for the analysis of the queueing models of SIP-servers with hysteretic control of arrivals. For many of the considered queuing models one can use well-known general techniques (like matrix-analytical methods for QBD, see [16] and [17] specifically related to the problem under consideration). Despite the presence of hysteretic loops the methods are applicable and from the theoretical point of view the problem can be considered (to a large extent) solved. But having looked a little deeper into the hysteretic mechanism and models considered by other authors, A. Pechinkin[1] suggested a new general method which allows the steady-state analysis of the whole class of systems with such hysteretic control in a unified way. It is easy implementable and allows the computation of the steady state distribution (and related quantities) for relatively high values of thresholds and not only in purely markovian systems. Below we present the idea of method by applying it to the $MAP/PH/1/r$ system introduced above. Of course, due to MAP arrivals and PH service times the method utilizes matrix analytic techniques and thus possesses the disadvantages inherent to matrix algebra.

Idea of the method

The idea of the method is based on the general property of the restricted Markov chains which says that if the Markov chain, say $X(t)$, with state space, say $\mathcal{X}$, is positive recurrent, then the stationary distribution of the same Markov chain but restricted to subset $F \subseteq \mathcal{X}$ differs from the stationary distribution of the original chain by a constant.

The operation of the $MAP/PH/1/r$ queueing system with bi-level hysteretic control of arrivals can be completely described by continuous-time Markov chain

$$\mathbf{X}(t) = (a(t); s(t); m(t); n(t))$$

with four components: $a(t)$ — MAP generation phase at time t, $s(t)$ — PH service phase at time t, $m(t)$ — system's mode at time t, $n(t)$ — total number of customers in the system at time t. Clearly, when $n(t) = 0$ the second component $s(t)$ is omitted. The state space of $\mathbf{X}(t)$ can be represented as $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \cup \mathcal{X}_2$, where $\mathcal{X}_0$ is the set of states of "normal" mode, $\mathcal{X}_1$ is the set of states of "overload" mode, and $\mathcal{X}_2$ is the set of states of "blocking" mode i.e.

$$\begin{aligned}\mathcal{X}_0 = \{(k,0,0) : 1 \le k \le N\} \cup \\ \{(k,0,n) : 1 \le k \le NM, 1 \le n \le H-1\}, \\ \mathcal{X}_1 = \{(k,1,n) : 1 \le k \le NM, L \le n \le r\}, \\ \mathcal{X}_2 = \{(k,2,n) : 1 \le k \le NM, H+1 \le n \le r+1\}.\end{aligned}$$

Here k represents the state of the background (arrival and service) processes. Indeed the state (k, m, n), $n > 0$ means that there are n customers in the system, system's mode is m, and arrival and service phases are i and j, but such that $(i-1)M + j = k$; the state $(k, 0, 0)$ means that the system is empty and the arrival phase is k. Let $p_{k,m,n}$ be the stationary probability of the state (k, m, n). Introduce the row vectors $\vec{p}_{0,0} = (p_{1,0,0}, \ldots, p_{N,0,0})$ and $\vec{p}_{m,n} = (p_{1,m,n}, \ldots, p_{NM,m,n})$. The idea of the method is to use the property of the restricted Markov chains for the sequential computation of probabilities $\vec{p}_{m,n}$ which is based only on the previously computed values of $\vec{p}_{k,l}$, $0 < k < m$, $0 < l < n$. This can be done as follows. Consider another $MAP/PH/1/0$ queueing system with bi-level hysteretic control of arrivals working in parallel and fed with completely the same MAP flow and PH service times but with no queue. This means that the state set of this new system is

$$\begin{aligned}F = \{(k,0,0) : 1 \le k \le N\} \cup \\ \{(k,0,1) : 1 \le k \le NM\}, \ F \subset \mathcal{X}.\end{aligned}$$

Assume also that in the new system the following rule applies: whenever a new customers arrives and sees server busy, it pre-empts the customer which is in service and occupies the server. Its service phase remains the same as of the pre-empted customer but the pre-empted customer leaves the system. Denote by $q_{k,0,n}$ the stationary distribution of the state $(k, 0, n)$, $n = 0, 1$, of the new system and introduce vectors $\vec{q}_{0,0} = (q_{1,0,0}, \ldots, q_{N,0,0})$ and $\vec{q}_{0,1} = (q_{1,0,1}, \ldots, q_{NM,0,1})$. Note now that in order to apply the property of the restricted Markov chains which guarantees that the stationary distributions $\vec{q}_{0,n}$ and $\vec{p}_{m,n}$ differ only by a constant one must make sure that the behaviour of the chains is probabilistically the same. This means that whenever there is an arrival at the new system when it is busy, the arrival and service phases must take the same values as the arrival and service phases of the original system sometime in the future, when the number of customers in it drops down back to 1. Indeed, assume that at instant when *both* systems are in the state $(i_1, 0, 1)$ an arrival occurs. Denote by A_1 the matrix of size $NM \times NM$. The $(i,j)^{th}$ entry of A_1 is the probability that at the moment of time when the total number of customers in the system equals 1 for the first time, the state of the background processes is j, given that at initial moment of time it was i and there were 2 customers in the system. In other words the entries of A_1 are the taboo probabilities, i.e.

$$\begin{aligned}[\mathrm{A}_1]_{(i,j)} = \mathbf{P}\{\mathbf{X}(\tau) = (j,0,1); \\ \mathbf{X}(t) \notin \cup_k (k,0,1), t \in (0,\tau) | \mathbf{X}(0) = (i,0,2)\},\end{aligned}$$

where $\tau = \inf\{t > 0 : n(t) = 1\}$. Then in the original system we have a transition to some state $(i_2, 0, 2)$, but in the new system the transition happens to the state, $(i_2, 0, 1)$. It is clear that the number of customers in the original system will evolve somehow and eventually the original system will come back to some state $(i_3, 0, 1)$ (without being empty in the meanwhile). Thus the two chains will behave probabilistically the same if and only if the state $(i_3, 0, 1)$ of the original system turns out to be the same as the state $(i_2, 0, 1)$ of the new system. This event has a probability, which in the described case is exactly equal to $[\mathrm{A}_1]_{(i_2,i_2)}$. Now we can write the balance equations for $\vec{q}_{0,1}$ from rate-in-rate-out principle. Denoting by $\mathbf{E}$ the identity matrix, and noting that $\vec{q}_{0,n}$ and $\vec{p}_{m,n}$ differ only by a constant, we have:

$$\begin{aligned}0 = \vec{p}_{0,0}(\mathbf{D_1} \otimes \vec{f}) + \\ + \vec{p}_{0,1}(\mathbf{D_0} \otimes \mathbf{E} + \mathbf{E} \otimes \mathbf{G}) + \vec{p}_{0,1}(\mathbf{D_1} \otimes \mathbf{E})\mathrm{A}_1.\end{aligned}$$

The next step is to sequentially increase the queue capacity of the new system each time by 1. Thus one considers step-

[1] Alexander Vladimirovich Pechinkin (1948-2014), professor, doctor of Sciences in Physics and Mathematics, principal scientist at the Institute of Informatics Problems of the Russian Academy of Sciences.

by-step $MAP/PH/1/j$ queueing systems but with queue of size $j = 1, 2, \ldots, L-3$. The set of states of such j-system is

$$F = \{(k,0,0) : 1 \leq k \leq N\} \cup \\ \{(k,0,n) : 1 \leq k \leq NM, 1 \leq n \leq j+1\}, \; F \subset \mathcal{X}.$$

Clearly the probability, which was denoted above by A_1, will be different now for each system $MAP/PH/1/j$. Denote it by A_{j+1}. Then applying the same argumentation as above we find that

$$0 = \vec{p}_{0,j}(\mathbf{D_1} \otimes \vec{f}) + \\ + \vec{p}_{0,j+1}(\mathbf{D_0} \otimes \mathbf{E} + \mathbf{E} \otimes \mathbf{G}) + \vec{p}_{0,j+1}(\mathbf{D_1} \otimes \mathbf{E})\mathrm{A}_{j+1},$$

for each $1 \leq j \leq L-3$. Starting from $j = L-2$ the argumentation becomes a little more complicated due to the fact that hysteretic loops allows additional transitions (see Fig.1). But the argumentation remains the same. Matrices A_j (as well as other matrices which are needed to compute all the stationary probabilities) can be calculated in a recursive manner (requiring multiple matrix inversions) using the first step analysis.

We note that the proposed method is also suitable for the calculation of such performance characteristics like moments of system's sojourn time in different modes.

Coming back to the application side of the problem, one should mention that hysteretic control has already been introduced and successfully used in the Signalling System No.7 protocols. Its main objective was to reduce the number of times the control system switches between the operating modes (see [11, 12]). The problem was solved by the choice of thresholds (similar to L and H) in order to minimize the mean time it takes system to switch from the "overload" mode to the "normal" mode. Of independent interest is the problem of the analysis of a queueing network with hysteretic arrival control implemented in the nodes.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (grants 15-07-03406 and 15-07-03007).

References

[1] Rosenberg, J. Requirements for Management of Overload in the Session Initiation Protocol. RFC 5390. 2008.

[2] Baker, F., Fairhurst, G. IETF Recommendations Regarding Active Queue Management. RFC 7567. 2015.

[3] Hilt, V., Noel, E., Shen, C., Abdelal, A. Design Considerations for Session Initiation Protocol (SIP) Overload Control. RFC 6357. 2011.

[4] Ohta, M. Overload Control in a SIP Signalling Network. International Journal of Electrical and Electronics Engineering, 2009. pp. 87–92.

[5] Hilt, V., Widjaja, I. Controlling Overload in Networks of SIP Servers. IEEE International Conference on Network Protocols, 2008. pp. 83–93.

[6] Shen, C., Schulzrinne, H., Nahum, E. Session Initiation Protocol (SIP) Server Overload Control: Design and Evaluation. Lecture Notes in Computer Science, Springer, 2008. Vol. 5310, pp. 149–173.

[7] Garroppo, R. G., Giordano, S., Spagna, S., Niccolini, S. Queueing Strategies for Local Overload Control in SIP Server. IEEE Global Telecommunications Conference, 2009. pp. 1–6.

[8] Montagna, S., Pignolo, M. Load Control techniques in SIP signalling servers using multiple thresholds. 13-th International Telecommunications Network Strategy and Planning Symposium, 2008. pp. 1–17.

[9] Garroppo, R. G., Giordano, S., Niccolini, S., Spagna, S. A Prediction-Based Overload Control Algorithm for SIP Servers. IEEE Transactions on Network and Service Management, Vol. 8, 2011. No 1, pp. 39–51.

[10] Takagi, H. Analysis of a Finite-Capacity M/G/1 Queue with a Resume Level. Performance Evaluation, Vol. 5, 1985. pp. 197–203.

[11] Takshing, Yum, P., Hung-Ming, Yen. Design algorithm for a hysteresis buffer congestion control strategy. IEEE International Conference on Communication. Boston, Massachusetts, 1983. pp. 499–503.

[12] Brown, P., Chemouil P., Delosme B. A Congestion Control Policy for Signalling Networks. Proceedings of 7th International Conference on Computer Communications, Sydney, Australia, 1984. pp.717–724.

[13] Abaev, P., Gaidamaka, Y., Samouylov, K. Modeling of Hysteretic Signalling Load Control in Next Generation. Lecture Notes in Computer Science: Proc. of the 12-th International Conference on Next Generation Wired/Wireless Networking (Winter Session), Germany, Heidelberg: Springer, 2012.

[14] Abaev, P., Gaidamaka, Y., Pechinkin, A., Razumchik, R., Shorgin, S. Simulation of overload control in SIP server networks. Proceedings of the 26th European Conference on Modelling and Simulation, ECMS, 2012. pp. 533–539.

[15] Abaev, P., Razumchik, R., Uglov, I. Statistical analysis of message delay in SIP proxy server. Journal of Telecommunications & Information Technology. 2014. Issue 4. pp. 79–87.

[16] Latouche, G., Ramaswami V. Introduction to Matrix Geometric Methods in Stochastic Modeling. ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia PA, 1999.

[17] Ye J., Li S. Analysis of Multi-Media Traffic Queues with Finite Buffer and Overload Control - Part 1: Algorithm. INFOCOM, 1991. pp. 1464–1474.

Waiting Times in BMAP/BMAP/1 Queues

[Extended Abstract]

Nail Akar
Electrical and Electronics Engineering Dept.
Bilkent University
Ankara, Turkey
akar@ee.bilkent.edu.tr

ABSTRACT

Recently, [1] proposed a computationally efficient and stable numerical algorithm to calculate the steady-state actual waiting time distribution for an infinite-capacity single-server semi-Markov queue with the auto-correlation in inter-arrival and service times modeled by a Markov Renewal Process with a Matrix Exponential kernel, called the MRP-ME process. In this paper, we study the distribution of the waiting time arising in a BMAP/BMAP/1 queue by writing a BMAP as an MRP-ME process, and subsequently using the same numerical algorithm of [1]. Moreover, we generalize a BMAP by GBMAP (Generalized BMAP) by allowing batch sizes to be of more general discrete PH-type as opposed to finite batch sizes. A GBMAP is also shown to be a MRP-ME process through which the waiting times for a GBMAP/GBMAP/1 queue may also be obtained using the same numerical algorithm.

Keywords

BMAP, waiting times, ordered Schur decomposition

1. INTRODUCTION

The Batch Markovian Arrival Process (BMAP) is represented with parameter matrices $D_k, 0 \leq k \leq K$ of size $m \times m$ where the matrix D_0 has negative diagonal elements and nonnegative off-diagonal elements, and the matrices $D_k, k \neq 0$ have non-negative elements; see [2],[7],[8]. A MAP (Markovian Arrival Process) is a special case of BMAP with $D_k = 0, k > 1$ corresponding to the case of a single arrival at a given epoch. BMAPs are known for their versatility and have been used effectively in modeling arrival processes; see the references [2]. If the same structure is used for modeling a service process, we then call it a BMAP (Batch Markovian Service Process). A BMAP can be generalized to a GBMAP (generalized BMAP) if the batch sizes are not finite but instead governed by a discrete PH-type distribution. A GBMAP can be defined accordingly.

In this paper, we provide an all matrix-analytical numerical algorithm to find the distribution of the waiting time in a GBMAP/GBMAP/1 queue. The idea is to represent a GBMAP as a Markov renewal process with matrix exponential kernels and then use the numerical algorithm proposed in [1] based on these representations to obtain the waiting time distribution. Actually, the waiting time can be written in matrix exponential form through which the moments of the waiting time can be obtained in a straightforward manner. The conventional approach for this line of problems is the matrix-analytical approach pioneered by Neuts which does not rely on calculating polynomial roots or eigenvectors. In this approach, the queue occupancy is observed at certain embedded epochs and a structured Markov chain (of M/G/1 or G/M/1 type) is constructed for the queue length [7],[8]. The most computation intensive part of the matrix analytical approach is the solution to a nonlinear matrix equation. Once a solution is obtained for this equation, one can find the queue length distribution recursively [9]. Given the steady-state queue length probabilities, the waiting time distribution and its moments can be obtained although not in a very compact form [6]. The current paper aims at proving an alternative, still matrix-analytical, algorithm to write the waiting time distribution in a BMAP/BMAP/1 queue in a compact matrix-exponential form.

2. MARKOV RENEWAL PROCESSES OF ME-TYPE

We first define a Markov renewal process based on [5]. We define, for each $k \in \mathbb{N}$, a random variable X_k taking values in a finite set $E = \{1, 2, \ldots, n\}$ and a random variable T_k taking values in $\mathbb{R}_+ = [0, \infty)$ such that $0 = T_0 \leq T_1 \leq T_2 \leq \cdots$. The stochastic process $(X, T) = \{X_k, T_k; k \in \mathbb{N}\}$ is called a Markov renewal process (MRP) with state space E provided that

$$P\{X_{k+1} = j, T_{k+1} - T_k \leq t \mid X_0, \ldots, X_k; T_0, \ldots, T_k\}$$

$$= P\{X_{k+1} = j, T_{k+1} - T_k \leq t \mid X_k\},$$

for all $k \in \mathbb{N}$, $1 \leq j \leq n$, and $t \in \mathbb{R}_+$. The time-homogeneous case is te focus of this paper for which the probability

$$F_{ij}(t) = P\{X_{k+1} = j, \Delta_k \leq t \mid X_k = i\}$$

is independent of the customer number k. The matrix $F(t) = \{F_{ij}(t)\}$ is called the semi-Markov kernel of the MRP. Defining $F_{ij} = \lim_{t\to\infty} F_{ij}(t)$, $F_{ij} = P\{X_{n+1} = j \mid X_n = i)$ is the state transition probability from state i to j and we assume

$F = \{F_{ij}\}$ is irreducible. Let π be the stationary solution of this discrete-time Markov chain (DTMC) such that

$$\pi F = \pi, \pi e = 1. \tag{1}$$

We also note that the quantity

$$F_{ij}(t)/F_{ij} = P\{T_{k+1} - T_k \le t \mid X_{k+1} = j, X_k = i\} \tag{2}$$

is the sojourn time distribution in state i when the next state is j. An MRP is of ME-type, or shortly MRP-ME, if $F(t)$ can be written in the following matrix-exponential form:

$$F(t) = Ve^{t\,T}U + F, t \ge 0, \tag{3}$$

and equals zero elsewhere. Here, T is square and of size m and all its eigenvalues have negative real parts. Moreover, V and U are $n \times m$ and $m \times n$, respectively. The kernel density $G(t), t \ge 0$ is defined as follows:

$$\begin{aligned} G(t) = \frac{d}{dt}F(t) &= Ve^{t\,T}TU + (F + V\,U)\delta(t), && (4)\\ &= Ve^{t\,T}H + D\delta(t), t \ge 0 && (5)\end{aligned}$$

where $\delta(t)$ is the Dirac delta function and $H = TU$ and $D = F + VU$. We also define the Laplace transform $G^*(s)$ of the kernel density matrix:

$$G^*(s) = \int_{0^-}^{\infty} e^{-ts}G(t)dt = V(sI - T)^{-1}H + D. \tag{6}$$

An MRP-ME is then characterized by the quadruple (V, T, H, D). In general, one uses the sojourn times of the MRP, i.e., $T_{k+1} - T_k, k \in \mathbb{N}$, to model inter-arrival or service times in queueing systems. We note that the moments of the sojourn times satisfy

$$E[(T_{k+1} - T_k)^i] = (-1)^{i+1}\, i!\, \pi V T^{-(i+1)} He, i > 0. \tag{7}$$

It is clear that a phase-type renewal process is MRP-ME with one state, i.e., $n = 1$, but with multiple modes, i.e., $m \ge 1$. On the other hand, the Markovian Arrival Process (MAP) is characterized with two square matrices D_0 and D_1 with D_0 having negative diagonal elements and non-negative off-diagonal elements, D_1 being non-negative, and $D = D_0 + D_1$ being an irreducible infinitesimal generator [8]. This MAP is actually an MRP-ME process with a kernel

$$F(t) = (e^{D_0 t} - I)D_0^{-1}D_1$$

and is therefore characterized by the quadruple $(I, D_0, D_1, 0)$ [6]. If the above model is used to describe a service process, then we refer to that as a Markovian Service Process (MSP). The Rational Arrival Process (RAP) of [3] may also be viewed as an MRP-ME characterized by the quadruple $(I, D_0, D_1, 0)$ similar to a MAP but the matrices D_0 and D_1 do not necessarily possess the probabilistic interpretation available for MAPs.

3. MRP-ME/MRP-ME/1 QUEUE

The following Lindley equation in continuous-time is relevant to the current paper:

$$W_{k+1} = (W_k + B_k - A_k)^+ = \max(0, W_k + B_k - A_k), k \ge 0, \tag{8}$$

where A_k and B_k are MRP-MEs and a numerical method to algorithmically obtain the distribution of $W = \lim_{k\to\infty} W_k$ when it exists is proposed in [1]. Relating (8) to a queuing system, B_k and A_k denote the service time of customer k and the inter-arrival times between customers k and $k+1$, respectively, and W_k denotes the kth customer's waiting time in the queue. The sojourn times of MRP-MEs are used to model the processes A_k and B_k. The MRP-ME for the arrival process is characterized with the quadruple

$$(V_A, T_A, H_A, D_A),$$

where T_A is square of size m_A and D_A is square of size n_A. Similarly, the MRP-ME for the service process is characterized with the quadruple

$$(V_B, T_B, H_B, D_B),$$

with T_B being square of size m_B and D_B being square of size n_B. Let the steady-state waiting time density be denoted by $f_W(t)$. The reference [1] shows that W has a matrix-exponential density in the form

$$f_W(t) = ve^{t\,T}h + d\delta(t), \tag{9}$$

with T being square of size $n_A m_B$ and all the factors of this expression can be obtained with a matrix-analytical algorithm. The most computation intensive part of the proposed numerical algorithm in [1] is the ordered Schur decomposition of a coupling matrix of size $n_A m_B + m_A n_B$. We note that obtaining the ordered Schur form is known to be backward stable and has a complexity of $O(n^3)$ [4].

4. (G)BMAP AS AN MRP-ME

A BMAP with characterizing matrices $D_k, 0 \le k \le K$ of size m each can be shown to an MRP-ME characterized with the quadruple

$$(V, T, H, D)$$

where

$$V = \begin{bmatrix} I_{m\times m} \\ 0_{m(K-1)\times m(K-1)} \end{bmatrix}, H = \begin{bmatrix} D_1 & D_2 & \cdots & D_K \end{bmatrix},$$

and

$$T = D_0, D = \begin{bmatrix} 0_{m\times m(K-1)} & 0_{m\times m} \\ I_{m(K-1)\times m(K-1)} & 0_{m(K-1)\times m} \end{bmatrix}.$$

In a BMAP, an event corresponding to the parameter matrix D_k is bound to a batch arrival with size k. In a GBMAP, we allow an event corresponding to $D_k, k \ge 1$ is a batch arrival corresponding to class-k traffic and we further assume that the batch size governed by class-k arrivals is discrete PH-type distributed with matrix pair (α_k, S) where the substochastic matrix S of size $n \times n$ is shared by all classes. We show that this GBMAP has the quadruple representation (V, T, H, D) where

$$V = \begin{bmatrix} I_{m\times m} \\ 0_{mn\times m} \end{bmatrix}, H = \begin{bmatrix} \sum_{k=1}^K D_k\gamma_k & \sum_{k=1}^K \beta_k \otimes D_k, \end{bmatrix}$$

$$T = D_0, D = \begin{bmatrix} 0_{m\times m} & 0_{mn\times m} \\ s \otimes I_m & S \otimes I_m \end{bmatrix},$$

and

$$s = (I - S)1_{n\times 1}, \gamma_k = \alpha_k s, \beta_k = \alpha_k S.$$

It is clear that, BMAP is a special case of GBMAP. Moreover, once the MRP-ME characterization is available for a GBMAP, then the steady-state waiting time distribution of a GBMAP/GBMAP/1 queue can be obtained in matrix exponential form by employing the algorithm provided in [1].

5. ACKNOWLEDGMENTS

The author would like to thank the Science and Research Council of Turkey (TUBITAK) for supporting this study under project grant EEEAG-115E360.

6. REFERENCES

[1] N. Akar and K. Sohraby. System-theoretical Algorithmic Solution to Waiting Times in semi-Markov Queues. *Perform. Eval.*, 66(11):587–606, nov 2009.

[2] J. Artalejo, A. Gomez-Corral, and Q. He. Markovian arrivals in stochastic modelling: a survey and some new results. *Statistics and Operations Research Transactions*, 34:101–144, 2011.

[3] S. Asmussen and M. Bladt. Point processes with finite-dimensional conditional probabilities. *Stoch. Proc. Appl.*, 82(1), 1999.

[4] Z. Bai and J. W. Demmel. On swapping diagonal blocks in real Schur form. *Lin. Alg. Appl.*, 186:73–95, 1993.

[5] E. Çınlar. *Introduction to Stochastic Processes.* Prentice Hall, 1975.

[6] A. Heindl. *Traffic-Based Decomposition of General Queueing Networks with Correlated Input Processes.* Shaker Verlag, Aachen, Germany, 2001. Ph. D. Thesis.

[7] D. M. Lucantoni. The BMAP/G/1 queue: A tutorial. In L. Donatiello and R. Nelson, editors, *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, pages 330–358. Springer-Verlag, 1993.

[8] M. F. Neuts. *Structured Stochastic Matrices of $M/G/1$ Type and TheirApplications.* Marcel Dekker, Inc., New York, 1989.

[9] V. Ramaswami. A stable recursion for the steady state vector in Markov chains of M/G/1 type. *Commun. Statist.- Stochastic Models*, 4:183–263, 1988.

An *MAP/PH/K* Queue with Constant Impatient Time

Qi-Ming He
University of Waterloo
200 University Avenue West
Waterloo, Ontario, Canada
1-519-888-4567 x. 35907
q7he@uwaterloo.ca

ABSTRACT

In this paper, we describe the formatting guidelines for ACM SIG Proceedings. This paper analyzes an *MAP/PH/K* queue with customer abandonment. Customer impatient time is assumed to be constant. Using the Laplace-Stieltjes transform method, a set of equations is established for the age process of the first customer waiting in queue. Based on the equations, for a few special cases, computational procedures are developed for computing performance measures such as loss probability, and distributions and moments of waiting times and queue lengths. (Joint work with Drs. Qishu Cai and Zhuo Huang)

Keywords

Queues, impatience customers, abandonment, matrix-analytic methods

Analysis of tandem fluid queues

Małgorzata M. O'Reilly[*]
School of Mathematics
University of Tasmania
Tas 7001, Australia
malgorzata.oreilly@utas.edu.au

Werner Scheinhardt
Department of Applied Mathematics
University of Twente
The Netherlands
w.r.w.scheinhardt@utwente.nl

ABSTRACT

We consider a model consisting of two fluid queues driven by the same background continuous-time Markov chain, such that the rates of change of the fluid in the second queue depend on whether the first queue is empty or not. We analyse this tandem model using operator-analytic methods.

Keywords: tandem, stochastic fluid model, Markov chain, Laplace-Stieltjes transform, transient analysis, limiting distribution.

1. INTRODUCTION

Stationary distributions of Markov-modulated fluid queues have been studied extensively, first using spectral methods [3], later via more efficient matrix-analytic methods [8, 9, 10, 11, 13, 17]. The analysis of networks of fluid queues is much harder, and only for a few special two-node cases the stationary joint distribution of both queue contents and the regulating Markov chain could be obtained.

However, a promising approach to find further results is the use of operator-analytic methods, studied in Bean and O'Reilly [4, 5], where a tandem model is considered, and also in Margolius and O'Reilly [16], where a time-varying queue is analysed. The operator-analytic methods generalise the matrix-analytic methods for single queues. In this work we show that this can indeed lead to good results.

The main difference with the tandem model in [4] is that here we consider fluid queues that have a lower bound, i.e., they can become empty but the content cannot become negative. The tandem model in [5] also considers queues with a lower bound, but the assumptions are slightly different and the results derived there are largely theoretical. Here, we derive numerical methods for a generalization of the tandem model in [14], for which the analytical results could be obtained by considering an embedded $M/G/1$ queue.

2. TANDEM FLUID QUEUE: MODEL AND PRELIMINARIES

In this section we first describe the model of interest and then give the stability condition. We end with some preliminary statements about the sample paths that can be taken by the model, and the implications for the shape (in particular the support) of the stationary distribution.

2.1 Model description

We consider two fluid queues, collecting fluid in buffers X and Y, with level variables recording the content at time t denoted by $X(t)$ and $Y(t)$, respectively, that are being driven by the same background continuous-time Markov chain $\{\varphi(t) : t \geq 0\}$ with some finite state space S and irreducible generator $\mathbf{T}$. The first queue behaves as a standard fluid queue $\{(\varphi(t), X(t)) : t \geq 0\}$ studied in [10], with a lower boundary at level 0, and real-valued fluid rates r_i collected in a diagonal matrix $\mathbf{R} = diag(r_i)_{i \in \mathcal{S}}$. Thus, the content $X(t)$ increases at rate r_i when $\varphi(t) = i$, unless r_i is negative and $X(t)$=0. More precisely,

$$\frac{d}{dt}X(t) = r_{\varphi(t)} \qquad \text{when } X(t) > 0,$$
$$\frac{d}{dt}X(t) = \max(0, r_{\varphi(t)}) \qquad \text{when } X(t) = 0.$$

We partition the state space $\mathcal{S}$ as $\mathcal{S} = \mathcal{S}_+ \cup \mathcal{S}_- \cup \mathcal{S}_{\bigcirc}$, where $r_i > 0$ when $i \in \mathcal{S}_+$ (states in $\mathcal{S}_+$ will be called upstates), $r_i < 0$ when $i \in \mathcal{S}_-$ (states in $\mathcal{S}_-$ will be called downstates), and $r_i = 0$ when $i \in \mathcal{S}_{\bigcirc}$ (states in $\mathcal{S}_{\bigcirc}$ will be called zero-states). With the behaviour at $X(t) = 0$ in mind it will sometimes be helpful to use additional notation $\mathcal{S}_{\ominus} = \mathcal{S}_- \cup \mathcal{S}_{\bigcirc}$ for the set of "zero-states at $X(t) = 0$". After appropriately ordering the states in $\mathcal{S}$ we can write $\mathbf{T}$ as 3×3 block matrix,

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{++} & \mathbf{T}_{+-} & \mathbf{T}_{+\bigcirc} \\ \mathbf{T}_{-+} & \mathbf{T}_{--} & \mathbf{T}_{-\bigcirc} \\ \mathbf{T}_{\bigcirc+} & \mathbf{T}_{\bigcirc-} & \mathbf{T}_{\bigcirc\bigcirc} \end{bmatrix}. \tag{1}$$

Further, we assume that the behaviour of the second fluid queue depends on both $\varphi(t)$ and $X(t)$ in the following way. Assuming fluid rates $\widehat{c}_i > 0$ and $\check{c}_i < 0$ for all $i \in \mathcal{S}$, collected in $\widehat{\mathbf{C}} = diag(\widehat{c}_i)_{i \in \mathcal{S}}$ and $\check{\mathbf{C}} = diag(\check{c}_i)_{i \in \mathcal{S}}$, we have

$$\frac{d}{dt}Y(t) = \widehat{c}_{\varphi(t)} > 0 \qquad \text{when } X(t) > 0,$$
$$\frac{d}{dt}Y(t) = \check{c}_{\varphi(t)} < 0 \qquad \text{when } X(t) = 0, Y(t) > 0,$$
$$\frac{d}{dt}Y(t) = \widehat{c}_{\varphi(t)} \cdot 1\{\varphi(t) \in \mathcal{S}_+\} \qquad \text{when } X(t) = 0, Y(t) = 0.$$

Thus, the fluid level $Y(t)$ increases when $X(t) > 0$, and decreases when $X(t) = 0$, unless both levels are at 0; in the latter case $Y(t)$ (and $X(t)$) increases as soon as $\varphi(t)$ makes a transition from $\mathcal{S}_{\ominus}$ to $\mathcal{S}_+$.

[*]This research is supported by Australian Research Council Linkage Project LP140100152.

Throughout we denote by $\mathbf{1}$, $\mathbf{0}$, $\mathbf{I}$ and $\mathbf{O}$ a column vector of ones, a row vector of zeros, an identity matrix, and a zero matrix of appropriate sizes, respectively. Also, for any matrix $\mathbf{A} = [A_{ij}]$, we use notation $|\mathbf{A}|$ for a matrix collecting absolute values of the elements of $\mathbf{A}$, with $|\mathbf{A}| = [\ |A_{ij}|\ ]$.

2.2 Stability condition

The stability condition for the first queue, $\{(\varphi(t), X(t)) : t \geq 0\}$, is well-known to be

$$\sum_{i \in \mathcal{S}} r_i P(\varphi = i) < 0, \tag{2}$$

where the random variable φ is distributed according to the stationary distribution of $\varphi(t)$. Assuming this condition is satisfied, the second queue (buffer Y) will be stable when the expected increase rate of $Y(t)$ is less than the expected decrease rate, i.e.,

$$\sum_{i \in \mathcal{S}} \widehat{c}_i P(\varphi = i, X > 0) < \sum_{i \in \mathcal{S}_\ominus} |\breve{c}_i| P(\varphi = i, X = 0), \tag{3}$$

where the random vector (φ, X) is distributed according to the stationary distribution of $(\varphi(t), X(t))$.

2.3 Qualitative behaviour

In this subsection we give a short discussion of how the process $\{(\varphi(t), X(t), Y(t)) : t \geq 0\}$ behaves and what the stationary distribution looks like. Here, and in the sequel, we will sometimes write e.g. 'the process hits $x = 0$', which will be short for 'the process $(\varphi(t), X(t), Y(t))$ hits the set $\mathcal{S} \times \{0\} \times [0, \infty)$', or we will speak of 'the probability mass at $x = 0, y > 0$' meaning 'the stationary probability that the process $(\varphi(t), X(t), Y(t))$ is in the set $\mathcal{S} \times \{0\} \times (0, \infty)$'.

Typically the process alternates, between:

(i) periods on $x = 0$, with $Y(t)$ decreasing, possibly being halted at $x = 0, y = 0$, and $\varphi(t)$ in $\mathcal{S}_\ominus$; such a period starts at $x = 0, y > 0$, with $\varphi(t)$ in $\mathcal{S}_-$ and ends at $x = 0, y > 0$ or at $x = 0, y = 0$ as soon as $\varphi(t)$ makes a transition from $\mathcal{S}_\ominus$ to $\mathcal{S}_+$;

(ii) periods on $x > 0$, with $Y(t)$ increasing, while $X(t)$ can either increase and decrease. Such a period starts where the previous type (i) period ended with $\varphi(t) \in \mathcal{S}_+$ and $X(t)$ increasing, and ends at $x = 0, y > 0$ with $\varphi(t)$ in $\mathcal{S}_-$ as soon as $X(t)$ decreases to 0.

Note that in stationarity, the process can not be at $y = 0$, $x > 0$, since $Y(t) = 0$ implies $X(t) = 0$ (or alternatively, $X(t) > 0$ implies $Y(t) > 0$). In fact when a type (ii) period starts from $x = 0, y = 0$, due to a transition of $\varphi(t)$ to some phase $i \in \mathcal{S}_+$, the process will move with $\frac{d}{dt}X(t) = r_i > 0$ and $\frac{d}{dt}Y(t) = \widehat{c}_i > 0$, so it will stay on the line $\{(x, y) : y = x\widehat{c}_i/r_i\}$ until some future transition of $\varphi(t)$ to some other state i'. Note that the slope of any such path leaving the origin is at least $\min_{i \in \mathcal{S}_+}\{\widehat{c}_i/r_i\}$, and also after the path has been left, the slope of the ensuing path can never be less than this value (assuming $i' \in \mathcal{S}_+$, otherwise $X(t)$ will not increase). Thus, after the process has hit the origin for the first time (which it will, due to stability), the set $\{(x, y) : y < x \cdot \min_{i \in \mathcal{S}_+}\{\widehat{c}_i/r_i\}\}$ can never be reached.

As a consequence of the above, the stationary distribution will have the following form.

- Corresponding to (i), there will be a (one-dimensional) density at $x = 0, y > 0$, denoted by $\boldsymbol{\pi}(0, y)$, and a probability point mass at $(0, 0)$, denoted by $\mathbf{p}(0, 0)$.
- Corresponding to (ii), there will be a two-dimensional density on $\{(x, y) : x > 0, y > x \cdot \min_{i \in \mathcal{S}_+}\{\widehat{c}_i/r_i\}\}$, denoted as $\boldsymbol{\pi}(x, y)$, and there will be one-dimensional densities on each of the lines $y = x\widehat{c}_i/r_i$, $i \in \mathcal{S}_+$, denoted as $\pi^i(x, x\widehat{c}_i/r_i)$. Also, define $\boldsymbol{\pi}^j(x, x\widehat{c}_j/r_j) = [\delta_{ij}\pi^j(x, x\widehat{c}_j/r_j)]_{i \in \mathcal{S}}$ for all $j \in \mathcal{S}$. There will be no other probability masses or densities, in particular there is no density at $y = 0, x > 0$.

It is important to realize that the one- and two-dimensional densities just mentioned, as well as the point mass at (0,0), are all *vectors* with $|\mathcal{S}|$ components, where the i-th component corresponds to $\varphi(t) = i$. Some of these components will be zero; in particular for $i \in \mathcal{S}_+$ we will have $[\mathbf{p}(0, 0)]_i = 0$ and $[\boldsymbol{\pi}(0, y)]_i = 0$. Also $[\boldsymbol{\pi}^j(x, x\widehat{c}_j/r_j)]_i = 0$ for all $i \neq j$.

In the next section we show how to proceed to find the stationary distribution.

3. TANDEM FLUID QUEUE: ANALYSIS

Roughly speaking, our analysis is based on the alternation between (i) stages during which $X(t) = 0$ and hence $Y(t)$ decreases, and (ii) stages during which $X(t) > 0$ and hence $Y(t)$ increases, as detailed in Section 2.3. For (parts of) both of these stages we will apply ideas from [4, 18], in order to keep track of the amount by which $Y(t)$ increases (or decreases), in much the same way as we can keep track of the amount of time that passes. We will review this in Section 3.1. In Section 3.2 we will look at the state $(\varphi(t), X(t))$ when the process hits the line $x = 0$, so that with these building blocks we can in Section 3.3 establish expressions for the stationary distribution.

3.1 Replacing time by shift

We are interested in certain behaviour of buffer X, not during some amount of *time*, but while buffer Y experiences a certain (downward/upward, virtual) *shift*. For a motivation of the expressions below we refer to [4], where the concept of shift was introduced, as well as to [18], where a generalization of this idea is discussed. We will consider two cases.

(i) The behaviour at $x = 0$, when the level in buffer Y is strictly decreasing, according to the rates in $\breve{\mathbf{C}}$;

(ii) The behaviour at $x > 0$, when the level in buffer Y is strictly increasing, according to the rates in $\widehat{\mathbf{C}}$.

First, consider the behaviour at $x = 0$, when the level in buffer Y is strictly decreasing, according to the rates in $\breve{\mathbf{C}}$. Below we define matrices $\breve{\mathbf{Q}}_{\ominus\ominus}$ and $\mathbf{Q}_{\ominus+}$ which are the key components of the analysis for this case.

Suppose $X(0) = 0$ and $\varphi(u) \in \mathcal{S}_\ominus$ for $0 \leq u \leq t$. Define the random variable $D(t)$,

$$D(t) = \int_{u=0}^{t} |\breve{c}_{\varphi(u)}| du, \tag{4}$$

interpreted as the total *downward shift* $Y(0) - Y(t)$ in buffer Y at time t when $Y(t) > 0$. Also, for any $z > 0$ define

$$t_z = \inf\{t > 0 : D(t) = z\}, \tag{5}$$

which we interpret, for any $y \geq 0$, as the first time at which the level in the buffer Y shifts from level $Y(0) = y + z$ to y.

Denote

$$\mathbf{T}_{\ominus\ominus} = \left[\begin{array}{cc} \mathbf{T}_{--} & \mathbf{T}_{-\bigcirc} \\ \mathbf{T}_{\bigcirc -} & \mathbf{T}_{\bigcirc\bigcirc} \end{array} \right] \tag{6}$$

and

$$\mathbf{T}_{\ominus +} = \left[\begin{array}{c} \mathbf{T}_{-+} \\ \mathbf{T}_{\bigcirc +} \end{array} \right], \quad \mathbf{T}_{\pm\bigcirc} = \left[\begin{array}{c} \mathbf{T}_{+\bigcirc} \\ \mathbf{T}_{-\bigcirc} \end{array} \right], \tag{7}$$

and let $\breve{\mathbf{C}}_{\ominus} = diag(\breve{c}_i)_{i \in \mathcal{S}_{\ominus}}$ be a diagonal matrix partitioned according to $\mathcal{S}_{\ominus} = \mathcal{S}_{-} \cup \mathcal{S}_{\bigcirc}$.

We define the generator matrix

$$\breve{\mathbf{Q}}_{\ominus\ominus} = (|\breve{\mathbf{C}}_{\ominus}|)^{-1}\mathbf{T}_{\ominus\ominus}, \tag{8}$$

which has the following physical interpretation. By the analysis in [10, Lemmas 1-2], for $i, j \in \mathcal{S}_{\ominus}$, and $z > 0$, we have

$$\begin{aligned} [e^{\breve{\mathbf{Q}}_{\ominus\ominus} z}]_{ij} &= P(\varphi(t_z) = j, \varphi(u) \in \mathcal{S}_{\ominus}, 0 \leq u \leq t_z \\ &\quad \mid \varphi(0) = i, X(0) = 0), \end{aligned} \tag{9}$$

which, for any $y > 0$, we interpret as the probability that the process is in phase j at time t_z and the phase remains in the set $\mathcal{S}_{\ominus}$ at least until time t_z, given the process starts from phase i with empty buffer X and level $y+z$ in buffer Y.

Also, define

$$\breve{\mathbf{Q}}_{\ominus +} = (|\breve{\mathbf{C}}_{\ominus}|)^{-1}\mathbf{T}_{\ominus +}, \tag{10}$$

which by [10, Lemma 2], is a matrix of transition rates, w.r.t. level, to phases in $\mathcal{S}_{+}$, corresponding to the moments at which the level in buffer Y begins to increase.

Second, consider the behaviour at $x > 0$, when the level in buffer Y is strictly increasing according to the rates in $\widehat{\mathbf{C}}$. The key components of the analysis are matrices $\widehat{\mathbf{Q}}(s)$ and $\widehat{\mathbf{\Psi}}(s)$ to be defined below and interpreted afterwards.

Let

$$\theta = \inf\{t > 0 : X(t) = 0\}, \tag{11}$$

be the first time at which the level in buffer X reaches 0.

Suppose $X(0) > 0$, or $X(0) = 0$ and $\varphi(0) \in \mathcal{S}_{+}$; and $t \leq \theta$. Define the random variable $U(t)$,

$$U(t) = \int_{u=0}^{t} \widehat{c}_{\varphi(u)} du, \tag{12}$$

interpreted as the total *upward shift* $Y(t) - Y(0)$ in buffer Y at time t.

We define the key generator matrix $\widehat{\mathbf{Q}}(s)$,

$$\widehat{\mathbf{Q}}(s) = \left[\begin{array}{cc} \widehat{\mathbf{Q}}(s)_{++} & \widehat{\mathbf{Q}}(s)_{+-} \\ \widehat{\mathbf{Q}}(s)_{-+} & \widehat{\mathbf{Q}}(s)_{--} \end{array} \right], \tag{13}$$

with

$$\begin{aligned} &\widehat{\mathbf{Q}}(s)_{++} = (\mathbf{R}_{+})^{-1} \left(\mathbf{T}_{++} - s\widehat{\mathbf{C}}_{+} - \mathbf{T}_{+\bigcirc}(\mathbf{T}_{\bigcirc\bigcirc} - s\widehat{\mathbf{C}}_{\bigcirc})^{-1}\mathbf{T}_{\bigcirc +}\right), \\ &\widehat{\mathbf{Q}}(s)_{+-} = (\mathbf{R}_{+})^{-1} \left(\mathbf{T}_{+-} - \mathbf{T}_{+\bigcirc}(\mathbf{T}_{\bigcirc\bigcirc} - s\widehat{\mathbf{C}}_{\bigcirc})^{-1}\mathbf{T}_{\bigcirc -}\right), \\ &\widehat{\mathbf{Q}}(s)_{-+} = (|\mathbf{R}_{-}|)^{-1} \left(\mathbf{T}_{-+} - \mathbf{T}_{-\bigcirc}(\mathbf{T}_{\bigcirc\bigcirc} - s\widehat{\mathbf{C}}_{\bigcirc})^{-1}\mathbf{T}_{\bigcirc +}\right), \\ &\widehat{\mathbf{Q}}(s)_{--} = (|\mathbf{R}_{-}|)^{-1} \left(\mathbf{T}_{--} - s\widehat{\mathbf{C}}_{-} - \mathbf{T}_{-\bigcirc}(\mathbf{T}_{\bigcirc\bigcirc} - s\widehat{\mathbf{C}}_{\bigcirc})^{-1}\mathbf{T}_{\bigcirc -}\right), \\ &\widehat{\mathbf{C}}_{+} = diag(\widehat{c}_i)_{i \in \mathcal{S}_{+}}, \widehat{\mathbf{C}}_{-} = diag(\widehat{c}_i)_{i \in \mathcal{S}_{-}}, \widehat{\mathbf{C}}_{\bigcirc} = diag(\widehat{c}_i)_{i \in \mathcal{S}_{\bigcirc}}. \end{aligned}$$

The physical interpretation of $\widehat{\mathbf{Q}}(s)$ was established in [4, Theorem 2]. For completeness, we state this result in Theorem 1 below. Now, for any $s > 0$, we can find the minimum nonnegative solution $\widehat{\mathbf{\Psi}}(s)$ of the Riccati equation

$$\widehat{\mathbf{Q}}(s)_{+-} + \widehat{\mathbf{Q}}(s)_{++}\widehat{\mathbf{\Psi}}(s) + \widehat{\mathbf{\Psi}}(s)\widehat{\mathbf{Q}}(s)_{--} + \widehat{\mathbf{\Psi}}(s)\widehat{\mathbf{Q}}(s)_{-+}\widehat{\mathbf{\Psi}}(s) = \mathbf{O}, \tag{14}$$

which has the following interpretation, by the analysis in [4, Theorem 3]. For all $i \in \mathcal{S}_{+}$ and $j \in \mathcal{S}_{-}$,

$$[\widehat{\mathbf{\Psi}}(s)]_{ij} = E(e^{-sU(\theta)} 1\{\varphi(\theta) = j\} \mid \varphi(0) = i, X(0) = 0), \tag{15}$$

is the Laplace-Stieltjes transform of the distribution of the upward shift in buffer Y at the moment the level in buffer X first returns to 0 and does so in phase j, given start from phase i and empty buffer X. We can write

$$\widehat{\mathbf{\Psi}}(s) = \int_{z=0}^{\infty} e^{-sz}\widehat{\boldsymbol{\psi}}(z)dz, \tag{16}$$

where the entry $[\widehat{\boldsymbol{\psi}}(z)]_{ij}$, for $i \in \mathcal{S}_{+}$ and $j \in \mathcal{S}_{-}$, is the corresponding probability density, which can be derived by numerically inverting $[\widehat{\mathbf{\Psi}}(s)]_{ij}$ using the algorithm by Abate and Whitt [1], for any $z > 0$. That is, the matrix $\widehat{\boldsymbol{\psi}}(z)$ is an $|\mathcal{S}_{+}| \times |\mathcal{S}_{-}|$ matrix of densities, the (i, j)-th component of which records the density of an upward shift of z in the buffer Y, from some y to $y + z$, during a busy period of the buffer X, ending in phase $j \in \mathcal{S}_{-}$, starting at phase $i \in \mathcal{S}_{+}$.

In the remainder of this section we will give a slightly enhanced proof of Theorem 2 in [4]. This theorem gives the matrix recording the Laplace-Stieltjes transforms of the distribution of the shift in buffer Y, during the time that an amount x has flown into or out of the buffer X, ending up in phase j given that it starts in i. In [4] this matrix was called[1] $\tilde{\Delta}^{y}(s)$, while in the current paper we will write it as $\mathbf{U}^{(x)}(s)$. But more importantly, we will modify its definition somewhat, to reflect the fact that the value of the shift in buffer Y does not only depend on the initial phase i, the ending phase j, and the time duration, but on the whole sample path of $\varphi(t)$ in between. For the moment we will assume that, in our current context, $Y(t)$ can only increase, so that the shift in buffer Y, expressed as $Y(t) - Y(0)$, is always nonnegative[2].

Let, as in [4], $f(t) = \int_0^t |r_{\varphi(u)}|du$ be the total amount of fluid that flowed into or out of buffer X during $(0, t)$, referred to as the *in-out fluid* of X, and let $\omega(x) = \inf\{t > 0 : f(t) = x\}$ be the first time this in-out fluid reaches level x. Moreover, let now $V^x = \{\varphi(u), 0 \leq u \leq \omega(x)\}$ denote the whole path of $\varphi(t)$ during this interval, and let V_i^x be the set of all such paths that can be taken, starting from $\varphi(0) = i$, such that the total in-out fluid in buffer X is precisely x.

Denoting the duration of any path v by $|v|$, let $U(|v|)$ be the total shift in the second buffer during $(0, |v|)$; note that this random variable is completely determined by the path v. Then we formally define the matrix $\mathbf{U}^{(x)}(s)$ via its (i, j)-th

[1] with superscript y rather than x, since unfortunately the monotonously increasing (or decreasing) buffer, in which the shift is measured, was there called X, so the notations for X and Y are interchanged.

[2] i.e. we only consider the case $X(t) > 0$; the case $X(t) = 0$ is similar, except that we should replace the word 'shift' by 'virtual shift', as if the buffer Y had no lower boundary at 0.

entry as follows,

$$[\mathbf{U}^{(x)}(s)]_{ij} = \int_{v \in V_i^x} e^{-sU(|v|)} 1\{\varphi(|v|) = j\} dP(V = v), \tag{17}$$

where the integral incorporates the (countable) number of all possible successive states that $\varphi(t)$ visits, as well as all the corresponding sojourn times during all of these visits (adding up to $\omega(x)$). Using this definition we can prove the following result.

Theorem 1. *(Theorem 2 in Bean and O'Reilly [4])*

$$\mathbf{U}^{(x+h)}(s) = \mathbf{U}^{(x)}(s)\mathbf{U}^{(h)}(s),$$

from which it follows that

$$\mathbf{U}^{(x)}(s) = e^{\widehat{\mathbf{Q}}(s)x}.$$

Proof. First note that any path $v \in V_i^{x+h}$ can be seen as a concatenation of two paths, $v_1 \in V_i^x$, ending in some phase k, and $v_2 \in V_k^h$ representing the in/outflow increase in buffer X from x to $x+h$. Due to the Markov property these paths are independent, conditional on v_2 starting in the same phase k as where v_1 finished. Since in that case clearly we also have $U(|v|) = U(|v_1|) + U(|v_2|)$, we arrive at

$$\begin{aligned} & e^{-sU(|v|)} 1\{\varphi(|v|) = j\} dP(V = v) \\ = & \sum_k e^{-sU(|v_1|)} 1\{\varphi(|v_1|) = k\} dP(V = v_1) \\ & \times e^{-sU(|v_2|)} 1\{\varphi(|v_2|) = j\} dP(V = v_2), \end{aligned}$$

from which we find

$$\begin{aligned} & \int_{v \in V_i^{x+h}} e^{-sU(|v|)} 1\{\varphi(|v|) = j\} dP(V = v) \\ = & \sum_k \int_{v \in V_i^x} e^{-sU(|v|)} 1\{\varphi(|v|) = k\} dP(V = v) \\ & \times \int_{v \in V_k^h} e^{-sU(|v|)} 1\{\varphi(|v|) = j\} dP(V = v), \end{aligned}$$

and hence the first statement follows. For the proof of the second statement we can simply refer to [4]. □

3.2 Embedded discrete-time Markov chain

Let θ_k be the k-th time that $(\varphi(t), X(t), Y(t))$ hits the line $x = 0$, and let the discrete-time Markov chain $J_k = (\varphi(\theta_k), Y(\theta_k))$ with discrete/continuous state space $\mathcal{S}_- \times (0, \infty)$, record the position of $(\varphi(t), Y(t))$ at time θ_k. Also, let $\tau_k > \theta_k$ be the k-th time the process *leaves* the boundary $x = 0$.

Lemma 1. *The transition kernel of J_k is given by*

$$\begin{aligned} \mathbf{P}_{z,y} \;=\; & \int_{u=[z-y]^+}^{z} \begin{bmatrix} \mathbf{I} & \mathbf{O} \end{bmatrix} e^{\check{\mathbf{Q}}_{\ominus\ominus} u} \check{\mathbf{Q}}_{\ominus +} \widehat{\boldsymbol{\psi}}(y - z + u) du \\ & + \begin{bmatrix} \mathbf{I} & \mathbf{O} \end{bmatrix} e^{\check{\mathbf{Q}}_{\ominus\ominus} z} (-\check{\mathbf{Q}}_{\ominus\ominus})^{-1} \check{\mathbf{Q}}_{\ominus +} \widehat{\boldsymbol{\psi}}(y). \end{aligned} \tag{18}$$

where $[x]^+$ denotes $\max(0, x)$*, and $\begin{bmatrix} \mathbf{I} & \mathbf{O} \end{bmatrix}$ is a $|\mathcal{S}_-| \times |\mathcal{S}_\ominus|$ matrix.*

Proof. We apply the physical interpretations of the quantities analysed in Section 3.1. Essentially, the process J_k satisfies a Lindley-type recursion, since for its second component $Y(\theta_k)$ we can write

$$Y(\theta_{k+1}) = [Y(\theta_k) - D_k]^+ + U_k, \tag{19}$$

where

$$D_k = \int_{u=\theta_k}^{\tau_k} |\check{c}_{\varphi(u)}| du, \quad U_k = \int_{u=\tau_k}^{\theta_{k+1}} \widehat{c}_{\varphi(u)} du \tag{20}$$

are appropriately chosen random variables. More precisely, starting from time θ_k, with $X(\theta_k) = 0$ and $\varphi(\theta_k) = i \in \mathcal{S}_-$, we recall the two consecutive stages described in Section 2.3.

First, (i) the process $Y(t)$ will make a negative shift of size $-D$, say, as long as $\varphi(t) \in \mathcal{S}_\ominus$ (while $X(t)$ remains at zero during this stage). Then, after a transition of $\varphi(t)$ from $\mathcal{S}_\ominus$ to $\mathcal{S}_+$, the second stage (ii) commences, during which the process $Y(t)$ will make a positive shift of size U, say, during a busy period of the first queue (i.e., during a first return time of $X(t)$ back to level zero, starting at level zero).

There are two alternatives. The first alternative is that the chain J_k transitions from (i, z) to (j, y) without the level in the buffer Y returning to 0 during time interval (θ_k, θ_{k+1}). Assume $y \geq z$. In this case,

- first the phase remains in the set $\mathcal{S}_\ominus$ at least until the level in buffer Y shifts down by u units (from z to $z - u$), for some u with $0 \leq u \leq z$; this occurs according to the probability matrix $e^{\check{\mathbf{Q}}_{\ominus\ominus} u}$;
- then the process makes a transition to some phase in $\mathcal{S}_+$, which starts the busy period in buffer X; this occurs according to the rate matrix $\check{\mathbf{Q}}_{\ominus +}$;
- finally, the busy process in buffer X ends and the level y is observed in buffer Y; this occurs according to the density matrix $\widehat{\psi}(y - z + u)$ since the shift in buffer Y during the busy period in X must be exactly $y - (z - u) = y - z + u$.

The transition kernel of the first alternative, when $y \geq z$, is therefore

$$I(y \geq z) \begin{bmatrix} \mathbf{I} & \mathbf{O} \end{bmatrix} \int_{u=0}^{z} e^{\check{\mathbf{Q}}_{\ominus\ominus} u} \check{\mathbf{Q}}_{\ominus +} \widehat{\boldsymbol{\psi}}(y - z + u) du, \tag{21}$$

and by analogous argument, when $y < z$,

$$I(y < z) \begin{bmatrix} \mathbf{I} & \mathbf{O} \end{bmatrix} \int_{u=z-y}^{z} e^{\check{\mathbf{Q}}_{\ominus\ominus} u} \check{\mathbf{Q}}_{\ominus +} \widehat{\boldsymbol{\psi}}(y - z + u) du. \tag{22}$$

The second alternative is that the chain J_k transitions from (i, z) to (j, y) with the level in the buffer Y returning to 0 some time during time interval (θ_k, θ_{k+1}). In this case,

- first the phase remains in the set $\mathcal{S}_\ominus$ at least until the level in buffer Y shifts down by z units (from z to 0); this occurs according to the probability matrix $\int_{u=z}^{\infty} e^{\check{\mathbf{Q}}_{\ominus\ominus} u} du = e^{\check{\mathbf{Q}}_{\ominus\ominus} z} (-\check{\mathbf{Q}}_{\ominus\ominus})^{-1}$;
- then the process makes a transition to some phase in $\mathcal{S}_+$, which starts the busy period in buffer X; this occurs according to the rate matrix $\check{\mathbf{Q}}_{\ominus +}$;
- finally, the busy process of buffer X ends at level y; this occurs according to the density matrix $\widehat{\psi}(y)$.

The transition kernel of the second alternative is

$$\begin{bmatrix} \mathbf{I} & \mathbf{O} \end{bmatrix} e^{\check{\mathbf{Q}}_{\ominus\ominus} z} (-\check{\mathbf{Q}}_{\ominus\ominus})^{-1} \check{\mathbf{Q}}_{\ominus +} \widehat{\boldsymbol{\psi}}(y), \tag{23}$$

and so the result follows by summing (21)–(23). □

We could work with the above directly, but in Section 4 we prefer to determine the following Laplace-Stieltjes transforms, which can then be inverted using the algorithm in Abate and Whitt [1]. We note that $\mathbf{P}_{z,y}$ is continuous w.r.t. $y > 0$, and it is easy to check that $\int_{y=0}^{\infty} \mathbf{P}_{z,y} dy \mathbf{1} = \mathbf{1}$.

Corollary 1. *The Laplace-Stieltjes transform of $\mathbf{P}_{z,y}$ w.r.t. y is given by the matrix*

$$\begin{aligned}\mathbf{P}_{z,\cdot}(s) \;=\; & \left[\begin{array}{cc}\mathbf{I} & \mathbf{O}\end{array}\right] e^{-sz}\left(\check{\mathbf{Q}}_{\ominus\ominus}+s\mathbf{I}\right)^{-1}\left(e^{(\check{\mathbf{Q}}_{\ominus\ominus}+s\mathbf{I})z}-\mathbf{I}\right)\\ & \times\check{\mathbf{Q}}_{\ominus+}\widehat{\mathbf{\Psi}}(s)\\ & +\left[\begin{array}{cc}\mathbf{I} & \mathbf{O}\end{array}\right] e^{\check{\mathbf{Q}}_{\ominus\ominus}z}(-\check{\mathbf{Q}}_{\ominus\ominus})^{-1}\check{\mathbf{Q}}_{\ominus+}\widehat{\mathbf{\Psi}}(s). \qquad (24)\end{aligned}$$

Proof. By straightforward computation of $\int_{y=0}^{\infty} e^{-sy}\mathbf{P}_{z,y}dy$, or by using (19) directly as follows. Letting $Y_k = Y(\theta_k)$ and $\varphi_k = \varphi(\theta_k)$ for notational convenience, we have

$$\begin{aligned}& E[e^{-sY_{k+1}}1\{\varphi_{k+1}=j\} \mid Y_k = z, \varphi_k = i]\\ &= E[e^{-s(z-D_k+U_k)}1\{\varphi_{k+1}=j\}1\{D_k \le z\} \mid Y_k = z, \varphi_k = i]\\ &\quad + E[e^{-sU_k}1\{\varphi_{k+1}=j\}1\{D_k > z\} \mid Y_k = z, \varphi_k = i].\end{aligned}$$

By conditioning on the phases m and ℓ just before and after the time when the process leaves $x = 0$, we rewrite the first term as

$$\begin{aligned}& E[e^{-s(z-D_k+U_k)}1\{\varphi_{k+1}=j\}1\{D_k \le z\} \mid Y_k = z, \varphi_k = i]\\ &= \sum_{m\in S_\ominus}\sum_{\ell\in\mathcal{S}_+} e^{-sz}\cdot E[e^{sD_k}1\{\varphi(\tau_k-)=m\}\\ &\qquad \times 1\{D_k \le z\} \mid Y_k = z, \varphi_k = i]\\ &\qquad \times E[1\{\varphi(\tau_k)=\ell\} \mid \varphi(\tau_k-)=m]\\ &\qquad \times E[e^{-sU_k}1\{\varphi_{k+1}=j\} \mid \varphi(\tau_k)=\ell]\\ &= \sum_{m\in\mathcal{S}_\ominus}\sum_{\ell\in\mathcal{S}_+} e^{-sz}\int_{u=0}^{z}\left[\left[\begin{array}{cc}\mathbf{I} & \mathbf{O}\end{array}\right]e^{\check{\mathbf{Q}}_{\ominus\ominus}u}\right]_{im}\\ &\qquad \times e^{su}du\,[\check{\mathbf{Q}}_{\ominus+}]_{m\ell}\,[\widehat{\mathbf{\Psi}}(s)]_{\ell j}\\ &= \left[\left[\begin{array}{cc}\mathbf{I} & \mathbf{O}\end{array}\right]e^{-sz}\left(\check{\mathbf{Q}}_{\ominus\ominus}+s\mathbf{I}\right)^{-1}\right.\\ &\qquad \left.\times\left(e^{(\check{\mathbf{Q}}_{\ominus\ominus}+s\mathbf{I})z}-\mathbf{I}\right)\check{\mathbf{Q}}_{\ominus+}\widehat{\mathbf{\Psi}}(s)\right]_{ij}.\end{aligned}$$

A similar expression can be given for the second term, by which the statement follows. □

We denote the stationary distribution of J_k by a row vector $\boldsymbol{\xi}_z = [\xi_{i,z}]_{i\in\mathcal{S}_-}$ of densities, satisfying

$$\left\{\begin{array}{lcl}\int_{z=0}^{\infty}\boldsymbol{\xi}_z\mathbf{P}_{z,y}dz & = & \boldsymbol{\xi}_y\\ \int_{y=0}^{\infty}\boldsymbol{\xi}_y dy\mathbf{1} & = & 1,\end{array}\right. \qquad (25)$$

and proceed in the next section to express the stationary distribution of the process $(\varphi(t), X(t), Y(t))$ at level $x = 0$ in terms of $\boldsymbol{\xi}_z$.

Remark 1. Instead of (19) we could also have worked with the true Lindley recursion

$$Y(\tau_{k+1}) = [Y(\tau_k) + U_k - D_{k+1}]^+. \qquad (26)$$

This is the approach that was followed in [14]. There, the stationary distribution of the chain, embedded at these times, in fact gave immediately also the stationary distribution of the whole process at $x = 0$, due to a PASTA-like argument related to the workload in an $M/G/1$ queue. However, in the more general model at hand, with possibly multiple phases being visited while $X(t) = 0$, this need not be true; e.g. there may be phases in $\mathcal{S}_\ominus$ from which it is impossible to jump to a state in $\mathcal{S}_+$. Moreover, one disadvantage would be that the stationary distribution of the embedded Markov chain besides having a density for $y > 0$, also has a mass at $y = 0$. Hence we decided to embed at *hitting times* of $x = 0$, in a manner similar to the analysis in [5].

3.3 Stationary distribution

In the following subsections we show how to find the various densities and probability masses that define the joint stationary distribution of the process.

3.3.1 *Density at* $\mathbf{x} = \mathbf{0}, \mathbf{y} > \mathbf{0}$ *and mass at* $\mathbf{x} = \mathbf{0}, \mathbf{y} = \mathbf{0}$

Recall from Section 2.3 that we need expressions for the vectors $\boldsymbol{\pi}(0, y)$ and $\mathbf{p}(0, 0)$, which we give in the following.

Lemma 2. *We have $\boldsymbol{\pi}(0,y) = \left[\begin{array}{cc}\mathbf{0} & \boldsymbol{\pi}(0,y)_\ominus\end{array}\right]$, where*

$$\boldsymbol{\pi}(0,y)_\ominus = \alpha\int_{z=y}^{\infty}\left[\begin{array}{cc}\boldsymbol{\xi}_z & \mathbf{0}\end{array}\right]e^{\check{\mathbf{Q}}_{\ominus\ominus}(z-y)}(|\check{\mathbf{C}}_\ominus|)^{-1}dz, \qquad (27)$$

and $\mathbf{p}(0,0) = \left[\begin{array}{cc}\mathbf{0} & \mathbf{p}(0,0)_\ominus\end{array}\right]$, where

$$\mathbf{p}(0,0)_\ominus = \alpha\int_{z=0}^{\infty}\left[\begin{array}{cc}\boldsymbol{\xi}_z & \mathbf{0}\end{array}\right]e^{\check{\mathbf{Q}}_{\ominus\ominus}z}dz(-\mathbf{T}_{\ominus\ominus})^{-1}. \qquad (28)$$

Here, α is a normalization constant that satisfies

$$\begin{aligned}1 \;=\; & \mathbf{p}(0,0)\mathbf{1} + \int_{y=0}^{\infty}\boldsymbol{\pi}(0,y)dy\mathbf{1} + \sum_{j\in\mathcal{S}_+}\int_{x=0}^{\infty}\pi^j(x, x\widehat{c}_j/r_j)dx\\ & + \int_{x=0}^{\infty}\int_{y=0}^{\infty}\boldsymbol{\pi}(x,y)dydx\mathbf{1}, \qquad (29)\end{aligned}$$

given by

$$\begin{aligned}\alpha \;=\; & \Bigg\{\left[\begin{array}{cc}\boldsymbol{\xi} & \mathbf{0}\end{array}\right](-\mathbf{T}_{\ominus\ominus})^{-1}\Bigg(\mathbf{1}\\ & +\mathbf{T}_{\ominus+}\mathbf{K}^{-1}\left[\begin{array}{cc}(\mathbf{R}_+)^{-1} & \mathbf{\Psi}(|\mathbf{R}_-|)^{-1}\end{array}\right]\\ & \times\left(\mathbf{1}+\mathbf{T}_{\pm\bigcirc}(-\mathbf{T}_{\bigcirc\bigcirc})^{-1}\mathbf{1}\right)\Bigg)\Bigg\}^{-1}, \qquad (30)\end{aligned}$$

where, $\boldsymbol{\xi} = \int_{z=0}^{\infty}\boldsymbol{\xi}_z dz$, $\mathbf{\Psi} = \widehat{\mathbf{\Psi}}(s)|_{s=0}$ and $\mathbf{K} = \widehat{\mathbf{K}}(s)|_{s=0}$ with

$$\widehat{\mathbf{K}}(s) = \widehat{\mathbf{Q}}(s)_{++} + \widehat{\mathbf{\Psi}}(s)\widehat{\mathbf{Q}}(s)_{-+}. \qquad (31)$$

Proof. In *(i)–(iii)* we prove (27)–(30) respectively.

(i) Observe the process whenever the level in buffer X hits 0. Denote by α the corresponding rate such that $E^* = \alpha^{-1}$ is the average time between two hits.

Let $E^*_{z,i}(j, 0, u)$ be the derivative w.r.t. y of the expected time in phase j, $x = 0$ and $y \le u$ until the next hit given start from state $(i, 0, z)$.

Consider the process $\{(\varphi(t), X(t), Y(t)) : t \ge 0\}$ in stationarity. By the argument analogous to [15, Theorem 4.1],

$$P(\phi = j, X = 0, Y \in dy) \;=\; \alpha\sum_{i\in\mathcal{S}_-}\int_{z=y}^{\infty}\xi_{z,i}E^*_{z,i}(j,0,y)dz\cdot dy, \qquad (32)$$

where the integral starts at y since for $z < y$ it is not possible to reach $(j, 0, y)$ from $(i, 0, z)$ without leaving $x = 0$ in between. Since, by adapting the argument in [2, Theorem 3.2.1] to the analysis here,

$$E^*_{z,i}(j,0,y) = 1 \cdot [e^{\breve{\mathbf{Q}}_{\ominus\ominus}(z-y)}]_{ij}/|\breve{c}_j|, \tag{33}$$

equation (27) for $\boldsymbol{\pi}(0,y)$ follows.

(ii) Similar arguments show the expression for (28); for ending up in $(j,0,0)$ from $(i,0,z)$ with $i \in \mathcal{S}_-$ and $z \geq 0$, the process $\varphi(t)$ now needs to stay in $\mathcal{S}_\ominus$ for an amount of 'shift' (rather than time) of $z+w$ for some $w \geq 0$, and end up in phase $j \in \mathcal{S}_\ominus$. We have

$$\begin{aligned}[\mathbf{p}(0,0)]_j &= \alpha \sum_{i\in\mathcal{S}_-} \int_{z=y}^{\infty} \xi_{z,i} \int_{w=0}^{\infty} [e^{\breve{\mathbf{Q}}_{\ominus\ominus}(z+w)}]_{ij}/\breve{c}_j dwdz \\ &= \alpha \sum_{i\in\mathcal{S}_-} \int_{z=y}^{\infty} \xi_{z,i}[e^{\breve{\mathbf{Q}}_{\ominus\ominus}z}(-\mathbf{T}_{\ominus\ominus})^{-1}]_{ij}dz.\end{aligned} \tag{34}$$

(iii) To find α, since this is a constant that does not depend on buffer Y, we only need to consider the process $(\varphi(t), X(t))$, together with the distribution of $\{\varphi(t)\}$ upon hitting $x=0$, which is $\boldsymbol{\xi} = \int_{z=0}^{\infty} \boldsymbol{\xi}_z dz$. The vector $\boldsymbol{\xi}$ is the stationary distribution of the corresponding discrete-time Markov chain with state space $\mathcal{S}_-$ which records the position of $\varphi(t)$ at time θ_k. The vector $\boldsymbol{\xi}$ is is the unique solution of the set of equations

$$\begin{aligned}\left[\begin{array}{cc}\boldsymbol{\xi} & \mathbf{0}\end{array}\right](-\mathbf{T}_{\ominus\ominus})^{-1}\mathbf{T}_{\ominus+}\boldsymbol{\Psi} &= \boldsymbol{\xi}, \\ \boldsymbol{\xi}\mathbf{1} &= 1.\end{aligned} \tag{35}$$

The stationary distribution for the SFM has been derived in the literature in [5, 6, 11, 13, 15, 17] in slightly different contexts. For completeness, we summarize here the results required for the derivation of the stationary distribution of $(\varphi(t), X(t))$, including the probability mass vector at level zero, $\mathbf{p} = \left[\begin{array}{ccc}\mathbf{0} & \mathbf{p}_- & \mathbf{p}_\bigcirc\end{array}\right]$, and the probability density vector, $\boldsymbol{\pi}(x) = \left[\begin{array}{ccc}\boldsymbol{\pi}(x)_+ & \boldsymbol{\pi}(x)_- & \boldsymbol{\pi}(x)_\bigcirc\end{array}\right]$, for all $x>0$. By conditioning on the last time the SFM $(\varphi(t), X(t))$ hits level zero from above, in a manner similar to [5, Theorem 2],

$$\left[\begin{array}{cc}\mathbf{p}_- & \mathbf{p}_\bigcirc\end{array}\right] = \alpha\left[\begin{array}{cc}\boldsymbol{\xi} & \mathbf{0}\end{array}\right](-\mathbf{T}_{\ominus\ominus})^{-1}, \tag{36}$$

and

$$\begin{aligned}\left[\begin{array}{cc}\boldsymbol{\pi}(x)_+ & \boldsymbol{\pi}(x)_-\end{array}\right] &= \left[\begin{array}{cc}\mathbf{p}_- & \mathbf{p}_\bigcirc\end{array}\right]\mathbf{T}_{\ominus+}e^{\mathbf{K}x} \\ &\quad \times\left[\begin{array}{cc}(\mathbf{R}_+)^{-1} & \boldsymbol{\Psi}(|\mathbf{R}_-|)^{-1}\end{array}\right], \\ \boldsymbol{\pi}(x)_\bigcirc &= \left[\begin{array}{cc}\boldsymbol{\pi}(x)_+ & \boldsymbol{\pi}(x)_-\end{array}\right]\mathbf{T}_{\pm\bigcirc} \\ &\quad \times(-\mathbf{T}_{\bigcirc\bigcirc})^{-1}.\end{aligned} \tag{37}$$

Alternatively, (36) can be found by integrating (27) w.r.t. y and adding to (28). Similarly, (37) can be found by integrating $\boldsymbol{\pi}(x,y)$ w.r.t. y and adding $\sum_{j\in\mathcal{S}_+}\boldsymbol{\pi}^j(x, x\widehat{c}_j/r_j)$; the expressions for these quantities will be derived in sections that follow.

Since α is a normalizing constant that solves

$$\mathbf{p}\mathbf{1} + \int_{x=0}^{\infty}\boldsymbol{\pi}(x)dx\mathbf{1} = 1, \tag{38}$$

we have

$$\begin{aligned}\alpha^{-1} &= \left[\begin{array}{cc}\boldsymbol{\xi} & \mathbf{0}\end{array}\right](-\mathbf{T}_{\ominus\ominus})^{-1}\Big(\mathbf{1} \\ &\quad +\mathbf{T}_{\ominus+}\mathbf{K}^{-1}\left[\begin{array}{cc}(\mathbf{R}_+)^{-1} & \boldsymbol{\Psi}(|\mathbf{R}_-|)^{-1}\end{array}\right] \\ &\quad \times\left(\mathbf{1}+\mathbf{T}_{\pm\bigcirc}(-\mathbf{T}_{\bigcirc\bigcirc})^{-1}\mathbf{1}\right)\Big),\end{aligned} \tag{39}$$

and so the expression (30) for α follows. □

Note that α can also be interpreted as the total (stationary) rate of leaving $x=0$, since by (36),

$$\begin{aligned}\left[\begin{array}{cc}\mathbf{p}_- & \mathbf{p}_\bigcirc\end{array}\right]\mathbf{T}_{\ominus+}\mathbf{1} &= -\left[\begin{array}{cc}\mathbf{p}_- & \mathbf{p}_\bigcirc\end{array}\right]\mathbf{T}_{\ominus\ominus}\mathbf{1} \\ &= \alpha\left[\begin{array}{cc}\boldsymbol{\xi} & \mathbf{0}\end{array}\right]\mathbf{1} \\ &= \alpha,\end{aligned} \tag{40}$$

and also as the total (stationary) rate of hitting $x=0$, since by (37) and $\boldsymbol{\Psi}\mathbf{1} = \mathbf{1}$,

$$\begin{aligned}\lim_{x\to 0^+}\boldsymbol{\pi}(x)_-|\mathbf{R}_-|\mathbf{1} &= \left[\begin{array}{cc}\mathbf{p}_- & \mathbf{p}_\bigcirc\end{array}\right]\mathbf{T}_{\ominus+}\mathbf{1} \\ &= \alpha,\end{aligned} \tag{41}$$

with the two forms equivalent, as expected in stationarity.

For the Laplace-Stieltjes transform vector of the density part, denoted as $\boldsymbol{\pi}(0,\cdot)(s) = \int_{z=0}^{\infty} e^{-sy}\boldsymbol{\pi}(0,y)dy$, we have the following.

Corollary 2. *We have* $\boldsymbol{\pi}(0,\cdot)(s) = \left[\begin{array}{cc}\mathbf{0} & \boldsymbol{\pi}(0,\cdot)(s)_\ominus\end{array}\right]$, *where*

$$\begin{aligned}\boldsymbol{\pi}(0,\cdot)(s)_\ominus &= \alpha\int_{z=0}^{\infty}\left[\begin{array}{cc}\boldsymbol{\xi}_z & \mathbf{0}\end{array}\right]e^{\breve{\mathbf{Q}}_{\ominus\ominus}z}(\breve{\mathbf{Q}}_{\ominus\ominus}+s\mathbf{I})^{-1} \\ &\quad \times\left(\mathbf{I}-e^{-(\breve{\mathbf{Q}}_{\ominus\ominus}+s\mathbf{I})z}\right)(|\breve{\mathbf{C}}_\ominus|)^{-1}dz.\end{aligned} \tag{42}$$

Proof. Since

$$\begin{aligned}\boldsymbol{\pi}(0,\cdot)(s)_\ominus &= \int_{y=0}^{\infty}e^{-sy}\alpha\int_{z=y}^{\infty}\left[\begin{array}{cc}\boldsymbol{\xi}_z & \mathbf{0}\end{array}\right] \\ &\quad \times e^{\breve{\mathbf{Q}}_{\ominus\ominus}(z-y)}(|\breve{\mathbf{C}}_\ominus|)^{-1}dzdy \\ &= \alpha\int_{z=0}^{\infty}\left[\begin{array}{cc}\boldsymbol{\xi}_z & \mathbf{0}\end{array}\right]e^{\breve{\mathbf{Q}}_{\ominus\ominus}z} \\ &\quad \times\int_{y=0}^{z}e^{-(\breve{\mathbf{Q}}_{\ominus\ominus}+s\mathbf{I})y}(|\breve{\mathbf{C}}_\ominus|)^{-1}dydz \\ &= \alpha\int_{z=0}^{\infty}\left[\begin{array}{cc}\boldsymbol{\xi}_z & \mathbf{0}\end{array}\right]e^{\breve{\mathbf{Q}}_{\ominus\ominus}z} \\ &\quad \times\left(-e^{-(\breve{\mathbf{Q}}_{\ominus\ominus}+s\mathbf{I})y}(\breve{\mathbf{Q}}_{\ominus\ominus}+s\mathbf{I})^{-1}\Big|_{y=0}^{z}\right)(|\breve{\mathbf{C}}_\ominus|)^{-1}dz,\end{aligned}$$

the result follows. □

3.3.2 *Density at* $\mathbf{x} > \mathbf{0}, \mathbf{y} > \mathbf{0}$

We now proceed to the density vector $\boldsymbol{\pi}(x,y)$ as a function of y for fixed value of x.

Define the Laplace-Stieltjes transform $\boldsymbol{\pi}(x,\cdot)(s)$ such that, $[\boldsymbol{\pi}(x,\cdot)(s)]_i = \int_{y=0}^{\infty} e^{-sy}[\boldsymbol{\pi}(x,y)]_i dy$ for $i \in \mathcal{S}_\ominus$, and $[\boldsymbol{\pi}(x,\cdot)(s)]_i = \int_{y=0}^{\infty} e^{-sy}[\boldsymbol{\pi}(x,y)]_i dy + e^{-sx\widehat{c}_i/r_i}\pi^i(x, x\widehat{c}_i/r_i)$ for $i \in \mathcal{S}_+$.

Lemma 3. *We have*

$$\boldsymbol{\pi}(x,\cdot)(s) = \left[\begin{array}{ccc}\boldsymbol{\pi}(x,\cdot)(s)_+ & \boldsymbol{\pi}(x,\cdot)(s)_- & \boldsymbol{\pi}(x,\cdot)(s)_\bigcirc\end{array}\right]$$

with

$$\left[\begin{array}{cc} \boldsymbol{\pi}(x,\cdot)(s)_+ & \boldsymbol{\pi}(x,\cdot)(s)_- \end{array}\right] = (\boldsymbol{\pi}(0,\cdot)(s)_\ominus + \mathbf{p}(0,0)_\ominus) \times \mathbf{T}_{\pm\bigcirc} e^{\widehat{\mathbf{K}}(s)x} \times \left[\begin{array}{cc} (\mathbf{R}_+)^{-1} & \widehat{\boldsymbol{\Psi}}(s)(|\mathbf{R}_-|)^{-1} \end{array}\right], \quad (43)$$

and

$$\boldsymbol{\pi}(x,\cdot)(s)_\bigcirc = \left[\begin{array}{cc} \boldsymbol{\pi}(x,\cdot)(s)_+ & \boldsymbol{\pi}(x,\cdot)(s)_- \end{array}\right] \times \mathbf{T}_{\pm\bigcirc}(s\widehat{\mathbf{C}}_\bigcirc - \mathbf{T}_{\bigcirc\bigcirc})^{-1}. \quad (44)$$

Proof. The result follows immediately by a partitioning of the sample paths argument, analogous to the one used in the derivation of (37). □

Corollary 3. *Letting* $\boldsymbol{\pi}(\cdot,\cdot)(v,s) = \int_{x=0}^{\infty} e^{-vx}\boldsymbol{\pi}(x,\cdot)(s)dx$, *we have*

$$\boldsymbol{\pi}(\cdot,\cdot)(v,s) = \left[\begin{array}{ccc} \boldsymbol{\pi}(\cdot,\cdot)(v,s)_+ & \boldsymbol{\pi}(\cdot,\cdot)(v,s)_- & \boldsymbol{\pi}(\cdot,\cdot)(s)_\bigcirc \end{array}\right]$$

with

$$\left[\begin{array}{cc} \boldsymbol{\pi}(\cdot,\cdot)(v,s)_+ & \boldsymbol{\pi}(\cdot,\cdot)(v,s)_- \end{array}\right] = (\boldsymbol{\pi}(0,\cdot)(s)_\ominus + \mathbf{p}(0,0)_\ominus) \times \left[\begin{array}{c} \mathbf{T}_{-+} \\ \mathbf{T}_{\bigcirc+} \end{array}\right] (-\widehat{\mathbf{K}}(s) + v\mathbf{I})^{-1} \left[\begin{array}{cc} (\mathbf{R}_+)^{-1} & \widehat{\boldsymbol{\Psi}}(s)(|\mathbf{R}_-|)^{-1} \end{array}\right], \quad (45)$$

and

$$\boldsymbol{\pi}(\cdot,\cdot)(s)_\bigcirc = \left[\begin{array}{cc} \boldsymbol{\pi}(\cdot,\cdot)(s)_+ & \boldsymbol{\pi}(\cdot,\cdot)(s)_- \end{array}\right] \mathbf{T}_{\pm\bigcirc} \times (s\widehat{\mathbf{C}}_\bigcirc - \mathbf{T}_{\bigcirc\bigcirc})^{-1}. \quad (46)$$

3.3.3 Density at $y = x\widehat{c}_i/r_i$

Finally, we state the result for the one-dimensional densities on each of the lines $y = x\widehat{c}_i/r_i$, $i \in \mathcal{S}_+$.

Lemma 4. *For all* $i \in \mathcal{S}_+$,

$$\pi^i(x, x\widehat{c}_i/r_i) = \sum_{j\in\mathcal{S}_\ominus} \mathbf{p}_j(0,0)T_{ji}\exp(-(T_{ii}/r_i)x)/r_i. \quad (47)$$

Proof. This result essentially follows by arguments analogous to the proof of the first equation in (37), in a slightly different environment.

By conditioning on the most recent time the process leaves the point $(0,0)$, in order to observe the process in stationarity at the point $(x, x\widehat{c}_i/r_i)$, the following must occur.

- First, the process starts from state $(j,0,0)$ for some $j \in \mathcal{S}_\ominus$, with probability $\mathbf{p}_j(0,0)$, and instantaneously transitions to phase i at a rate T_{ji}.
- Next, the process remains in phase i at least for the duration of time x/r_i, with probability $\exp(-(T_{ii}/r_i)x)$.

Denote by $E(i, x, x\widehat{c}_i/r_i)$ the expected number of visits to state $(i, x, x\widehat{c}_i/r_i)$ given the process starts in state $(i,0,0)$ and avoids returning to level 0 in both buffer X and Y. Clearly, $E(i, x, x\widehat{c}_i/r_i) = 1 \cdot \exp(-(T_{ii}/r_i)x)$.

Further, we note that, by [2, Theorem 3.2.1],

$$\pi^i(x, x\widehat{c}_i/r_i) = \sum_{j\in\mathcal{S}_\ominus} \mathbf{p}_j(0,0)T_{ji}E(i, x, x\widehat{c}_i/r_i)/r_i, \quad (48)$$

and the result (47) follows. □

3.4 Main Result

We now summarize the results for the stationary distribution of the process $\{(\varphi(t), X(t), Y(t)) : t \geq 0\}$.

Theorem 2. *The probability mass components of the stationary distribution, corresponding to* $x = 0$, *are*

$$\boldsymbol{\pi}(0,y) \quad \textit{and} \quad \mathbf{p}(0,0),$$

given in Lemma 2. The Laplace-Stieltjes transforms of $\boldsymbol{\pi}(0,y)$ *w.r.t.* y *are given in Corollary 2.*

The one-dimensional density components of the stationary distribution, corresponding to $y = x\widehat{c}_j/r_j$, *are*

$$\boldsymbol{\pi}^j(x, x\widehat{c}_j/r_j) = [\delta_{ij}\pi^j(x, x\widehat{c}_j/r_j)]_{i\in\mathcal{S}}, \quad j \in \mathcal{S}_+,$$

given in Lemma 4.

The Laplace-Stieltjes transforms of the two-dimensional density components of the stationary distribution, $\boldsymbol{\pi}(x,y)$, *corresponding to* $x > 0$, *w.r.t.* y, *are*

$$[\boldsymbol{\pi}(x,\cdot)(s)]_i, \quad i \in \mathcal{S}_\ominus$$

and

$$[\boldsymbol{\pi}(x,\cdot)(s)]_i - e^{-sx\widehat{c}_i/r_i}\pi^i(x, x\widehat{c}_i/r_i), \quad i \in \mathcal{S}_+,$$

given in Lemma 3. The corresponding Laplace-Stieltjes transforms w.r.t. x *and* y *are given in Corollary 3.*

4. TANDEM FLUID QUEUE: NUMERICAL TREATMENT

In order to evaluate the stationary distribution of the model using the theoretical results of Section 3, we apply discretization and truncation with appropriate parameters Δu, and L, $\ell = 0,1,2,\ldots L$. The key points of the methodology are summarized below.

Step 1. Construct discretized version of the process J_k discussed in Section 3.2, with a truncated level variable as follows.

Fix some small $\Delta u > 0$ and some large integer $L > 0$, and consider a discrete-time Markov chain $\{\bar{J}_k : k = 0,1,2,\ldots\}$ with state space $\{(i,\ell) : i \in \mathcal{S}_-, \ell = 0,1,2,\ldots L\}$, with the interpretation that when $J_k = (j,z)$ for some z with $\ell\Delta u \leq z < (\ell+1)\Delta u$, $\ell = 0,1,2,\ldots(L-1)$, then we have $\bar{J}_k = (j,\ell)$, and when $J_k = (j,z)$ with $z \geq L\Delta u$, we let $\bar{J}_k = (j,L)$.

(i). Approximate the corresponding one-step transition probabilities $P_{i,\ell;j,m} = P(\bar{J}_{k+1} = (j,m)|\bar{J}_k = (i,\ell))$, which are collected in matrix $\mathbf{P} = [\mathbf{P}_{\ell m}]_{\ell,m=0,1,2,\ldots,L}$ made of block matrices $\mathbf{P}_{\ell m} = [P_{i,\ell;j,m}]_{i,j\in\mathcal{S}_-}$ as follows.

First, for $\ell, m = 0,1,2,\ldots L$, evaluate

$$\tilde{\mathbf{P}}_{\ell m} = \int_{y=m\Delta u}^{(m+1)\Delta u} \mathbf{P}_{\ell\Delta u,y}dy, \quad (49)$$

and then normalize $\tilde{\mathbf{P}}_{\ell m}$ to obtain $\mathbf{P}_{\ell m}$ so that

$$\sum_{m=0}^{L} \mathbf{P}_{\ell m}\mathbf{1} = \mathbf{1}. \quad (50)$$

(ii). Next, with the notation $\lim_{k\to\infty} P(\bar{J}_k = (j,\ell)) = \bar{\xi}_{j;\ell}$ whenever the limits exist, denote by $\bar{\boldsymbol{\xi}} = [\bar{\boldsymbol{\xi}}_\ell]_{\ell=0,1,2,\ldots L}$, $\bar{\boldsymbol{\xi}}_\ell = [\bar{\xi}_{j;\ell}]_{j\in\mathcal{S}_-}$, the stationary distribution vector of the process $\{\bar{J}_k : k = 0, 1, 2, \ldots\}$. Derive $\bar{\boldsymbol{\xi}}$ by solving the set of equations, using standard methods,

$$\bar{\boldsymbol{\xi}}\mathbf{P} = \bar{\boldsymbol{\xi}}, \quad \bar{\boldsymbol{\xi}}\mathbf{1} = \mathbf{1}. \tag{51}$$

Step 2. Approximate the values of stationary distribution of the process $\{(\varphi(t), X(t), Y(t)) : t \geq 0\}$ as follows.

(i). For any z with $\ell\Delta u \leq z < (\ell{+}1)\Delta u$, $\ell = 0, 1, 2, \ldots L$, approximate

$$\boldsymbol{\xi}_z \approx \frac{\bar{\boldsymbol{\xi}}_\ell}{\Delta u}. \tag{52}$$

(ii). Using (28), apply

$$\begin{aligned}
\mathbf{p}(0,0)_\ominus &= \alpha \int_{z=0}^{\infty} \boldsymbol{\xi}_z e^{\check{\mathbf{Q}}_{\ominus\ominus} z} dz (-\mathbf{T}_{\ominus\ominus})^{-1} \\
&= \alpha \sum_{\ell=0}^{\infty} \int_{z=\ell\Delta u}^{(\ell+1)\Delta u} \boldsymbol{\xi}_z e^{\check{\mathbf{Q}}_{\ominus\ominus} z} dz (-\mathbf{T}_{\ominus\ominus})^{-1} \\
&\approx \alpha \sum_{\ell=0}^{L} \int_{z=\ell\Delta u}^{(\ell+1)\Delta u} \frac{\bar{\boldsymbol{\xi}}_\ell}{\Delta u} e^{\check{\mathbf{Q}}_{\ominus\ominus} z} dz (-\mathbf{T}_{\ominus\ominus})^{-1} \\
&\approx \alpha \sum_{\ell=0}^{L} \bar{\boldsymbol{\xi}}_\ell e^{\check{\mathbf{Q}}_{\ominus\ominus} \ell\Delta u} (-\mathbf{T}_{\ominus\ominus})^{-1}. \qquad (53)
\end{aligned}$$

Apply analogous approximation idea to calculating $\boldsymbol{\pi}(0, y)$, $y > 0$, and $\boldsymbol{\pi}(x, y)$, $x > 0$, $y > 0$, using (42), (45)-(47) and the inversion method of Abate and Whitt in [1].

Work on the numerical application of the above methodology is in progress.

5. CONCLUSION

We considered a tandem fluid queue model consisting of two queues, in which the first queue, $\{(\varphi(t), X(t)) : t \geq 0\}$, is a standard stochastic fluid model with a finite buffer and real rates r_i, and the second queue, $\{(\varphi(t), Y(t)) : t \geq 0\}$, is a stochastic fluid model with a finite buffer and rates $\widehat{c}_i > 0$ and $\check{c}_i < 0$, such that the rates of change of level depend on whether the first queue is empty or not. Specifically, we assumed that the rates of change of level in the second queue are negative ($dY(t)/dt = \check{c}_i$) when the first queue is empty, and positive ($dY(t)/dt = \widehat{c}_i$) otherwise.

We derived theoretical results for the stationary analysis of such tandem fluid queue, and summarized the key points of the methodology for the numerical evaluation of the stationary distribution of the process based on these results.

As future work we are also interested in the analysis of a dual tandem fluid queue model, with the difference that the rates of change of level in the second queue are positive ($dY(t)/dt = \widehat{c}_i$) when the first queue is empty, and negative ($dY(t)/dt = \check{c}_i$) otherwise. Work on the theoretical analysis of the dual model is in progress.

6. REFERENCES

[1] J. Abate and W. Whitt. Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal of Computing*, 7(1):36–43, 1995.

[2] S. Ahn and V. Ramaswami. Transient analysis of fluid models via elementary level-crossing arguments. *Stochastic Models*, 22(1):129–147, 2006.

[3] D. Anick, D. Mitra and M.M. Sondhi. Stochastic theory of data handling system with multiple sources. *Bell System Technical Journal*, 61:1871–1894, 1982.

[4] N.G. Bean and M.M. O'Reilly. A stochastic two-dimensional fluid model. *Stochastic Models*, 29(1):31–63, 2013.

[5] N.G. Bean and M.M. O'Reilly. The Stochastic Fluid-Fluid Model: A Stochastic Fluid Model driven by an uncountable-state process, which is a Stochastic Fluid Model itself. *Stochastic Processes and Their Applications*, 124(5):1741–1772, 2014.

[6] N.G. Bean, M.M. O'Reilly and J. Sargison. A stochastic fluid flow model of the operation and maintenance of power generation systems. *IEEE Transactions on Power Systems*, 25(3):1361–1374, 2010.

[7] N.G. Bean, M.M O'Reilly and P.G. Taylor. Hitting probabilities and hitting times for stochastic fluid flows. *Stochastic Processes and Their Applications*, 115(9):1530–1556, 2005.

[8] N.G. Bean, M.M O'Reilly and P.G. Taylor. Algorithms for return probabilities for stochastic fluid flows. *Stochastic Models*, 21(1):149–184, 2005.

[9] N.G. Bean, M.M O'Reilly and P.G. Taylor. Algorithms for the Laplace-Stieltjes transforms of first return times for stochastic fluid flows. *Methodology and Computing in Applied Probability*, 10(3):381–408, 2008.

[10] N.G. Bean, M.M. O'Reilly, P.G. Taylor, Hitting probabilities and hitting times for stochastic fluid flows. *Stochastic Processes and their Applications*, 115(9):1530–1556, 2005.

[11] A. Da Silva Soares, "Fluid Queues - Building upon the analogy with QBD Processes," Doctoral Dissertation, Universite Libre de Bruxelles, Belgium, 2005.

[12] A. Da Silva Soares and G. Latouche, Fluid queues with level dependent evolution, *European Journal of Operational Research*, 196(3):1041–1048, 2009.

[13] A. Da Silva Soares and G. Latouche, Matrix-analytic methods for fluid queues with finite buffers, *Performance Evaluation*, 63:295–314, 2006.

[14] D.P. Kroese and W.R.W. Scheinhardt. Joint Distributions for Interacting Fluid Queues. *Queueing Systems*, 37:99–139, 2001.

[15] G. Latouche and P.G. Taylor. A stochastic fluid model for an ad hoc mobile network. *Queueing Systems*, 63:109–129, 2009.

[16] B. Margolius and M.M. O'Reilly. The analysis of cyclic stochastic fluid flows with time-varying transition rates. *Queueing Systems*, 82(1-2):43–73, 2016.

[17] V. Ramaswami, Matrix analytic methods for stochastic fluid flows, *Proceedings of the 16th International Teletraffic Congress*, Edinburgh, 7-11 June 1999, pages 1019–1030, 1999.

[18] A. Samuelson, M.M. O'Reilly and N.G. Bean. Generalised reward generator for stochastic fluid models. Submitted to the 9th International Conference on Matrix-Analytic Methods in Stochastic Models, 2016.

Asymptotic error bounds for truncated buffer approximations of a 2-node tandem queue

Eleni Vatamidou[⋆]
evatamid@gmail.com

Ivo Adan[⋆,†]
i.j.b.f.adan@tue.nl

Maria Vlasiou[⋆,‡]
m.vlasiou@tue.nl

Bert Zwart[⋆,‡]
Bert.Zwart@cwi.nl

[⋆]Eurandom and Department of Mathematics & Computer Science, Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[†]Department of Mechanical Engineering, Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[‡]Centrum Wiskunde & Informatica (CWI), P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

ABSTRACT

We consider the queue lengths of a tandem queueing network. The number of customers in the system can be modelled as QBD with a doubly-infinite state-space. Due to the infinite phase-space, this system does not have a product-form solution. A natural approach to find a numerical solution with the aid of matrix analytic methods is by truncating the phase-space; however, this approach imposes approximation errors. The goal of this paper is to study these approximation errors mathematically, using large deviations and extreme value theory. We obtain a simple asymptotic error bound for the approximations that depends on the truncation level. We test the accuracy of our bound numerically.

Keywords

Matrix-analytic methods; tandem queues; batch arrivals; queue length approximations; asymptotic error bound; large deviations theory; renewal theory; extreme value analysis

1. INTRODUCTION

The algorithmic evaluation of performance measures in stochastic networks is a central topic in applied probability. Indeed, many processes of interest can be modelled as Markov chains on a product space of the form $\mathbb{N} \times P$; the main coordinate of the Markov chain, called the level, is integer-valued and the phase-space P carries supplementary information. This partitioning is one of the key underlying ideas connecting phase-type distributions with algorithms that are often summarised as *Matrix-Analytic Methods* (MAM).

MAM are widely studied in the literature (see for example [7, 8, 14, 16, 24, 26, 27, 28]) and can be effective when the phase-space P is a finite set. This restriction on P limits the applicability of MAM. For example, it prevents the usage of heavy-tailed distributions as models for service times and it prevents the analysis of queueing networks with infinite waiting rooms that do not have a product form solution. Though the mathematics behind MAM can be extended to this setting using connections with the general theory of Markov additive processes [2, 25], this does not seem to lead to concrete numerical algorithms.

A natural idea to overcome this issue is simply the truncation of the phase-space P so MAM become applicable. In the examples mentioned above, this entails the approximation of heavy-tailed distributions by phase-type distributions, truncating the waiting room of a station in a queueing network, or approximating output processes by Markovian arrival processes. These ideas have in fact been applied in many engineering-oriented studies, a small sample of references being [1, 10, 13, 17, 19, 20, 30].

Somewhat surprisingly, the impact of such approximations on the accuracy of the resulting numerical algorithms is not well investigated mathematically. Classical bounds on truncation errors in Markov chains, as in [32], do not offer much insight. They are not aimed at the type of structured Markov chains encountered in queueing networks, where, for example, there is no reason to truncate the level space. The goal of this paper is to analyse mathematically the impact of truncation by means of a rigorous analysis.

Motivated by this, we consider the queue lengths of the $\mathrm{M}^X/\mathrm{M}/1 \to \bullet/\mathrm{M}/1$ tandem queueing network, where customers arrive in batches in the first queue (abbreviated as Q_1). This tandem network is a useful example of a non-product form queueing network (for non-trivial batch sizes). The number of customers in the system can be modelled as a two-dimensional Markov chain, where the marginal distribution of the number of jobs in the downstream queue (Q_2) is the hardest to obtain. For this reason, this coordinate will be chosen to be the level. A numerical solution for this model can be found by using MAM only if the buffer size of either queue is finite. For this specific model, we shall derive error bounds, with a particular emphasis on the regime where the truncation level is large, so that the resulting error is (hopefully) small.

Within the MAM literature, there have been several related works. The model we consider in this paper can be modelled as a *Quasi-Birth and Death* process (QBD) with infinite phase-space. Formally, the invariant distribution of such processes can be written as $\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 \boldsymbol{R}^i$, with $\boldsymbol{R}$ an infinite matrix [33]. A natural question is whether truncating the phase-space to a size N leads to a matrix $\boldsymbol{R}_N$ with the property that $\boldsymbol{R}_N \to \boldsymbol{R}$. This is also related to the question how the phase-space should be truncated: the transition probabilities of the approximating Markov chain should be augmented in such a way that the transition matrix becomes stochastic. Background on this procedure can be found in [6]. In our paper, we consider the *Partial Batch Acceptance Strategy* (PBAS), which is called *last-column augmentation* in [6]. Remarkably, this procedure does not always imply that the invariant distribution of the approximating Markov chain converges to the original one, as illustrated by Example 4.1 of [6].

Even when the invariant distribution of the approximating Markov chain converges to the original invariant distribution, one would like to know more, such as the speed of convergence. Ideally, one would like to have analytical guidelines on choosing the truncation level in such a way that a pre-described accuracy level is met. We are not aware of any analytic result in this domain. The results that seem to come closest relate to the robustness of large deviations approximations, which are in turn related to the spectral radius $\nu(N)$ of the matrix $\boldsymbol{R}_N$. There are studies showing that $\nu(N)$ does not always converge to the spectral radius ν of $\boldsymbol{R}$ [22, 31] and that the way the model is truncated actually plays a role [23].

The question examined in the present paper is closest to [6], which is to analyse the accuracy of the invariant distribution after the truncation and analyse how the error decreases when the truncation level increases. Unlike the above-mentioned works, our asymptotic techniques are based on large-deviations theory and extreme value theory, as well as Markov renewal theory. We believe that such asymptotic techniques are promising and natural to consider in this domain and have the potential to provide useful insight in the quality of numerical algorithms. This has been observed and exploited in the simulation literature (especially rare event simulation), but less so in the literature on MAM.

Specifically, our approach is as follows. Using uniformisation, we model our tandem network as a discrete time Markov chain, of which the state $(0,0)$ will be taken as regeneration point. Let $T_{(0,0)}$ be the length of a cycle and let $M^{T_{(0,0)}}$ be the maximum number of customers in the first queue during a cycle. Our first step is to show that

$$\begin{aligned} 0 &\le \mathbb{P}\big(X_\infty \ge x, Y_\infty \ge y\big) - \mathbb{P}\big(X_\infty^{(N)} \ge x, Y_\infty^{(N)} \ge y\big) \\ &\le \mathbb{E}\big[T_{(0,0)} \bigm| M^{T_{(0,0)}} \ge N\big] \frac{\mathbb{P}\big(M^{T_{(0,0)}} \ge N\big)}{\mathbb{E}T_{(0,0)}}, \qquad (1) \end{aligned}$$

where X_∞ and Y_∞ denote the number of customers in the upstream and downstream queue in steady state, while $\mathbb{P}\big(X_\infty \ge x, Y_\infty \ge y\big)$, $\mathbb{P}\big(X_\infty^{(N)} \ge x, Y_\infty^{(N)} \ge y\big)$ are the steady-state probabilities of the original and the truncated system, respectively. The first inequality is derived using a so-called Markov reward approach. The second inequality is based on arguments from regenerative process theory and essentially exploits that the original and approximating process only differ in cycles where the first queue has at least N customers. These are rather standard arguments. The main work is to analyse the asymptotic behaviour of each of the three factors on the right hand side as $N \to \infty$.

The behaviour of $\mathbb{P}\big(M^{T_{(0,0)}} \ge N\big)$ can be reduced to studying the maximum queue length during a *small* cycle, corresponding to the busy period of the first queue in isolation. This reduction is possible using extreme value theory for regenerative processes, as surveyed in [3]. We show that we are allowed to do this by relying on ideas that date back to [18], which we adapt to the lattice case. Moreover, the term $\mathbb{E}T_{(0,0)}$ is treated in conjunction with $\mathbb{P}\big(M^{T_{(0,0)}} \ge N\big)$.

The behaviour of $\mathbb{E}\big[T_{(0,0)} \bigm| M^{T_{(0,0)}} \ge N\big]$ is more challenging to derive. In this paper, we give a heuristic treatment, using intuition from large deviations theory. For a formal proof we have to decompose $T_{(0,0)}$ into several (up to four) pieces, each of which we analyse using different methods. Key ingredients are optional stopping, the key Markov renewal theorem (for Markov additive processes with countable background state space) and various estimates of stopped (Markov) random walks; see e.g. [12]. Details of the proof, which is omitted for space considerations, can be found in the PhD thesis [34].

Our analysis results in a simple asymptotic estimate of the error of the form $KNe^{-\gamma N}$, where K and γ can be described explicitly in terms of the basic parameters of the model. The error is sharp, in the sense that our expression for γ in the leading term $e^{-\gamma N}$ is optimal, though we do not exclude that the linear term N may be removed using different arguments that are beyond the scope of this study. A numerical investigation shows that our bound may be overly conservative. Still our study seems the first to establish an asymptotic error estimate in this context.

The rest of the paper is organised as follows. In Section 2, we introduce the model under consideration and we present some additional preliminary results. In Section 3, we derive the error bounds. Furthermore, in Section 4, we derive a Cramér-Lundberg approximation for the probability $\mathbb{P}\big(M^{T_{(0,0)}} \ge N\big)$, which we treat together with the mean cycle length $\mathbb{E}T_{(0,0)}$. We explain intuitively in Section 5 its asymptotic behaviour. Furthermore, in Section 6, we perform numerical experiments to check the quality of the asymptotic error bound and we summarise our conclusions.

2. MODEL DESCRIPTION AND PRELIMINARIES

We consider an M^X/M/1→ •/M/1 tandem queueing network. Customers arrive in batches according to a Poisson stream with rate λ and join Q_1. A customer that finishes service in Q_1 moves to Q_2. The service times for each queue are exponential with rates μ_1 and μ_2, respectively. The customer leaves the system after finishing his service in Q_2. We describe the system by a two-dimensional Markov chain $(X_n, Y_n) \in \mathbb{N}^2$, where X_n and Y_n are the queue lengths at the nth jump epoch of Q_1 and Q_2, respectively, including customers in service in either queue. For this system, we are interested in evaluating the distribution of its weak limit (X_∞, Y_∞).

We denote by B a generic r.v. of the batch sizes and we assume its mean $\mathbb{E}B = \sum_{i=1}^\infty ib_i < \infty$, where $b_i = \mathbb{P}(B = i)$, $i = 1, 2, \ldots$ Furthermore, for stability reasons, we assume that $\lambda\mathbb{E}B/\mu_i < 1$, $i = 1, 2$. In addition, w.l.o.g., we consider a uniformised version of this chain: $\lambda + \mu_1 + \mu_2 = 1$ and we denote the netput between the $(n-1)$st and the nth jump epoch in the 1st and 2nd queue as Z_n and W_n, respectively. Namely,

$$Z_n = \begin{cases} 0, & \text{w.p.}\mu_2, \\ -1, & \text{w.p. } \mu_1, \\ m, & \text{w.p. } \lambda b_m, m = 1, 2, \ldots, \end{cases} \tag{2}$$

and

$$W_n = \begin{cases} -1, & \text{if } Z_n = 0, \\ 1, & \text{if } Z_n = -1 \text{ and } X_{n-1} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Recall that due to uniformisation, $\lambda, \mu_1, \mu_2 < 1$ and the rates λ, μ_1, μ_2 can be seen as probabilities.

The number of customers X_n in Q_1 satisfies the following Lindley recursion

$$X_0 = 0, \quad X_{n+1} = \left(X_n + Z_{n+1}\right)^+, \quad n = 0, 1, \ldots \tag{4}$$

Thus, $\{X_n\}_{n=0,1,\ldots}$ evolves as a reflected at 0 discrete version of a random walk with increments $Z_1, Z_2, \ldots$ Similarly, the number of customers Y_n in Q_2 satisfies

$$Y_0 = 0, \quad Y_{n+1} = \left(Y_n + W_{n+1}\right)^+, \quad n = 0, 1, \ldots \tag{5}$$

The initial state of the system is $(X_0, Y_0) = (0, 0)$ and we define the first return time to the origin as $T_{(0,0)} = \inf\{n \geq 1 : X_n = Y_n = 0 \mid X_0 = Y_0 = 0\}$, which is also called *cycle length*. Therefore, since we have a two-dimensional positive recurrent irreducible Markov chain, it is known that

$$\begin{aligned} &\mathbb{P}\left(X_\infty \geq x, Y_\infty \geq y\right) \\ &= \frac{1}{\mathbb{E}T_{(0,0)}} \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}} \mathbb{1}\left(X_n \geq x, Y_n \geq y\right)\right]. \end{aligned}$$

From Eqs. (2) and (3), we can easily verify that the two-dimensional Markov chain (X_n, Y_n) is a QBD with an infinite phase-space $P = \{0, 1, \ldots\}$, which does not admit a product form solution according to Theorem 15.1.1 of [24] unless $B = 1$.

State space truncation

As we mentioned in Section 1, the number of customers in Q_1 and Q_2 correspond to the phase and level, respectively, of the QBD introduced earlier. Thus, we truncate the buffer size of Q_1 at level N, which we call *truncation level*. More precisely, the arriving customers are admitted in the system by applying the PBAS; i.e. if the batch size is larger than the number of available free positions in the buffer (which has capacity $N - 1$), then we accept only so many customers until there are in total N customers waiting in front of Q_1 and we dismiss the remaining ones.

Moreover, we denote by $\left(X_n^{(N)}, Y_n^{(N)}\right) \in (\mathbb{N}_N \times \mathbb{N})$ the approximate Markov chain associated with the truncation level N and by $\left(Z_n^{(N)}, W_n^{(N)}\right)$ the corresponding netput process, where $\mathbb{N}_n = \{0, 1, \ldots, n\} \subset \mathbb{N}$. Observe that definitions (3)–(5) are still valid (but with the notation adapted to the truncated system) for the processes $X_n^{(N)}$, $Y_n^{(N)}$, and $W_n^{(N)}$, respectively. However, the definition of $Z_{n+1}^{(N)}$ takes two alternative forms depending on the value of $X_n^{(N)}$. More precisely, if $X_n^{(N)} = N$, then

$$Z_{n+1}^{(N)} = \begin{cases} 0, & \text{w.p. } \lambda + \mu_2, \\ -1, & \text{w.p. } \mu_1, \end{cases} \tag{6}$$

while in case $X_n^{(N)} = N - m$, $m \in \{1, \ldots, N\}$

$$Z_{n+1}^{(N)} = \begin{cases} 0, & \text{w.p. } \mu_2, \\ -1, & \text{w.p. } \mu_1, \\ k, & \text{w.p. } \lambda b_k \text{ for } k < m, \\ m, & \text{with probability } \lambda \sum_{i=m}^{\infty} b_i. \end{cases} \tag{7}$$

We also define $T_{(0,0)}^{(N)} = \inf\{n \geq 1 : X_n^{(N)} = Y_n^{(N)} = 0 \mid X_0^{(N)} = Y_0^{(N)} = 0\}$ as the first return time to the origin for the truncated system. Finally, we denote by $\boldsymbol{m} = (m_1, m_2)$ the two-dimensional states of the Markov chain (X_n, Y_n), where m_1 and m_2 are non-negative integers. If $\boldsymbol{P}$ is the transition probability matrix of the Markov chain and $\boldsymbol{P}^{(N)}$ its truncation, then $\forall \boldsymbol{m}, \boldsymbol{n}$ with $m_1, n_1 \in \mathbb{N}_{N-1}$ we have that

$$\boldsymbol{P}^{(N)}\left(\boldsymbol{m}, \boldsymbol{n}\right) = \boldsymbol{P}\left(\boldsymbol{m}, \boldsymbol{n}\right). \tag{8}$$

In other words, the entries in the two matrices $\boldsymbol{P}^{(N)}$ and $\boldsymbol{P}$ coincide as long as both two-dimensional Markov chains (original and truncated) live within the boundaries. This property is very useful in Section 3, where our error bounds for the approximation of the joint queue length distribution stem from this truncation.

Note that to analyse the terms $\mathbb{P}\left(M^{T_{(0,0)}} \geq N\right)$ and $\mathbb{E}\left[T_{(0,0)} \mid M^{T_{(0,0)}} \geq N\right]$ (see Sections 4–5), an exponential change of measure is first required. Thus, we conclude this section by providing some results with respect to such an exponential change of measure.

Exponential change of measure

We define the *cumulant generating function* (c.g.f.) of the r.v.'s $Z_1, Z_2, \ldots$ as

$$\begin{aligned} \kappa(\alpha) :&= \ln \mathbb{E}e^{\alpha Z_1} = \ln\left(\mu_2 + \mu_1 e^{-\alpha} + \lambda \mathbb{E}e^{\alpha B}\right) \\ &= \ln\left(\mu_2 + \mu_1 e^{-\alpha} + \lambda M_B(\alpha)\right), \end{aligned} \tag{9}$$

where $M_B(\alpha)$ is the *moment generating function* (m.g.f.) of the batch sizes. We assume that there exists a solution $\gamma > 0$ to the *Lundberg equation* $\kappa(\gamma) = 0$ such that $\kappa'(\gamma) < \infty$. The parameter γ is called the *adjustment coefficient* and conditions for its existence can be found in [5].

If F is the distribution of the $Z \overset{\mathfrak{D}}{=} Z_n$, we define $\breve{F}$ to be the probability distribution with density $e^{\gamma x}$ w.r.t. F, i.e. $\breve{F}(dx) = e^{\gamma x} F(dx)$ (obvious notations like $\breve{\kappa}(\alpha)$, $\breve{\mathbb{P}}$, $\breve{\mathbb{E}}$, etc, are used for quantities under the exponential change of measure). It can easily be verified that, under this exponential change of measure, the arrival rate of the batches is equal to $\breve{\lambda} = \lambda + (1 - e^{-\gamma})\mu_1$, the batch size distribution is equal to

$$\breve{\mathbb{P}}(B = n) = \frac{e^{\gamma n}}{M_B(\gamma)} \mathbb{P}(B = n), \quad n = 1, 2, \ldots, \tag{10}$$

and the customers are served with rates $\breve{\mu}_1 = e^{-\gamma}\mu_1$ and $\breve{\mu}_2 = \mu_2$ in each server, respectively. In addition, it holds that $\breve{\mathbb{E}}Z = \breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1 > 0$.

We continue in the next section by providing the main results of the paper.

3. MAIN RESULTS

In this section, we present error bounds for the probability $\mathbb{P}\left(X_\infty \geq x, Y_\infty \geq y\right)$. In particular, we prove the two inequalities in Eq. (1). The left hand side of Eq. (1) shows that $\mathbb{P}\left(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\right)$ always underestimates the exact probability. We formulate this result in the following proposition.

Proposition 1. *If N is the truncation level of the buffer size of Q_1, then $\forall (x,y) \in \mathbb{N}^2$ it holds:*

$$\mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big) \leq \mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big). \quad (11)$$

Proof. The proof is based on Markov reward techniques and is omitted for space considerations, for details see Section 5.3 of [34]. □

To prove the right hand side of Eq. (1), we split the steady state probability as follows

$$\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) = \frac{1}{\mathbb{E}T_{(0,0)}}\big(\mathbb{I} + \mathbb{II}\big), \quad (12)$$

$$\mathbb{I} = \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}} \mathbb{1}\left(X_n \geq x, Y_n \geq y\right) \cdot \mathbb{1}\left(\max_{1 \leq l \leq T_{(0,0)}} X_l < N\right)\right],$$

$$\mathbb{II} = \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}} \mathbb{1}\left(X_n \geq x, Y_n \geq y\right) \cdot \mathbb{1}\left(\max_{1 \leq l \leq T_{(0,0)}} X_l \geq N\right)\right].$$

Let $M^{T_{(0,0)}} = \max_{1 \leq n \leq T_{(0,0)}} X_n$ be the maximum queue length of the first queue before the first return time to the state $(0,0)$. We show in Proposition 2 that term $\mathbb{I}$ is related to $\mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$, while term $\mathbb{II}$ evolves in some sense like $M^{T_{(0,0)}}$. With the aid of Eq. (12), we derive an upper bound for $\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big)$.

Proposition 2. *An upper bound for the probability $\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big)$ is as follows:*

$$\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) \leq \mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$$
$$+\mathbb{E}\big[T_{(0,0)} \bigm| M^{T_{(0,0)}} \geq N\big] \frac{\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)}{\mathbb{E}T_{(0,0)}}.$$

Proof. We discuss the terms $\mathbb{I}$ and $\mathbb{II}$ separately. *Term $\mathbb{I}$:* If we set $\zeta = \inf\{n \geq 0 : X_n \geq N\}$ and $\zeta^{(N)} = \inf\{n \geq 0 : X_n^{(N)} \geq N\}$, then from Eq. (8) it holds that $(X_n : n < \zeta) \overset{\mathfrak{D}}{=} (X_n^{(N)} : n < \zeta^{(N)})$. Observe that $T_{(0,0)} = T_{(0,0)}^{(N)}$ when $\mathbb{1}\left(\max_{1 \leq l \leq T_{(0,0)}} X_l < N\right) = 1$. Thus, since term $\mathbb{I}$ contains the sample paths of the truncated system, we obtain:

$$\begin{aligned}\mathbb{I} =& \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}^{(N)}} \mathbb{1}\left(X_n^{(N)} \geq x, Y_n^{(N)} \geq y\right)\right.\\ &\left.\times\, \mathbb{1}\left(\max_{1 \leq l \leq T_{(0,0)}^{(N)}} X_l^{(N)} < N\right)\right]\\ \leq& \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}^{(N)}} \mathbb{1}\left(X_n^{(N)} \geq x, Y_n^{(N)} \geq y\right)\right]\\ =& \mathbb{E}T_{(0,0)}^{(N)} \mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big). \end{aligned} \quad (13)$$

Term $\mathbb{II}$: For the second term, we have

$$\begin{aligned}\mathbb{II} &= \mathbb{E}\left[\sum_{n=1}^{T_{(0,0)}} \mathbb{1}\left(X_n \geq x, Y_n \geq y\right) \cdot \mathbb{1}\left(\max_{1 \leq l \leq T_{(0,0)}} X_l \geq N\right)\right]\\ &\leq \mathbb{E}\left[T_{(0,0)} \cdot \mathbb{1}\left(\max_{1 \leq l \leq T_{(0,0)}} X_l \geq N\right)\right]\\ &= \mathbb{E}\left[T_{(0,0)} \cdot \mathbb{1}\left(M^{T_{(0,0)}} \geq N\right)\right]\\ &= \mathbb{E}\big[T_{(0,0)} \bigm| M^{T_{(0,0)}} \geq N\big]\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big). \end{aligned} \quad (14)$$

Combining Eqs. (12), (13), and (14), we obtain

$$\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) \leq \frac{\mathbb{E}T_{(0,0)}^{(N)}}{\mathbb{E}T_{(0,0)}} \mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$$
$$+\frac{\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)}{\mathbb{E}T_{(0,0)}} \mathbb{E}\big[T_{(0,0)} \bigm| M^{T_{(0,0)}} \geq N\big]. \quad (15)$$

Finally, we need to show that $\mathbb{E}T_{(0,0)} \geq \mathbb{E}T_{(0,0)}^{(N)}$. Observe that $\mathbb{E}T_{(0,0)}$ and $\mathbb{E}T_{(0,0)}^{(N)}$ are by definition the expected first return times to the state $(0,0)$ in the original and the truncated system, respectively. By the strong law of large numbers for ergodic Markov chains [21], $\mathbb{E}T_{(0,0)}^{(N)} = 1/\mathbb{P}\big(X_\infty^{(N)} = 0, Y_\infty^{(N)} = 0\big)$ and $\mathbb{E}T_{(0,0)} = 1/\mathbb{P}\big(X_\infty = 0, Y_\infty = 0\big)$. Therefore, it is sufficient to show that the inequality $\mathbb{P}\big(X_\infty^{(N)} = 0, Y_\infty^{(N)} = 0\big) \geq \mathbb{P}\big(X_\infty = 0, Y_\infty = 0\big)$ holds. This inequality follows from a cost structure approach; for details see Section 5.5 of [34]. □

Observe that the term $\mathbb{E}\big[T_{(0,0)} \bigm| M^{T_{(0,0)}} \geq N\big] \times \mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)/\mathbb{E}T_{(0,0)}$ is involved in the upper bound of the steady state probability, according to Proposition 2. All factors involved in this term are hard to evaluate exactly. Instead, we examine the behaviour of these factors as $N \to \infty$.

With the aid of the exponential change of measure presented in the previous section, in Section 4, we provide asymptotic results for $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$, which is treated in conjunction with the factor $\mathbb{E}T_{(0,0)}$. Asymptotic results for the conditional expectation $\mathbb{E}\big[T_{(0,0)} \bigm| M^{T_{(0,0)}} \geq N\ \big]$ are derived in Section 5. The expression for the asymptotic upper bound is then formulated in Theorem 1. With $f(N) \lesssim g(N)$ we denote $\limsup_{N \to \infty} f(N)/g(N) \leq 1$.

Theorem 1. *As $N \to \infty$,*

$$\mathbb{P}\big(X_\infty \geq x, Y_\infty \geq y\big) - \mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$$
$$\lesssim KNe^{-\gamma N},$$

where

$$K = \left(\frac{1}{\mu_2 - \lambda\mathbb{E}B} \cdot \Big(\frac{(\breve{\mu}_1 - \mu_2)^+}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1} + \frac{(\mu_1 - \mu_2)^+}{\mu_1 - \lambda\mathbb{E}B}\Big)\right.$$
$$\left.+\frac{1}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1} + \frac{1}{\mu_1 - \lambda\mathbb{E}B}\right) \times NC_1e^{\gamma}\left(1 - \frac{\lambda\mathbb{E}B}{\mu_1}\right),$$

and C_1 is a constant calculated from Proposition 3.

We devote Sections 4–5 to the proof of Theorem 1.

4. ASYMPTOTIC APPROXIMATION FOR THE MAXIMUM

In this section, we derive an asymptotic approximation for $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ with the aid of extreme value theory. Observe that the number of customers in the first queue $\{X_n\}_{n=0,1,\ldots}$ forms a one-dimensional Markov chain on its own. Therefore, we denote as $T_0 = \inf\{n \geq 1 : X_n = 0 \mid X_0 = 0\}$ the return time to the origin of the first queue only and we define $M^{T_0} = \max_{1 \leq n \leq T_0} X_n$. We show that the probability $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ exhibits a similar tail behaviour with the probability $\mathbb{P}\big(M^{T_0} \geq N\big)$. Thus, we first discuss the behaviour of $\mathbb{P}\big(M^{T_0} \geq N\big)$ as $N \to \infty$.

We define $\tau_1 = \inf\{n : X_n \geq N\}$. Observe that $\mathbb{P}\big(M^{T_0} \geq N\big) = \mathbb{P}(\tau_1 < T_0)$. Moreover, the Lindley process X_n has the same transition mechanism as the random walk $U_n = Z_1 + \cdots + Z_n$, with $U_0 = 0$, until T_0, because X_n does not hit zero before T_0. Thus, it also holds that $\{\tau_1 < T_0\} = \{\tau(N-1) < \tau_-\}$, and consequently $\mathbb{P}\big(M^{T_0} \geq N\big) = \mathbb{P}(\tau(N-1) < \tau_-)$, where $\tau(N) = \inf\{n \geq 1 : U_n > N\}$ is the time of *first passage* to level $N \geq 0$ and $\tau_- = \inf\{n \geq 1 : U_n \leq 0\}$ is the first (weak) *descending ladder epoch*. We also denote the first (strict) *ascending ladder epoch* as $\tau_+ = \inf\{n \geq 1 : U_n > 0\}$. If $B(N) = U_{\tau(N)} - N$ is the *overshoot* of N, then a variant of the *Cramér-Lundberg approximation* is already known for the probability $\mathbb{P}\big(M^{T_0} \geq N\big)$ by Corollary XIII.5.9 in [4]. Therefore, we provide the following lemma without proof.

Lemma 1. *If $B(N)$ converges in $\breve{\mathbb{P}}$ as $N \to \infty$, say to $B(\infty)$, then*

$$e^{\gamma(N-1)}\mathbb{P}\big(M^{T_0} \geq N\big) = \breve{\mathbb{E}}e^{-\gamma B(N-1)}\mathbb{1}\left(\tau_1 < T_0\right) \to C_1,$$

where $C_1 = \breve{\mathbb{P}}(\tau_- = \infty)C_0$ and $C_0 = \breve{\mathbb{E}}e^{-\gamma B(\infty)}$.

We continue by showing that the tail behaviour of $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ is similar to the tail behaviour of $\mathbb{P}\big(M^{T_0} \geq N\big)$. For this purpose, note that both $T_{(0,0)}$ and T_0 are regeneration cycles for the Markov chain X_n. Thus, if we denote $M_i^{T_0} \overset{\mathfrak{D}}{=} M^{T_0}$ as the maximum of X_n in the ith cycle T_0, where M^{T_0} is the generic cycle maximum, and similarly $M_i^{T_{(0,0)}} \overset{\mathfrak{D}}{=} M^{T_{(0,0)}}$ as the maximum of X_n in the ith cycle $T_{(0,0)}$, we have that [3, 18, 29]

$$\max_{i=1,\ldots,\frac{n}{\mathbb{E}T_{(0,0)}}} M_i^{T_{(0,0)}} \approx \max_{i=1,\ldots,n} X_i \approx \max_{i=1,\ldots,\frac{n}{\mathbb{E}T_0}} M_i^{T_0}. \tag{16}$$

We now make this precise. From Lemma 1, we know the tail behaviour of M^{T_0}. Therefore, we can derive asymptotics for the maximum $\max_{i=1,\ldots,n} X_i$. As such, Eq. (16) indicates that in order to study the asymptotic behaviour of $M^{T_{(0,0)}}$, we first need to study the asymptotics of $\max_{i=1,\ldots,n} X_i$, as $n \to \infty$.

Classically, extreme value theory focuses on finding constants a_n, b_n, such that

$$\frac{\max_{i=1,\ldots,n} X_i - a_n}{b_n} \xrightarrow{\mathfrak{D}} H, \tag{17}$$

where H is some non-degenerate r.v. and $\xrightarrow{\mathfrak{D}}$ denotes convergence in distribution. This is equivalent to showing that the probability $\mathbb{P}\big(\max_{i=1,\ldots,n} X_i \leq a_n x + b_n\big)$ has a limit, for any x. In our case, we prove that given the tail behaviour of M^{T_0} from Lemma 1, there exist constants a_n, b_n, such that (17) holds with H following the *Gumbel* function $\Lambda(x) = e^{-e^{-x}}$, $x \in \mathbb{R}$ [15].

The asymptotic behaviour of $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$ is given in the following theorem. To establish this asymptotic result, we use Eq. (16) to first derive the asymptotics of $\max_{i=1,\ldots,n} X_i$, as $n \to \infty$, and later connect these asymptotics with $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)$.

Theorem 2. *It holds that*

$$\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big) \sim \frac{\mathbb{E}T_{(0,0)}}{\mathbb{E}T_0} C_1 e^{-\gamma(N-1)}, \quad N \to \infty,$$

where C_1 is defined in Lemma 1.

Proof. The proof is based on the above-mentioned approach and is omitted for space limitations; see Section 5.5 of [34] for details. □

Observe that only the constants C_0 and C_1 are missing in order to find a closed-form asymptotic relation for the fraction $\mathbb{P}\big(M^{T_{(0,0)}} \geq N\big)/\mathbb{E}T_{(0,0)}$ that appears in Eq. (1). We can find explicit expressions for these constants by using properties of lattice random walks. Thus, we conclude this section by providing explicit expressions for C_0 and C_1. We also calculate $\mathbb{E}T_0$.

Observe that both C_0 and C_1 require the evaluation of the limiting distribution of the overshoot $B(\infty)$, which can be found through the *ladder height distribution* with respect to the probability measure $\breve{\mathbb{P}}$.

Let now $\breve{H}_+$ be the distribution function of the ascending ladder height with respect to $\breve{\mathbb{P}}$ and $\breve{l}_+$ be its corresponding mean. In addition, we denote by $\breve{H}_-$ the (weak) descending ladder height distribution with respect to $\breve{\mathbb{P}}$. We have the following result.

Lemma 2. *For a discrete-time lattice random walk, $B(\infty)$ exists with respect to $\breve{\mathbb{P}}$. In this case, C_0 is given in terms of the ladder height distributions by*

$$C_0 = \breve{\mathbb{E}}e^{-\gamma B(\infty)} = \frac{\big(1 - \|H_+\|\big)\big(1 - \|\breve{H}_-\|\big)}{(e^{\gamma} - 1)\kappa'(\gamma)},$$

where $\|H_+\| = \mathbb{P}(\tau_+ < \infty)$ and $\|\breve{H}_-\| = \breve{\mathbb{P}}(\tau_- < \infty)$.

Proof. To prove this lemma, we need the limiting distribution of the overshoot, which can be obtained by adapting the renewal theorem to the lattice case, and we use Wald's equation; see Section 5.5 of [34] for details. □

Proposition 3. *For a downward skip-free (or left-continuous) random walk, the constant C_1 in Lemma 1 is equal to*

$$C_1 = -\frac{\mathbb{E}Z}{\breve{\mathbb{E}}Z}(1 - e^{-\gamma})e^{-\gamma}\mu_1 = -\frac{\kappa'(0)}{\kappa'(\gamma)}(1 - e^{-\gamma})e^{-\gamma}\mu_1.$$

Proof. From Lemma 2, it is evident that we need to find exact values for $1 - \|H_+\|$ and $1 - \|\breve{H}_-\|$. Observe that U_n is downward skip-free random walk.

We start with the evaluation of $1 - \|H_+\|$. We set $f_n = \mathbb{P}(Z = n)$. Under the probability measure $\mathbb{P}$, it holds that $\mathbb{E}Z = \kappa'(0) < 0$. Therefore, according to Corollary VIII.5.6 [4], $\|H_+\| = 1 + \mathbb{E}Z/f_{-1}$, where from Eq. (3) we know that $f_{-1} = \mathbb{P}(Z = -1) = \mu_1$.

By the definition of the descending ladder height distribution, we have that

$$1 - \|\breve{H}_-\| = \breve{\mathbb{P}}(\tau_- = \infty) = \breve{\mathbb{P}}(U_n \geq 1 \text{ for all } n \geq 1).$$

We set now $\breve{f}_n = \breve{\mathbb{P}}(Z = n)$ and $T_1 = \inf\{n : U_n = -1\}$. Since U_n is a downward skip-free random walk with an upward drift under the probability measure $\breve{\mathbb{P}}$, it holds from Proposition 11 in [9] that

$$1 - \|\breve{H}_-\| = \breve{f}_{-1} \cdot \frac{1 - \breve{\mathbb{P}}(T_1 < \infty)}{\breve{\mathbb{P}}(T_1 < \infty)}.$$

Thus, it is left to find the probability $\breve{\mathbb{P}}(T_1 < \infty)$, which according to Lemma 2 in [9] is equal to the unique value $s \in (0, 1)$ that satisfies the equation $\breve{\mathbb{E}}s^Z = 1$. Using $\kappa(\alpha) = 0$, we get from Proposition XIII.1.1 in [4] that $\breve{\mathbb{E}}e^{\alpha Z_1} = e^{\kappa(\alpha+\gamma)}$. Therefore, $\breve{\mathbb{E}}e^{-\gamma Z} = e^{\kappa(0)} = 1$, and consequently $s = e^{-\gamma} \in (0, 1)$ is the unique solution to the equation $\breve{\mathbb{E}}s^Z = 1$. As a result, $\breve{\mathbb{P}}(T_1 < \infty) = e^{-\gamma}$. We also find $\breve{f}_{-1} = \breve{\mathbb{P}}(Z = -1) = e^{-\gamma}\mu_1$. Combining all the above and Lemma 1, the result is immediate. □

We turn now our attention to the evaluation of $\mathbb{E}T_0$. Observe that $\mathbb{E}T_0 = 1/\mathbb{P}\big(X_\infty = 0\big)$. By applying Little's law for a busy server we find that $\rho_1 = \lambda\mathbb{E}B/\mu_1$, with $\lambda\mathbb{E}B$ being the average number of customers entering the system per time unit. Consequently, $\mathbb{P}\big(X_\infty = 0\big) = 1 - \rho_1 = 1 - \lambda\mathbb{E}B/\mu_1$. Thus, we have proven:

Lemma 3. $\mathbb{E}T_0 = \big(1 - \lambda\mathbb{E}B/\mu_1\big)^{-1}$.

5. THE CONDITIONAL MEAN RETURN TIME

Our last goal is to study the asymptotic behaviour of the conditional expectation $\mathbb{E}\big[T_{(0,0)} \bigm| M^{T_{(0,0)}} \geq N\big]$. More precisely, we study the limit $\lim\limits_{N\to\infty} \frac{1}{N}\mathbb{E}\big[T_{(0,0)} \bigm| M^{T_{(0,0)}} \geq N\big]$. We take a heuristic approach, using intuition from large deviations theory. A formal proof can be found in Section 5.6 of [34]. Define

τ_1: the time at which Q_1 reaches or exceeds level N. Recall that it was defined in Section 4 as $\tau_1 = \inf\{n : X_n \geq N\}$.

τ_2: the return time to 0 in Q_1 after τ_1. Formally, $\tau_2 = \inf\{n > \tau_1 : X_n = 0\}$.

τ_3: the first time Q_2 empties after τ_2. Formally, $\tau_3 = \inf\{n > \tau_2 : Y_n = 0\}$. The time τ_3 can either coincide with or happen before $T_{(0,0)}$.

We describe heuristically how both queues behave, given that the number of customers in Q_1 has reached a very high level before the first return time $T_{(0,0)}$ to the empty state $(0,0)$. Our description is based on intuition from large deviations theory and fluid limits. We write $a \approx b$ to denote that a is approximately equal to b, without explicitly determining the degree of accuracy. Denote by $\#Q_1$ and $\#Q_2$ the number of customers in Q_1 and Q_2.

Observe that the behaviour of Q_1 is not affected by what happens in Q_2. On the other hand, we recognise three different cases for the behaviour of Q_2 that arise from the relation between the rates μ_1, μ_2, and $\breve{\mu}_1$. We summarise all cases in Figure 1. We start by discussing the behaviour of Q_1.

To describe the behaviour of Q_1 until time $T_{(0,0)}$, given that $\#Q_1$ reached or exceeded level N, we apply arguments from large deviations theory. According to Section 2, this event happens by a change of measure, from $\mathbb{P}$ to $\breve{\mathbb{P}}$. Since $N \to \infty$, the time its takes Q_1 from τ_1 to reach its maximum value (something above N) before $T_{(0,0)}$ is negligible (compared to τ_1). Moreover, until τ_1, the departure rate of the customers is asymptotically equal to $\breve{\mu}_1$ because the system is overloaded ($\breve{\lambda}\breve{\mathbb{E}}B > \breve{\mu}_1$). On the other hand, after τ_1, all the rates are back to normal. As we have already mentioned, τ_2 is the point at which the Q_1 reaches 0 after reaching its maximum value within cycle $T_{(0,0)}$. Since during the time interval $[\tau_1, \tau_2]$ Q_1 is always full, the departure rate of customers equals μ_1.

Next, we describe the behaviour of Q_2 before $T_{(0,0)}$.

Case 1: $\mu_1 < \mu_2$

It always holds that $\breve{\mu}_1 < \mu_1$ (see Section 2 for the definition of $\breve{\mu}_1$). Therefore, in this case, Q_2 behaves asymptotically as a stable M/M/1 queue in all time intervals (but with different arrival rates of customers). Thus, the number of customers in Q_2 is bounded by the number of customers in a stable M/M/1 queue until $T_{(0,0)}$. Consequently, the time interval $[\tau_2, T_{(0,0)}]$ is negligible compared to $[0, \tau_2]$ and we expect that $\mathbb{E}\big[T_{(0,0)} \bigm| M^{T_{(0,0)}} \geq N\big] \approx \mathbb{E}\big[\tau_2 \bigm| M^{T_{(0,0)}} \geq N\big]$, where from Euclidean geometry we can easily verify that (see Figure 1)

$$\tau_1 \approx \frac{N}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1}, \qquad \tau_2 - \tau_1 \approx \frac{N}{\mu_1 - \lambda\mathbb{E}B}. \tag{18}$$

Case 2: $\breve{\mu}_1 < \mu_2 < \mu_1$

Since $\breve{\mu}_1 < \mu_2$, Q_2 behaves asymptotically as a stable M/M/1 queue with arrival rate $\breve{\mu}_1$ and service rate μ_2 until time τ_1. This means that the number of customers in Q_2 at time τ_1 is bounded by the number of customers in the latter M/M/1 queue. From τ_1 onwards, the arrival rate of customers in Q_2 is equal to μ_1, which is greater than the service rate μ_2. Therefore, the number of customers in Q_2 grows linearly with rate $\mu_1 - \mu_2$ up until τ_2. After τ_2, the output rate from Q_1 is equal to $\lambda\mathbb{E}B$ and the customers in Q_2 reduce linearly with rate $\lambda\mathbb{E}B - \mu_2$ until the queue empties at time τ_3. We calculate (see Figure 1)

$$\begin{aligned} h_2 &\approx (\mu_1 - \mu_2)\frac{N}{\mu_1 - \lambda\mathbb{E}B}, \\ \tau_3 - \tau_2 &\approx \frac{h_2}{\mu_2 - \lambda\mathbb{E}B} = \frac{\mu_1 - \mu_2}{\mu_2 - \lambda\mathbb{E}B} \cdot \frac{N}{\mu_1 - \lambda\mathbb{E}B}. \end{aligned} \tag{19}$$

Obviously, in this case $\mathbb{E}\big[T_{(0,0)} \bigm| M^{T_{(0,0)}} \geq N\big] \approx \mathbb{E}\big[\tau_3 \bigm| M^{T_{(0,0)}} \geq N\big]$, because the interval $[\tau_3, T_{(0,0)}]$ is negligible compared to $[0, \tau_3]$.

Case 3: $\mu_2 < \breve{\mu}_1 < \mu_1$

Since $\breve{\mu}_1 > \mu_2$, the number of customers in Q_2 grows linearly with rate $\breve{\mu}_1 - \mu_2$ up until time τ_1. For the remaining time intervals, Q_2 behaves in a similar manner as in Case 2. Therefore, $\mathbb{E}\big[T_{(0,0)} \bigm| M^{T_{(0,0)}} \geq N\big] \approx \mathbb{E}\big[\tau_3 \bigm| M^{T_{(0,0)}} \geq N\big]$, where (see Figure 1)

$$\begin{aligned} h_1 &\approx (\breve{\mu}_1 - \mu_2)\frac{N}{\breve{\lambda}\breve{\mathbb{E}}B - \breve{\mu}_1}, \\ h_2 &\approx h_1 + (\mu_1 - \mu_2)\frac{N}{\mu_1 - \lambda\mathbb{E}B}, \\ \tau_3 - \tau_2 &\approx \frac{h_2}{\mu_2 - \lambda\mathbb{E}B}. \end{aligned} \tag{20}$$

To prove rigorously the behaviour of Q_2 in $[0, \tau_2]$, we use renewal theory arguments and the relation between $\mathbb{P}$ and $\breve{\mathbb{P}}$. For the time interval $[\tau_2, \tau_3]$, the idea is to see our two-dimensional Markov chain as a *Markov Additive Process* (MAP) [11]. Finally, for $[\tau_3, T_{(0,0)}]$, we use that the hitting time of the origin is finite since the latter is a recurrent state for our ergodic Markov chain.

6. NUMERICAL EXPERIMENTS

We perform now numerical experiments to check the quality of our asymptotic upper error bound ($a.u.e.b.$) in Theorem 1. As an example, we use geometric batch sizes, where we calculate the exact queue lengths through simulation.

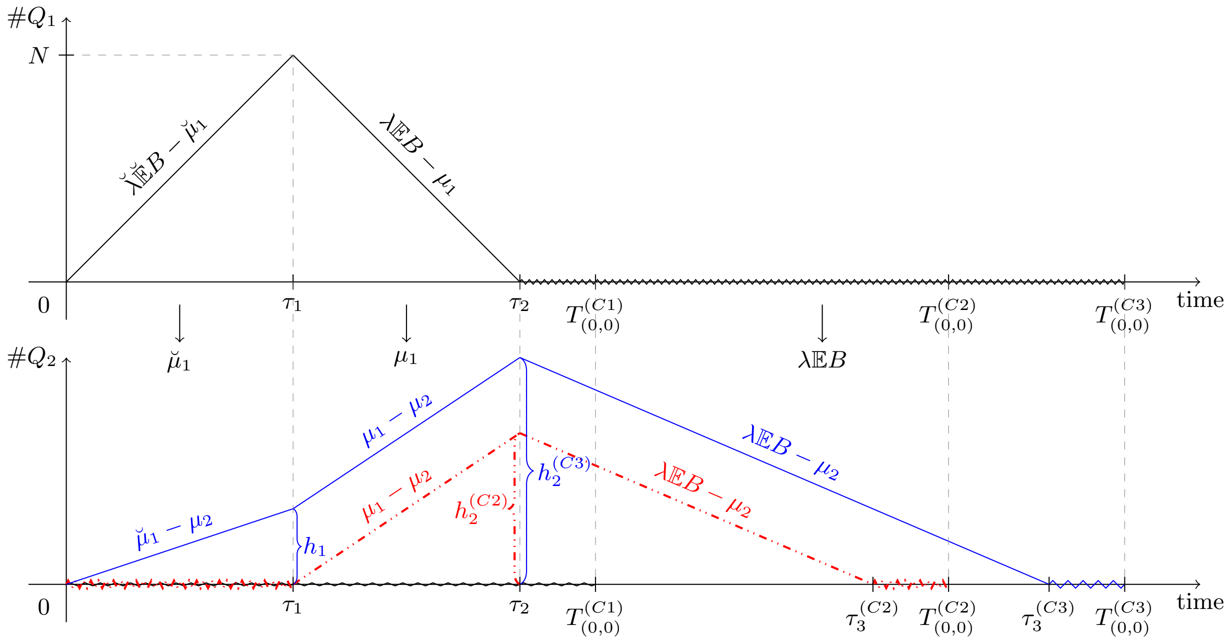


Figure 1: The asymptotic behaviours of Q_1 and Q_2, given that $\#Q_1$ before $T_{(0,0)}$ exceeded the truncation level N, for all 3 different cases; solid black for Case 1, dash-dotted red for Case 2, and solid blue for Case 3.

y	$N = 10$	$N = 20$	$N = 30$	$N = 50$
5	0.128921	0.025536	0.005539	0.000755
10	0.123171	0.029763	0.006556	0.000517
15	0.086761	0.026535	0.006317	0.000349
20	0.054454	0.020534	0.005432	0.000237
25	0.032516	0.014616	0.004358	0.000221
30	0.018948	0.009835	0.003276	0.000195
a.u.e.b.	0.617191	0.243018	0.018839	0.004636

Table 1: Observed errors between the original marginal distribution of Q_2 and its QBD approximation for $\rho_1 = 0.7$ and $\rho_2 = 0.8$. The last line corresponds to the $a.u.e.b$ for each N.

Suppose that $\mathbb{P}(B = n) = \beta(1-\beta)^{n-1}$, $n = 1, 2, \ldots$ We find $\gamma = -\ln\big((\lambda + \mu_1 - \beta\mu_1)/\mu_1\big)$ and the rates with respect to the measure $\breve{\mathbb{P}}$ take the form $\breve{\lambda} = \beta\mu_1$ and $\breve{\mu}_1 = \lambda + \mu_1 - \beta\mu_1$. We also find that $\breve{\mathbb{E}}B = (\lambda + \mu_1 - \beta\mu_1)/\lambda$. Finally, using Proposition 3, we also calculate that $C_1 = (\beta\mu_1 - \lambda)\lambda/\beta\mu_1$. Combining these expressions, we calculate the *a.u.e.b.* in Theorem 1.

For our numerical experiments, we focus on the marginal distribution of Q_2. We performed extensive numerical experiments for various combinations of the parameters. We present here the combinations $\{\beta = 0.5, \rho_1 = 0.7, \rho_2 = 0.8\}$ (Case 2),since the qualitative results are similar among the various combinations we tested. Observe that due to the uniformisation $\lambda + \mu_1 + \mu_2 = 1$ of the rates, there exists a unique combination of $\{\lambda, \mu_1, \mu_2\}$ given a combination $\{\beta, \rho_1, \rho_2\}$. For this combination, we choose a number of truncation levels and we calculate for each N the truncated approximation $\mathbb{P}\big(Y_\infty^{(N)} \geq y\big)$, $y \geq 0$, with MAM.

To check the quality of our *a.u.e.b.*, we compare it with the differences between the exact and the truncated approximation of the marginal distribution of Q_2. We summarise our findings in Table 1.

From the table, we observe that the truncated approximations become more accurate as N increases, which is in accordance with our expectations. The same also holds for the asymptotic bound. However, the bound is at least 5 times greater than the observed error, which makes it rather conservative.

Similar results were derived in [34], where we performed additional numerical experiments for the special case of single arrivals of customers.

7. CONCLUSIONS

The conclusions we can draw for the asymptotic upper bound are summarised as follows: (i) The bound depends only on the truncation level and the parameters of the model; i.e. it is uniform in the values x and y of $\mathbb{P}\big(X_\infty^{(N)} \geq x, Y_\infty^{(N)} \geq y\big)$. (ii) The bound is rather conservative. Moreover, the bound becomes more conservative as the truncation level increases. (iii) Given the fact that the expression for γ in the leading term $e^{-\gamma N}$ is optimal, the conservative behaviour that our bound exhibits is probably attributed to the factor N.

The above observations indicate that further modifications are important to improve the accuracy of the asymptotic upper bound. One possible direction is to make the bound dependent on the values x and y. Most importantly, since the factor N of the bound seems to be more responsible for the latter's conservative behaviour, further improvements should be sought towards the removal of this factor from the bound. Nonetheless, the advantage of our bound is clear, in that it makes the procedure of truncating the background state rigorous, thereby reducing concerns raised in [6].

8. ACKNOWLEDGMENTS

The work of Maria Vlasiou and Eleni Vatamidou is supported by Netherlands Organisation for Scientific Research (NWO) through project number 613.001.006. The work of Bert Zwart is supported by an NWO VICI grant.

9. REFERENCES

[1] A. Alfa, B. Liu, and Q. He. Discrete-time analysis of $MAP/PH/1$ multiclass general preemptive priority queue. *Naval Research Logistics*, 50(6):662–682, 2003.

[2] E. Arjas and T. P. Speed. Symmetric Wiener-Hopf factorisations in Markov additive processes. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 26:105–118, 1973.

[3] S. Asmussen. Extreme value theory for queues via cycle maxima. *Extremes. Statistical Theory and Applications in Science, Engineering and Economics*, 1(2):137–168, 1998.

[4] S. Asmussen. *Applied Probability and Queues.* Springer-Verlag, New York, 2003.

[5] S. Asmussen and H. Albrecher. *Ruin Probabilities.* Advanced Series on Statistical Science & Applied Probability, 14. World Scientific, Second edition, 2010.

[6] N. Bean and G. Latouche. Approximations to quasi-birth-and-death processes with infinite blocks. *Advances in Applied Probability*, pages 1102–1125, 2010.

[7] D. A. Bini, G. Latouche, and B. Meini. *Numerical Methods for Structured Markov Chains.* Numerical Mathematics and Scientific Computation. Oxford University Press, 2005.

[8] L. Breuer and D. Baum. *An Introduction to Queueing Theory: and Matrix-Analytic Methods.* Springer, 2005.

[9] M. Brown, E. A. Peköz, and S. M. Ross. Some results for skip-free random walk. *Probability in the Engineering and Informational Sciences*, 24(4):491–507, 2010.

[10] G. Casale, E. Z. Zhang, and E. Smirni. KPC-Toolbox: Best recipes for automatic trace fitting using Markovian Arrival Processes. *Performance Evaluation*, 67(9):873–896, 2010.

[11] E. Çinlar. Markov additive processes. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 24(2):85–93, 1972.

[12] R. Doney, R. Maller, and M. Savov. Renewal theorems and stability for the reflected process. *Stochastic Processes and their Applications*, 119(4):1270 – 1297, 2009.

[13] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31(3–4):245–279, 1998.

[14] H. R. Gail, S. L. Hantler, and B. A. Taylor. Spectral analysis of $M/G/1$ and $G/M/1$ type Markov chains. *Advances in Applied Probability*, 28(1):114–165, 1996.

[15] E. J. Gümbel. *Statistics of extremes.* Columbia University Press, 1958.

[16] Q. M. He. *Fundamentals of Matrix-Analytic Methods.* Springer, 2014.

[17] A. Horváth, G. Horváth, and M. Telek. A joint moments based analysis of networks of MAP/MAP/1 queues. *Performance Evaluation*, 67(9):759–778, 2010.

[18] D. L. Iglehart. Extreme values in the GI/G/1 queue. *The Annals of Mathematical Statistics*, 43(2):627–635, 1972.

[19] E. Kao and K. Narayanan. Modeling a multiprocessor system with preemptive priorities. *Management Science*, 37(2):185–197, 1991.

[20] A. Kapadia, M. Kazmi, and A. Mitchell. Analysis of a finite capacity non preemptive priority queue. *Computers & Operations Research*, 11(3):337–343, 1984.

[21] M. Kijima. *Markov processes for stochastic modeling.* Springer, 1997.

[22] D. P. Kroese, W. R. W. Scheinhardt, and P. G. Taylor. Spectral properties of the tandem Jackson network, seen as a Quasi-Birth-and-Death process. *The Annals of Applied Probability*, 14(4):2057–2089, 2004.

[23] G. Latouche, G. Nguyen, and P. Taylor. Queues with boundary assistance: the effects of truncation. *Queueing Systems*, 69(2):175–197, 2011.

[24] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* SIAM, 1999.

[25] M. Miyazawa and B. Zwart. Wiener-Hopf factorizations for a multidimensional Markov additive process and their applications to reflected processes. *Stochastic Systems*, 2(1):67–114, 2012.

[26] M. F. Neuts. *Structured Stochastic Matrices of $M/G/1$ Type and their Applications*, volume 5 of *Probability: Pure and Applied.* Marcel Dekker Inc., 1989.

[27] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models.* Dover Publications Inc., 1994. Corrected reprint of the 1981 original.

[28] A. Ost. *Performance of communication systems: a model-based approach with matrix-geometric methods.* Springer, 2001.

[29] H. Rootzén. Maxima and exceedances of stationary Markov chains. *Advances in Applied Probability*, 20(2):371–390, 1988.

[30] R. Sadre. *Decomposition-Based Analysis of Queueing Networks.* PhD thesis, University of Twente, 2007.

[31] Y. Sakuma and M. Miyazawa. On the effect of finite buffer truncation in a two-node Jackson network. *Journal of Applied Probability*, 42(1):199–222, 2005.

[32] E. Seneta. *Nonnegative Matrices and Markov Chains.* Springer Series in Statistics. Springer-Verlag, Second edition, 1981.

[33] R. L. Tweedie. Operator-geometric stationary distributions for Markov chains, with application to queueing models. *Advances in Applied Probability*, 14(2):368–391, 1982.

[34] E. Vatamidou. *Error analysis of stuctured Markov chains.* PhD thesis, Eindhoven University of Tehcnology, 2015.

A Matrix-Analytic Approximation for Closed Queueing Networks with General FCFS Nodes

Giuliano Casale*
Imperial College London
United Kingdom
g.casale@imperial.ac.uk

Gábor Horváth
Budapest University of Technology and Economics
Hungary
ghorvath@hit.bme.hu

Juan F. Pérez†
University of Melbourne
Australia
juan.perez@unimelb.edu.au

1. INTRODUCTION

Multiclass closed queueing networks are established stochastic models used in performance evaluation of computer and communication systems. In the presence of nodes with first-come first-served (FCFS) scheduling, these models become difficult to analyze, since they can be handled by product-form theory only under restrictive assumptions, namely exponential service times and, locally to each node, identical mean service times for all classes [1]. In this paper, we propose a new method, called *decay rate approximation* (DRA), to heuristically analyze closed networks of FCFS nodes with phase-type distributed (PH) service times. Mean service times are assumed to be chosen arbitrarily.

DRA may be seen as a multiclass extension of the work in [2], described in the next section, which applies only to single-class networks. The central idea is to approximate joint queue-length distributions at multiclass FCFS nodes by multinomial distributions. The latter are obtained by studying each node as an MMAP[R]/PH[R]/1 queue in isolation, in order to extract an asymptotic decay rate of the joint queue-length distribution, which is then used to parameterize the multinomials. This multinomial approximation can be mapped into a product-form closed queueing network that approximates the original model and that can be efficiently analyzed. DRA also features an optimization program that refines the MMAP[R] input processes based on the estimated utilization at each queue. The goal of this program is to make the DRA predictions maximally consistent, in the sense of reconciling certain utilization predictions that are partly misaligned due to the heuristic nature of DRA.

By validating the method on networks of increasing size, we find that DRA incurs in less than 15% error on queue-lengths in all our tests, whereas state-of-the-art methods such as AMVA-FCFS [1, §10.2] and MVA incur errors that often exceed 20%-25% error.

2. BACKGROUND

Model. We consider a closed queueing network composed of M FCFS queues. The network processes R service classes, with a population of K_r jobs each ($r = 1, \dots, R$). Queue i ($i = 1, \dots, M$) processes jobs of class r with mean service demand $\theta_{ir} = v_{ir}s_{ir}$, where v_{ir} is the mean number of visits of jobs to the resource and s_{ir} is the mean service time per visit. We model the service time distribution of class r jobs at node i as a PH. Below we describe two heuristics to analyze this kind of models.

AMVA-FCFS Method. The AMVA-FCFS method considers a simple iterative solution scheme similar to the Bard-Schweitzer AMVA algorithm [1], but in which the arrival theorem is corrected to describe multiclass FCFS scheduling. Recall that the (exact) arrival theorem states that the waiting time of class-r jobs a queue i may be written as $W_{ir} = \theta_{ir} + \theta_{ir} \sum_{s=1}^{R} A_{is}^{(r)}$, where $A_{is}^{(r)}$ is the mean class-s queue-length at i seen by an arriving job of class r. This expression is exact for several queueing systems, in particular for processor-sharing (PS) queues. AMVA-FCFS corrects this expression at FCFS nodes as $W_{ir} \approx \theta_{ir} + \sum_{s=1}^{R} \theta_{is} A_{is}^{(r)}$, which accounts for the fact that jobs found ahead in the queue impose delays that depend on their classes.

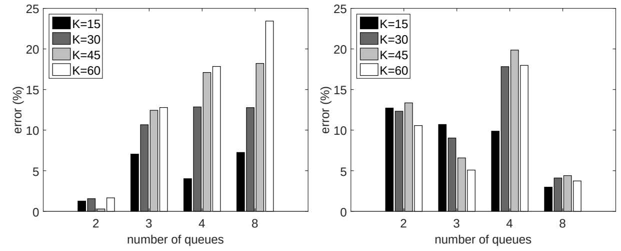
Despite its conceptual simplicity, AMVA-FCFS offers good performance on small models. However, it is easy to show cases where AMVA-FCFS incurs non-negligible errors as the model size grows. Moreover, the method is insensitive to moments of the service time distribution other than the mean, which limits its ability to account for PH service times. Figure 1 shows two such cases, the details of which are described in Section 4. This example indicates that the error of AMVA-FCFS can easily exceed 15-20%.

Single-Class Decay Rate Approximation. Recently, [2] considers single-class networks of FCFS nodes with PH service times and proposes a heuristic solution inspired by matrix-analytic methods. The authors first observe that the stationary queue-length distribution of a MAP/PH/1 queue may be quite accurately approximated in many cases by the expression

$$p_i(n) = \begin{cases} (1-\rho_i) & n = 0 \\ \rho_i(1-\eta_i)\eta_i^n & n \geq 1 \end{cases}$$

where i is the node label, ρ_i is the utilization, and η_i is the *caudal characteristic* of the node, i.e., the dominant eigenvalue

*The work of G. Casale has received funding from the Engineering and Physical Sciences Research Council (EPSRC) under grant agreement No. EP/M009211/1 (OptiMAM).
†The research of J. F. Pérez is supported by the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).



(a) Case 1

(b) Case 3

Figure 1: Approximation error with the AMVA-FCFS method.

of the rate matrix $\mathbf{R}$ associated to the quasi-birth-and-death process used to solve the MAP/PH/1 queue. This may be regarded as a heavy-traffic approximation, since η_i corresponds to the limiting decay rate of the exact queue-length probability as the number of jobs grows large. A heuristic product-form probability expression is then formulated, i.e.,

$$p(n_1, \ldots, n_M) = \frac{1}{C} \prod_{i=1}^{M} p_i(n_i) \qquad (1)$$

where C is a normalizing constant relatively to the state space of the closed network. In order to parameterize the above product-form approximation, the ρ_i and η_i values are required for all $i = 1, \ldots, M$. The approach proposed in [2] obtains these values by iteratively analyzing each MAP/PH/1 queue in isolation. Each iteration step assumes that the input flow of jobs to each node i is known. To this end, a utilization-based scaling of the service process of the feeding queues is computed and a MAP of their superposition is then used as input process for each of the MAP/PH/1 queues. Matrix-geometric solutions are then computed, from which new values for the decay rates η_i, $i = 1, \ldots, M$, are obtained.

New values for the utilizations are obtained by the expression $\rho_i = X\theta_i$, where X estimates the network throughput and θ_i is the mean service demand at node i. The network throughput is obtained as an average of the node throughputs predicted by the product-form expression (1), appropriately scaled by the visits v_i at each node. Averaging here helps convergence to a solution where the values $\rho_i = X\theta_i$ and the utilizations computed by summing (1) over states with $n_i > 0$ are identical and thus the model is self-consistent.

3. DECAY RATE APPROXIMATION

Although the heuristic in [2] offers good accuracy on single-class models, its generalization to the multiclass case is difficult for several reasons. First, the notion of caudal characteristic is difficult to define in the multiclass case, where there exist multiple ways to increase the job populations inside a queue. Moreover, the analysis of load-dependent multiclass models is computationally difficult, making it challenging to use probabilistic expressions such as (1) to model the network performance. In the next sections, we therefore examine and address these problems.

3.1 The MMAP[R]/PH[R]/1 queue

We begin by considering the problem of characterising the caudal characteristics in multiclass MMAP[R]/PH[R]/1 queues. An MMAP[R] process is ruled by R positive matrices $\mathbf{D}_r$ and a sub-generator matrix $\mathbf{D_0}$ such that $\sum_{r=0}^{R} \mathbf{D}_r$ is a Markov-chain generator matrix. In [4] Sengupta introduced semi-Markovian queues and a novel solution technique based on the age process analysis. The analysis of the age process turns out to be useful for the solution of many other queueing models, including the multiclass FCFS queue (see [3]).

If $\mathcal{A}(t)$ denotes the age and $\mathcal{J}(t)$ the phase of the background process (including the phase of the MMAP, the phase and the class of the current job in service), then the joint process $\{(\mathcal{A}(t), \mathcal{J}(t)) \geq 0\}$ is a continuous time Markov process on a continuous state space. It is proven in [4] that the stationary solution is matrix-exponential, thus the density $\pi_i(x) = \lim_{t\to\infty} \frac{d}{dx} P(\mathcal{A}(t) < x, \mathcal{J}(t) = i)$ is given by

$$\pi(x) = \pi(0) \cdot e^{\mathbf{T}x}, \qquad x > 0, \qquad (2)$$

where $\pi(x) = [\pi_i(x)]$. Matrix $\mathbf{T}$ in (2) is the solution of a non-linear matrix equation, and vector $\pi(0)$ is obtained by the solution of a set of linear equations. Both are discussed in detail in [3, Example 3.1].

3.2 The exact queue-length distribution

To express the distribution of the number of jobs in the system we first need to express the distribution of the number of jobs waiting in the queue. At time t, the age of the job in the server is $\mathcal{A}(t)$, thus all jobs waiting in the queue arrived in $(t - \mathcal{A}(t), t)$. Let us now introduce matrix $\mathbf{L}(\mathbf{n})$, which is related to the probability that $\mathbf{n} = (n_1, \ldots, n_R)$ jobs arrive over the (stationary) age of the current job in the service. In [3, Example 5.2] it is proven that $\mathbf{L}(\mathbf{n})$ can be recursively calculated by Sylvester matrix equations.

The joint distribution of the number of waiting jobs is

$$w(\mathbf{n}) = \begin{cases} 1 - \rho + \pi(0)\mathbf{L}(\mathbf{0})\mathbb{1}, & \mathbf{n} = \mathbf{0}, \\ \pi(0)\mathbf{L}(\mathbf{n})\mathbb{1}, & \mathbf{n} \neq \mathbf{0}, \end{cases} \qquad (3)$$

where $\mathbf{0}$ is the vector of zeros and $\mathbb{1}$ is the column vector of ones. To obtain the total number of jobs, the job in the server has to be taken into consideration as well. Let vector $\mathbf{h_r}$ hold 1 for states where a class-r job is in service and 0 otherwise. The distribution of the number of jobs in the system is

$$p(\mathbf{n}) = \begin{cases} 1 - \rho, & \mathbf{n} = \mathbf{0}, \\ \pi(0) \sum_{k=1, n_r>0}^{R} \mathbf{L}(\mathbf{n} - \mathbf{e}_r)\mathbf{h}_r, & \mathbf{n} \neq \mathbf{0}. \end{cases} \qquad (4)$$

where $\mathbf{e_r}$ is the rth unit vector.

3.3 A multinomial approximation

In the simplest special case, when the arrival processes are Poisson processes, and the service times are exponentially distributed with the same parameters, the joint distribution is a multinomial distribution as

$$p(\mathbf{n}) = (1 - \rho)\frac{(n_1+n_2+\ldots+n_R)!}{n_1!n_2!\cdots n_R!} \left(\frac{\lambda_1}{\mu}\right)^{n_1} \left(\frac{\lambda_2}{\mu}\right)^{n_2} \cdots \left(\frac{\lambda_R}{\mu}\right)^{n_R}.$$

To be able to use some of methods of product-form theory in the analysis of closed queueing networks composed by MMAP[R]/PH[R]/1 nodes, we are looking for an approximation which has a similar form. This is because the product-form solution involves a product of multinomial distributions [1]. According to [2] the queue length distribution of a MAP-driven single class queue with PH service times can be approximated by a geometric distribution reasonably well, with the geometric decay of the distribution given by the

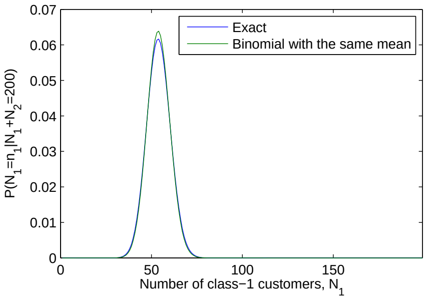


Figure 2: Distribution of the number of class-2 jobs given that the total number of jobs is 200

caudal characteristic. Unfortunately, such a straightforward approximation is not known for the multiclass variant of this family of queues. However, by numerical experiments we have found that, if the total number of jobs in the system is fixed, the distribution of the number of jobs belonging to different job classes is reasonably close to a multinomial distribution. A 4-state example with two job classes is shown in Figure 2, where the density of the number of class-2 jobs (N_2) is shown conditional on a queue-size of 200 jobs.

This and similar experiments we have performed in other cases suggest that this multinomial approximation is often accurate. For given per-class decay rates $\eta_r, r = 1, \dots, R$, a multinomial approximation of the queue length would be

$$p(\mathbf{n}) \approx p^*(\mathbf{n}) = \Big(1 - \sum_{i=1}^{R} \eta_i\Big) \frac{(n_1 + \cdots + n_R)!}{n_1! \cdots n_R!} \eta_1^{n_1} \cdots \eta_R^{n_R}.$$

Unfortunately, we were not able to analytically determine the decay rates of the exact distribution in order to minimize the distance from the multinomial expression. Therefore we give a numerical method to determine decay rates.

In order to uniquely define the caudal characteristic in a multiclass setting, we propose to consider a decay rate along the direction of queue-length increase that preserves the steady-state class mix of the MMAP[R]/PH[R]/1 queue, initially determined under a given MMAP[R] arrival process $\mathbf{D} = (\mathbf{D}_0, \mathbf{D}_1, \dots, \mathbf{D}_R)$. Let $y_r(q|\mathbf{D}), r = 1, \dots, R$, $q > 0$, be the decay of the joint queue length distribution along a direction (i.e., mix) corresponding to the ratios of the mean queue lengths under arrival process $\mathbf{D}$, i.e.,

$$y_r(q|\mathbf{D}) = \hat{p}(q\boldsymbol{\beta})/\hat{p}(q\boldsymbol{\beta} - \mathbf{e_r}), \tag{5}$$

where $\hat{p}(\mathbf{n}) = \frac{n_1! \cdots n_R!}{(n_1 + \cdots + n_R)!} p^*(\mathbf{n})$, $\boldsymbol{\beta} = \lfloor \mathbf{Q}/(\mathbf{Q}\mathbb{1}) \rfloor$ is the class mix, q is a scale factor on the total queue-length size, $\mathbf{Q}$ is the size R row vector of the mean queue lengths in the MMAP[R]/PH[R]/1 queue. We then propose an approximation for the decay rates for class k given by $\eta_r = \lim_{q \to \infty} y_r(q|\mathbf{D})$.

A drawback of this approach is that the stationary distribution has to be calculated up to a given threshold, which can be computationally demanding. Although in this paper we report results for large thresholds, we have empirically observed on the experiments in Section 4 that setting $q = 1 + \sum_r K_r$ preserves the accuracy of the method.

3.4 The DRA method

The decay rate approximation assumes the availability of an AMVA-FCFS implementation, a solver for continuous non-linear optimization programs (e.g., an interior point method), and an AMVA solver for product-form closed queueing networks. Given that many AMVA solvers exist[1], e.g., Bard-Schweitzer and Linearizer [1], which differ for the trade-off between accuracy and speed, the chosen product-form solver is generically referred to as AMVA-PF.

Let $\mathbf{X} = (X_1, \dots, X_R)$ be a vector of estimates of mean throughputs. We assume the initial value of $\mathbf{X}$ to be given by the throughput estimates returned by AMVA-FCFS for the multiclass FCFS network, which provides a reasonable initial guess of the closed network performance. DRA seeks to optimize, locally to the neighborhood of this initial point, the following non-convex non-linear program

$$\min f(\mathbf{X}) = \sum_{i=1}^{M} \sum_{r=1}^{R} |\tilde{\rho}_{ir} - \rho_{ir}| \text{ subject to } \mathbf{X} \geq \mathbf{X}^-.$$

Here $\mathbf{X}^-$ is a throughput lower bound, which we set in the experiments to a small positive quantity ($\epsilon = 10^{-3}$). The term $\rho_{ir} = X_r \theta_{ir}$, $r = 1, \dots, R$, is the mean per-class utilization at each queue when the throughputs are given by $\mathbf{X}$. The term $\tilde{\rho}_{ir}$ is the mean per-class utilization given by the product-form solver AMVA-PF, when the model is evaluated with service demands η_{ir}/X_r and populations K_r.

The decay rate η_{ir} corresponds to the decay rate η_r for a MMAP[R]/PH[R]/1 queue that represents node i in isolation. Such rates are obtained at each iteration of the nonlinear solver by the procedure outlined in Section 3.3, but where $\mathbf{Q}$ is replaced by the mean queue-length provided by AMVA-PF at the last iteration of the interior point method. At the first iteration, $\mathbf{Q}$ is obtained by AMVA-FCFS.

The service distributions of the MMAP[R]/PH[R]/1 queues are set equal to the PH service distributions of the FCFS queue. For a given queue i, the MMAP[R] is instead obtained by superposing the service processes PH[R] of all the queues j that feed i. Such superposition is carried out class-wise, i.e., only PH rates of the same job class r are superposed to define $\mathbf{D}_r$. Thus, the resulting MMAP[R] has the same number of classes R as the queueing network model. Note that prior to applying the superposition, the rates of the PH of queue j for class r are multiplied by ρ_{jr}, in order to account for idle periods. This provides a basic approximation of the class-r departure process of j; it would be interesting in future work to examine whether more sophisticated departure process models could improve accuracy.

Upon finding a local optimum, the DRA method returns as mean performance metrics those obtained at the last evaluation of the AMVA-PF solver. As observed before, the net-effect of the above approximation is to seek for a maximally consistent solution for a multinomial approximation of the joint queue-length distribution at multiclass FCFS nodes parameterized with the decay rates.

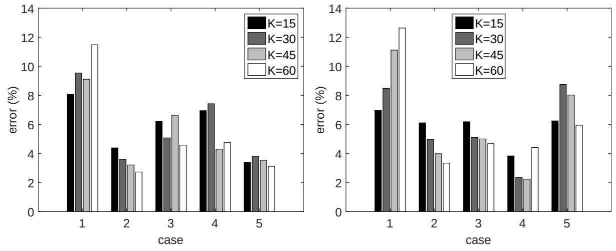
4. VALIDATION

We now evaluate the accuracy of the approximation proposed in Section 3.4. To this end we consider networks with $R = 2$ job classes and $M \in \{2, 3, 4, 8\}$ FCFS nodes in tandem. To account for different job population sizes, we

[1] Note that it is possible to combine Bard-Schweitzer and AMVA-FCFS in a single implementation, provided that different waiting time expressions are used at PS and FCFS nodes.

Table 1: Validation test cases - Mean service times ($s_{i,r}$)

	Queue							
Case	**1**		**2**		**3**		**4**	
1	1	0.5	0.8	0.6	0.4	0.7	0.7	0.8
2	1	0.5	0.8	0.2	0.4	0.1	0.7	0.2
3	1	0.5	0.8	0.8	0.4	0.4	0.7	0.7
4	1	1	1	1	1	1	1	1
5	1	0.5	1.2	0.6	1.4	0.7	1.6	0.8



(a) 2 queues

(b) 3 queues

Figure 3: DRA error: networks with 2 and 3 queues.

vary the total number of jobs K in the set $\{15, 30, 45, 60\}$, while keeping the ratio of class-2 to class-1 jobs constant and equal to 2. In addition, we consider 5 different cases for the demands, varying the relative demands across job classes and stations. Table 1 summarizes the mean demands for the network instances with 4 stations. The scenarios with 2 and 3 stations use the demands in this table corresponding to the first 2 and 3 queues, respectively. The scenarios with 8 queues use the demands in Table 1 for the first 4 queues while the demands for the remaining queues are set similarly. Finally, we assume all demands are exponentially distributed, except for the last one (e.g., the fourth one in the case with 4 stations), where the demands follow PH distributions. For this last station we select PH distributions that, using the method in [5] for hyper-exponential distributions, match the mean demand and a given squared coefficient of variation, which we set to 2 and 5 for jobs of class 1 and 2, respectively.

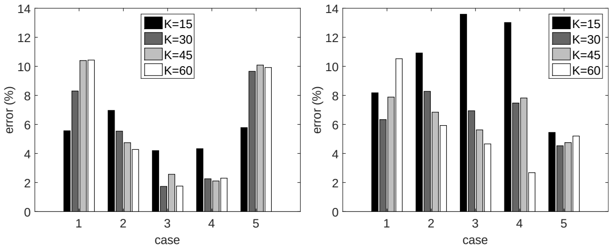
4.1 Results

Experiments have been carried out in MATLAB, using the fmincon interior-point method. The results for the scenarios with 2 and 3 queues are summarized in Figure 3. Here we report the mean absolute error in the mean queue length across all classes and stations, computed as

$$\text{error} = \frac{1}{2K}\sum_{i=1}^{M}\sum_{r=1}^{R}|Q_{i,r} - \hat{Q}_{i,r}|,$$

where $Q_{i,r}$ is the mean queue length for class-r jobs in station i obtained with the approximation, while $\hat{Q}_{i,r}$ is the same quantity obtained from an event-driven simulation. This error can be interpreted as the amount of (queue length) mass misplaced by the approximation, proportional to the total number of jobs K. The scaling factor in front of the summations ensures that the error is in $[0, 1]$. To ensure convergence, the simulation runs until the half-width of the confidence interval of the estimates is below 1% of the mean.

Figure 3 shows that DRA obtains errors mostly below 10% and always below 15%. In many cases the errors are actually between 2% and 5%. Similar results are obtained for 4 and 8 queues, as depicted in Figure 4, although the errors for the 8-queue case are somewhat larger. We do not observe any pattern in the approximation error as a function of the number of jobs K, nor across the different demand cases. Comparing with the results of the AMVA-FCFS method in Figure 1, we see that for case 1 AMVA-FCFS is better than DRA for a network with 2 queues, but for 3 and more queues the errors of AMVA-FCFS increase rapidly, while DRA remains below 13%. In fact, we have observed that while the AMVA-FCFS approximation may present large errors in some cases, especially with many queues, DRA offers errors that remain below 15% in all the cases tested. This is further illustrated in Table 2, where we depict the distribution of the errors achieved across all test cases by three methods: DRA, AMVA-FCFS, and the standard AMVA. There we observe how AMVA-FCFS improves upon plain AMVA, but still in over 35% of the scenarios the error is above 10%. With our proposed method, this percentage is reduced to just over 12%, all of which is concentrated in scenarios with error between 10% and 15% as none of the cases display errors beyond 15%.



(a) 4 queues

(b) 8 queues

Figure 4: DRA error: networks with 4 and 8 queues.

Table 2: Distribution of errors for different methods across all test cases

	Method		
Error (%)	DRA	AMVA-FCFS	AMVA
0 - 5	42.5%	33.75%	20%
5 - 10	45%	30%	38.75%
10 - 15	12.5%	27.5%	26.25%
15 - 20	-	7.5%	11.25%
20 - 25	-	1.25%	3.75%

5. REFERENCES

[1] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi. *Queueing Networks and Markov Chains*. Wiley, 2006.

[2] G. Casale and P. Harrison. A class of tractable models for run-time performance evaluation. In *3rd ACM/SPEC ICPE*, pages 63–74. ACM, 2012.

[3] Q. He. Analysis of a continuous time SM[K]/PH[K]/1/FCFS queue: Age process, sojourn times, and queue lengths. *Journal of Systems Science and Complexity*, 25(1):133–155, 2012.

[4] B. Sengupta. The semi-Markovian queue: theory and applications. *Stochastic Models*, 6(3):383–413, 1990.

[5] W. Whitt. Approximating a point process by a renewal process, I: two basic methods. *Oper Res*, 30:125–147, 1982.

On a 2-Class Polling Model with Class-dependent Reneging, Switchover Times, and Phase-type Service

[Extended Abstract] *

Steve Drekic
Department of Statistics and Actuarial Science
University of Waterloo
Canada
sdrekic@uwaterloo.ca

Kevin Granville
Department of Statistics and Actuarial Science
University of Waterloo
Canada
kgranville@uwaterloo.ca

ABSTRACT

In this talk, we analyze a 2-class, single-server polling model operating under a k_i-limited service discipline with class-dependent switchover times. Arrivals to each class are assumed to follow a Poisson process with phase-type distributed service times. Within each queue, customers are impatient and renege (i.e., abandon the queue) if the time before entry into service exceeds an exponentially distributed patience time. We model the queueing system as a level-dependent quasi-birth-and-death process, and the steady-state joint queue length distribution as well as the per-class waiting time distributions are computed via the use of matrix analytic techniques. The impact of reneging customers and choice of service time distribution is investigated through a series of numerical examples, with particular emphasis on the determination of (k_1, k_2) which minimizes a cost function involving the expected time a customer spends waiting in the queue and an additional penalty cost should reneging take place.

Keywords

Queueing theory; Polling model; Phase type distribution;

*A full version of this paper will be submitted elsewhere for publication

Matrix-analytic solution of second order Markov fluid models by using matrix-quadratic equations

[Extended Abstract]

Gábor Horváth
Budapest Univ. of Technology and Economics
Department of Networked Systems and Services
ghorvath@hit.bme.hu

Miklós Telek
MTA-BME Information Systems Research Group
Magyar Tudósok krt. 2, 1117 Budapest, Hungary
telek@hit.bme.hu

ABSTRACT

For the first-order Markovian fluid models, where the (state dependent) fluid rates are constant, there are various solution methods available. Among these solution methods the recently developed matrix-analytic method provides an efficient, numerically stable way to determine the stationary fluid level distribution even if the number of states is high.

In second-order Markovian fluid models the process determining the change of the fluid level is a Brownian motion with state-dependent drift and variance parameters. This paper presents a matrix-analytic solution of second-order fluid models where the matrix parameter of the matrix-exponential solution is obtained as a minimal non-negative solution of a matrix-quadratic equation.

1. INTRODUCTION

First-order Markovian fluid flow models are popular modeling tools with many practical applications. The differential equation providing the steady state distribution of the fluid level has been solved by eigenvalue decomposition based methods (see [3]). Later, more efficient procedures appeared that can solve larger models without the need of the numerically demanding eigenvalue decomposition and complex arithmetic. Such a method is the matrix-analytic solution (appearing in [4]), that provides the stationary distribution in a matrix-exponential form. The crucial step of this procedure is obtaining the minimal non-negative solution of a matrix-Riccati equation. In [4], this step is reduced to the solution of the matrix-quadratic equation.

In this paper we consider second-order Markovian fluid flows, which are Markov-modulated Brownian motions with a boundary at level 0. The differential equations governing the system are provided in [1], where an eigenvalue-based solution is also provided. As the matrix-analytical approach turned out to be more capable than the eigenvalue-based one in the first-order case, the aim to generalize it to the second-order case is natural. The contribution of the paper is the introduction of a matrix-quadratic equation whose minimal non-negative solution provides the matrix parameter of the matrix-exponentially distributed stationary fluid level distribution.

2. SECOND-ORDER FLUID MODELS

Second-order fluid flows are two-dimensional processes $\{\mathcal{X}(t), \mathcal{Z}(t),\ t \geq 0\}$, where $\mathcal{X}(t)$ is a continuous time background Markov chain (CTMC) with generator $\mathbf{Q}$ and state space $\mathcal{S}$, and $\mathcal{X}(t)$ represents the level of the fluid in a buffer. While the CTMC is in state i, the increment of $\mathcal{X}(t)$ is normally distributed with mean r_i and variance σ_i^2. Diagonal matrices $\mathbf{R}$ and $\mathbf{S}$ contain the drift and variance parameters, hence $\mathbf{R} = \text{diag}\langle r_i \rangle$ and $\mathbf{S} = \text{diag}\langle \sigma_i^2/2 \rangle$ (note that the variances are multiplied by 2 in order to make the arising expressions simpler).

The fluid buffer has a boundary at level 0. Two typical boundary behaviors are distinguished in the literature: the absorbing and the reflecting boundaries.

Let us denote the stationary fluid level density by vector $f(x) = [f_i(x), i \in \mathcal{S}]$, defined by $f_(x) = \lim_{t\to\infty} \frac{d}{dx} P(\mathcal{X}(t) < x, \mathcal{Z}(t) = i)$. The probability mass accumulating at level 0 and state i is denoted by $p_i = \lim_{t\to\infty} P(\mathcal{X}(t) = 0, \mathcal{Z}(t) = i)$. According to [2] $f(x)$ satisfies the differential equation

$$\frac{d}{dx} f(x)\mathbf{R} - \frac{d}{dx} f(x)\mathbf{S} = f(x)\mathbf{Q}. \tag{1}$$

Two different boundary behaviors are frequently distinguished in the literature, the *reflecting*, and the *absorbing* boundary.

- In case of a *reflecting* boundary, the density at level 0 satisfies
$$f(0)\mathbf{R} - f'(0)\mathbf{S} = f(x)\mathbf{Q}, \tag{2}$$
and for the probability mass at zero $p_i = 0,\ \forall i : r_i > 0$ or $\sigma_i > 0$.

- In case of an *absorbing* boundary, equation (2) still holds. Additionally, the density at level 0 is zero in the second order states ($f_i(0) = 0,\ \forall i : \sigma_i^2 > 0$), and the probability mass at 0 is zero in the positive states with zero variance ($p_i = 0,\ \forall i : r_i > 0$ and $\sigma_i^2 = 0$).

3. THE STATIONARY SOLUTION

The state space $\mathcal{S}$ is partitioned according to the sign of the rates and variances as follows:

- $\mathcal{S}^{+} = \{i : r_i > 0, \sigma_i^2 = 0\}$, $\mathcal{S}^{-} = \{i : r_i < 0, \sigma_i^2 = 0\}$,
- $\mathcal{S}^{\sigma+} = \{i : r_i > 0, \sigma_i^2 > 0\}$, $\mathcal{S}^{\sigma-} = \{i : r_i < 0, \sigma_i^2 > 0\}$,

and in this extended abstract we do not allow $r_i = 0$.

Hence, the set of states are $\mathcal{S} = \mathcal{S}^{+} \cup \mathcal{S}^{\sigma+} \cup \mathcal{S}^{\sigma-} \cup \mathcal{S}^{-} = \mathcal{S}^{\bullet} \cup \mathcal{S}^{-}$, where $\mathcal{S}^{\bullet} = \mathcal{S}^{+} \cup \mathcal{S}^{\sigma+} \cup \mathcal{S}^{\sigma-}$. In the rest of the paper it is assumed that the states of the CTMC are numbered according to the $\mathcal{S}^{+}, \mathcal{S}^{\sigma+}, \mathcal{S}^{\sigma-}, \mathcal{S}^{-}$ order of subsets.

From [1] (and from other sources as well) is known that $f(x)$ can be expressed in a matrix-exponential form. The order of this matrix-exponential equals $|\mathcal{S}^{\bullet}|$ ([1, Theorem 4.]). Taking this fact into consideration the solution can be transformed into the following form

$$f(x) = \pi e^{\mathbf{K}x} \begin{bmatrix} \mathbf{I} & \mathbf{\Psi} \end{bmatrix}, \tag{3}$$

where the size of $\mathbf{K}$ is $|\mathcal{S}^{\bullet}| \times |\mathcal{S}^{\bullet}|$ and the size of $\mathbf{\Psi}$ is $|\mathcal{S}^{\bullet}| \times |\mathcal{S}^{-}|$. Hence, the form of the solution is the same as in first order fluid models, therefore we used the same notations for the matrices. It is important to note, however, that matrices $\mathbf{K}$ and $\mathbf{\Psi}$ do not have the same elegant probabilistic interpretations as they have in [4] for the first order case.

In order to fully characterize the stationary behavior, it remains to solve

- matrices $\mathbf{K}$ and $\mathbf{\Psi}$,
- vector π,
- and the vector of probability masses at level 0 p.

3.1 Computing matrices $\mathbf{K}$ and $\mathbf{\Psi}$

Substituting the solution (3) into the differential equation (1) gives

$$\mathbf{K}\mathbf{R}_{\bullet} - \mathbf{K}^2\mathbf{S}_{\bullet} = \mathbf{Q}_{\bullet\bullet} + \mathbf{\Psi}\mathbf{Q}_{-\bullet}, \tag{4}$$

$$\mathbf{K}\mathbf{\Psi}\mathbf{R}_{-} \underbrace{-\mathbf{K}^2\mathbf{\Psi}\mathbf{S}_{-}}_{\mathbf{0}} = \mathbf{Q}_{\bullet-} + \mathbf{\Psi}\mathbf{Q}_{--}, \tag{5}$$

where $\mathbf{S}_{-} = \mathbf{0}$ has been exploited.

Let us now define diagonal matrixes with strictly positive diagonal elements $\mathbf{C}_{\bullet} = \begin{bmatrix} \mathbf{R}_{+} & & \\ & \mathbf{R}_{\sigma+} & \\ & & -\mathbf{R}_{\sigma-} \end{bmatrix}$ and $\mathbf{C}_{-} = -\mathbf{R}_{-}$, and choose an arbitrary constant c such that

$$c > \max\left(\max_{i \in \mathcal{S}^{+}} \frac{|q_{ii}|}{r_i}, \max_{i \in \mathcal{S}^{\sigma-} \bigcup \mathcal{S}^{\sigma+}} \frac{1}{2s_i}\left(\sqrt{r_i^2 + 4s_i|q_{ii}|} - r_i\right) \right). \tag{6}$$

Furthermore, by introducing matrices $\hat{\mathbf{K}} = \frac{1}{c}\mathbf{C}_{\bullet}^{-1}\mathbf{K}\mathbf{C}_{\bullet}$, $\hat{\mathbf{\Psi}} = \frac{1}{c}\mathbf{C}_{\bullet}^{-1}\mathbf{K}\mathbf{C}_{-}$, $\hat{\mathbf{S}}_{\bullet} = c\mathbf{C}_{\bullet}^{-1}\mathbf{S}_{\bullet}$, and $\hat{\mathbf{Q}} = \frac{1}{c}\mathbf{C}^{-1}\mathbf{Q}$ equations (4) and (5) simplify to

$$\hat{\mathbf{K}}\hat{\mathbf{I}}_{\bullet} - \hat{\mathbf{K}}^2\hat{\mathbf{S}}_{\bullet} = \hat{\mathbf{Q}}_{\bullet\bullet} + \hat{\mathbf{\Psi}}\hat{\mathbf{Q}}_{-\bullet}, \tag{7}$$

$$-\hat{\mathbf{K}}\hat{\mathbf{\Psi}} = \hat{\mathbf{Q}}_{\bullet-} + \hat{\mathbf{\Psi}}\hat{\mathbf{Q}}_{--}, \tag{8}$$

where $\hat{\mathbf{I}}_{\bullet} = \begin{bmatrix} \mathbf{I}_{+} & & \\ & \mathbf{I}_{\sigma+} & \\ & & -\mathbf{I}_{\sigma-} \end{bmatrix}$.

In the first-order case, when $\mathcal{S}^{\sigma+} = \mathcal{S}^{\sigma-} = \emptyset$, identity $\hat{\mathbf{I}}_{\bullet} = \mathbf{I}$ holds, which makes equations (7) and (8) easy to solve: expressing $\mathbf{K}$ from (7) and inserting the result into (8) leads to the well-known matrix Riccati equation for matrix $\mathbf{\Psi}$. In the second-order case, however, $\mathbf{\Psi}$ and $\mathbf{K}$ can not be obtained this way. Instead, a special quasi birth-death Markov chain (QBD) is introduced, and the fundamental matrix of this QBD will provide matrices $\mathbf{\Psi}$ and $\mathbf{K}$.

The regular part of the block-tri-diagonal generator of QBDs are characterized by three matrices: the transition rates corresponding to level-forward ($\mathbf{F}$), local ($\mathbf{L}$) and level-backward ($\mathbf{B}$) transitions. We define these matrices as follows

$$\mathbf{F} = \begin{bmatrix} \hat{\mathbf{Q}}_{\bullet\bullet} + \hat{\mathbf{I}}_{\bullet} + \hat{\mathbf{S}}_{\bullet} & \hat{\mathbf{Q}}_{\bullet-} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{9}$$

$$\mathbf{L} = \begin{bmatrix} -\hat{\mathbf{I}}_{\bullet} - 2\hat{\mathbf{S}}_{\bullet} & \mathbf{0} \\ \hat{\mathbf{Q}}_{-\bullet} & \hat{\mathbf{Q}}_{--} - \mathbf{I} \end{bmatrix}, \tag{10}$$

$$\mathbf{B} = \begin{bmatrix} \hat{\mathbf{S}}_{\bullet} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{11}$$

Observe that these matrices define a proper QBD, since due to (6)

$$\hat{\mathbf{Q}}_{++} + \mathbf{I}_{+} > 0, \tag{12}$$

$$\hat{\mathbf{Q}}_{\sigma+\sigma+} + \mathbf{I}_{\sigma+} + \hat{\mathbf{S}}_{\sigma+} > 0, \tag{13}$$

$$\hat{\mathbf{Q}}_{\sigma-\sigma-} - \mathbf{I}_{\sigma-} + \hat{\mathbf{S}}_{\sigma-} > 0 \tag{14}$$

$$\tag{15}$$

hold, hence $\hat{\mathbf{Q}}_{\bullet\bullet} + \hat{\mathbf{I}}_{\bullet} + \hat{\mathbf{S}}_{\bullet}$ (and therefore $\mathbf{F}$) is non-negative, furthermore,

$$\mathbf{I}_{\sigma-} - 2\hat{\mathbf{S}}_{\sigma-} < 0 \tag{16}$$

holds, hence $-\hat{\mathbf{I}}_{\bullet} - 2\hat{\mathbf{S}}_{\bullet}$ (and therefore $\mathbf{L}$) is a valid sub-generator. The non-negativity of $\mathbf{B}$ is straightforward. It can be checked that the row-sum of $\mathbf{F} + \mathbf{L} + \mathbf{B}$ is zero as well.

THEOREM 1. *The minimal non-negative solution of the matrix-quadratic equation* $\mathbf{F} + \mathbb{R}\mathbf{L} + \mathbb{R}^2\mathbf{B} = \mathbf{0}$ *is*

$$\mathbb{R} = \begin{bmatrix} \hat{\mathbf{K}} + \mathbf{I} & \hat{\mathbf{\Psi}} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{17}$$

PROOF. Substituting the solution gives identity for the matrix equation. When the eigenvalues of $\mathbb{R}$ are in the unit disk the eigenvalues of $\hat{\mathbf{K}}$ as well as the eigenvalues of $\mathbf{K}$ have negative real part. □

Second-order fluid flow models: reflected Brownian motion in a random environment

3.2 Computing vectors π and p

3.2.1 *Reflecting boundary*

If the boundary is reflecting, $p_{\bullet} = 0$ holds. Inserting the matrix-exponential solution into (2) and taking the state partitioning into account leads to equations

$$\pi\mathbf{R}_{\bullet} - \pi\mathbf{K}\mathbf{S}_{\bullet} = p_{-}\mathbf{Q}_{-\bullet}, \tag{18}$$

$$\pi\mathbf{\Psi}\mathbf{R}_{-} = p_{-}\mathbf{Q}_{--}, \tag{19}$$

since $\mathbf{S}_{-} = 0$.

After some manipulation vectors π and p_{-} are expressed by

$$\pi(\mathbf{R}_{\bullet} - \mathbf{K}\mathbf{S}_{\bullet} - \mathbf{\Psi}|\mathbf{R}_{-}|(-\mathbf{Q}_{--})^{-1}\mathbf{Q}_{-\bullet}) = 0, \tag{20}$$

$$p_{-} = \pi|\mathbf{R}_{-}|(-\mathbf{Q}_{--})^{-1}. \tag{21}$$

In order to obtain a unique solution, the normalization condition has to be taken into consideration as well, thus the additional equation is

$$\pi\left((-\mathbf{K})^{-1}\begin{bmatrix}\mathbf{I} & \boldsymbol{\Psi}\end{bmatrix}\mathbb{1} + |\mathbf{R}_-|(-\mathbf{Q}_{--})^{-1}\mathbb{1}\right) = 1. \qquad (22)$$

3.2.2 Absorbing boundary

In case of the absorbing boundary, the density is zero in the second order states, hence, $f_{\sigma+}(0) = 0$ and $f_{\sigma-}(0) = 0$. Since the density at zero is expressed by $f(0) = \pi\begin{bmatrix}\mathbf{I} & \boldsymbol{\Psi}\end{bmatrix}$, this implies that $\pi_{\sigma+} = 0$ and $\pi_{\sigma-} = 0$, given that $\pi = \begin{bmatrix}\pi_+ & \pi_{\sigma+} & \pi_{\sigma-}\end{bmatrix}$. With such a π vector the terms in (2) are

$$f(0)\mathbf{R} = \begin{bmatrix}\pi_+\mathbf{R}_+ & 0 & 0 & \pi_+\boldsymbol{\Psi}_{+-}\mathbf{R}_-\end{bmatrix}, \qquad (23)$$

$$f'(0)\mathbf{S} = \begin{bmatrix}0 & \pi_+\mathbf{K}_{+,\sigma_+}\mathbf{S}_{\sigma_+} & \pi_+\mathbf{K}_{+,\sigma_-}\mathbf{S}_{\sigma_-} & 0\end{bmatrix}, \qquad (24)$$

since $\mathbf{S}_+ = \mathbf{S}_- = 0$. Hence, for our partitioned vectors and block matrices (2) can be rewritten as

$$\begin{aligned}
\pi_+\mathbf{R}_+ &= p_{\sigma+}\mathbf{Q}_{\sigma+,+} + p_{\sigma-}\mathbf{Q}_{\sigma-,+} + p_-\mathbf{Q}_{-,+},\\
-\pi_+\mathbf{K}_{+,\sigma_+}\mathbf{S}_{\sigma_+} &= p_{\sigma+}\mathbf{Q}_{\sigma+,\sigma_+} + p_{\sigma-}\mathbf{Q}_{\sigma-,\sigma_+} + p_-\mathbf{Q}_{-,\sigma_+},\\
-\pi_+\mathbf{K}_{+,\sigma_-}\mathbf{S}_{\sigma_-} &= p_{\sigma+}\mathbf{Q}_{\sigma+,\sigma_-} + p_{\sigma-}\mathbf{Q}_{\sigma-,\sigma_-} + p_-\mathbf{Q}_{-,\sigma_-},\\
\pi_+\boldsymbol{\Psi}_{+-}\mathbf{R}_- &= p_{\sigma+}\mathbf{Q}_{\sigma+,-} + p_{\sigma-}\mathbf{Q}_{\sigma-,-} + p_-\mathbf{Q}_{-,-},
\end{aligned}$$

which, in matrix form, defines

$$\begin{bmatrix}\pi_+ & p_{\sigma+} & p_{\sigma-} & p_-\end{bmatrix}\cdot
\begin{bmatrix}
-\mathbf{R}_+ & \mathbf{K}_{+,\sigma_+}\mathbf{S}_{\sigma_+} & \mathbf{K}_{+,\sigma_+}\mathbf{S}_{\sigma_-} & -\boldsymbol{\Psi}_{+-}\mathbf{R}_-\\
\mathbf{Q}_{\sigma+,+} & \mathbf{Q}_{\sigma+,\sigma_+} & \mathbf{Q}_{\sigma+,\sigma_-} & \mathbf{Q}_{\sigma+,-}\\
\mathbf{Q}_{\sigma-,+} & \mathbf{Q}_{\sigma-,\sigma_+} & \mathbf{Q}_{\sigma-,\sigma_-} & \mathbf{Q}_{\sigma-,-}\\
\mathbf{Q}_{-,+} & \mathbf{Q}_{-,\sigma_+} & \mathbf{Q}_{\sigma-,-} & \mathbf{Q}_{-,-}
\end{bmatrix} = 0. \qquad (25)$$

Finally, the normalization condition to be added to the above set of equations is

$$\begin{bmatrix}\pi_+ & p_{\sigma+} & p_{\sigma-} & p_-\end{bmatrix}\cdot
\begin{bmatrix}
(-\mathbf{K})^{-1}_{+\bullet}\begin{bmatrix}\mathbf{I} & \boldsymbol{\Psi}\end{bmatrix}\mathbb{1}\\
\mathbb{1}\\
\mathbb{1}\\
\mathbb{1}
\end{bmatrix} = 1. \qquad (26)$$

Acknowledgments

This research is partially supported by the OTKA K101150 project.

4. REFERENCES

[1] R. L. Karandikar and V.G. Kulkarni. Second-order fluid flow models: reflected brownian motion in a random environment. *Operations Research*, 43:77–88, 1995.

[2] Samuel Karlin and Howard E Taylor. *A second course in stochastic processes.* Elsevier, 1981.

[3] V. G. Kulkarni. Fluid models for single buffer systems. In J. H. Dshalalow, editor, *Models and Applications in Science and Engineering*, Frontiers in Queueing, pages 321–338. CRC Press, 1997.

[4] V. Ramaswami. Matrix analytic methods for stochastic fluid flows. In *International Teletraffic Congress*, pages 1019–1030, Edinburg, 1999.

Markov-modulated Brownian motion and the flip-flop fluid queue

A symbiotic relationship

Guy Latouche
Université libre de Bruxelles
Département d'informatique
1050, Bruxelles, Belgium
latouche@ulb.ac.be

Giang T. Nguyen
The University of Adelaide
School of Mathematical Sciences
SA 5005, Australia
giang.nguyen@adelaide.edu.au

ABSTRACT

Fluid queues are a particular family of Markov-modulated Brownian motions (MMBMs) characterized by the fact that the evolution is piece-wise linear without Brownian noise. A key difference is that stochastic fluid queues are amenable to Kolmogorov-type analysis.

Ramaswami has shown in 2013 that standard Brownian motion is the limit of a family of two-phase fluid queues with increasing rates of transition from one phase to the other. This has been extended to Markov-modulated Brownian motions and we have coined the nomenclature *flip-flop fluid queue* as a reminder of the fast transitions between up and down movements.

This approximation has proved to give us a versatile tool to analyse properties of MMBMs, in the spirit of Matrix-Analytic methods. In particular, we have been able to use a semi-regenerative approach to analyse the stationary properties of various processes related to MMBMs.

We present here a summary of the idea behind the technique and of the results obtained so far.

Keywords

Fluid queues, Markov-modulated Brownian motion, regenerative analysis

1. INTRODUCTION

We start from $\{Y(t), \kappa(t)\}_{t\geq 0}$, a Markov-modulated Brownian motion where the phase process κ is a Markov chain with state space $\mathcal{M} = \{1, \ldots, m\}$, and Y is a Brownian motion with drift μ_i and variance σ_i^2 whenever $\kappa(t) = i \in \mathcal{M}$. We denote by Δ_μ the drift matrix $\mathrm{diag}(\mu_1, \ldots, \mu_m)$, by Δ_σ^2 the variance matrix $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$, and by Q the generator of κ, and we assume that Q is irreducible.

The family of fluid processes $\{X_\lambda(t), \beta_\lambda(t), \varphi_\lambda(t)\}_{t\geq 0}$ is constructed as follows: the phase process $(\beta_\lambda(t), \varphi_\lambda(t))$ is a two-dimensional Markov chain with state space $\mathcal{S} = \{(k, i) : k \in \{1, 2\}$ and $i \in \mathcal{M}\}$ and generator

$$T_\lambda = \begin{bmatrix} Q - \lambda I & \lambda I \\ \lambda I & Q - \lambda I \end{bmatrix},$$

where the entries of T_λ follow the lexicographic ordering of $\{1, 2\} \times \mathcal{M}$, and I denotes an appropriately-sized identity matrix. Whenever ambiguity might arise, we write I_n to denote the $n \times n$ identity matrix. The fluid rate matrix C_λ is given by

$$C_\lambda = \begin{bmatrix} \Delta_\mu + \sqrt{\lambda}\Delta_\sigma & \\ & \Delta_\mu - \sqrt{\lambda}\Delta_\sigma \end{bmatrix}.$$

Intuitively speaking, we duplicate the state space $\mathcal{M}$ in the Markov-modulated Brownian motion $\{Y(t), \kappa(t)\}$, and the auxiliary process $\beta_\lambda(t)$ keeps track of which copy is in use. Note that for λ sufficiently large, the phases in the copy with $\beta_\lambda(t) = 1$ have all positive rates while the phases in the other copy have all negative rates. We use the term *flip-flop* processes to characterize the triplet $\{X_\lambda(t), \beta_\lambda(t), \varphi_\lambda(t)\}$.

The idea originated with Ramaswami [8], where it is shown that standard Brownian motion arises as the limit of a family of Markov-modulated linear fluid processes. We extended in [5] the argument of Ramaswami and showed that the flip-flop fluid queue converges weakly, as λ goes to infinity, to the Markov-modulated Brownian motion. We proved that the stationary distribution of a Markov-modulated Brownian motion regulated at zero is the limit from the well-analyzed stationary distribution of fluid queues, and so provided a new approach for obtaining the stationary distribution of a regulated MMBM.

Our results opened the way to the analysis of more complex processes and we give in this presentation a brief summary of some processes that have been analysed, with emphasis on two approaches which proved to be very fruitful:

- We identify a set of regeneration points, and so are able to confirm old results, and provide new ones, without having to rely on time-reversal;
- We use techniques developed for fluid queues to supplement, wherever needed, known results from traditional MMBM analysis.

2. MMBM WITH STICKY BOUNDARY

Systems in real life are designed with feedback loops. This is our reason for studying stochastic processes with *reactive* boundaries, that is, processes that change behaviour upon hitting some boundary. In [6] we focus on regulated MMBMs with a *sticky boundary* at level 0.

Brownian motions with a sticky boundary were introduced by Feller [1] in the 1950s. Briefly stated, the regulated Brownian motion is slowed down when it is at level 0, in such a way that, without actually staying at zero for any interval of time of positive length, it does spend in that level an amount of time with positive Lebesgue measure.

We extend in two ways the construction in Harrison and Lemoine [2]. First, we define a straightforward generalization based on the regulator $R^*(t) = |\inf_{0 \leq s \leq t} X^*(s)|$ and a change of clock. The new clock increases at rate 1 while the regulated process is strictly positive, at a slower rater when the process is equal to 0. In a further extension, the phase is controlled with a different Markov process when the MMBM is at level 0.

To determine the stationary distribution of our processes, we choose points of regeneration forming a subset of the epochs when the process hits level 0: once the process hits the boundary, we start an exponential timer and we do not register the instantaneous returns to 0 by the Brownian motion until the timer has expired.

We use a flip-flop approximation of the MMBM with sticky boundary to determine the expected time spent in various states during an interval between regeneration, and we obtain an expression for the stationary distribution, expression which has a very simple physical interpretation.

3. TWO-SIDED MMBM WITH BOUNDARY CONTROL

We consider in [3] a Markov-modulated Brownian motion with two boundaries at 0 and $b > 0$, and we allow for the controlling Markov chain to instantaneously undergo a change of phase upon hitting either of the two boundaries at *semi-regenerative epochs.* These are defined to be the first time the process reaches a boundary since it last hits the other boundary. To give one example, assume that losses occur whenever the buffer gets full; to reduce the losses, one might instantaneously increase the speed at which the buffer is emptied. When the buffer content drops to level 0, operations may resume under normal conditions.

To determine the stationary distribution, the key ingredients needed are the expected time spent in an interval $[0, x]$ and in a given phase during an excursion from 0 to b for the process regulated at 0, and from b to 0 for the process regulated at b, as well as the distribution of the phase upon reaching a boundary. To obtain these quantities, we use connections obtained in [4] between MMBMs and their flip-flop approximations, and we prove new ones.

4. CONCLUSION

The synergy of the semi-regenerative analysis of stochastic processes and the flip-flop fluid approximation to MMBMs is a powerful one. It is being put to use in Latouche and Simon [7], and it will prove to be useful in many circumstances beyond the models already mentioned.

5. ACKNOWLEDGMENTS

The authors thank the Ministère de la Communauté française de Belgique for supporting this research through the ARC grant AUWB-08/13–ULB 5, they acknowledge the financial support of the Australian Research Council through the Discovery Grant DP110101663. Giang Nguyen also acknowledges the support of ARC Centre of Excellence for Mathematical and Statistical Frontiers.

6. REFERENCES

[1] W. Feller. The parabolic differential equations and the associated semi-groups of transformations. *Ann. of Math.*, 55:468–519, 1952.

[2] J. M. Harrison and A. J. Lemoine. Sticky Brownian motion as the limit of storage processes. *J. Appl. Probab.*, 18:216–226, 1981.

[3] G. Latouche and G. T. Nguyen. Feedback control: two-sided markov-modulated brownian motion with instantaneous change of phase at boundaries. *In preparation*, 2015.

[4] G. Latouche and G. T. Nguyen. Fluid approach to two-sided Markov-modulated Brownian motion. *Queueing Systems*, 80:105–125, 2015. doi: 10.1007/s11134-014-9432-8, arXiv:1403.2522.

[5] G. Latouche and G. T. Nguyen. The morphing of fluid queues into Markov-modulated Brownian motion. *Stochastic Systems*, 5:62–86, 2015. doi: 10.1214/13-SSY133.

[6] G. Latouche and G. T. Nguyen. Slowing time: Markov-modulated Brownian motion with a sticky boundary. *Submitted*, 2015. arXiv:1508.00922.

[7] G. Latouche and M. Simon. Regulated Markov-modulated Brownian motion with temporary change of regime upon visits to level zero. *In preparation*, 2016.

[8] V. Ramaswami. A fluid introduction to Brownian motion and stochastic integration. In G. Latouche, V. Ramaswami, J. Sethuraman, K. Sigman, M. Squillante, and D. Yao, editors, *Matrix-Analytic Methods in Stochastic Models*, volume 27 of *Springer Proceedings in Mathematics & Statistics*, chapter 10, pages 209–225. Springer Science, New York, NY, 2013.

Componentwise accurate numerical methods for Markov-modulated Brownian motion[*]

Giang T. Nguyen
School of Mathematical Sciences
The University of Adelaide
Adelaide, Australia
giang.nguyen@adelaide.edu.au

Federico Poloni
Dipartimento di Informatica
Università di Pisa
Pisa, Italy
federico.poloni@unipi.it

ABSTRACT

We describe a componentwise accurate algorithm to find the stationary distribution of a Markov-modulated Brownian motion $\{Y(t), \varphi(t)\}_{t\geq 0}$. The algorithm is based on finding a suitable invariant pair (X, U), with $U = \begin{bmatrix} I & \Psi \end{bmatrix}$, satisfying $X^2UV - XUD + UQ = 0$, where Q is the rate matrix of the driving continuous-time Markov chain $\varphi(t)$ on state space $\mathcal{M}$, and the diagonal matrices D and V contain, respectively, the drifts d_i and parameters $2\sigma_i^2$, where σ_i^2 is the variance, for $i \in \mathcal{M}$. The algorithm is based on a componentwise accurate variant of Cyclic Reduction; a special treatment based on the shift technique is needed in the case when V is singular.

CCS Concepts

•**Mathematics of computing** → **Queueing theory; Computations on matrices;** *Markov processes;*

Keywords

Markov-modulated Brownian motion; Cyclic Reduction; componentwise accurate computation; numerical linear algebra; matrix equations

1. INTRODUCTION

Markov-modulated Brownian motion [6, 11] is a popular tool in modeling fluid processes evolving with a stochastic behavior. The process is defined by a continuous-time Markov chain $\varphi(t)$ with transition matrix Q on state space $\mathcal{M} = \{1, \dots, n\}$, and by a level process $Y(t)$ which evolves as a Brownian motion with drift d_i and variance σ_i^2 whenever $\varphi(t) = i$, for $i \in \mathcal{M}$. Here, we consider the case in which the boundary conditions are absorbing at zero, that is, we set $Y(t) = 0$ whenever it would become negative.

The stationary density $\boldsymbol{p}(x) : (0, \infty) \mapsto \mathbb{R}^{1\times n}$ is a vector-valued function such that

$$(\boldsymbol{p}(x))_i = \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}[Y(t) \leq x, \varphi(t) = i].$$

It was proved [6] that it satisfies the differential equation

$$\ddot{\boldsymbol{p}}(x)V - \dot{\boldsymbol{p}}(x)D + \boldsymbol{p}(x)Q = 0, \tag{1}$$

with $V = \operatorname{diag}(2\sigma_i^2)_{i\in\mathcal{M}}$ and $D = \operatorname{diag}(d_i)_{i\in\mathcal{M}}$. Moreover, the stationary distribution $\boldsymbol{p}$ is matrix-exponential:

$$\boldsymbol{p}(x) = \boldsymbol{v}\exp(Xx)U$$

for suitable $\boldsymbol{v} \in \mathbb{R}^{1\times k}$, $X \in \mathbb{R}^{k\times k}$, and $U \in \mathbb{R}^{k\times n}$, with $k = |\mathcal{M}_+|$, where $\mathcal{M}_+ = \{i : v_i \neq 0 \text{ or } d_i > 0\}|$. Without loss of generality, we assume the states are ordered so that $\mathcal{M}_+ = \{1, 2, \dots, k\}$.

2. NUMERICAL METHODS

A number of different methods have been suggested in the literature to find the parameters $\boldsymbol{v}, X, U$. The method in [6] is based on the explicit computation of the eigenvectors and eigenvalues of a suitable matrix constructed from Q, D, V; the one in [1] is based instead on a block diagonal decomposition, which is computed through an iterative method. Other algorithms stem from the fact that

$$X^2UV - XUD + UQ = 0, \tag{2}$$

or, in linear algebra terms, (X, U) is a *left invariant pair* of the matrix polynomial $P(z) := Vz^2 - Dz + Q$. The method in [2] to compute invariant pairs can be used for this problem. If V is nonsingular, then (2) reduces to a matrix equation

$$X^2V - XD + Q = 0;$$

the method [8] uses a Cayley transform to reduce this equation to a version commonly studied in the setting of quasi-birth-death (QBD) processes, the most efficient algorithm to solve it is then Cyclic Reduction [4].

Moreover, it is shown in [5] that we can take $U = \begin{bmatrix} I & \Psi \end{bmatrix}$ for a suitable matrix $\Psi \geq 0$ (here and in the following, inequalities on matrices are intended in the componentwise sense), and that the associated X is a subgenerator matrix, that is, $X_{ij} \leq 0$ for $i \neq j$ and $X\mathbf{1} \leq \mathbf{0}$, where $\mathbf{1}$ is the vector

[*] Giang Nguyen acknowledges the financial support of the Australian Research Council through the Discovery Grant DP110101663, and the support of ARC Centre of Excellence for Mathematical and Statistical Frontiers. Federico Poloni acknowledges the financial support of a PRA research project of the University of Pisa, and of a INDAM-GNCS young researcher grant.

of all ones. These matrices can be interpreted as the probabilities of first return to the starting level and the generator of the process of downward records for the time-reversed process, respectively.

3. COMPONENTWISE ACCURACY

The papers [9, 12] present algorithms to solve the special case in which $V = 0$, which are stable in a strong sense: we say that $\widetilde{M} \in \mathbb{R}^{m\times n}$ is a *componentwise accurate* approximation (within a threshold ε) of $M \in \mathbb{R}^{m\times n}$ if

$$\frac{|\widetilde{M}_{ij} - M_{ij}|}{|M_{ij}|} \leq \varepsilon \qquad \text{for all } i \text{ and } j.$$

With this definition, these algorithms produce quantities $X, U, \boldsymbol{v}$, and $\boldsymbol{p}(x)$ that are componentwise accurate, within a threshold that is a small multiple $C\mathrm{u}$ of the machine precision u, for some constant C. This guarantee is stronger than what is traditionally provided by numerical linear algebra algorithms, that is, $\|\widetilde{M}-M\|/\|M\| \leq \varepsilon$. In particular, tiny entries are computed with a high number of significant digits, and quantities that have an interpretation in terms of probability are always nonnegative, a property highly valued in models of this kind.

Componentwise stability is obtained by ensuring that the algorithms contain (almost) no subtractions between two quantities with the same sign (*subtraction-free algorithms*), which is made possible by the special sign structure of the problem. The main building block is the *GTH algorithm* [10], which allows one to solve accurately linear systems with an M-matrix M if two vectors $\boldsymbol{u} > \boldsymbol{0}, \boldsymbol{v} \geq \boldsymbol{0}$ such that $M\boldsymbol{u} = \boldsymbol{v}$ are known accurately. The triple $(M, \boldsymbol{u}, \boldsymbol{v})$ is known as a *triplet representation.*

4. THE CASE OF POSITIVE VARIANCES

In this extended abstract and the associated presentation, we describe an extension of the techniques introduced in the previous section to deal with the Brownian motion case. The simplest case is the one in which $V > 0$. In this setting, we need only a few modifications to the strategy in [8] to ensure componentwise accuracy. We have $k = n$, and we can choose $U = I$. We take $h > 0$ small enough such that $v_i + d_i h + Q_{ii}h^2 \geq 0$ for all $i = 1, 2, \ldots, n$; then, direct verification shows that $R = I + hX$ satisfies the equation $R^2A - RB + C = 0$, where

$$A := V, \quad B := 2V + hD, \quad C := V + hD + h^2Q. \qquad (3)$$

These coefficients satisfy $A \geq 0, C \geq 0$, and $(A - B + C)\mathbf{1} = \mathbf{0}$. These matrices can be interpreted (possibly after a scaling) as the transition matrices of a QBD process; hence, we can compute R with the following iteration, known as *cyclic reduction* [4].

$$\begin{aligned}
&A_0 = A,\, B_0 = \hat{B}_0 = B,\, C_0 = C,\\
&A_{k+1} = A_k B_k^{-1} A_k,\\
&B_{k+1} = B_k - A_k B_k^{-1} C_k - C_k B_k^{-1} A_k,\\
&C_{k+1} = C_k B_k^{-1} C_k,\\
&\hat{B}_{k+1} = \hat{B}_k - C_k B_k^{-1} A_k.
\end{aligned}$$

Indeed, $\hat{B}_k$ converges to a matrix $\hat{B}_\infty$, and $R = C_0\hat{B}_\infty^{-1}$. The required inversions can be performed with the GTH algorithm, using the triplet representation $B_k\mathbf{1} = (A_k + C_k)\mathbf{1}$ and $\hat{B}_\infty\mathbf{1} = C_0\mathbf{1} + \boldsymbol{w}$, with $\boldsymbol{w} = \lim_{k\to\infty} A_k\mathbf{1}$. The former triplet has been already used in the implementation of Cyclic Reduction (see, for example, [3]) but the latter, to the best of our knowledge, is new. A triplet representation for the M-matrix $-X^\top$ can be derived explicitly as well.

In the previous algorithm, subtractions are needed only in the computation of $\operatorname{diag}(C)$, but choosing h suitably one can make sure that they cause no significant cancellation.

5. THE GENERAL CASE

When $k \neq n$, there are a number of difficulties in the previous algorithm; the most apparent one is that we cannot choose $h > 0$ to satisfy the required inequalities. Further ones appear if one considers the location of eigenvalues and the minimality of the computed solutions, which we have not done in this very brief note.

To work around these issues, let us subdivide (1) into blocks corresponding to S and its complementary set

$$\begin{bmatrix}\ddot{\boldsymbol{p}}_1 & \ddot{\boldsymbol{p}}_2\end{bmatrix}\begin{bmatrix}V_1 & 0\\ 0 & 0\end{bmatrix} - \begin{bmatrix}\dot{\boldsymbol{p}}_1 & \dot{\boldsymbol{p}}_2\end{bmatrix}\begin{bmatrix}D_1 & 0\\ 0 & D_2\end{bmatrix} + \begin{bmatrix}\boldsymbol{p}_1 & \boldsymbol{p}_2\end{bmatrix}\begin{bmatrix}Q_{11} & Q_{12}\\ Q_{21} & Q_{22}\end{bmatrix} = \mathbf{0}.$$

We differentiate the equation corresponding to the second column, $-\dot{\boldsymbol{p}}_2 D_2 + \boldsymbol{p}_1 Q_{12} + \boldsymbol{p}_2 Q_{22} = \mathbf{0}$, obtaining

$$-\ddot{\boldsymbol{p}}_2 D_2 + \dot{\boldsymbol{p}}_1 Q_{12} + \dot{\boldsymbol{p}}_2 Q_{22} = \mathbf{0}.$$

Multiplying the latter equation by $-h$ and summing it to the original second column equation leads to

$$\begin{bmatrix}\ddot{\boldsymbol{p}}_1 & \ddot{\boldsymbol{p}}_2\end{bmatrix}\begin{bmatrix}V_1 & 0\\ 0 & -hD_2\end{bmatrix} - \begin{bmatrix}\dot{\boldsymbol{p}}_1 & \dot{\boldsymbol{p}}_2\end{bmatrix}\begin{bmatrix}D_1 & -hQ_{12}\\ 0 & D_2 - hQ_{22}\end{bmatrix} + \begin{bmatrix}\boldsymbol{p}_1 & \boldsymbol{p}_2\end{bmatrix}\begin{bmatrix}Q_{11} & Q_{12}\\ Q_{21} & Q_{22}\end{bmatrix} = \mathbf{0}. \qquad (4)$$

We now continue with the transformation (3), replacing Q, D, V with the corresponding coefficients of (4). This technique of differentiating equations is common in the field of differential-algebraic equations [7], where it is used for different reasons: not to preserve sign structures, but to obtain a system with a nonsingular highest-order factor for numerical integration. Alternatively, it can be understood as a shift technique [4] to relocate some of the infinite eigenvalues of $P(z)$.

From the solution R of the resulting matrix equation, one can recover (X, U) as $X = h^{-1}(Y - I)$, $U = \begin{bmatrix}I & \Psi\end{bmatrix}$, where

$$\begin{aligned}
&\Psi = -B_{12}B_{22}^{-1} \geq 0,\\
&Y = (C_{11} + \Psi C_{21})(B_{11} + \Psi B_{21})^{-1} \geq 0.
\end{aligned}$$

Again, it is possible to perform the whole algorithm in a subtraction-free way with the help of triplet representations, apart from the subtractions in $\operatorname{diag}(C)$ which are not problematic. A triplet representation for $-X^\top$ can be returned.

6. CONCLUSIONS

This informal description aims to present the main points of the algorithm. A full treatment, including proofs that these formulas work, a full discussion of drifts, location of the eigenvalues and minimality of the solutions, a componentwise stability analysis, and numerical experiments, will be available in a future paper.

7. REFERENCES

[1] M. Agapie and K. Sohraby. Algorithmic solution to second-order fluid flow. In *Proceedings IEEE INFOCOM 2001, The Conference on Computer Communications, Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies, Twenty years into the communications odyssey, Anchorage, Alaska, USA, April 22-26, 2001*, pages 1261–1270, 2001.

[2] T. Betcke and D. Kressner. Perturbation, extraction and refinement of invariant pairs for matrix polynomials. *Linear Algebra Appl.*, 435(3):574–536, 2011.

[3] D. Bini, B. Meini, and S. Steffè. SMCSolver (structured Markov chain solver) 2.1, 2009. Available at http://bezout.dm.unipi.it/SMCSolver/.

[4] D. A. Bini, G. Latouche, and B. Meini. *Numerical Methods for Structured Markov Chains*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2005.

[5] J. Ivanovs. Markov-modulated Brownian motion with two reflecting barriers. *J. Appl. Probab.*, 47(4):1034–1047, 2010.

[6] R. L. Karandikar and V. Kulkarni. Second-order fluid flow models: Reflected Brownian motion in a random environment. *Oper. Res*, 43:77–88, 1995.

[7] P. Kunkel and V. Mehrmann. *Differential-algebraic equations*. EMS Textbooks in Mathematics. European Mathematical Society (EMS), Zürich, 2006. Analysis and numerical solution.

[8] G. T. Nguyen and G. Latouche. The morphing of fluid queues into Markov-modulated Brownian motion. *Stochastic systems*. To appear. Available online at http://arxiv.org/abs/1311.3359.

[9] G. T. Nguyen and F. Poloni. Componentwise accurate fluid queue computations using doubling algorithms. *Numer. Math.*, 130(4):763–792, 2015.

[10] C. A. O'Cinneide. Entrywise perturbation theory and error analysis for Markov chains. *Numer. Math.*, 65(1):109–120, 1993.

[11] L. C. G. Rogers. Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. *Ann. Appl. Probab.*, 4:390–413, 1994.

[12] J. Xue, S. Xu, and R.-C. Li. Accurate solutions of M-matrix algebraic Riccati equations. *Numer. Math.*, 120:671–700, 2012.

On a class of dependent Sparre Andersen risk models with application.

[Extended Abstract] *

Florin Avram
Laboratoire de Mathématiques Appliquées,
University of Pau,
Pau, France.
florin.avram@univ-Pau.fr

Andrei L. Badescu
Department of Statistics,
University of Toronto,
100 St. George Street,
Toronto, Ontario, Canada.
badescu@utstat.toronto.edu

Martijn R. Pistorius
Faculty of Natural Science,
Department of Mathematics,
Imperial College,
London, UK.
m.pistorius@imperial.ac.uk

Landy Rabehasaina
Laboratoire de
Mathématiques, University of
Franche Comté,
16 route de Gray,
25030 Besançon, France.
lrabehas@univ-fcomte.fr

ABSTRACT

In this paper a one-dimensional surplus process is considered with a certain Sparre Andersen type dependence structure under general interclaim times distribution and correlated phase-type claim sizes. The Laplace transform of the time to ruin under such a model is obtained as the solution of a fixed point problem. An efficient algorithm for solving the fixed point problem is derived together with bounds that illustrate the quality of the approximation. A two-dimensional risk model is analyzed under a bailout type strategy with both fixed and variable costs and the proposed dependence structure.

Keywords

Bailout strategy, phase-type distribution, ruin probability, Sparre Andersen dependence structure

1. THE RISK PROCESS

We consider in this paper the following risk process. For a given initial surplus $u \in \mathbb{R}_+$, we denote by $X = \{X(t), t \in \mathbb{R}_+\}$ the insurer's surplus, whose evolution at $t \in \mathbb{R}_+$ is given by

$$X(t) = u + ct - \sum_{i=1}^{N(t)} J_k.$$

The premium rate c is assumed to be strictly positive. We denote by $N(t) = \max\{k \in \mathbb{N} : T_k \leq t\}$ for $t \in \mathbb{R}_+$ the number of claims by time t and we assume independence among each generic pair interclaim time-claim size $\{(T_k, J_k)\}_{k=1}^{\infty}$. Furthermore, we assume that the surplus process $X(t)$ has a Sparre Andersen type dependence structure, defined by

$$P(T_k \in \mathrm{d}t, J_k \in \mathrm{d}x) = \alpha(\mathrm{d}t)\, \mathrm{e}^{Rx}\, \underline{r}\, \mathrm{d}x \qquad t, x \in \mathbb{R}_+, \quad (1)$$

where $\alpha(\mathrm{d}t) \in \mathbb{R}^m$, is a $1 \times m$ distribution vector, R is an $m \times m$ sub-generator matrix, $\underline{r}$ an $m \times 1$ vector given by $\underline{r} = (-R)\underline{1}$, with $\underline{1}$ denoting the $m \times 1$ vector of ones. Note that within each pair interclaim time-claim size the random variables T_k and J_k are dependent, whereas the pairs $\{(T_k, J_k)\}_{k=1}^{\infty}$ are independent and identically distributed (iid) random variables. This dependence structure is a slight generalization of [3], with marginals J_k following a Phase Type distribution with parameters $(R, \alpha(\mathbb{R}_+))$. For this risk model, we assume that he safety loading condition for surplus $\{X_t,\ t \geq 0\}$ is satisfied, i.e. that $cE(T_1) > E(J_1)$. We let τ the time to ruin, defined as $\tau = \inf\{t \geq 0 : X(t) < 0\}$ and $\psi(t, u) = P(\tau < t)$ to be the finite time ruin probability, and denote it's associated Laplace transform by $\hat{\psi}(q, u) = \int_0^\infty q\mathrm{e}^{-qt}\psi(t, u)\, \mathrm{d}t$.

2. EQUATION SATISFIED BY THE LAPLACE TRANSFORM

This section relates $\hat{\psi}(q, u)$ to the solution of a matrix equation. We first extend the definition of the Laplace transform of $\alpha(\mathrm{d}t)$ to matrix arguments:

Definition 1. *For any $m \times m$ negative-definite matrix Q and $1 \times m$ sub-probability vector valued measure $\alpha(dt)$ on*

[*] A full version of this paper is available as *On a class of dependent Sparre Andersen risk models with applications*, available upon request

$\mathbb{R}_+\backslash\{0\}$, we denote by $\hat{\alpha}(Q)$ the $1 \times m$ vector

$$\hat{\alpha}(Q) := \int_0^\infty \alpha(dt)\mathrm{e}^{Qt}.$$

The main result of this section is the following:

Theorem 1. *For $q > 0$ and $u \geq 0$ the Laplace transform of the time to ruin is given by*

$$\hat{\psi}(q,u) \quad = \quad \hat{\rho}(q)\mathrm{e}^{\Gamma(q)u}\underline{1}, \tag{2}$$

where $\hat{\rho}(q)$ is a $1 \times m$ sub-probability vector satisfying the fixed point equation

$$\hat{\rho}(q) \quad = \quad \hat{\alpha}(cR + c\,\underline{r}\,\hat{\rho}(q) - qI), \tag{3}$$

and $\Gamma(q) = R + \underline{r}\,\hat{\rho}(q)$.

If $q = 0$ there exists a $1 \times m$ sub-probability vector $\hat{\rho}(0)$ verifying (3) such that expression (2) holds for $\hat{\psi}(0,u)$.

The above result obtained is new and extend the results obtained under the renewal risk model in [1] in Chapter 9, Theorem 4.4, and the fixed point problem in Proposition 4.3.

3. ALGORITHM FOR THE RUIN PROBABILITY

Determining $\hat{\rho}(q)$ as the solution to (3) turns out not to be practical, as this a non linear multidimensional equation. We propose here an algorithm easily implementable for any distributions $\alpha(\mathrm{d}t)$, provided that the following joint moments

$$M_k(\delta) := \int_0^\infty t^k \mathrm{e}^{-\delta t}\alpha(\mathrm{d}t) \in \mathbb{R}^{1\times m}, \tag{4}$$

are available for all $k \in \mathbb{N}$. We then define

$$\hat{\alpha}^N(Q) = \sum_{k=0}^N M_k(\delta)\frac{(Q+\delta I)^k}{k!}. \tag{5}$$

for $\delta > -\min_{i=1,\dots,m}(Q_{i,i})$. One can easily prove that $\hat{\alpha}^N(Q) \longrightarrow \hat{\alpha}(Q)$ as $N \to \infty$. The goal of this section is to approximate $\hat{\rho}(q)$ by $\hat{\rho}^N(q)$ that satisfies

$$\hat{\rho}^N(q) = \hat{\alpha}^N(cR + c\,\underline{r}\,\hat{\rho}^N(q) - qI) = \hat{\alpha}^N\left(cB(\hat{\rho}^N(q)) - qI\right), \tag{6}$$

$\hat{\rho}^N(q)$ is in turn obtained thanks to the following result:

Proposition 1. *For $q \geq 0$ and*

$$\delta \geq q - c\min_{i=1,\dots,m} R_{ii}, \tag{7}$$

the sequence $(\hat{\rho}_n^N(q))_{n\in\mathbb{N}}$ defined as

$$\left\{\begin{array}{lcl} \hat{\rho}_0^N(q) & = & (0,\dots,0) \\ \hat{\rho}_{n+1}^N(q) & = & \hat{\alpha}^N(cR + c\,\underline{r}\,\hat{\rho}_n^N(q) - qI), \quad n \geq 0, \end{array}\right. \tag{8}$$

converges to a solution of Equation (6) in the set of sub-probability vectors.

Finally, the following theorem justifies that $\hat{\rho}^N(q)$ provides the approximation for the ruin probability starting from 0, namely $\hat{\psi}(q,0)$, and also provides its accuracy for $q \geq q_0$ large enough. Remember that $\hat{\psi}(q,u)$ is also expressed in function of $\hat{\rho}(q)$ for all $u \geq 0$ thanks to (2).

Theorem 2. *For $q > 0$, if $\hat{\rho}^N(q)$ is a solution to the fixed point equation (6), then $\hat{\rho}^N(q)\underline{1}$ converges to $\hat{\psi}(q,0)$, as $N \to \infty$.*

For $q = 0$, if $\hat{\rho}^N(0)$ is the solution to the fixed point equation (6) defined in Proposition 1, then $\hat{\rho}^N(0)\underline{1}$ converges to $\hat{\psi}(0,0)$, as $N \to \infty$.

Finally, for δ and q satisfying conditions (7) and $q > c(e^{-1} + 5||R||)$ the following bound holds

$$\begin{aligned}\left|\hat{\rho}(q) - \hat{\rho}^N(q)\right|_m \quad &\leq \quad \frac{1}{1-C}\sum_{k=N+1}^\infty \frac{|M_k(\delta)|_m}{k!}\delta^k \\ &= \quad \frac{1}{1-C}\left[\hat{\alpha}(0).\underline{1} - \sum_{k=0}^N \frac{|M_k(\delta)|_m}{k!}\delta^k\right]\end{aligned}$$

where constant C is given by $C := \frac{c.e^{-1}||R||}{q-5c||R||} < 1$, and for all $m \in \mathbb{N}\setminus\{0\}$.

4. APPLICATION TO A BAILOUT PROBLEM

Although the dependence structure proposed in Section 1 has a lot of potential for many applications, we propose one dealing with a bailout type model. This model is quite recent and triggered some interesting development, see [2]. In economics, a bailout is an act of loaning or giving capital to a failing business in order to save it from bankruptcy, insolvency, or total liquidation and ruin. To this extent, we consider a main economic unit that replenishes the level of capital of a secondary economic unit when the last one faces financial difficulties. A possible interpretation of the actual problem from an insurance point of view is that the main unit that we generically call *the Central Branch* (CB) infuses capital into the secondary unit referred as *the Subsidiary*, whenever the level of the surplus in the subsidiary drops below level 0.

We start by introducing the bivariate risk model

$$U_i(t) = u_i + c_i\,t - S^{(i)}(t), \;\; S^{(i)}(t) = \sum_{j=1}^{N_i(t)} J_j^{(i)}, \;\; i \in \{0,1\},$$

where $\{N_i(t),\ t \geq 0\}$ is a counting process that describes the claim arrivals, and the claims sizes $J_j^{(i)}, j \geq 1$ are nonnegative i.i.d. random variables with arbitrary marginal distributions given by $F_i(x), i \in \{0,1\}$. We let $U_0(t)$ to describe the surplus of the CB, whereas $U_1(t)$ represents the surplus of the **independent** subsidiary. We assume that the loading condition for the subsidiary is satisfied, namely $c_1 \geq \frac{E[J^1]}{E[T^1]}$, where T^1 represents the generic interclaim random variable. It is assumed that at the ruin instants of the subsidiary, the central branch replenishes the shortfall of the ruined subsidiary back to a zero surplus level. Furthermore, we consider that the transaction cost associated to the generic replenishment amount $\zeta^{(1)}$ corresponding to the subsidiary, is given by $k_1\zeta^{(1)} + K_1$. Consequently, the cost associated to each replenishment $\zeta^{(1)}$ has two components: a "variable" one (that depends on the size of the deficit) introduced via the proportionality constant k_1, and a "fixed" one (for e.g. administration costs) described by the generic variable K_1. For mathematical tractability, K_1 is assumed to be a random variable independent on the replenishment levels $\zeta^{(1)}$. To avoid very complicated scenarios, we assume that the only feasible transactions are to be made from the CB to the subsidiary.

We are then interested in the Laplace transform $\psi_0(q, u_1)$ of the ruin time τ_0 of the Central Branch, and are set to determine this quantity thanks to the approximation procedure described in Section 3. As explained there, the crucial point is to identify, in the present scenario, the structure (1), get closed form expression for the corresponding joint moments (4), then use recursion (8) in order to get the corresponding approximating vector $\hat{\rho}^N(q)$.

5. REFERENCES

[1] S. Asmussen and H. Albrecher. Ruin Probabilities. New Jersey: World Scientific, 2010.

[2] E. Ivanovs and O. Boxma. A bivariate risk model with mutual deficit coverage. Insurance: Mathematics and Economics, 64:126–134, September 2015.

[3] G. Willmot and J. Woo. On the analysis of a general class of dependent risk processes. Insurance: Mathematics and Economics, 51(1):134–141, 2012.

Parisian ruin for Markov additive risk processes with phase-type claims.

[Extended Abstract]

Mogens Bladt
Institute for Applied Mathematics and Systems,
National University of Mexico
A.P. 20-726, 01000 Mexico
D.F., Mexico
bladt@sigma.iimas.unam.mx

Oscar Peralta
Department of Applied Mathematics and
Computer Science, Technical University of
Denmark
DK-2800
Kongens Lyngby, Denmark
osgu@dtu.dk

ABSTRACT

Parisian ruin (proposed in [5]) is defined as the event in which a risk process has an excursion below zero of duration larger than some possibly random $L > 0$. In this paper we apply a probabilistic methodology to calculate the probability of a parisian ruin for a Markov additive risk process with phase-type distributed claims.

CCS Concepts

•**Mathematics of computing** → **Stochastic processes;**

Keywords

parisian ruin; markov additive risk process; phase–type distributions; fluid flow models

1. THE MARKOV ADDITIVE RISK PROCESS

Consider a Markov additive risk process $\{R_t\}_{t\geq 0}$ on the form

$$R_t := u + \int_0^t r_{J_s} ds + \int_0^t \sigma_{J_s} dB_s - \sum_{k=1}^{N_t} U_k,$$

where $u > 0$ is the initial reserve, $\{J_t\}_{t\geq 0}$ is a Markov jump process with intensity matrix $\boldsymbol{\Gamma} = \boldsymbol{C} + \boldsymbol{D}$ and initial distribution $\boldsymbol{\mu}$, $r_i > 0$, $\sigma_i \geq 0$, $\{B_t\}_{t\geq 0}$ is a standard Brownian motion, and $\{N_t\}_{t\geq 0}$ is a Markovian Arrival Process MAP$_m(\boldsymbol{C}, \boldsymbol{D})$. Claim sizes $U_1, U_2, \ldots$ are phase–type distributed with representations $(\boldsymbol{\pi}^{ij}, \boldsymbol{T}^{ij})$ (each one having transient state space E^{ij}), where ij denotes the type of the transition governed by $\boldsymbol{D}$. This model is essentially a spectrally negative Markov modulated Lévy process with phase–type jumps (and no small jumps).

We shall employ a fluid embedding (see [2], [3] and [4]) in order to calculate different probabilities of interest. From here on, we will denote with $\boldsymbol{I}$ the identity matrix, $\boldsymbol{0}$ the matrix filled with 0's, $\boldsymbol{e}$ the column vector filled with 1's, and $\boldsymbol{e}_j^{(m)}$ the column vector with value 1 in its j-th entry and 0 everywhere else, all of them of appropriate dimension. If their dimension needs to be specified, we will indicate it with a superscript. In addition,we will partition the state space of the Markov jump process associated to the fluid model into $E^\sigma := \{i : \sigma_i > 0\}$, $E^+ := \{i : \sigma_i = 0, r_i > 0\}$ and $E^- := \{i : \sigma_i = 0, r_i < 0\}$ and order the space into these three blocks of states. Throughout this manuscript, we will usually use diagonal matrices (indicated by $\boldsymbol{\Delta}$) whose subscript indicates the entries on the diagonal, and whose superscript indicates the block we are referring to; for example, $\boldsymbol{\Delta}_{1/r}^{++} = diag(1/r_i : i \in E^+)$.

To create the fluid embedding of the risk process, we need to consider a fluid flow model (see [1] for more details) with the following characteristics. Let $E^\sigma \cup E^+ \cup E^-$ be the states underlying the fluid flow process, with drifts $\boldsymbol{r} := (r_1, \ldots, r_m, -1, \ldots, -1)$, diffusion parameters $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m, 0, \ldots, 0)$, initial distribution $(\boldsymbol{\mu}, 0, \ldots, 0)$ and intensity matrix given by

$$\boldsymbol{\Lambda} := \begin{pmatrix} \boldsymbol{\Lambda}^{\sigma\sigma} & \boldsymbol{\Lambda}^{\sigma+} & \boldsymbol{\Lambda}^{\sigma-} \\ \boldsymbol{\Lambda}^{+\sigma} & \boldsymbol{\Lambda}^{++} & \boldsymbol{\Lambda}^{+-} \\ \boldsymbol{\Lambda}^{-\sigma} & \boldsymbol{\Lambda}^{-+} & \boldsymbol{\Lambda}^{--} \end{pmatrix}, \tag{1}$$

with

$$\begin{pmatrix} \boldsymbol{\Lambda}^{\sigma\sigma} & \boldsymbol{\Lambda}^{\sigma+} \\ \boldsymbol{\Lambda}^{+\sigma} & \boldsymbol{\Lambda}^{++} \end{pmatrix} = \boldsymbol{C}, \quad \begin{pmatrix} \boldsymbol{\Lambda}^{\sigma-} \\ \boldsymbol{\Lambda}^{+-} \end{pmatrix} = diag(\boldsymbol{H}_i : \sum_j d_{ij} > 0)$$

$$\begin{pmatrix} \boldsymbol{\Lambda}^{-\sigma} & \boldsymbol{\Lambda}^{-+} \end{pmatrix} = col(\boldsymbol{t}^{ij} (\boldsymbol{e}_j^{(m)})' : d_{ij} > 0),$$

$$\boldsymbol{\Lambda}^{--} = diag(\boldsymbol{T}^{ij} : d_{ij} > 0),$$

where $\boldsymbol{H}_i = row(d_{ij}\boldsymbol{\pi}^{ij} : d_{ij} > 0)$ and $\boldsymbol{t}^{ij} = -\boldsymbol{T}^{ij}\boldsymbol{e}$. The next result is a direct consequence of this embedding and [1].

THEOREM 1. *Consider the Markov additive risk process which starts in level $u > 0$ and is either transient (in the sense of [6]) or goes to $+\infty$ a.s.. Then, its probability of* *ruin is given by*

$$\begin{aligned}&\boldsymbol{\mu}\begin{pmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{\alpha}^{(+\sigma)} & \boldsymbol{\alpha}^{(+-)}\end{pmatrix}\exp\left(\begin{pmatrix}\boldsymbol{U}^{(\sigma\sigma)} & \boldsymbol{U}^{(\sigma-)}\\ \boldsymbol{U}^{(-\sigma)} & \boldsymbol{U}^{(--)}\end{pmatrix}u\right)\boldsymbol{e}\\ &\quad=\boldsymbol{\mu}\boldsymbol{P}(u)\boldsymbol{e}\end{aligned}$$

where

$$\begin{aligned}&\left(\left(\boldsymbol{U}^{\sigma\sigma},\boldsymbol{U}^{\sigma-}\right),\boldsymbol{U}^{-\sigma},\boldsymbol{U}^{--},\left(\boldsymbol{\alpha}^{+\sigma},\boldsymbol{\alpha}^{+-}\right)\right)\\ &\quad=\lim_{n\to\infty}f^{(n)}\left(\left(-\boldsymbol{\Delta}_{\eta}^{\sigma\sigma},\boldsymbol{0}\right),\boldsymbol{\Lambda}^{-\sigma},\boldsymbol{\Lambda}^{--},(\boldsymbol{0},\boldsymbol{0})\right),\end{aligned}$$

and $f^{(n)}$ denotes the n-th application of the operator f defined by

$$\begin{aligned}&f\left(\left(\boldsymbol{U}^{\sigma\sigma},\boldsymbol{U}^{\sigma-}\right),\boldsymbol{U}^{-\sigma},\boldsymbol{U}^{--},\left(\boldsymbol{\alpha}^{+\sigma},\boldsymbol{\alpha}^{+-}\right)\right)\\ &\quad=\Bigg(\left(\boldsymbol{\Delta}_{\eta}^{\sigma\sigma},\boldsymbol{0}^{\sigma-}\right)+\sum_{i\in E^{\sigma}}\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{F}_1(\omega_i\boldsymbol{I}-\boldsymbol{U})^{-1},\\ &\qquad\boldsymbol{\Lambda}^{-\sigma}+\boldsymbol{\Lambda}^{-+}\boldsymbol{\alpha}^{+\sigma},\\ &\qquad\boldsymbol{\Lambda}^{--}+\boldsymbol{\Lambda}^{-+}\boldsymbol{\alpha}^{+-},\\ &\qquad\sum_{i\in E^{+}}\boldsymbol{e}_i\boldsymbol{e}_i'\left(f_2(\boldsymbol{\alpha}^{+\sigma},\boldsymbol{\alpha}^{+-})\right)\left(\frac{\lambda_i}{|r_i|}\boldsymbol{I}-\boldsymbol{U}\right)^{-1}\Bigg),\end{aligned}$$

where

$$\boldsymbol{F}_1=\boldsymbol{\Delta}_{2\lambda/\sigma^2}^{\sigma\sigma}\left((\boldsymbol{I}+\boldsymbol{\Delta}_{1/r}^{\sigma\sigma}\boldsymbol{\Lambda}^{\sigma\sigma}),\boldsymbol{\Delta}_{1/r}^{\sigma\sigma}\boldsymbol{\Lambda}^{\sigma-}\right),$$

$$\begin{aligned}&f_2(\boldsymbol{\alpha}^{+\sigma},\boldsymbol{\alpha}^{+-})\\ &\quad=\boldsymbol{\Delta}_{1/r}^{++}\left(\left(\boldsymbol{\Delta}_{\lambda}^{++}+\boldsymbol{\Lambda}^{++}\right)\left(\boldsymbol{\alpha}^{+\sigma},\boldsymbol{\alpha}^{+-}\right)+\left(\boldsymbol{\Lambda}^{+\sigma},\boldsymbol{\Lambda}^{+-}\right)\right),\end{aligned}$$

and

$$\eta_i=\frac{r_i}{\sigma_i^2}+\sqrt{\frac{r_i^2}{\sigma_i^4}+\frac{2\lambda_i}{\sigma_i^2}},\quad\omega_i=-\frac{r_i}{\sigma_i^2}+\sqrt{\frac{r_i^2}{\sigma_i^4}+\frac{2\lambda_i}{\sigma_i^2}},\quad\lambda_i=-\lambda_{ii}.$$

Moreover, in the case the associated MAP is transient, then the probability that the maximum of the risk reserve process reaches level $s>u$ before termination is

$$\boldsymbol{\mu}\exp\left(\begin{pmatrix}\boldsymbol{V}^{(\sigma\sigma)} & \boldsymbol{V}^{(\sigma+)}\\ \boldsymbol{V}^{(+\sigma)} & \boldsymbol{V}^{(++)}\end{pmatrix}(s-u)\right)\boldsymbol{e}$$

where

$$\begin{aligned}&\left(\left(\boldsymbol{V}^{\sigma\sigma},\boldsymbol{V}^{\sigma+}\right),\boldsymbol{V}^{+\sigma},\boldsymbol{V}^{++},\left(\boldsymbol{\beta}^{-\sigma},\boldsymbol{\beta}^{-+}\right)\right)\\ &\quad=\lim_{n\to\infty}g^{(n)}\Big(\left(-\boldsymbol{\Delta}_{\omega}^{\sigma\sigma},\boldsymbol{0}\right),\boldsymbol{\Delta}_{1/r}^{++}\boldsymbol{\Lambda}^{+\sigma},\boldsymbol{\Delta}_{1/r}^{++}\boldsymbol{\Lambda}^{++},(\boldsymbol{0},\boldsymbol{0})\Big),\end{aligned}$$

and $g^{(n)}$ denotes the n-th application of the operator g defined by

$$\begin{aligned}&g\left(\left(\boldsymbol{V}^{\sigma\sigma},\boldsymbol{V}^{\sigma+}\right),\boldsymbol{V}^{+\sigma},\boldsymbol{V}^{++},\left(\boldsymbol{\beta}^{-\sigma},\boldsymbol{\beta}^{-+}\right)\right)\\ &\quad=\Bigg(\left(\boldsymbol{\Delta}_{\omega}^{\sigma\sigma},\boldsymbol{0}^{\sigma+}\right)+\sum_{i\in E^{\sigma}}\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{G}_1(\eta_i\boldsymbol{I}-\boldsymbol{V})^{-1},\\ &\qquad\boldsymbol{\Delta}_{1/r}^{++}\left(\boldsymbol{\Lambda}^{+\sigma}+\boldsymbol{\Lambda}^{+-}\boldsymbol{\beta}^{-\sigma}\right),\\ &\qquad\boldsymbol{\Delta}_{1/r}^{++}\left(\boldsymbol{\Lambda}^{++}+\boldsymbol{\Lambda}^{+-}\boldsymbol{\beta}^{-+}\right),\\ &\qquad\sum_{i\in E^{-}}\boldsymbol{e}_i\boldsymbol{e}_i'\left(g_2(\boldsymbol{\beta}^{-\sigma},\boldsymbol{\beta}^{-+})\right)(\lambda_i\boldsymbol{I}-\boldsymbol{V})^{-1}\Bigg),\end{aligned}$$

where

$$\boldsymbol{G}_1=\boldsymbol{\Delta}_{2\lambda/\sigma^2}^{\sigma\sigma}\left((\boldsymbol{I}+\boldsymbol{\Delta}_{1/r}^{\sigma\sigma}\boldsymbol{\Lambda}^{\sigma\sigma}),\boldsymbol{\Delta}_{1/r}^{\sigma\sigma}\boldsymbol{\Lambda}^{\sigma+}\right),$$

and

$$\begin{aligned}&g_2(\boldsymbol{\beta}^{-\sigma},\boldsymbol{\beta}^{-+})\\ &\quad=\left(\boldsymbol{\Delta}_{\lambda}^{--}+\boldsymbol{\Lambda}^{--}\right)\left(\boldsymbol{\beta}^{-\sigma},\boldsymbol{\beta}^{-+}\right)+\left(\boldsymbol{\Lambda}^{-\sigma},\boldsymbol{\Lambda}^{-+}\right).\end{aligned}$$

2. PARISIAN RUIN

To compute the probability of not getting ruined in a parisian way, we need to calculate the probability that for every $i\geq 1$, the duration of the i-th excursion below zero of $\{R_t\}_{t\geq 0}$ is smaller than L_i, where $\{L_n\}_{n\geq 1}$ is a sequence of i.i.d. phase-type distributed r.v.'s (referred to as parisian clocks) with parameters $(\boldsymbol{\kappa},\boldsymbol{K})$ of dimension ℓ. If there is at least one genuine Brownian component in the Markov additive risk model, then we need to define the ϵ-parisian ruin and ϵ-recoveries in the following way: the risk process undergoes an ϵ-recovery whenever it downcrosses level $-\epsilon<0$ (at which point the parisian clock is started) and is capable of reaching level 0 before the parisian clock rings. If at least one of those ϵ-recoveries is not succesful, then we declare ϵ-parisian ruin.

In order to calculate the probability that once the risk process downcrosses level $-\epsilon$ it will reach level 0 before the parisian clock rings, we kill the underlying MAP at an independent $\mathrm{PH}_\ell(\boldsymbol{\kappa},\boldsymbol{K})$-distributed random time, which according to [6], leads us to the transient MAP process with parameters $(\boldsymbol{C}\oplus\boldsymbol{K},\boldsymbol{D}\otimes\boldsymbol{I})$. Then, by applying the fluid embedding method, we end up working with a terminal fluid flow process with intensity matrix given by

$$\boldsymbol{\Lambda}^*:=\begin{pmatrix}\boldsymbol{\Lambda}^{\sigma\sigma}\oplus\boldsymbol{K} & \boldsymbol{\Lambda}^{\sigma+}\otimes\boldsymbol{I} & \boldsymbol{\Lambda}^{\sigma-}\otimes\boldsymbol{I}\\ \boldsymbol{\Lambda}^{+\sigma}\otimes\boldsymbol{I} & \boldsymbol{\Lambda}^{++}\oplus\boldsymbol{K} & \boldsymbol{\Lambda}^{+-}\otimes\boldsymbol{I}\\ \boldsymbol{\Lambda}^{-\sigma}\otimes\boldsymbol{I} & \boldsymbol{\Lambda}^{-+}\otimes\boldsymbol{I} & \boldsymbol{\Lambda}^{--}\otimes\boldsymbol{I}^{(\ell)}\end{pmatrix},$$

with drifts $\boldsymbol{r}^*:=((r_1,\dots,r_m)\otimes(\boldsymbol{e}^{(\ell)})',-1,\dots,-1)$, diffusion parameters $\boldsymbol{\sigma}^*:=((\sigma_1,\dots,\sigma_m)\otimes(\boldsymbol{e}^{(\ell)})',0,\dots,0)$ and initial distribution $(\boldsymbol{\mu}\otimes(\boldsymbol{e}^{(\ell)})',0,\dots,0)$. This is the very same idea of "freezing time" during downwards movements of a fluid flow process proposed by [2]. Also, notice that the underlying state space of the fluid flow model is an augmented one, where each block's cardinality is ℓ-times bigger than the original one.

This construction implies that the original risk process makes a parisian ϵ-recovery if and only if this associated fluid flow model which starts in level $-\epsilon<0$ ever upcrosses 0. With this in mind, we can state the next theorem.

THEOREM 2. *Suppose that the risk process downcrosses $-\epsilon<0$ according to a probability row-vector $\boldsymbol{\xi}$, with the possible states being $E^-=E^\sigma\cup\left(\cup_{i,j}E^{ij}\right)$. Then, the probability that the process ϵ-recovers is*

$$\begin{aligned}&\boldsymbol{\xi}(\boldsymbol{I}^-\otimes\boldsymbol{\kappa})\begin{pmatrix}\boldsymbol{I}^{\sigma} & \boldsymbol{0}\\ \gamma^{(-\sigma)} & \gamma^{(-+)}\end{pmatrix}\exp\left(\begin{pmatrix}\boldsymbol{W}^{(\sigma\sigma)} & \boldsymbol{W}^{(\sigma+)}\\ \boldsymbol{W}^{(+\sigma)} & \boldsymbol{W}^{(++)}\end{pmatrix}\epsilon\right)\boldsymbol{e}\\ &\quad=\boldsymbol{\xi}\boldsymbol{Q}(\epsilon)\boldsymbol{e},\end{aligned}$$

where $\boldsymbol{\gamma}^{(-\sigma)},\boldsymbol{\gamma}^{(-+)}$ and $\boldsymbol{W}$ can be recursively approximated in the same way $\boldsymbol{\beta}^{(-\sigma)},\boldsymbol{\beta}^{(-+)}$ and $\boldsymbol{V}$ were computed in Corollary 1, with $(\boldsymbol{\Lambda},\boldsymbol{r},\boldsymbol{\sigma})$ replaced with $(\boldsymbol{\Lambda}^*,\boldsymbol{r}^*,\boldsymbol{\sigma}^*)$.

Finally, to compute the ϵ-probability of ruin we just need to construct an adequate Markov chain whose space state contains $E^{\sigma\uparrow}\cup E^+$, $E^{\sigma\downarrow}\cup\left(\cup_{i,j}E^{i,j}\right)$, and two absorbing states Δ_{NR} and Δ_R. The elements in $E^{\sigma\downarrow}\cup\left(\cup_{i,j}E^{i,j}\right)$ are the underlying states in which the risk process can downcross level $-\epsilon$, and $E^{\sigma\uparrow} \cup E^{+}$ are the underlying states in which the killed risk process can upcross level 0. The states Δ_{NR} and Δ_{R} mark the events {No further downcrossings of $-\epsilon$} and {Failed ϵ-recovery}, respectively. The distribution of this Markov chain is characterized by the initial distribution

$$(\mathbf{0}, \boldsymbol{\mu}\boldsymbol{P}(u+\epsilon), 1-\boldsymbol{\mu}\boldsymbol{P}(u+\epsilon)\boldsymbol{e}, 0)$$

and the transition matrix

$$\begin{pmatrix} \mathbf{0} & \boldsymbol{P}(\epsilon) & \boldsymbol{e}-\boldsymbol{P}(\epsilon)\boldsymbol{e} & \mathbf{0} \\ \boldsymbol{Q}(\epsilon)(\boldsymbol{I}\otimes\boldsymbol{e}^{(\ell)}) & \mathbf{0} & \mathbf{0} & \boldsymbol{e}-\boldsymbol{Q}(\epsilon)\boldsymbol{e} \\ \mathbf{0} & \mathbf{0} & 1 & 0 \\ \mathbf{0} & \mathbf{0} & 0 & 1 \end{pmatrix}$$

This way, we have the next result.

Theorem 3. *The probability that a Markov additive risk process gets ruined in an ϵ-parisian way on its n-th downcrossing of $-\epsilon$ is given by*

$$\psi_{\epsilon}^{(n)}(u) = \boldsymbol{\mu}\boldsymbol{P}(u+\epsilon)\left(\boldsymbol{Q}(\epsilon)(\boldsymbol{I}\otimes\boldsymbol{e}^{(\ell)})\boldsymbol{P}(\epsilon)\right)^{n-1}(\boldsymbol{e}-\boldsymbol{Q}(\epsilon)\boldsymbol{e}).$$

Moreover, the probability of an ϵ-parisian ruin occurring at all is given by

$$\psi_{\epsilon}(u) = \boldsymbol{\mu}\boldsymbol{P}(u+\epsilon)\left(\boldsymbol{I}-\boldsymbol{Q}(\epsilon)(\boldsymbol{I}\otimes\boldsymbol{e}^{(\ell)})\boldsymbol{P}(\epsilon)\right)^{-1}(\boldsymbol{e}-\boldsymbol{Q}(\epsilon)\boldsymbol{e}).$$

To compute the 0-parisian probability of ruin (the one that was originally proposed in [5]) for risk processes with genuine Brownian components, we only need to compute the probability of being ϵ_n-parisianly-ruined for some sequence $\epsilon_n \downarrow 0$ and calculate the limit of such a sequence of probabilities. If there are no Brownian components at all, we can compute it directly with the formulae found in the previous theorems by substituting $\epsilon = 0$. Erlangization provides a tool to compute parisian ruin with deterministic parisian clocks.

3. REFERENCES

[1] S. Asmussen. *Stationary distributions via first passage times.* . Theory, Methods, and Open Problems. CRC PressI Llc, 1995.

[2] S. Asmussen and F. Avram. Erlangian approximations for finite-horizon ruin probabilities. *Astin Bulletin*, 2005.

[3] A. Badescu, L. Breuer, A. da Silva Soares, G. Latouche, M.-A. Remiche, and D. Stanford. Risk processes analyzed as fluid queues. *Scandinavian Actuarial Journal*, 2005(2):127–141, Apr. 2005.

[4] L. Breuer. A quintuple law for Markov additive processes with phase-type jumps. *Draft*, pages 1–25, Jan. 2011.

[5] A. Dassios and S. Wu. Parisian ruin with exponential claims. 2008.

[6] G. Latouche, M.-A. Remiche, and P. Taylor. Transient markov arrival processes. *Ann. Appl. Probab.*, 13(2):628–640, 05 2003.

Estimation of discretely observed Markov Jump Processes with applications in survival analysis

Salim Serhan[*]
DTU Compute
Anker Engelunds Vej 1
Kgs. Lyngby, Denmark
sase@dtu.dk

Bo Friis Nielsen[†]
DTU Compute
Anker Engelunds Vej 1
Kgs. Lyngby, Denmark
bfni@dtu.dk

Mogens Bladt[‡]
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 København Ø, Denmark
mogens.bladt@icloud.com

1. INTRODUCTION

Consider a Markovian Arrival Process $\mathrm{MAP}_k(\boldsymbol{\pi}, \mathbf{C}, \mathbf{D})$. Let $X(t)$ be the Markov Jump Process $\{X(t)\}_{t\geq 0}$, with intensity matrix $\mathbf{Q} = \mathbf{C} + \mathbf{D}$ of dimension $k \times k$ and initial probability vector $\boldsymbol{\pi}$, which generates the Markovian Arrival Process. We observe the Markov Jump process at certain discrete time points, as well as at the times of all arrivals. Thus the states have physical interpretations, as opposed to most models involving MAPs. This type of data is frequently encountered in survival analysis in models for long-term disease development. For example, the states could indicate the different phases of a disease development, where the current state of a patient is monitored periodically, but certain acute transitions might occur as well, e.g. heart events or psychiatric relapses. Hence transitions can be hidden or observable. We shall present maximum likelihood methods for estimation, and a somewhat simplified example based on a well-known model of breast cancer.

2. ESTIMATION

The estimation problem can be seen as an incomplete data problem, since we do not observe the complete trajectories of the Markov Jump processes. The current model comprises three different scenarios, two of which have been considered previously. Estimation in phase-type distributions was considered in [1], while [2] dealt with discretely observed Markov Jump Processes. These works employed the EM algorithm and MCMC methods. In survival analysis, similar problems have been treated, by direct optimization of the likelihood function [3]. We employ the EM-algorithm, which makes use of the complete-data likelihood as well. First, we consider the case of complete data. Hence we observe the complete trajectory of the Markov jump process, and the arrivals of the MAP. Figure 1 displays a sample path of a MAP. Without loss of generality, we may assume that the number of independently observed MAPs is one. The complete-data likelihood function is

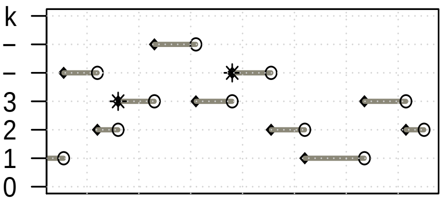


Figure 1: A sample path of MAP with k states. The stars indicate transitions associated with an arrival.

$$\begin{aligned} L(\boldsymbol{\theta}) = &\prod_{i=1}^{k} \pi_i^{b_i} \\ &\cdot \prod_{i=1}^{k} \prod_{j\neq i} c_{ij}^{n_{ij}} \exp(-c_{ij} z_i) \\ &\cdot \prod_{i=1}^{k} \prod_{j=1}^{k} d_{ij}^{\overline{n}_{ij}} \exp(-d_{ij} z_i), \end{aligned}$$

where

- b_i, the number of processes that start in state i,
- z_i, the total time spent in state i,
- n_{ij}, the total number of transitions from state i to state j not associated with an arrival,
- $\overline{n}_{ij}$, the total number of transitions from state i to state j associated with an arrival,

constitutes a sufficient statistic. These parameters are collected in vectors and matrices respectively, as $\mathbf{B} = \{b_i\}_{i=1,..,k}$, $\mathbf{Z} = \{z_i\}_{i=1,..,k}$, $\mathbf{N} = \{n_{ij}\}_{i=1,..,k,j=1,..,k}$ and $\overline{\mathbf{N}} = \{\overline{n}_{ij}\}_{i=1,..,k,j=1,..,k}$. The maximum likelihood estimator is given by

$$\hat{\boldsymbol{\pi}} = \mathbf{B}, \quad \hat{c}_{ij} = \frac{n_{ij}}{z_i}, \quad \hat{d}_{ij} = \frac{\overline{n}_{ij}}{z_i}. \tag{1}$$

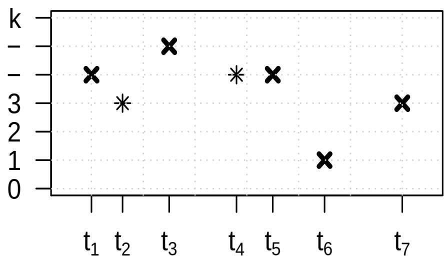


Figure 2: An illustration of the discrete observation sampling scheme. The stars are arrivals while the crosses are discrete observations.

Now, consider the case of incomplete data. We observe a vector of states $(x_{t_1}, x_{t_2}, \ldots, x_{t_n})$, with n being the total number of observations and where $t_1 < t_2 < \ldots < t_n$. We also observe a vector of indicators $(i_1, i_2, \ldots, i_n)$. Here, i_h equals 1 if the h'th observation is an arrival, 0 otherwise. The likelihood function for the discrete data is a product of the transition probabilities, and a closed-form expression for the maximum likelihood estimator is not easily found. As mentioned, we instead use the EM-algorithm to find the maxmimum likelihood estimator. The EM-algorithm requires a specification of the conditional expectations of the sufficient statistics given the incomplete-data. Introduce the following notation.

- $M^k_{ij}(t) = E(Z_k|X(0) = i, X(t) = j)$ = the expected sojourn time in state k, given that the process was initialised in state i and is in state j at time t.
- $f^{kl}_{ij}(t) = E(N_{kl}|X(0) = i, X(t) = j)$ = the expected number of jumps not caused by an event from k to l, given that X was initialised in state i and is in state j after time t.
- $\overline{f}^{kl}_{ij}(t) = E(\overline{N}_{kl}|X(0) = i, X(t) = j)$ = same as for $f^{kl}_{ij}(t)$, but for the number of jumps *caused* by an event.

These variables can been seen as the individual contributions of each observation-interval to the sufficient statistics. The sufficient statistics are then given by

$$\begin{aligned}
E(Z_k|\mathbf{x}) &= M^k_{\cdot x_1}(t_1) + \sum_{h=2}^{n} M^k_{x_{h-1}x_h}(\Delta_h),\\
E(N_{ij}|\mathbf{x}) &= f^{ij}_{\cdot x_1}(t_1) + \sum_{h=2}^{n} f^{ij}_{x_{h-1}x_h}(\Delta_h),\\
E(\overline{N}_{ij}|\mathbf{x}) &= \overline{f}^{ij}_{\cdot x_1}(t_1) + \sum_{h=2}^{n} \overline{f}^{ij}_{x_{h-1}x_h}(\Delta_h)\\
E(B_i|\mathbf{x}) &= E(B_i|X(t_1) = x_1),
\end{aligned}$$

Here, we have separated the contribution of the interval between time $t = 0$ and the first observation, because only a distribution over the states is know at $t = 0$, and therefore this contribution requires a slightly different approach. We write $M^k_{\cdot j}(t)$ instead of $M^k_{ij}(t)$, and similar for f and $\overline{f}$, to indicate that no previous state is known. The remainder is hence to evaluate $M^k_{ij}(t)$, $f^{kl}_{ij}(t)$, $\overline{f}^{kl}_{ij}(t)$ and $E(B_k|X(t_1) = x_1)$.

To shorten the equations that follow, define the matrices

$$\mathbf{M}^{kk'}(t) = \int_0^t \exp(\mathbf{C}u)\mathbf{e}_k\mathbf{e}'_k \exp(\mathbf{Q_0}(t-u))\mathrm{d}u, \qquad (2)$$

$$\mathbf{M}^{kl'}(t) = \int_0^t \exp(\mathbf{C}u)\mathbf{e}_k\mathbf{e}'_l \exp(\mathbf{Q_0}(t-u))\mathrm{d}u. \qquad (3)$$

A fast way to calculate these integrals is with the method described in [5]:

$$\mathbf{M}^{kl'}(t) = \begin{pmatrix} I & \mathbf{0} \end{pmatrix} \exp\left(\begin{bmatrix} \mathbf{C} & \mathbf{e}_k\mathbf{e}'_l \\ \mathbf{0} & \mathbf{C} \end{bmatrix} t\right) \begin{pmatrix} \mathbf{0} \\ I \end{pmatrix}, \qquad (4)$$

where I is the identity matrix of dimension $k \times k$ and $\mathbf{0}$ is a matrix of zeroes of dimension $k \times k$. The contribution of the first interval is

$$M^k_{\cdot j}(t) = \frac{\boldsymbol{\pi}\mathbf{M}^{kk'}\mathbf{D}^{i_1}\mathbf{e}_j}{\boldsymbol{\pi}\exp(\mathbf{C}t)\mathbf{D}^{i_1}\mathbf{e}_j}, \quad f^{kl}_{\cdot j}(t) = q_{0,kl}\frac{\boldsymbol{\pi}\mathbf{M}^{kl'}\mathbf{D}^{i_1}\mathbf{e}_j}{\boldsymbol{\pi}\exp(\mathbf{C}t)\mathbf{D}^{i_1}\mathbf{e}_j},$$

$$\overline{f}^{kl}_{\cdot j} = 0, \quad E(B_i|X(t_1)) = \frac{\pi_i\mathbf{e}'_i\exp(\mathbf{C}t_1)\mathbf{D}^{i_1}\mathbf{e}_{x_{t_1}}}{\pi\exp(\mathbf{C}t_1)\mathbf{D}^{i_1}\mathbf{e}_{x_{t_1}}}.$$

If the h'th observation is caused by an arrival, the contribution of the interval is

$$M^k_{ij}(t) = \frac{\mathbf{e}_i\mathbf{M}^{kk'}\mathbf{D}\mathbf{e}_j}{\mathbf{e}_i\exp(\mathbf{C}t)\mathbf{D}\mathbf{e}_j}, \quad f^{kl}_{ij}(t) = q_{0,kl}\frac{\mathbf{e}_i\mathbf{M}^{kl'}\mathbf{D}\mathbf{e}_j}{\mathbf{e}_i\exp(\mathbf{C}t)\mathbf{D}\mathbf{e}_j},$$

$$\overline{f}^{kl}_{ij} = 0 \text{ for } l \neq j, \quad \overline{f}^{kl}_{ij} = \frac{\mathbf{e}_i\exp(\mathbf{C}t)\mathbf{e}_k q_{1,kj}}{\mathbf{e}_i\exp(\mathbf{C}t)\mathbf{D}\mathbf{e}_j} \text{ for } l = j.$$

If the h'th observation is caused by a discrete observation, the contribution of the interval is

$$M^k_{ij}(t) = \frac{\mathbf{e}_i\mathbf{M}^{kk'}\mathbf{e}_j}{\mathbf{e}_i\exp(\mathbf{C}t)\mathbf{e}_j}, \quad f^{kl}_{ij}(t) = q_{0,kl}\frac{\mathbf{e}_i\mathbf{M}^{kl'}\mathbf{D}\mathbf{e}_j}{\mathbf{e}_i\exp(\mathbf{C}t)\mathbf{e}_j}$$

$$\overline{f}^{kl}_{ij} = 0.$$

3. AN APPLICATION IN SURVIVAL ANALYSIS

While usually not identified as such, many so-called multi-state models used in survival analysis can be seen as MAPs. One such model is used to analyse late-term effects of breast cancer surgery. [4] This model has 5 states:

State 1: Post-surgery. The patient is in this state following the surgery.

State 2: Local reccurrence. The tumor reappears in the vicinity of the operated tumor.

State 3: Distant metastatis. A tumor appears at a distant location from the original tumor.

State 4: Local reccurence and distant metastasis. A tumor has occured both in vicinity of and distant to the original tumor. This can happen in any order, although local occurence happening first is most common.

State 5: Dead. The absorbing state.

Transitions into state 2, 3 and 4 are not observed when they happen, but only at screenings in the doctor's office. The time at which transitions into the state *death* occur is however known very precisely, and we label such transitions as events. Since the initial state is known, estimating the initial distribution $\boldsymbol{\pi}$ is not relevant for this example. The matrix structure, with some chosen values, is illustrated below.

$$\mathbf{C} = \begin{bmatrix} - & 0.3 & 0.1 & 0 & 0 \\ 0 & - & 0 & 0.3 & 0 \\ 0 & 0 & - & 0.25 & 0 \\ 0 & 0 & 0 & - & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix} - & 0 & 0 & 0 & 0.025 \\ 0 & - & 0 & 0 & 0.04 \\ 0 & 0 & - & 0 & 0.03 \\ 0 & 0 & 0 & - & 0.1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

We simulated 100 women until absorption occurred, and discrete observations were taken every 5 years. The model was then fitted using the EM algorithm to 3-decimal points. Estimating for multiple series amounts to summing the sufficient statistics for each series, while using the same maximum likelihood estimator (The $\boldsymbol{\pi}$-vector would need to be scaled by the number of series). The estimated values are shown below.

$$\hat{\mathbf{C}} \begin{bmatrix} - & 0.2 & 0.118 & 0 & 0 \\ 0 & - & 0 & 0.342 & 0 \\ 0 & 0 & - & 0.419 & 0 \\ 0 & 0 & 0 & - & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\hat{\mathbf{D}} \begin{bmatrix} - & 0 & 0 & 0 & 0.030 \\ 0 & - & 0 & 0 & 0.035 \\ 0 & 0 & - & 0 & 0.004 \\ 0 & 0 & 0 & - & 0.104 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

And these appear reasonably close to the chosen values.

4. REFERENCES

[1] S. Asmussen, O. Nerman, and M. Olsson. Fitting Phase-type Distributions via the EM Algorithm. *Scandinavian Journal of Statistics*, 23(4):419–441, 1996.

[2] M. Bladt and M. Sørensen. Statistical inference for discretely observed markov jump processes. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 37(3):395–410, 2005.

[3] C. H. Jackson et al. Multi-state models for panel data: the msm package for r. *Journal of Statistical Software*, 38(8):1–29, 2011.

[4] H. Putter, J. van der Hage, G. H. de Bock, R. Elgalta, and C. J. van de Velde. Estimation and prediction in a multi-state model for breast cancer. *Biometrical journal*, 48(3):366–380, 2006.

[5] C. Van Loan. Computing integrals involving the matrix exponential. Technical report, Cornell University, 1977.

Dependence patterns related to the BMAP

[Extended Abstract]

Rosa Lillo
Department of Statistics
Universidad Carlos III de Madrid
Spain
rosaelvira.lillo@uc3m.es

Joanna Rodriguez
Department of Statistics
Universidad Carlos III de Madrid
Spain
jvrcesar@est-econ.uc3m.es

Pepa Ramírez-Cobo
Department of Statistics and Operational Research
Universidad de Cádiz
Spain
pepa.ramirez@uca.es

ABSTRACT

It is well known that the Batch Markovian Arrival Process permits dependent inter-event times and batch sizes. In this work, the characterization of the dependence structure related to this model is analyzed for the general stationary BMAPm(k), with event occurrences up to size k. In the case of two states, it is proven that both auto-correlation functions (inter-event times and batch sizes) decrease geometrically as the time lag increases. More rich patterns can be found when more than two states are considered in the embedded Markov process. It is also shown how the dependence associated to the model affects to the quantities and distributions that describe the reliability of the process.

Keywords

Batch Markovian Arrival Process; Dependence structures; Auto-correlations functions.

On Order Statistics of Matrix–Exponential distributions

[Extended Abstract]

Azucena Campillo N.
Technical University of Denmark
Richard Petersens Plads, Building 324, DTU
Lyngby, Denmark
azca@dtu.dk

Bo Friis Nielsen
Technical University of Denmark
Richard Petersens Plads, Building 321, DTU
Lyngby, Denmark
bfni@dtu.dk

ABSTRACT

Order statistics play an important role in many areas of probability and statistics such as reliability (waiting for the k–th failure before a system breaks down), robust statistics (replace mean estimator by mean central order statistic) and tail prediction (how does order statistics increase and how we can predict yet unobserved higher orders). For a few more examples and details we refer to [2] and [3].

Much of the literature of order statistics is concerned with the (numerical) calculation of (possibly higher–order) moments and cross–moments, and in particular the more general models with independent non–identically distributed random variables (i.n.i.d.) has turned out to be rather challenging (see e.g. [4] or [2]).

We will consider the class of matrix–exponential distributions, which is an extension of the class of phase–type distributions with a rational Laplace transform. The class of phase–type distributions is dense in the class of distributions on the positive real axis (this means that any non–negative distribution may be approximated arbitrarily close by a phase–type distribution), and they are well known for their probabilistic attractive properties which allows for explicit and exact solutions in even complex stochastic models.

We work with the case when the order statistics are both identically and non–identically, independently distributed random variables having either ME or PH distributions, and we derive expressions for both marginal and joint distributions respectively. While it has been mentioned in several places in the literature that order statistics of phase–type distributions are again of phase–type, it has not been possible for us to retrieve any particular representation other than for the cases of the minimum and maximum. We shall provide specific representations for any order statistic of matrix–exponential distributed random variables, and consequently for the case of phase–type distributed random variables as well. Also we shall present formulas for calculating joint distributions and higher order cross moments which can be calculated in an efficient and numerically stable way by standard methods involving matrix inversions. At the end, we will be able to calculate fractional moments of order statistics by using a recent method of functional calculus, which in turn will provide an explicit formula for their Mellin transform (see [5] for details).

Previous work on order statistics within the classes of phase–type and matrix–exponential distributions seems to be limited to the papers by [1] and [8]. The two papers are somewhat related in that they use similar recursive methods for calculating the Laplace transform and moments respectively, however, neither of the papers appear to take advantage of the probabilistic interpretation of the phase–type distributions.

1. ORDER STATISTICS OF MATRIX–EXPONENTIAL DISTRIBUTIONS

Let $X_1, ..., X_n$ be independent random variables with $X_i \sim \mathrm{ME}_{p_i}(\boldsymbol{\alpha}_i, \boldsymbol{S}_i, \mathbf{s}_i)$, $i = 1, 2, ..., n$. We assume that the representations $\mathrm{ME}_{p_i}(\boldsymbol{\alpha}_i, \boldsymbol{S}_i, \mathbf{s}_i)$ are such that $\mathbf{s}_i = -\boldsymbol{S}_i\mathbf{e}$. We write $\mathrm{ME}(\boldsymbol{\alpha}_i, \boldsymbol{S}_i)$. This assumption is not essential but simplifies the derivations and the expressions.

Define the block diagonal matrices

$$\boldsymbol{S}_{(k)} = \mathrm{diag}\left(\left(\boldsymbol{S}_{j_1} \oplus \cdots \oplus \boldsymbol{S}_{j_{n-k}}\right)_{(j_1,\ldots,j_{n-k}) \in \mathcal{I}(n:k)}\right), \quad (1)$$

where $\mathcal{I}(n:k)$ denotes the set of $n!/((n-k+1)!(k-1)!)$ lexicographically ordered $n-k$ tuples $(j_1, ..., j_{n-k})$ such that $j_1 < j_2 < \cdots < j_{n-k}$.

Theorem 1. *Let $X_1, ..., X_n$ be matrix–exponential distributed random variables with representation $ME(\boldsymbol{\alpha}_i, \boldsymbol{S}_i), i = 1, \ldots, n$. Then the k-th order statistic $X_{(k)}$, with $1 \le k \le n$, has a matrix–exponential distribution with representation $ME(\boldsymbol{\pi}_{(k)}, \boldsymbol{T}_{(k)})$, where $\boldsymbol{\pi}_{(k)} = (\boldsymbol{\alpha}_1 \otimes \cdots \otimes \boldsymbol{\alpha}_n, \mathbf{0}, ..., \mathbf{0})$ ($k-1$ blocks of zeros of appropriate dimension) and*

$$\boldsymbol{T}_{(k)} = \begin{pmatrix} \boldsymbol{S}_{(1)} & \boldsymbol{S}^0_{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{(2)} & \boldsymbol{S}^0_{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{S}_{(k)} \end{pmatrix}.$$

The matrices $\boldsymbol{S}^0_{(j)}$ are uniquely determined in terms of the ordering of $\boldsymbol{S}_{(j)}$ and $\boldsymbol{S}_{(j+1)}$. For example, in the case when

$n = 3$, we have that

$$\begin{aligned}
\boldsymbol{S}_{(1)} &= \boldsymbol{S}_1 \oplus \boldsymbol{S}_2 \oplus \boldsymbol{S}_3, \\
\boldsymbol{S}^0_{(1)} &= \left(\boldsymbol{I} \otimes \boldsymbol{I} \otimes \mathbf{s}_3, \boldsymbol{I} \otimes \boldsymbol{s}_2 \otimes \boldsymbol{I}, \mathbf{s}_1 \otimes \boldsymbol{I} \otimes \boldsymbol{I}\right), \\
\boldsymbol{S}_{(2)} &= \begin{pmatrix} \boldsymbol{S}_1 \oplus \boldsymbol{S}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_1 \oplus \boldsymbol{S}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{S}_2 \oplus \boldsymbol{S}_3 \end{pmatrix}, \\
\boldsymbol{S}^0_{(2)} &= \begin{pmatrix} \boldsymbol{I} \otimes \mathbf{s}_2 & \mathbf{s}_1 \otimes \boldsymbol{I} & \mathbf{0} \\ \boldsymbol{I} \otimes \mathbf{s}_3 & \mathbf{0} & \mathbf{s}_1 \otimes \boldsymbol{I} \\ \mathbf{0} & \boldsymbol{I} \otimes \mathbf{s}_3 & \mathbf{s}_2 \otimes \boldsymbol{I} \end{pmatrix}, \\
\boldsymbol{S}_{(3)} &= \begin{pmatrix} \boldsymbol{S}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{S}_3 \end{pmatrix}.
\end{aligned}$$

Then

$$\boldsymbol{T}_{(1)} = \boldsymbol{S}_1 \oplus \boldsymbol{S}_2 \oplus \boldsymbol{S}_3, \tag{2}$$

$$\boldsymbol{T}_{(2)} = \begin{pmatrix} \boldsymbol{S}_{(1)} & \boldsymbol{S}^0_{(1)} \\ \mathbf{0} & \boldsymbol{S}_{(2)} \end{pmatrix}, \tag{3}$$

$$\boldsymbol{T}_{(3)} = \begin{pmatrix} \boldsymbol{S}_{(1)} & \boldsymbol{S}^0_{(1)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{(2)} & \boldsymbol{S}^0_{(2)} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{S}_{(3)} \end{pmatrix}. \tag{4}$$

We conclude that

$$X_{(i)} \sim \mathrm{ME}\left(\boldsymbol{\pi}_{(i)}, \boldsymbol{T}_{(i)}\right), \quad i = 1, 2, 3,$$

where $\boldsymbol{T}_{(1)}$, $\boldsymbol{T}_{(2)}$ and $\boldsymbol{T}_{(3)}$ are given by (2), (3) and (4) respectively, and , $\boldsymbol{\pi}_{(1)} = \boldsymbol{\alpha}_1 \otimes \boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_3$, $\boldsymbol{\pi}_{(2)} = (\boldsymbol{\alpha}_1 \otimes \boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_3, \mathbf{0})$ and $\boldsymbol{\pi}_{(3)} = (\boldsymbol{\alpha}_1 \otimes \boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_3, \mathbf{0}, \mathbf{0})$.

Next we consider the joint distribution of $X_{(r)}$ and $X_{(u)}$.

Theorem 2. *If $X_1, ..., X_n$ are independent and $X_i \sim$ ME($\boldsymbol{\alpha}_i \boldsymbol{S}_i$), then the distribution of $(X_{(r)}, X_{(u)})$ is given by the equation* (5)*:*

$$f_{(r,u)}(x,y) = \left(\bigotimes_{i=1}^{n} \boldsymbol{\alpha}_i, \mathbf{0}\right) \exp\left(\begin{pmatrix} \boldsymbol{S}_1 & \boldsymbol{S}^0_1 & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{S}_r \end{pmatrix} x\right) \times \begin{pmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{S}^0_r & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} \exp\left(\begin{pmatrix} \boldsymbol{S}_{r+1} & \boldsymbol{S}^0_{r+1} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{r+2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{S}_u \end{pmatrix} (y-x)\right) \begin{pmatrix} \mathbf{0} \\ \boldsymbol{S}^0_u \mathbf{e} \end{pmatrix} \tag{5}$$

The joint distributions of order statistics of matrix exponential and phase type distributions are of the MME* and MPH* types respectively, see [6]. The natural representation uses the matrix corresponding to the highest order statistic involved and a reward matrix R with columns of ones and zeros. The simplest example is the joint distribution of the maximum and minimum of two variables. This representation will be $((\boldsymbol{\alpha}_1 \otimes \boldsymbol{\alpha}_2, \mathbf{0}, \mathbf{0}), \boldsymbol{T}_{(2)}, \boldsymbol{R})$, where $\boldsymbol{R}$ is given by

$$\boldsymbol{R} = \begin{pmatrix} \mathbf{e} & \mathbf{e} \\ \mathbf{e} & \mathbf{0} \\ \mathbf{e} & \mathbf{0} \end{pmatrix}.$$

The first column of the matrix $\boldsymbol{R}$ refers to the maximum of the variables while the second column refers to the minimum.

The formulation of joint order statistics as MME* or MPH* distributions leads to stable formulas for the calculation of their cross moments. The cross moments can be calculated using the following theorem, which is a reformulation of a Theorem 4.2 from [7].

Theorem 3. *Let $\mathbf{Y} \sim$ MME*$(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{R})$. Then with $\boldsymbol{U} = -\boldsymbol{S}^{-1}$ and $\boldsymbol{W}_i = \boldsymbol{U}\boldsymbol{\Delta}(\mathbf{R}\mathbf{e}_i)$ we have that*

$$\mathbb{E}\left(\prod_{j=1}^{n} Y_j^{h_j}\right) = \boldsymbol{\alpha} \; mper \begin{vmatrix} \boldsymbol{W}_1 & \boldsymbol{W}_2 & \boldsymbol{W}_3 & \cdots & \boldsymbol{W}_n \\ \boldsymbol{W}_1 & \boldsymbol{W}_2 & \boldsymbol{W}_3 & \cdots & \boldsymbol{W}_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \underbrace{\boldsymbol{W}_1}_{h_1} & \underbrace{\boldsymbol{W}_2}_{h_2} & \underbrace{\boldsymbol{W}_3}_{h_3} & \cdots & \underbrace{\boldsymbol{W}_n}_{h_n} \end{vmatrix} \mathbf{e},$$

where mper is like the permanent but with entrances being square matrices of the same dimension. The mper is calculated as the usual permanent in terms of sums of products of matrices taking into account the possible non–commutativity and expanded by the first block row only. The notation of $\boldsymbol{\Delta}(\boldsymbol{A})$ refers to the diagonal of the matrix $\boldsymbol{A}$.

The matrix–exponential methodology allows us to calculate moments, cross moments and Mellin transforms (fractional moments) in an efficient and numerically stable way for reasonably sized matrices $\boldsymbol{T}_{(k)}$. The computationally demanding operations in terms of time consumption and memory allocation are matrix inversion for the case of moments of integer order, and matrix–logarithm and exponentials for general fractional moments. The exit rate vector corresponding to $\boldsymbol{T}_{(i)}$ is given by $\mathbf{t}_{(i)} = -\boldsymbol{T}_{(i)}\mathbf{e}$. By calculating $\boldsymbol{U}_{(i)} = -\boldsymbol{T}^{-1}_{(i)}$, we then obtain Mellin transforms for the $X_{(i)}$'s by

$$M_{X_{(i)}}(\alpha) = \mathbb{E}\left(X^{\alpha-1}_{(i)}\right) = \Gamma(\alpha)\; \boldsymbol{\pi}_{(i)} \boldsymbol{U}^{\alpha}_{(i)} \mathbf{t}_{(i)}.$$

The fractional moments of $\boldsymbol{U}_{(i)}$ can be obtained using the method of functional calculus, see [5].

2. REFERENCES

[1] Y. H. Abdelkader. A Laplace transform method for order statistics from nonidentical random variables and its application in Phase-type distribution. *Statistics & Probability Letters*, 81(8):1143–1149, Aug. 2011.

[2] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A First Course in Order Statistics.* SIAM, 1992.

[3] D. Assaf and B. Levikson. Closure of phase type distributions under operations arising in reliability theory. *The Annals of Probability*, 1982.

[4] N. Balakrishnan. Permanents, order statistics, outliers, and robustness. *Revista matemática complutense*, 20(1):7–107, Feb. 2007.

[5] M. Bladt, A. Campillo Navarro, and B. Nielsen. *On the use of functional calculus for phase-type and related distributions.* DTU Compute-Technical Report-2014. Technical University of Denmark (DTU), 2014.

[6] M. Bladt and B. F. Nielsen. Multivariable Matrix–exponential distributions. *Stochastic Models*, 2010.

[7] M. Bladt and B. F. Nielsen. Multivariate Matrix-Exponential Distributions. *Stochastic Models*, 26(1):1–26, Feb. 2010.

[8] X. Zhang and Z. Hou. Computing the moments of order statistics from nonidentically distributed phase-type random variables. *Journal of Computational and Applied Mathematics*, 235(9):2897–2903, Mar. 2011.

A Stochastic EM algorithm for construction of Mortality Tables

Full paper

Luz Judith R. Esparza
Facultad de Ciencias Exactas
Universidad Juarez del Estado de Durango
Av. Veterinaria 210, Valle del Sur, 34120
Durango, Dgo. Mexico.
ljre@ujed.mx

F. Baltazar-Larios
Facultad de Ciencias
Universidad Nacional Autónoma de México
A.P. 20-726
01000 México, CDMX
fernandobaltazar@ciencias.unam.mx

ABSTRACT

In this paper we propose to use the concept of physiological age to modelling the process of aging by using phase-type distributions to calculate the probability of death.

We propose a finite-state Markov jump process to model the hypothetical aging process. Since the Markov process has only a single absorbing state (the death state), the death time (the absorbing time) follows a phase-type distribution. We assume an specific evolution for the aging process that determine the structure of the intensity matrix of the Markov process. Thus, to build a mortality table the challenge is to estimate this matrix based on the records of the aging process.

Considering the nature of the data, we consider two cases: first, having continuous time information of the aging process (hypothetical case), and the more interesting case, where we have reports of the process just in determined times.

If the aging process is only observed at discrete time points we have a missing data problem, thus, we use a stochastic EM algorithm for finding the maximum likelihood estimator of the intensity matrix. And in order to do that, we build Markov bridges which are sampled using the Bisection method.

The theory is illustrated by a simulation study and finally we used our model to fit real data.

Keywords

Mortality, Physiological Age, Phase-type distributions, Stochastic EM algorithm, Bisection, Markov Bridges.

1. INTRODUCTION

Throughout the history of mankind the study and measurement of mortality risk has been very important. Nowadays it is essential to have a way to measure this in the valuation of insurance, contingent rentals, and management of social security systems and pensions. Since the solvency and financial stability of institutions depends, among other things, of the availability of suitable tools that reflect a proper measurement of the sinister rate, they will face possible deviations. Moreover, mortality rates play an important role in calculus of risk premium and risk reserves.

Studies on the behavior of mortality in humans is traced back to the early seventeenth century. In 1693 E. Halley ([14]) built the first annuity table based on the lifespan. On another hand, A. de Moivre ([24]) in 1725 proposed the first mathematical mortality model, and in 1817, E. Sang ([11]) proposed a model based on that the number of survivors in a group must decrease geometrically together with a different age factor.

However, it was not until 1825 that B. Gompertz ([13]) proposed that a law of geometric progression pervades in mortality after a certain age. He obtained the following expression known as *the Gompertz equation*:

$$\mu_x = \alpha e^{\beta x},$$

where μ_x denotes the mortality rate at age x, and α and β are constants. This equation represents a force of mortality that increases progressively with the age in such a manner that $\log \mu_x$ grows linearly. In this equation α is known as the baseline mortality and the term β describes the actuarial aging rate.

In 1860, W. M. Makeham ([22]) extended the Gompertz model by adding a constant:

$$\mu_x = \sigma + \alpha e^{\beta x},$$

where σ represents all random factors with no willingness to death, for example accidents, epidemics, etc. This model is known as *Gompertz-Makeham law of mortality.*

In 1980, Heligman and Pollard ([15]) proposed the following formula that fits Australian mortality rates fairly well at all ages:

$$\frac{q_x}{1-q_x} = A^{(x+\alpha)^\beta} + D\exp\left[-E\left\{\log\left(\frac{x}{F}\right)\right\}^2\right] + GH^x,$$

where q_x is the probability that a person at aged x will die within a year. This model reflects the exponential pattern of mortality at adult ages, the hump at age 23 that is found in many mortality tables, and the fall in mortality during childhood.

Currently, the actuaries use mortality tables as a basic tool to describe the mortality through an age structure.

However, a mortality table is only an option when there is no mathematical law available. Several factors can alter this probability, the more considered factor is the age but there are other important characteristics such as sex, clinical history, smoking, age policy, etc. Actuaries have used two mortality models widely, the Heligman-Pollard model and Lee-Carter model ([21]), however, in these models, we cannot determine the distribution of the time of death explicitly. In this paper we will study a model in which the time until death is explicit and has a phase-type distribution.

Sheldon and Xiaoming [29] used phase-type distributions to model human mortality. In their model a health index called physiological age was introduced and modeled by a Markov process, however, they only considered the possibility to go to the next state with certain parameters, estimating those using the simplex algorithm. In this work, we will consider a more general case for the form of the intensity matrix for modeling a hypothetical aging process, giving another interpretation of the states (physiological ages), and using a stochastic EM algorithm for the estimation part when the Markov process associated with the phase-type distribution is observed at discrete times. The advantage of using this model is to have a closed-form expression for calculating premiums of all life insurance based on a model that considers smooth transitions in the human's aging process.

The rest of the paper is organized as follows. In section 2 we introduce the physiological age concept. As background of phase-type distributions in section 3 we discuss some analytic aspects of these distributions. In section 4 we propose our model. In section 5 we discuss the estimation for the continuous and discrete cases. Finally in section 6 we used our model to fit real data. The paper is concluded in section 7.

2. PHYSIOLOGICAL AGE

The properties and functional capacities of the body are changing as the accumulated time increases from birth. We mean by *aging process* to simultaneous degradation of multiple organ systems. In [18] the aging process is defined as "the progressive, and essentially irreversible diminution with the passage of time of the ability of an organism or one of its parts to adapt to its environment, manifested as diminution of its capacity to withstand the stresses to which it is subjected and culminating in the death of the organism".

To model the aging process, we consider the concept of physiological age, which can be interpreted as relative health index representing the degree of aging on the individual. Physiological age has been studied in [19] where they proposed a linear model of aging, which allowed a latent adjustment to be made to an individual's chronological age to give their physiological age. In [30] for example, studied the relation of the physiological indicators of organisms in a population to the age specific population mortality rate.

Physiological age is a relative index. Since the time spent at a certain physiological age can be considered within a unbounded time interval, it is enough to know the current physiological age of the individual (regardless of its evolution in the past) to determine its future evolution, i.e., this process presents a Markovian behavior, so is natural assume that the time spent in each state (physiological age) has an exponential behavior. Assuming that during the time that individuals remains in this state, their ability to adapt to their environment are the same, and when they move to another physiological age it is because these capacities were diminished. There are different factors affecting the deterioration of these capacities, the individuals may suffer an incident that make they pass to any physiological age where their abilities are less than those of the current physiological age.

Determine the age in which human life is segregated is not an easy task, since it is necessary to have information on the health status of the population under study.

3. PHASE-TYPE DISTRIBUTIONS

Let consider a Markov jump process $\{X_t\}_{t\geq 0}$ with state space $E = \{1, 2, \dots, n, n+1\}$, where the states $1, 2, \dots, n$ are transient, and the state $n+1$ is an absorbing one. Given these conditions the infinitesimal generator of the process can be written as follows

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{r} \\ \bar{0} & 0 \end{pmatrix} \tag{1}$$

where $\boldsymbol{Q}$ is a square matrix of dimension n, $\boldsymbol{r}$ a column vector of dimension n, and $\bar{0}$ is a row vector of dimension n with all its entries zero. $\boldsymbol{Q}$ is called the phase-type generator and $\boldsymbol{r}$ is called the exit vector since it contains the rates by which exit to the absorbing state takes place. Since the rows in an intensity matrix must sum 0, we also have that $\boldsymbol{r} = -\boldsymbol{Q}\boldsymbol{e}$, where $\boldsymbol{e}$ is a vector of dimension n with all of its entries one.

Let $\boldsymbol{\alpha}$ be the initial distribution associated to the process, i.e. $\mathbb{P}(X_0 = i)$. We assume that $\sum_{i=1}^{n} \alpha_i = 1$, this means that the process cannot start in the absorbing state.

Consider a random variable τ which models the time it takes for the process to reach the absorbing state $n+1$, i.e.:

$$\tau = \inf\{t \geq 0; X_t = n+1\}.$$

Note that the distribution of τ only depends on $\boldsymbol{\alpha}$ and $\boldsymbol{Q}$, since $\boldsymbol{r}$ is given in terms of $\boldsymbol{Q}$.

Definition 3.1. *Phase-type (PH) distributions with representation $(\boldsymbol{\alpha}, \boldsymbol{Q})$ is the distribution of a stopping time τ for a Markov jump process $\{X_t\}_{t\geq 0}$ on state space $E = \{1, 2, \dots, n, n+1\}$ with an absorbing state $n+1$. We denote this by $\tau \sim PH(\boldsymbol{\alpha}, \boldsymbol{Q})$.*

PH distributions were considered first by M. Neuts ([25, 26]).

The transition probability matrix of the Markov jump process at time $t \geq 0$ is given by

$$P(t) = \exp(t\boldsymbol{\Lambda}) = \begin{pmatrix} \exp(t\boldsymbol{Q}) & \boldsymbol{e} - \exp(t\boldsymbol{Q})\boldsymbol{e} \\ \bar{0} & 1 \end{pmatrix}. \tag{2}$$

Some of the properties of the PH distributions are below. Consider $\tau \sim PH(\boldsymbol{\alpha}, \boldsymbol{Q})$, then for all $t \geq 0$:

1. The survival function of τ is given by $\mathbb{P}(\tau > t) = \boldsymbol{\alpha} \exp(t\boldsymbol{Q})\boldsymbol{e}$.
2. The density of τ is $f_\tau(t) = \boldsymbol{\alpha} \exp(t\boldsymbol{Q})\boldsymbol{r}$.
3. The i-th moment of τ is given by
$$\mathbb{E}(\tau^i) = i!(-1)^i \boldsymbol{\alpha}(\boldsymbol{Q}^{-1})^i \boldsymbol{e}.$$

The demonstration of these properties can be found in [5].

Given the characteristics of the aging process (presented in the previous section) it is natural to consider that this is a Markov jump process with the death considered as an absorbing state and the remaining life time following a PH distribution.

Since mortality modeling is important for pension plans and annuity business, the use of parametric models to extrapolate the past into the future has changed over the years due to such as parameters have to have information of the aging process. The new models apart from adjusting data effectively show the connection between physiological age given by experts and the aging process. Markov processes and PH distributions have been used to model human mortality (e.g. see [1]). Reference [29] has shown that using PH distributions to model mortality data with the properties before mentioned is satisfactory, however, its model has an special structure of the transition matrix that in this work will be extended.

4. MODEL AND DATA

Let consider a finite-state Markov jump process (time-homogeneous) to model the hypothetical aging process. Assuming that the life of a person can be segregated into n physiological ages and death is represented by the state $n+1$, the state space of the Markov process associated to the aging process is $E = \{1, 2, \dots, n, n+1\}$. Since the Markov process has only a single absorbing state (the death state), the death time (the absorbing time) follows a PH distribution.

Seeing that we want a model to explain more precisely human mortality, we propose the following evolution for the aging process. If a person has physiological age i ($i \in \{1, 2 \dots, n\}$) we assume that the time spent at this age is exponentially distributed with parameter λ_i ($0 < \lambda_i < \infty$). Also, for $i \in \{1, \dots, n-1\}$, we consider three possible cases (additionally of the possibility of continue in the current state):

1. The person presents a natural development of the aging process, in this case, the person eventually transits to the next physiological age $i+1$. We suppose that the intensity of this transition is given by $\lambda_{i,i+1}$ ($0 < \lambda_{i,i+1} < \infty$).

2. The aging process of a person is affected by an unusual incident which causes a diminution of its capacity to continue alive, and then the person transits to some physiological age j, with $j \in \{i+2, i+3, \dots, n\}$. The intensity of this transition is denoted by λ_{ij} ($0 < \lambda_{ij} < \infty$).

3. The possibility of death for the person at that physiological status. The intensity of this transition is given by r_i ($0 < r_i < \infty$).

Being at the physiological age n, the only possibility of transition is the eventual transition to the death state, with intensity $\lambda_n = r_n$ ($0 < r_n < \infty$).

Note that each state represents a physiological age, and aging is described as a process of transitions from one physiological age to a higher physiological age. In section 6, we will present how to obtain the parameters λ_{ij} and r_i, which basically depend on risk factors.

Thus, the sub-intensity matrix is given by

$$\boldsymbol{Q} = \begin{pmatrix} -\lambda_1 & \lambda_{12} & \lambda_{13} & \dots & \lambda_{1n} \\ 0 & -\lambda_2 & \lambda_{23} & \dots & \lambda_{2n} \\ 0 & 0 & -\lambda_3 & \dots & \lambda_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -\lambda_n \end{pmatrix},$$

where

$$\lambda_i = \sum_{j=i+1}^{n} \lambda_{ij} + r_i, \quad i = 1, 2, \dots, n-1. \tag{3}$$

The initial distribution is $\boldsymbol{\alpha} = \boldsymbol{e}_i$, which denotes the i-th unit vector, and the exit vector is given by $\boldsymbol{r} = (r_1, r_2, \dots, r_n)'$.

We denote by $q_i(t)$ the probability of death for a person at physiological age $i \in \{1, 2 \dots, n\}$, in the interval $[0, t]$, which is given by

$$q_i(t) = \mathbb{P}(\tau_i \leq t) = 1 - \boldsymbol{e}_i \exp(t\boldsymbol{Q})\boldsymbol{e}, \tag{4}$$

where $\tau_i \sim PH(\boldsymbol{e}_i, \boldsymbol{Q})$. It is important to note that in our model, if a person at physiological age $i \in \{1, 2 \dots, n-1\}$ who died in a time interval $[0, t]$, then the transition from age i to the death state not necessarily happened, it is probably that the person visited some other states $\{i+1, \dots, n-1\}$ before die. Since our model proposes to use factors that affect the aging process, so it presents a better way to model the time to death.

In order to generate mortality tables based on this model, we need to estimate the parameters. Since by this model the time of death is PH distributed, a technical advantage is that, its density, survival function, and moments have simple analytical form. It is well known that PH distributions allow the use of matrix-analytic methods in stochastic models (see [26]). Note that we have an acyclic PH distribution (see [9, 8]) with a certain interpretation of the states. There are many software available for the estimation part of this distribution: PhFit, EMpht, G-FIT, Hyperstar, among others, but in our work, because of the variance reduction, we propose a new way of estimate these distributions using the Bisection algorithm ([2]).

Furthermore, the Heligman-Pollard model cannot identify the distribution of the time of death explicitly, and no analytical tool is available, most of the studies using this model focus on numerical or statistical experiments.

Now, we will consider the following two scenarios regarding the nature of the data: first when we have continuous time information of the aging process of the population and secondly considering a more realistic scene where there are reports of the development process only at determined moments, i.e., there are only discrete time observations of a Markov jump process. In both scenarios we consider a population that is subject to the same risk factors and the distribution of the aging process of each individual it is independent of the others. Also, persons at the same physiological age have the same distribution, and the risk factor only depends on the physiological age.

With these assumptions, we can consider that the aging process to death of each individual represents a path of a Markov jump process, indeed the paths of the individuals are independent.

Considering a population of size M and the time interval $[0, T]$, let $\boldsymbol{X}^m = \{X_t^m\}_{t \geq 0}^T, m = 1, \dots, M$, be independent Markov jump processes with the same finite state space E

and the same intensity matrix (infinitesimal generator) given in (1). Each $\boldsymbol{X}^m$ represents the aging process for each person in the population. In the next section we develop a stochastic EM algorithm for finding maximum likelihood estimators of $(\boldsymbol{\alpha}, \boldsymbol{Q})$ considering the two scenarios presented before.

5. STOCHASTIC EM ALGORITHM

It is well known that the EM algorithm is very useful for finding the maximum likelihood estimators of the PH distributions considering both cases: continuous and discrete.

Now, for the first case -continuous case- we just present the formulae and an example, we let the reader the reference [3] for more details about the estimation. For the second case, which is one contribution of the paper, we present an algorithm for estimating continuous time PH distributions for discrete observations using the Bisection algorithm, which in turn uses Markov bridges (see also [10, 27, 28] for other algorithms of this case).

5.1 Continuous case

Considering that the $\boldsymbol{X}^m$'s have been observed continuously in the time interval $[0, T]$, the complete likelihood function is given by (see [4], [17], or [20])

$$L_T^c(\boldsymbol{\theta}) = \prod_{i=1}^n \alpha_i^{B_i} \prod_{i=1}^n \prod_{j \neq i}^n \lambda_{ij}^{N_{ij}(T)} e^{-\lambda_i Z_i(T)} \prod_{i=1}^n r_i^{N_i(T)}, \qquad (5)$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{Q})$; B_i is the number of processes starting in state i; $N_{ij}(T) = \sum_{m=1}^M N_{ij}^m(T)$, where $N_{ij}^m(T)$ is the number of jumps from state i to j of the m-th process in $[0, T]$; $N_i(T) = \sum_{m=1}^M N_i^m(T)$ is the number of processes exiting from state i to the absorbing state in the time interval $[0, T]$, where

$$N_i^m(T) = \begin{cases} 1 & \text{if the } m\text{-th process is exiting from} \\ & \text{state } i \text{ to the absorbing state} \\ 0 & \text{other case;} \end{cases}$$

and $Z_i(T) = \sum_{m=1}^M Z_i^m(T)$, where $Z_i^m(T)$ is the total time spent in state i of the m-th process in $[0, T]$.

Because of (3) we can rewrite the complete likelihood function as

$$L_T^c(\boldsymbol{\theta}) = \prod_{i=1}^n \alpha_i^{B_i} \prod_{i=1}^n \prod_{j \neq i}^n \lambda_{ij}^{N_{ij}(T)} e^{-\lambda_{ij} Z_i(T)} \prod_{i=1}^n r_i^{N_i(T)} e^{-r_i Z_i(T)}.$$

The log-likelihood function is as follows

$$\begin{aligned} l_T^c(\boldsymbol{\theta}) &= \sum_{i=1}^n B_i \log(\alpha_i) + \sum_{i=1}^n \sum_{j \neq i}^n N_{ij}(T) \log(\lambda_{ij}) \\ -\sum_{i=1}^n \sum_{j \neq i}^n \lambda_{ij} Z_i(T) &+ \sum_{i=1}^n N_i(T) \log(r_i) - \sum_{i=1}^n r_i Z_i(T). \quad (6) \end{aligned}$$

It is immediately clear that the maximum likelihood estimators of the parameters are given by

$$\hat{\alpha}_i = \frac{B_i}{M}; \quad \hat{\lambda}_{ij} = \frac{N_{ij}(T)}{Z_i(T)}; \quad \hat{r}_i = \frac{N_i(T)}{Z_i(T)}.$$

In the following we will present an easy example (that will be also consider in the next case) to see the behaviour of the estimators in relation with the real parameters.

Example 5.1.

Let consider the initial vector $\boldsymbol{\alpha} = (0.25, 0.25, 0.25, 0.25)$, the subintensity matrix

$$\boldsymbol{Q} = \begin{pmatrix} -0.125 & 0.03125 & 0.03125 & 0.03125 \\ 0 & -0.09375 & 0.03125 & 0.03125 \\ 0 & 0 & -0.0625 & 0.03125 \\ 0 & 0 & 0 & -0.03125 \end{pmatrix}, \qquad (7)$$

and the exit vector $\boldsymbol{r} = \begin{pmatrix} 0.03125 \\ 0.03125 \\ 0.03125 \\ 0.03125 \end{pmatrix}$.

In Figure 1 we present the norm-1 of $\hat{\boldsymbol{Q}} - \boldsymbol{Q}$ calculated for values of T from 10 to 10 until $T = 200$. We also considered different values of M to see the behaviour of the norm.

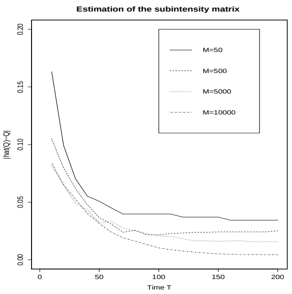


Figure 1: Norm-1 of $\hat{\boldsymbol{Q}} - \boldsymbol{Q}$ for $T = 10, 20, \ldots, 200$.

As we can see, for small values of T, the norm-1 is bigger, and as we increases the value of T this norm decreases. Note also that if we increase the value of M the estimator is closer to the value of the real parameter. At around $T = 150$ the value of the norm-1 is stabilized.

We found that the maximum likelihood estimators for $M = 10000$ and $T = 200$ were the following:

$$\hat{\boldsymbol{\alpha}} = (0.257, 0.249, 0.252, 0.242),$$

$$\hat{\boldsymbol{Q}} = \begin{pmatrix} -0.128 & 0.0332 & 0.0321 & 0.0306 \\ 0 & -0.0950 & 0.0315 & 0.0314 \\ 0 & 0 & -0.0599 & 0.0303 \\ 0 & 0 & 0 & -0.0309 \end{pmatrix}, \hat{\boldsymbol{r}} = \begin{pmatrix} 0.0318 \\ 0.0321 \\ 0.0295 \\ 0.0309 \end{pmatrix}.$$

5.2 Discrete case

In this case, the aging process of a person is recorded only at discrete times, hence it is necessary to have an algorithm to estimate the intensity matrix and build the corresponding mortality probabilities. We are interested in the inference about the intensity matrix $\boldsymbol{\Lambda}$ based on samples of observations of $\boldsymbol{X}^m$'s at discrete times points, i.e., we suppose that all processes have been observed only at K time points $0 = t_1 < \ldots < t_K = T$ denoted by $Y_k^m = X_{t_k}^m$. Assuming that the observation points are equidistant, i.e., $t_{k+1} - t_k = \Delta$ for all $k = 1, \ldots, K$, then $\boldsymbol{Y}^m = \{Y_k^m : k = 1, \ldots, K\}$ is the discrete time Markov chain associated with $\boldsymbol{X}^m$ with transition matrix $P(\Delta) = \exp(\Delta\boldsymbol{\Lambda})$, for $m = 1, \ldots, M$. We denote the observed values by $\boldsymbol{y} = \{\boldsymbol{y}^1, \ldots, \boldsymbol{y}^M\}$ where $\boldsymbol{y}^m = \{Y_1^m = y_1^m, \ldots, Y_K^m = y_K^m\}$.

In this case there are some difficulties that must be taken into account. First, from a finite number of samples it is impossible to tell if the underlying process is actually Markovian. Second, it is not clear if the observed data are originated indeed from discrete samples of a Markov jump process with some generator $\mathbf{\Lambda}$, or rather from a discrete-time Markov chain which cannot be embedded into a time-continuous counterpart (embedding problem), and finally, the fact that the matrix exponential function is not injective if the eigenvalues of the generator are complex, see [7].

The discrete log-likelihood function $L_T^d(\mathbf{\Lambda})$ of a time series $\boldsymbol{Y}^m$ is given in terms of the transition matrix $P(T)$ (see 2):

$$L_T^d(\mathbf{\Lambda}) = \prod_{m=1}^{M}\prod_{k=1}^{K-1} p_{y_k^m y_{k+1}^m}(\Delta), \tag{8}$$

where $p_{y_k^m y_{k+1}^m}(\Delta)$ is the probability that the m-th process makes a transition from the state y_k^m to the state y_{k+1}^m in the time Δ. The problem is that the derivative of (8) with respect to the entries has such a complicated form that the null cannot be found analytically. Hence no analytical expression for the maximum likelihood estimator with respect to $\mathbf{\Lambda}$ is available.

In this part, the data can be viewed as incomplete observations from a model with a tractable likelihood function. The full data set is a continuous time record of the Markov jump processes and the initial states. We can therefore find the maximum likelihood estimates by applying the Expectation-Maximization (EM) algorithm, see [12] and [23]. We need to find the likelihood function for the full data set and the conditional expectation of this function given the observations $\boldsymbol{Y} = \boldsymbol{y}$.

Let $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0, \boldsymbol{Q}_0)$ denote any initial value of parameters. Then the EM algorithm works as follows.

Algorithm 1 EM algorithm

```
1: (E-step) Calculate the function
       $h(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_0}(l_T^c(\boldsymbol{\theta})|\boldsymbol{Y} = \boldsymbol{y});$
2: (M-step)
       $\boldsymbol{\theta}_0 = argmax_{\boldsymbol{\theta}} h(\boldsymbol{\theta});$
3: Go to 1
```

The E-step and the M-step are repeated until convergence. In order to calculate the E-step, from (6) we have that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}_0}(l_T^c(\boldsymbol{\theta})|\boldsymbol{Y} = \boldsymbol{y}) \quad &= \quad \sum_{i=1}^{n}\log(\alpha_i)\mathbb{E}_{\boldsymbol{\theta}_0}(B_i|\boldsymbol{Y} = \boldsymbol{y}) \\ +\sum_{i=1}^{n}\sum_{j\neq i}^{n}\log(\lambda_{ij})\mathbb{E}_{\boldsymbol{\theta}_0}(N_{ij}(T)|\boldsymbol{Y} = \boldsymbol{y}) &- \sum_{i=1}^{n}\sum_{j\neq i}^{n}\lambda_{ij}\mathbb{E}_{\boldsymbol{\theta}_0}(Z_i(T)|\boldsymbol{Y} = \boldsymbol{y}) \\ +\sum_{i=1}^{n}\log(r_i)\mathbb{E}_{\boldsymbol{\theta}_0}(N_i(T)|\boldsymbol{Y} = \boldsymbol{y}) &- \sum_{i=1}^{n}r_i\mathbb{E}_{\boldsymbol{\theta}_0}(Z_i(T)|\boldsymbol{Y} = \boldsymbol{y}). \end{aligned} \tag{9}$$

Since (6) is a linear function of the sufficient statistics B_i, $Z_i(T)$, $N_i(T)$, and $N_{ij}(T)$, it is enough to calculate the corresponding conditional expectations of these statistics:

$$\mathbb{E}_{\boldsymbol{\theta}_0}(B_i|\boldsymbol{Y} = \boldsymbol{y}) = \sum_{m=1}^{M}\mathbf{1}_{\{Y_1^m = i\}}; \quad i = 1,\ldots,n, \tag{10}$$

where $\mathbf{1}_{\{.\}}$ is the indicator function,

$$\mathbb{E}_{\boldsymbol{\theta}_0}(N_{ij}(T)|\boldsymbol{Y} = \boldsymbol{y}) = \sum_{m=1}^{M}\sum_{k=2}^{K}\tilde{N}^{mij}_{y_{k-1}^m y_k^m}(t_k - t_{k-1}), \tag{11}$$

$$\mathbb{E}_{\boldsymbol{\theta}_0}(Z_i(T)|\boldsymbol{Y} = \boldsymbol{y}) = \sum_{m=1}^{M}\sum_{k=2}^{K}\tilde{Z}^{mi}_{y_{k-1}^m y_k^m}(t_k - t_{k-1}), \tag{12}$$

and

$$\mathbb{E}_{\boldsymbol{\theta}_0}(N_i(T)|\boldsymbol{Y} = \boldsymbol{y}) = \sum_{m=1}^{M}\sum_{k=2}^{K}\tilde{N}^{mi}_{y_{k-1}^m}(t_k - t_{k-1}), \tag{13}$$

where by the Markov property and the homogeneity of the processes, it is sufficient to evaluate for the states $i, j, r, s = 1,\ldots,n$, the processes $m = 1,\ldots,M$, and the times $0 \leq s_1 < s_2 < \infty$, the following

$$\tilde{N}^{mij}_{rs}(s_2 - s_1) = \mathbb{E}_{\boldsymbol{\theta}_0}(N^m_{ij}(s_2 - s_1)|X^m_{s_2} = s, X^m_{s_1} = r), \tag{14}$$

$$\tilde{Z}^{mi}_{rs}(s_2 - s_1) = \mathbb{E}_{\boldsymbol{\theta}_0}(Z^m_i(s_2 - s_1)|X^m_{s_2} = s, X^m_{s_1} = r), \tag{15}$$

$$\tilde{N}^{mi}_{s}(s_2 - s_1) = \mathbb{E}_{\boldsymbol{\theta}_0}(N^m_i(s_2 - s_1)|X^m_{s_2} = n+1, X^m_{s_1} = s). \tag{16}$$

The delicate part of using the EM algorithm rely on the E-step, since it is the computationally demanding one. In 2009, the reference [27] proposed two EM algorithms for fitting Markovian Arrival Processes (MAPs) and Markov Modulated Poisson Processes with generalized group data, the estimation is performed when exact arrival times are not known. Since in their work the size of the MAPs is limited due to the computational effort, they extended it in [28] considering also PH distributions and giving an improvement of reducing the time complexity of their algorithms using uniformization. Some of their sofware available are mapfit (estimation methods for PH distributions and MAPs from empirical data (point and grouped data)), and PHPACK (Phase-Type Analysis Package) which is a software package to use the PH distribution in stochastic modeling.

In 2012, the reference [10] proposed a new form of calculating the steps of the EM in MAPs based on the matrix exponential function.

Considering PH distributions, in simulation-based statistical estimation, one needs to generate a sample path of the Markov chain, the problem can, however, be translated to endpoint conditioned simulation. In [16] is suggested some algorithms for end-point conditional simulation from continuous time Markov chains (CTMCs): rejection sampling, uniformization, and direct simulation. Thus, as Asmussen mentioned in [2], we will use bridge sampling from CTMCs for the estimation of PH distributions.

To calculate (14), (15) and (16) we propose to generate L (where L is a fixed positive integer number) sample paths of the Markov bridge $X^m(r, s_1, s, s_2)$ using the parameter value $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0, \boldsymbol{Q}_0)$, i.e., a stochastic process defined on $[s_1, s_2]$ and having the same distribution of the Markov jump process $\{X^m_t\}_{t\in[s_1,s_2]}$ conditioned on $X^m_{s_1} = r$ and $X^m_{s_2} = s$ for $m = 1,\ldots,M$.

Let $X^{m(l)}(r, s_1, s, s_2)$ be the l-th path of the Markov bridge $X^m(r, s_1, s, s_2)$. For $l = 1,\ldots,L$ and $m = 1,\ldots,M$, we can calculate for the corresponding statistics by using: $N^{mij(l)}_{rs}(s_2 - s_1)$, the number of jumps from state i to j; $N^{mi(l)}_{s}(s_2 - s_1)$, the number of processes exiting from state i to the absorbing state; and $Z_{rs}^{mi(l)}(s_2 - s_1)$, the total time spent in state i.

Now, based on these bridges we approximate the conditional expectations (14), (15), and (16) by

$$\tilde{N}^{mij}_{y^m_{k-1}y^m_k}(t_k - t_{k-1}) \approx \frac{1}{L}\sum_{l=1}^{L} N^{mij(l)}_{y^m_{k-1}y^m_k}(t_k - t_{k-1}), \quad (17)$$

$$\tilde{Z}^{mi}_{y^m_{k-1}y^m_k}(t_k - t_{k-1}) \approx \frac{1}{L}\sum_{l=1}^{L} Z^{mi(l)}_{y^m_{k-1}y^m_k}(t_k - t_{k-1}), \quad (18)$$

$$\tilde{N}^{mi}_{y^m_{k-1}}(t_k - t_{k-1}) \approx \frac{1}{L}\sum_{l=1}^{L} N^{mi(l)}_{y^m_{k-1}}(t_k - t_{k-1}), \quad (19)$$

respectively, for $i, j = 1, \ldots, n$; $m = 1 \ldots, M$; and $k = 1, \ldots, K$.

Thus we rewrite the EM algorithm for the maximum likelihood estimation as follows:

Algorithm 2 EM algorithm

```
1: Let $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0, \boldsymbol{Q}_0)$ denote any initial value of parameters and we make $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Given $M$, $K$ for $m = 1, \ldots, M$, $k = 2, \ldots, K$:
2: Generate $L$ paths of the Markov bridge $X^m(y_{k-1}, t_{k-1}, y_k, t_k)$ using the parameter value $\boldsymbol{\theta}$.
3: E-step.
    • Using the equations (17), (18), and (19) calculate $\tilde{N}^{mij}_{y^m_{k-1}y^m_k}(t_k - t_{k-1})$, $\tilde{Z}^{mi}_{y^m_{k-1}y^m_k}(t_k - t_{k-1})$, and $\tilde{N}^{mi}_{y^m_{k-1}}(t_k - t_{k-1})$, respectively.
    • Using the equations (10), (11), (12), and (13) we can calculate $\mathbb{E}_{\boldsymbol{\theta}}(B_i|\boldsymbol{Y} = \boldsymbol{y})$, $\mathbb{E}_{\boldsymbol{\theta}}(N_{ij}(T)|\boldsymbol{Y} = \boldsymbol{y})$, $\mathbb{E}_{\boldsymbol{\theta}}(Z_i(T)|\boldsymbol{Y} = \boldsymbol{y})$, and $\mathbb{E}_{\boldsymbol{\theta}}(N_i(T)|\boldsymbol{Y} = \boldsymbol{y})$, respectively.
4: M-step Calculate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{Q}})$ by
    $\hat{\alpha}_i = \frac{\mathbb{E}_{\boldsymbol{\theta}}(B_i|\boldsymbol{Y} = \boldsymbol{y})}{M}; \quad \hat{\lambda}_{ij} = \frac{\mathbb{E}_{\boldsymbol{\theta}}(N_{ij}(T)|\boldsymbol{Y} = \boldsymbol{y})}{\mathbb{E}_{\boldsymbol{\theta}}(Z_i(T)|\boldsymbol{Y} = \boldsymbol{y})};$
    $\hat{r}_i = \frac{\mathbb{E}_{\boldsymbol{\theta}}(N_i(T)|\boldsymbol{Y} = \boldsymbol{y})}{\mathbb{E}_{\boldsymbol{\theta}}(Z_i(T)|\boldsymbol{Y} = \boldsymbol{y})}.$
5: $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ go to 2.
```

To implement this algorithm, the main issue is how to sample Markov bridges. We use the bisection method proposed by S. Asmussen and A. Hobolth ([2]) because of its potential for variance reduction. The idea of this algorithm is to formulate a recursive procedure where we finish off intervals with zero or one jumps and keep bisecting intervals with two or more jumps. The recursion ends when no intervals with two or more jumps are presented.

Considering a Markov bridge $X(a, 0, b, T)$ and using the bisection algorithm we have two type of scenarios:

1. If $a = b$ and there are no jumps. In this case we are done: $X_t = a$ (a is not an absorbing state) for $0 \leq t \leq T$.

2. If $a \neq b$ and there is one jump we are done: $X_t = a$ for $t \in [0, \tau)$, and $X_t = b$ for $t \in [\tau, T]$. Here τ is the jump time.

To determine τ we use the following lemma from [2].

Lema 5.1. *Considering an interval of length T, let $X_0 = a$, the probability that $X_T = b \neq a$ and there is only one single jump (from a to b) in the interval is given by*

$$R_{ab}(T) = \lambda_{ab} \begin{cases} \frac{e^{-\lambda_a T} - e^{-\lambda_b T}}{\lambda_a - \lambda_b} & \lambda_a \neq \lambda_b \\ T e^{-\lambda_a T} & \lambda_a = \lambda_b. \end{cases}$$

The density of the time of state change is

$$f_{ab}(t;T) = \frac{\lambda_{ab} e^{-\lambda_b T}}{R_{ab}(T)} e^{-(\lambda_a - \lambda_b)t}; \quad 0 \leq t \leq T.$$

Example 5.2. ***Bridges****. Considering the intensity matrix (7), we simulate some bridges which are presented in Figure 2. In this example, we suppose that a person is in the initial state is 1 at time 0 and at time 33.309 is in the state 4. Here we present four different scenarios that could happen to this person.*

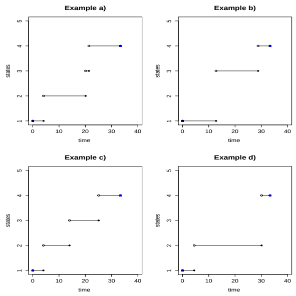


Figure 2: Example of Bridges.

As we can see in this figure, in the example b), the person passes directly from state 1 to the state 3 at the age around 12. This is not the case for the examples a) and c), where the person visits all the state and spends certain amount of time in each of them. Note also in the example d) that the person is in the state 2 for almost 26 units of time.

5.2.1 Fisher information matrix

Because we are interested in the variance reduction, based on [6] we have that the Fisher information matrix using our proposed EM algorithm is given by $\mathbf{I}(\hat{\boldsymbol{\theta}}) = -\left[\frac{\partial^2 \mathcal{Q}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})}{\partial \hat{\boldsymbol{\theta}}^2} + \frac{\partial^2 \mathcal{Q}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \hat{\boldsymbol{\theta}}}\right]_{\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}}$, where $\mathcal{Q}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}(l^c_T(\hat{\boldsymbol{\theta}})|\boldsymbol{Y} = \boldsymbol{y})$. Thus, in order to get this matrix, we just have to calculate the following:

$$\frac{\partial^2 \mathcal{Q}}{\partial \hat{\alpha}_i^2} = -\frac{1}{\hat{\alpha}_i^2} * eq(10), \quad \frac{\partial^2 \mathcal{Q}}{\partial \hat{\lambda}_{ij}^2} = -\frac{1}{\hat{\lambda}_{ij}^2} * eq(11),$$

$$\frac{\partial^2 \mathcal{Q}}{\partial \hat{r}_i^2} = -\frac{1}{\hat{r}_i^2} * eq(13).$$

Further, the inverse of the Fisher information matrix is an estimator of the asymptotic covariance matrix. Hence, the estimated standard deviation of the maximum likelihood estimates is given by:

$$SD(\hat{\boldsymbol{\theta}}_{ML}) = \frac{1}{\sqrt{\mathbf{I}(\hat{\boldsymbol{\theta}}_{ML})}}.$$

Now, we will present an example of estimation using the Algorithm 2 and calculating the standard errors of the estimators. We will see the convergence of the algorithm and thus we propose a way to find the final estimation. This methodology will be used in the next section for the estimators of the transition matrix that models mortality data.

Example 5.3. ***Estimation****. As real parameters we use the same as the Example 5.1, and $M = 500, T = 100, \Delta = 5, K = 20, L = 50$, with arbitrary initial parameters. Now, using the stochastic EM and the bisection algorithms, we present the results.*

In the following figure we present the norm-1 of the matrix $\hat{\boldsymbol{Q}} - \boldsymbol{Q}$ for 100 iterations.

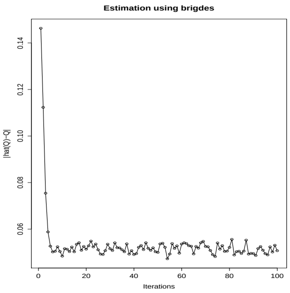


Figure 3: Norm-1 of $\hat{\boldsymbol{Q}} - \boldsymbol{Q}$ for 100 iterations, where the estimation was obtained using Bridges.

Considering a burn-in of $b = 25$ (see Figure 3) and maximum number of iterations $I = 100$, the stochastic EM estimate of $\boldsymbol{\theta}$ is given by

$$\frac{1}{I-b} \sum_{i=(b+1)}^{I} \hat{\boldsymbol{\theta}}_i,$$

where $\hat{\boldsymbol{\theta}}_i$ denotes the maximum likelihood estimator at the iteration i.

After finding the maximum likelihood estimators, the Fisher information matrix was obtained, so as the variance-covariance matrix. The results are presented in Table 1.

Table 1: Maximum likelihood estimators (MLEs) and Standard deviations (SDs) of the parameters presented in the Example 5.1.

Parameter	True value	MLE	SD
$\hat{\alpha}_1$	0.25	0.2680	0.0232
$\hat{\alpha}_2$	0.25	0.2820	0.0237
$\hat{\alpha}_3$	0.25	0.2480	0.0223
$\hat{\alpha}_4$	0.25	0.2020	0.0201
$\hat{\lambda}_{12}$	0.03125	0.0390	0.0064
$\hat{\lambda}_{13}$	0.03125	0.0320	0.0060
$\hat{\lambda}_{14}$	0.03125	0.0377	0.0060
$\hat{\lambda}_{23}$	0.03125	0.0252	0.0036
$\hat{\lambda}_{24}$	0.03125	0.0323	0.0041
$\hat{\lambda}_{34}$	0.03125	0.0345	0.0034
$\hat{r}_1$	0.03125	0.0210	0.0040
$\hat{r}_2$	0.03125	0.0354	0.0043
$\hat{r}_3$	0.03125	0.0325	0.0033
$\hat{r}_4$	0.03125	0.0334	0.0020

With this simple example we can corroborate that the EM algorithm using the Bisection works very well (see Table 1, column 3). Thus, we can apply it into a real scenario: Mortality.

6. MORTALITY TABLE

In this section we use detailed mortality and population data from the U.S.A. (check http://www.mortality.org) for the construction of hypothetical physiological ages. The database contains annual information on mortality and population from 1933 to 2013. In each year, the information is classified at 111 ages, from 0 to 110. Let $\mu_i(j)$ be the probability of death observed for a person at age $i \in \{0, \ldots, 110\}$ in the year $j \in \{1990, \ldots, 2013\}$ and $\mu_i = \frac{\sum_j \mu_i(j)}{24}$.

Since the biological age is the most important factor for the construction of physiological ages, we assume that we have the same number of physiological and biological ages. In order to construct these physiological ages, we need a way to establish the transition probabilities between them, so as to have the structure of our model. Since the information on the database is recorded annually, we just have to calculate the probability that the person passes from age i to age j in one year ($i = 0, 1, \ldots, 109$; $j = 0, \ldots, 110$), denoted by p_{ij}. These probabilities are given by:

$$p_{ij} = \begin{cases} (1-\mu_i)s_i & \text{if} \quad j = i, \\ (1-\mu_i)w_i & \text{if} \quad j = i+1, \\ (1-\mu_i)(1-w_i-s_i)t_{ij} & \text{if} \quad j \geq i+2, \end{cases}$$

where $s_i \in (0, 1)$ represents a greater ability to adapt to the environment for a person at physiological age i , $w_i \in (0, 1)$ represents the natural aging factor for a person at physiological age i (if we want to have an exactly or estimate of this value we need detailed information about medical record and lifestyle of each person - feeding habits, sports practice, etc.), and t_{ij} represents the aging factor caused by an incident which causes that a person at age i passes to age j in a period of one year (for obtaining this value we need history incidents for each physiological age).

Thus, we have a hypothetical population with 111 physiological ages and a death state, to which we can associate a discrete time Markov chain that models the aging process with transition matrix $P = \{p_{ij}\}_{i,j}$ (for one year). Since $\lambda_{ij} \propto p_{ij}$, for the corresponding infinitesimal generator we do $\lambda_{ij} = p_{ij}$ for all $i \neq j$.

We generate historical information of the aging process of the population of the United States observed at discrete times. We use the algorithm 2 to estimate the corresponding infinitesimal sub-generator $\hat{\boldsymbol{Q}}$ ($n = 111, , M = 1000, T = 50, \Delta = 1, K = 50, L = 20, I = 100, b = 20$, $s_i = 0.02, w_i = 0.8$, $\forall i$, and $t_{ij} = \frac{n-1+i+2-j}{(n-1)(n/2)-(i+1)(i+2)/2}$) and therefore the parameters of the PH distribution used for building the mortality tables.

Considering the year 2013, in figure 4 we plot the estimation of the mortality tables using the equation (4). The figure shows the increasing susceptibility with aging.

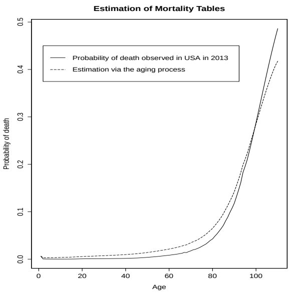


Figure 4: Estimation of Mortality Tables.

7. CONCLUSIONS

In this paper we considered the concept of physiological age to such a model in order to have a more realistic one, since it considers more risk factors. For the estimation part, in the case of missing data, we used a stochastic EM using Markov bridges. We have fitted real data from the United States using our algorithm, choosing arbitrarily the factors for the construction of the evolution of the aging process, and we consider the results are satisfactory. Based on the statistics, we would expect that s_i takes a small value for ages under 35 (at that age people start to lead a healthier life) and a little larger from 35, w_i is the same for all ages (close to one), and t_{ij} can take four different values: one for ages under of 2, another between 2 and 15, the third one for ages from 15 to 35, and the last one greater than 35; and thus the model presented in this paper would be a great tool for modeling human mortality. This work serves as motivation to propose a new method for calculating the risk factors mentioned in the last section. Furthermore, it is possible to extend this model if we assume that the risk factors are modeled by a stochastic process and seen the aging process as a stochastic process in random environment.

Acknowledgements
We would like to acknowledge the support from the Sistema Nacional de Investigadores of CONACYT, PRODEP UJED-PTC-100, and PAPIIT-IA105716.

References

[1] O. O. Aalen. On Phase Type Distributions in Survival Analysis. *Scandinavian Journal of Statistics*, 22:447–463, 1995.

[2] S. Asmussen and A. Hobolth. Markov Bridges, Bisection and Variance Reduction. Technical report, 2011.

[3] S. Asmussen, O. Nerman, and M. Olsson. Fitting phase-type distributions via the EM algorihtm. *Scandinavian Journal of Statistics*, 23:419–441, 1996.

[4] P. Billingsley. Statistical Inference for Markov Processes. *Chicago: University of Chicago Press*, 1961.

[5] M. Bladt. A Review on Phase–Type Distributions and Their Use in Risk Theory. *Astin Bulletin*, 35(1):145–161, 2005.

[6] M. Bladt, L. J. R. Esparza, and B. F. Friis. Fisher Information and Statistical Inference for Phase-type distributions. *J. Appl. Prob.*, 48A:277–293, 2011.

[7] M. Bladt and M. Sorensen. Statistical inference for discretely observed markov jump processes. *J. Roy. Statist. Soc.*, (67):395 – 410, 2005.

[8] A. Bobbio, A. Horvath, M. Scarpa, and M. Telek. Acyclic discrete phase type distributions: properties and a parameter estimation algorithm. *Performance evaluation An international Journal*, 54:1–32, 2003.

[9] A. Bobbio and M. Telek. A benchmark of Ph estimation algorithms: results for acyclic-PH. *Stochastic models*, 10:661–677, 1994.

[10] L. Breuer and A. Kume. An EM algorithm for Markovian Arrival Processes Observed at discrete times. *B. Müller-Clostermann et al. (Eds.): MMB & DFT. Springer, Berlin.*, LNCS 5987:242–258, 2010.

[11] A. Craik. Edward Sang (1805-1890): calculator extraordinary', special number of Newsletter in memory of J.G. Fauvel. *British Society for the History of Mathematics Newsletter*, 45:32–45, 2002.

[12] A. P. Dempster, D. B. Rubin, and N. M. Laird. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. B*, (39):1–38, 1977.

[13] B. Gompertz. On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. *Philosophical Transactions of the Royal Society of London*, 115:513–583, 1825.

[14] E. Halley. An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslau; With An Attempt to ascertain the prices of annuities upon lives. *Philosophical Transactions*, (196), 1693.

[15] M. A. L. Heligman and J. H. Pollard. The Age Pattern of Mortality. *Journal of the Institute of Actuaries*, (107):49–80, 1980.

[16] A. Hobolth and E. Stone. Efficient simulation from finite-state, continuous-time Markov chains with incomplete observations. *Ann. Appl. Statist.*, 3:1204–1231, 2009.

[17] M. Jacobsen. Statistical analysis of counting processes. *Lect. Notes Statist., 12*, 1982.

[18] H. B. Jones. A Special Consideration of the Aging Process, Disease and Life Expectancy. *In Advances in Biological and Medical Physics, ed. J. H. Lawrence and C. A. Tobias*, 4:281–337, 1956.

[19] D. A. Knowles, S. L. Part, D. Glass, and J. M. Winn. Inferring a measure of physiological age from multiple ageing related phenotypes. 2011.

[20] H. Küchler and M. Sorensen. *Exponential Families of Stochastic Processes.* New York, Springer, 1997.

[21] R. D. Lee and L. Carter. Modeling and forecasting the time series of u.s. mortality. *Journal of the American Statistical Association*, (87):659–671, 1992.

[22] W. M. Makeham. On the Law of Mortality and the Construction of Annuity Tables. *The Assurance Magazine, and Journal of the Institute of Actuaries*, 8, 1860.

[23] G. J. McLachlan and T. Krishnan. *The EM algorithm and Extensions.* Wiley, New York, 1997.

[24] A. D. Moivre. *Annuties upon Lives or The Valuation of Annuities upon any Number of Live; as alfo, of Reversions.* W.P. and fold by Francis Fayram, 1725.

[25] M. F. Neuts. Probability distributions of phase-type. *In Liber Amicorum Prof. Emeritus H. Florin*, pages 173–206, 1975.

[26] M. F. Neuts. *Matrix Geometric solutions in stochastic models*, volume 2. Johns Hopkins University Press, Baltimore, Md., 1981.

[27] H. Okamura, T. Dohi, and S. Trivedi. Markovian arrival process parameter estimation with group data. *IEEE/ACM Trans. Netw.*, 17(4):1326–1339, 2009.

[28] H. Okamura, T. Dohi, and S. Trivedi. Improvement of expectation-maximization algorithm for phase-type distributions with grouped and truncated data. *Appl. Stoch. Model. Bus. Ind.*, 29(2):141–156, 2012.

[29] L. X. Sheldon and L. Xiaoming. Markov Aging Process and Phase-Type Law of Mortality. *North American Actuarial Journal*, 11(4):92–109, 2008.

[30] S. M. Zuev, A. L. Yashin, K. G. Manton, E. Dowd, I. B. Pogojev, and R. Usmanov. Vitality index in survival modeling: how physiological aging influences mortality. *Journal of Gorontology*, 55 A:10–19, 2000.

Fitting Markovian binary trees using global and individual population data

[Extended Abstract]

Sophie Hautphenne
The University of Melbourne
Department of Mathematics and Statistics
Melbourne, 3010, Australia
Ecole Polytechnique fédérale de Lausanne
Institute of Mathematics
Lausanne, 1015, Switzerland
sophiemh@unimelb.edu.au
sophie.hautphenne@epfl.ch

Katharine Turner
Ecole Polytechnique fédérale de Lausanne
Institute of Mathematics
Lausanne, 1015, Switzerland
katharine.turner@epfl.ch

ABSTRACT

In this paper, we estimate the parameters of the transient Markovian arrival process controlling the individuals lifetime and reproduction epochs in a Markovian binary tree. The datasets used are population data containing information on age-specific fertility and mortality rates, and we apply a non-linear regression method or a maximum likelihood method, depending on the precision of the available data. We discuss the accuracy of the parameter estimates, as well as the optimal choice of the number of phases, and we provide confidence intervals for some model outputs.

Keywords

Markovian binary tree; transient Markovian arrival process; Markov modulated Poisson process; parameter estimation; non-linear regression; maximum likelihood

1. INTRODUCTION

Simple birth-an-death processes do not offer enough flexibility to model real biological populations in which the age of individuals impacts on their fertility and mortality rates. Indeed the memoryless property inherent to these models implies that individuals do not age. Despite this problem, these Markovian models have the very attractive property that they are tractable and amenable to efficient parameter estimation. We aim at generalising birth-and-death models, while keeping the desirable tractability property.

In this paper, we model the lifetime and reproduction epochs of individuals in a population using a *transient Markovian arrival process* (TMAP). Roughly speaking, a TMAP is a point process in which the event rate depends on the state transition of an underlying transient Markov chain with n transient states (also called *phases*), and one absorbing phase. Each event in the TMAP corresponds to the birth of a child, and the absorption in phase 0 corresponds to the individual's death. The resulting continuous-time branching process, called *Markovian binary tree* (MBT), is the matrix generalisation of the birth-and-death process, and allows for much more flexibility than the latter, while keeping an excellent computational tractability.

Performance measures of MBTs include the extinction probability of the population, the distributions of the time until extinction, the population size at any given time, and the total progeny size until any given time. All these model outputs are function of the initial phase (or the initial age-class) of the first individual in the population. The MBT model has already been used to efficiently compare demographic properties of female families in different countries, see [2]. The purpose of this paper is to develop the statistical tools necessary to fit an MBT to a population of endangered species such as the *black robins* of the Chatham Islands [5]. The model can then be used to compare different strategies of conservation for the species.

For that purpose, we fit a TMAP to different types of population datasets which may be available from demographic databases or may result from field research in biology. These datasets can have different degrees of accuracy: we distinguish between

- *global population data*, consisting in the *average* age-specific fertility and mortality *rates* over an entire population; this is the case in human fertility and mortality database for instance; and
- *individual population data*, consisting in data on age-specific fertility and mortality *counts* for each individual in a population; this is the case for closely monitored species such as the black robin population.

The parameter estimation method depends on the type of data which are available. The optimal number of phases n in the TMAP has to be chosen according to some validation methods, such as the cross-validation method, the mean squared error, or the Akaike information criterion. Once a value of n is determined and an estimator is found for the model parameters, we derive confidence intervals for the model outputs.

The paper is organised as follows: in the next section, we provide more details on TMAPs and describe one particularly simple case of such processes that we shall focus on. In Section 3 we perform the model parameter estimation based on the average age-specific fertility and mortality rates, using a non-linear regression method. In Section 4, we estimate the parameters based on individual age-specific fertility and mortality counts, using a maximum likelihood method. We compare the results of both methods in Section 5.

2. TMAPS

TMAPS are two-dimensional Markovian processes $\{(N(t), \varphi(t)) : t \in \mathbb{R}^+\}$ on the state space $\mathbb{N} \times \{0, 1, \ldots, n\}$, where n is finite. The states $(k, 0)$ are absorbing for all $k \geq 0$; the other states are transient. The process $\{N(t)\}$ counts the number of arrivals in $[0, t]$ and is called the *level* process. The process $\{\varphi(t)\}$ is a continuous-time Markov chain, called the *phase* process.

A TMAP is characterized by two $n \times n$ rate matrices D_0 and D_1 and a non-negative $n \times 1$ rate vector $\boldsymbol{d}$. Feasible transitions are from (k, i) to (k, j), for $k \geq 0$ and $1 \leq i \neq j \leq n$ at the rate $(D_0)_{ij}$, or from (k, i) to $(k + 1, j)$ for $1 \leq i, j \leq n$ at the rate $(D_1)_{ij}$, or from (k, i) to $(k, 0)$ at rate d_i. The first transitions (at rate $(D_0)_{ij}$) are *hidden*: the phase of the individual changes but not the count. The second transitions (at rate $(D_1)_{ij}$) are *observable*: a birth (arrival) is recorded, at which time the state of the individual may or may not change. Finally, the third transitions (at rate d_i) indicate the termination of the individual's life.

The matrix D_1 and the vector $\boldsymbol{d}$ are nonnegative, D_0 has nonnegative off-diagonal elements and strictly negative elements on the diagonal such that $D_0\,\mathbf{1} + D_1\,\mathbf{1} + \boldsymbol{d} = \mathbf{0}$, where $\mathbf{1}$ is an $n \times 1$ vector of ones. One also defines the initial probability vector $\boldsymbol{\alpha} = (\alpha_i)_{1 \leq i \leq n}$ such that $\alpha_i = \mathbb{P}(\varphi(0) = i)$; we assume that $\boldsymbol{\alpha}\mathbf{1} = 1$, so that $\varphi(0) \neq 0$ a.s. More details on TMAPs can be found in [3].

There is a total of $2n^2 + n - 1$ entries in the matrices $\boldsymbol{\alpha}, D_0, D_1, \boldsymbol{d}$ if no assumption is made on their structure. For efficiency purposes in our parameter estimation, we shall consider a particular case of TMAP called the *acyclic transient Markov modulated Poisson process* (ATMMPP). In this model we assume that

- individuals start their lifetime in phase 1 with probability one,
- they can only move from phase i to phase $i + 1$ or to phase 0, with respective rates γ_i for $1 \leq i \leq n - 1$ and μ_i for $1 \leq i \leq n$,
- while in phase i, they reproduce at rate λ_i and do not make any simultaneous phase transition at reproduction time.

With these assumptions, the matrices $\boldsymbol{\alpha}, D_0, D_1, \boldsymbol{d}$ have the following structure: $\boldsymbol{\alpha} = [1, 0, \ldots, 0]$, $\boldsymbol{d} = [\mu_1, \ldots, \mu_n]^\top$, D_1 =diag$(\lambda_1, \ldots, \lambda_n)$, and the only non-zero entries of D_0 are $(D_0)_{i,i+1} = \gamma_i$ and

$$(D_0)_{ii} = \begin{cases} \lambda_i - \mu_i - \gamma_i & 1 \leq i \leq n - 1 \\ -\lambda_i - \mu_i & i = n. \end{cases}$$

There is a total of $3n - 1$ parameters in an ATMMPP.

Note that the corresponding lifetime distribution is phase-type PH$(\boldsymbol{\alpha}, D_0 + D_1)$. Due to the particular structure of D_0 and D_1 in the ATMMPP case, this corresponds more precisely to a Coxian distribution. Such distributions are very important as any acyclic phase-type distribution has an equivalent Coxian representation. Therefore, in terms of the lifetime distribution, the ATMMPP does not impose too much restriction compared to the general TMAP.

3. GLOBAL POPULATION DATA

In this section, we assume that the available data are

- estimates of the expected age-specific fertility rates, $\hat{b}_x$, and
- estimates of the expected age-specific mortality rates, $\hat{d}_x$,

where $x \in \{0, 1, 2, \ldots, M\}$ denotes the age, that is, the period of time $[x, x + 1)$ during the lifetime, and M is the maximal age for which data are available. We denote by $\bar{d}(x)$ and $\bar{b}(x)$ the equivalent quantities computed from the TMAP model, for any $x \geq 0$. We can show that these functions have an explicit expression.

Proposition 1. *The age-specific mortality and fertility rates in a TMAP are respectively given by*

$$\begin{aligned} \bar{d}(x) &= \frac{\boldsymbol{\alpha} e^{Dx}(I - e^D)\mathbf{1}}{\boldsymbol{\alpha} e^{Dx}\mathbf{1}} \\ \bar{b}(x) &= \frac{\boldsymbol{\alpha} e^{Dx}(I - e^D)(-D)^{-1}D_1\mathbf{1}}{\boldsymbol{\alpha} e^{Dx}\mathbf{1}}, \end{aligned}$$

where $D = D_0 + D_1$ is the phase transition rate matrix.

We extend the approach developed in [4], in which only death rates were used to fit phase-type lifetime distributions, and we estimate the model parameters by minimizing the sum of weighted squared errors

$$F = \sum_{x=0}^{M} \left[(\hat{d}_x - \bar{d}(x))^2 + (\hat{b}_x - \bar{b}(x))^2\right] \hat{S}_x, \qquad (1)$$

where the weights $\hat{S}_x$ are the observed probability of survival until age x, computed as

$$\hat{S}_x = (1 - \hat{d}_0)(1 - \hat{d}_1) \cdots (1 - \hat{d}_{x-1}).$$

Since the functions $\bar{d}(x)$ and $\bar{b}(x)$ are non-linear in both the input variable x and in the parameters of the TMAP, we are dealing with a weighted non-linear regression.

4. INDIVIDUAL POPULATION DATA

In this section, we assume that the available data are individual age-specific fertility and mortality counts, that is, they consist of N vectors (one for each individual) of the type

$$\boldsymbol{v} = [6,\ 8,\ -2,\ 9,\ 0,\ 3,\ 3,\ -1], \qquad (2)$$

of variable length, whose entries $v_i, i \geq 1$ are interpreted as follows:

- $v_i = k \in \{0, 1, 2, \ldots\}$ means that the individual had k offspring while in the age interval $[i - 1, i)$ (that is, at age $x = i - 1$),

- $v_i = -1$ means that the individual died in the previous age interval $[i-2, i-1)$, possibly after producing some offspring, and

- $v_i = -2$ means that we do not have any information on the number of offspring generated by the individual in the age class $[i-1, i)$, but we know that the individual was still alive at the end of this age class. We therefore allows for missing information in the data.

We use a maximum likelihood estimation method based on a sample of i.i.d. individual count vectors $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(N)}\}$. The log-likelihood function is

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(N)}) = \sum_{j=1}^{N} \log p(\boldsymbol{v}^{(j)}|\boldsymbol{\theta}), \qquad (3)$$

where $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, D_0, D_1, \boldsymbol{d}\}$, and $p(\boldsymbol{v}^{(j)}|\boldsymbol{\theta})$ is the likelihood of the jth observation, that is, the probability of observing the individual population count vector $\boldsymbol{v}^{(j)}$, under the model parameter $\boldsymbol{\theta}$. The probabilities $p(\boldsymbol{v}^{(j)}|\boldsymbol{\theta})$ can be decomposed into products of some matrices $P(k)$ for $1 \le k \le K$, where

$$P_{ij}(k) = P[N(1) = k, \varphi(1) = j \,|\, N(0) = 0, \varphi(0) = i],$$

and where $K = \max_{i,j}\{v_i^{(j)} : 1 \le i, 1 \le j \le N\}$ is the maximum number of offspring per age-class among the individuals in the sample. The matrices $P(k)$ can be computed explicitly, as shown in the next proposition.

Proposition 2. *For $1 \le k \le K$,*

$$P(k) = (1/k!)(\boldsymbol{e}_k \otimes I)\exp(\mathcal{M})(\boldsymbol{e}_1^\top \otimes I),$$

where $\boldsymbol{e}_k$ is the kth unit row vector of size K, and

$$\mathcal{M} = \begin{bmatrix} D_0 & & & & \\ D_1 & D_0 & & & \\ & 2D_1 & D_0 & & \\ & & \ddots & & \\ & & & KD_1 & D_0 \end{bmatrix}.$$

This result generalises those of Davison and Ramesh [1] who considered the parameter estimation of Markov modulated Poisson processes in the binary data case $v_i = \mathbb{1}_{\{N(i)-N(i-1)\ge 1\}}$. To the best of our knowledge, other methods for the parameter estimation of such processes are based on the observation of the successive inter-event times rather than on the number of events within a given time interval, see for instance [6] and [7] .

5. NUMERICAL EXAMPLE

Consider an ATMMPP with $n = 3$ phases and parameter values given in Table 1.

γ_1	γ_2	μ_1	μ_2	μ_3	λ_1	λ_2	λ_3
0.25	0.25	0.2	0.4	0.9	6	3	2

Table 1: Parameter values of the ATMMPP

We simulated $N = 250$ trajectories of this process until time $T = 15$, and we counted the number of points falling in successive one-year intervals (age-classes). This produced a sample $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(250)}\}$ of 250 individual data vectors of the form (2). The average age-specific fertility and mortality rate vectors $\hat{\boldsymbol{b}} = (\hat{b}_x)$ and $\hat{\boldsymbol{d}} = (\hat{d}_x)$ were computed from that sample.

In a first experiment, we fix the value $n = 3$, and we use the Matlab function `fmincon` to minimize the sum of weighted squared errors (1) in the global population data case, or to maximize the log-likelihood function (3) in the individual population data case, under the constraint that the parameters are positive. The optimization algorithm requires an initial value for the parameters. A reasonable guess for the model parameters is

$$\begin{aligned} \gamma_i &= (M+1)^{-1}\, n, & 1 \le i \le n-1 \\ \lambda_i &= (M+1)^{-1} \sum_{x=0}^{M} \hat{b}_x, & 1 \le i \le n \\ \mu_i &= (M+1)^{-1} \sum_{x=0}^{M} \hat{d}_x, & 1 \le i \le n. \end{aligned}$$

We compare the fits obtained with the two different estimations methods in Figure 1, where the entries of the average age-specific mortality and fertility rate vector $\hat{\boldsymbol{d}}$ and $\hat{\boldsymbol{b}}$ are represented by the blue stars. In order to assess the accuracy of the parameter estimates, we re-sampled the data 50 times from the true model, and we show in Figure 2 the mean curves compared to the real ones, as well as the corresponding pointwise confidence intervals. On both figures, we see that, as expected, the fits corresponding to the individual population data are much closer to the real model than those corresponding to the global population data.

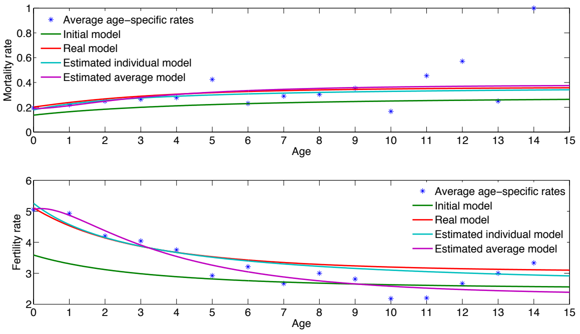


Figure 1: Comparison of the model fits obtained using global population data (estimated average model) and individual population data (estimated individual model). The initial model is the one used as a seed in the optimisation algorithms

In our next experiments, we aim to find the optimal value of the number of phases n by using different criteria and validation techniques. This is still on-going work.

6. REFERENCES

[1] A. C. Davison and N. I. Ramesh. Some models for discretized series of events. *Journal of the American Statistical Association*, 91(434): 601–609, 1996.

[2] S. Hautphenne and G. Latouche. The Markovian binary tree applied to demography. *Journal of mathematical biology*, 64(7):1109–1135, 2012.

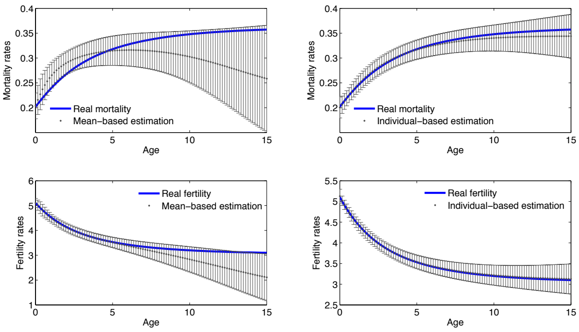


Figure 2: Mean and pointwise confidence intervals of the model fits corresponding to 50 simulations from the real model

[3] G. Latouche, M.-A. Remiche, and P. Taylor. Transient Markov arrival processes. *Annals of Applied Probability*, pages 628–640, 2003.

[4] X.S. Lin and X. Liu. Markov aging process and phase-type law of mortality. *North American Actuarial Journal*, 11:92–109, 2007.

[5] D. Merton. The Chatham Island Black Robin. *Forest and bird*, 21(3):14–19, 1990.

[6] N. I. Ramesh. Statistical analysis on Markov-modulated poisson processes. *Environmetrics*, 6(2):165–179, 1995.

[7] T. Rydén. Parameter estimation for Markov modulated Poisson processes. *Stochastic Models*, 10(4):795–829, 1994.

Lattice and Non-Lattice Markov Additive Models

Jevgenijs Ivanovs
Department of Actuarial Science
University of Lausanne
CH-1015 Lausanne
Switzerland
jevgenijs.ivanovs@unil.ch

Guy Latouche
Université Libre de Bruxelles
Département d'Informatique
CP 212, Boulevard du Triomphe
1050 Bruxelles, Belgium
latouche@ulb.ac.be

Peter Taylor
School of Mathematics and Statistics
University of Melbourne
Victoria, 3010
Australia
p.taylor@ms.unimelb.edu.au

ABSTRACT

Dating from the work of Neuts in the 1980s, the field of matrix-analytic methods has been developed to analyse discrete or continuous-time Markov chains with a two-dimensional state space in which the increment of a *level* variable is governed by an auxiliary *phase* variable. More recently, there has been considerable interest in general Markov additive models in which the level variable is continuous, and its dynamics have either bounded or unbounded variation.

From the Markov additive perspective, traditional matrix-analytic models can be viewed as special cases where increments in the level are constrained to be *lattice* random variables. For example, the class of $M/G/1$*-type Markov chains* in which transitions of the level variable are skip-free downward can have increments which are an integral multiple $k \geq -1$ of the distance between levels. The analogue of an $M/G/1$-type Markov chain in the non-lattice context is a one-sided Markov additive process that does not have negative jumps.

In this paper we discuss such one-sided lattice and non-lattice Markov additive processes side by side. Results that are standard in one tradition are interpreted in the other, and new perspectives emerge.

Keywords

One-Sided Markov Additive Processes, Markov-modulated Lévy Processes, Scale matrices.

Monotonicity, convexity and comparability of some functions associated with block-monotone Markov chains and their applications to queueing systems [*]

[Extended Abstract]

Hai-Bo YU
Beijing University of Technology
No.100 Ping Le Yuan Chaoyang District, Beijing
haibo@bjut.edu.cn

ABSTRACT

Motivated by various applications in queueing theory, this paper is devoted to the monotonicity, convexity and comparability of some functions associated with discrete-time or continuous-time denumerable Markov chains with a general block transition matrices. *First*, we introduce the notion of block-increasing convex ordering for probability vectors, and characterize the block-monotone matrices in the sense of the block-increasing ordering and block-increasing convex ordering. *Second*, we provide the general conditions under which we obtain stochastic monotonicity and comparability of Markov chains with general block structure and conditions for functions associated with block structure Markov chains to be monotone, convex, or concave. By using the uniformization technique, the results are extended to the case of continuous-time Markov chains. Those results can be applied to special Markov chain including GI/M/1 type, M/G/1 type and quasi-birth-and-death (QBD) both discrete-time or continuous-time cases. *Third*, the obtained results can be applied to a number of queueing systems such as the discrete-time $GI/Geo/1$ queue, the continuous-time $PH/M/c$ queue, those conditions for some functions of the queue length and phase of arrival to be monotone, convex, or concave are found.

Keywords

Block stochastic ordering; stochastically block-monotone matrices; block-monotone Markov chains; uniformization; queueing systems.

1. INTRODUCTION

One of the more recent researches within the field of matrix-analytic methods has been its generalization to discrete-time bivariate Markov chains. In this area, the classical work on Markov chains of $GI/M/1$ type and $M/G/1$ type by Neuts (for example, Neuts (1981, 1989)) has been well recognized. Many other researchers have also made important contributions in this area(see, e.g., Gail, Hantler and Taylor (1996, 1997); Zhao, Li and Braun (1998) Zhao, Li and Alfa (2000); Neuts (1998)). It is well-known that the transition matrix of the embbeded Markov chains for both $GI/M/1$ queue and $M/G/1$ queue is stochastic monotone. When a Markov chain has a monotone transition matrix, it often leads to new properties or/and makes analysis of the chain easier. For example, the discussion of approximating the stationary distribution of infinite Markov chains becomes easier and unified (Gibson and Seneta (1987)).

Various semi-Markovian queues and their state-dependent extensions can be analyzed through block-structured Markov chains characterized by an infinite number of block matrices, such as level-dependent quasi-birth-and-death processes (LD-QBDs), $M/G/1$-, $GI/M/1$- and $GI/G/1$-type Markov chains (see, e.g., He(2014)). Tweedie (1998) and Liu (2010) study the estimation of error caused by truncating (stochastically) monotone Markov chains (see, e.g., Daley (1968)). Tweedie (1998) presents error bounds for the last-column-augmented truncation of a monotone Markov chain with geometric ergodicity.

Unfortunately, block-structured Markov chains are not monotone in general. Li and Shaked (1994) introduced the notation of the stochastic block-monotone of stochastic vectors based on $\mathcal{F}$-orderings. Li and Zhao (2000) extend the notion of monotonicity to block-structured Markov chains. The new notion is called (stochastic) block-monotonicity. Block-monotone Markov chains (BMMCs) arise from queues in Markovian environments, such as queues with batch Markovian arrival process (BMAP) (Lucantoni 1991). Li and Zhao (2000) prove that if an original Markov chain is block-monotone, then the stationary distributions of its augmented truncations converge to that of the original Markov chain, which motivates this study. Masuyama (2015) provide error bounds for augmented truncations of discrete-time block-monotone Markov chains under geometric drift conditions; Masuyama (2016) consider continuous-time block-monotone Markov chains and their block-argument truncations.

In this paper, we first give the notion of stochastically

[*]Supported by the National Social Science Fund Mega Project of China(No.11&ZD40) and the National Natural Science Fund Emergency Management Project of China(No.11541006)

block-increasing ordering of stochastic vectors introduced by Li and Zhao (2000) and introduced the notion of stochastically block-increasing convex ordering for stochastic vectors, provide some general conditions to obtain stochastic monotonicity and comparability of block structure Markov chain and conditions for functions associated with block-montone Markov chains to be monotone, convex, or concave.

Throughout this the paper,we use $\mathbb{N}_0$ to denote the set of nonnegative integers and $\mathbb{R}_+$ for the set of nonnegative real numbers. We use $\boldsymbol{a}$ to denote a row vector and $\boldsymbol{a}^T$ is a column vector. e is a column vector with all components equal to 1. $\boldsymbol{0}$ denotes a row vector when its order is clear from the context. We use $c(i)$ to denote the ith component of a vector $\boldsymbol{c}$ and $c(i,j)$ to denote the (i,j)th entry of a matrix C. For two row vectors $\boldsymbol{c}$ and $\boldsymbol{d}$ of dimension m, let $\boldsymbol{c} = (c(1), c(2), \ldots, c(m))$ and $\boldsymbol{d} = (d(1), d(2), \ldots, d(m))$, if $c(i) \le d(i)$ hold for all $i = 1, 2, \ldots, m$, we say $\boldsymbol{c} \le \boldsymbol{d}$ element-wise, denoted by $\boldsymbol{c} \le_{el} \boldsymbol{d}$. Similarly, for two $m \times m$ dimension matrices $C = (c(i,j))$ and $D = (d(i,j))$, if $c(i,j) \le d(i,j)$ hold for all $i, j \in \{1, 2, \ldots, m\}$, we say $C \le D$ element-wise, denoted by $C \le_{el} D$. The terms "increasing" and "decreasing" mean "non-decreasing" and "non-increasing", respectively. For convenience, we shall use $A \Longrightarrow B$ to denote that A implies B; $A \Longleftrightarrow B$ to denote that A is equivalent to B.

2. BLOCK-MONOTONICITY AND CONVEXITY PROPERTIES

2.1 Block-Monotonicity of Probability Vectors and Stochastic Matrices

In this section, we provide a notion of stochastically block-increasing ordering and introduce a notion of stochastically block-increasing convex ordering for vectors partitioned into blocks. We give the concept of block-monotonicity stochastic dominance and discuss properties of probability vectors and stochastic matrices in sense of two class of block-monotone stochastic orderings.

Let $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$, $\mathbb{D}_m = \{1, 2, \ldots, m\}$. Denoted by I_m the $m \times m$ identity matrix and I the identity matrix when its order is clear from the context. Let O denote a zero matrix of dimension $m \times m$. Furthermore, let

$$\boldsymbol{E}_m = \begin{pmatrix} I_m & O & O & O & \cdots \\ I_m & I_m & O & O & \cdots \\ I_m & I_m & I_m & O & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}. \tag{1}$$

It is easy to see that

$$\boldsymbol{E}_m^{-1} = \begin{pmatrix} I_m & O & O & O & \cdots \\ -I_m & I_m & O & O & \cdots \\ O & -I_m & I_m & O & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}, \tag{2}$$

and

$$\boldsymbol{E}_m^{-2} = \begin{pmatrix} I_m & O & O & O & \cdots \\ -2I_m & I_m & O & O & \cdots \\ I_m & -2I_m & I_m & O & \cdots \\ O & I_m & -2I_m & O & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}. \tag{3}$$

The matrix $\boldsymbol{E}_m$ in Eq.(1) and its inverses in Eq.(2) and (3) will be used throughout the paper to simplify notation.

Stochastic block-ordering is defined based on $\mathcal{F}$-orderings (see, e.g.„ Li and Shaked (1994)). This definition includes stochastic matrices for study and also leads to a mathematically tractable analysis. Throughout the paper, let $\mathcal{P}(\mathbb{N}_0)$ denote the set of all probability vectors defined on $\mathbb{N}_0$. A vector $\boldsymbol{a} \in \mathcal{P}(\mathbb{N}_0)$ is called to be a probability vector with block size m, if it can be written as $\boldsymbol{a} = (\boldsymbol{a}_0, \boldsymbol{a}_1, \boldsymbol{a}_2, \ldots)$, where $\boldsymbol{a}_k = (a_{k,1}, a_{k,2}, \ldots, a_{k,m})$ are probability vectors having m elements. Similarly, a stochastic matrix $\boldsymbol{P} = (A_{k,l})$ is called to be with block size m, where all entries $A_{k,l}$ are sub-matrices of size $m \times m$, $k, l \in \mathbb{N}_0$.

We first provide the definitions associated with the block monotonicity and block-wise dominance relation for probability vectors and stochastic matrices in the sense of block-increasing ordering.

Definition 2.1 Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two probability vectors of block size m defined on $\mathcal{P}(\mathbb{N}_0)$; $\boldsymbol{P} = (A_{k,l})$ is a stochastic matrix of block size m. $\boldsymbol{E}_m$ and $\boldsymbol{E}_m^{-1}$ are given in Eq.(1) and (2) respectively. $\boldsymbol{O}$ be a zero matrix with block size m.

(i) (Masuyama (2016) Definition 2.1) $\boldsymbol{a}$ is said to be stochastically block-less than $\boldsymbol{b}$ with block size m (written as $\boldsymbol{a} \le_{m-st} \boldsymbol{b}$) if $\sum_{k=v}^{\infty} \boldsymbol{a}_k \le_{el} \sum_{k=v}^{\infty} \boldsymbol{b}_k$(i.e., $\sum_{k=v}^{\infty} a_{k,j} \le \sum_{k=v}^{\infty} b_{k,j}$, $j = 1, 2, \ldots, m$) for all $v \in \mathbb{N}_0$, or equivalently, $\boldsymbol{a}\boldsymbol{E}_m \le_{el} \ \boldsymbol{b}\boldsymbol{E}_m$.

(ii) (Masuyama (2015) Definition 2.1 and Proposition 2.1) $\boldsymbol{P}$ is said to be block-monotone with respect to block-increasing ordering $\le_{m-st}$ (denoted by $\boldsymbol{P} \in \mathcal{M}_{m-st}$) if $\sum_{l=v}^{\infty} A_{k,l} \le_{el} \sum_{l=v}^{\infty} A_{k+1,l}$ for all $k, v \in \mathbb{N}_0$, or equivalently, $\boldsymbol{E}_m^{-1} \ \boldsymbol{P}\boldsymbol{E}_m \ge \boldsymbol{O}$.

(iii) Let $\boldsymbol{P} = (A_{k,l})$ and $\tilde{\boldsymbol{P}} = (\tilde{A}_{k,l})$ are two block-increasing stochastic matrices of block size m, i.e., $\boldsymbol{P}, \tilde{\boldsymbol{P}} \in \mathcal{M}_{m-st}$. $\boldsymbol{P}$ is said to be less than $\tilde{\boldsymbol{P}}$ with respect to block-increasing ordering $\le_{m-st}$ (denoted by $\boldsymbol{P} \le_{m-st} \tilde{\boldsymbol{P}}$) if $\sum_{l=v}^{\infty} A_{k,l} \le_{el} \sum_{l=v}^{\infty} \tilde{A}_{k,l}$ for all $k, l \in \mathbb{N}_0$, or equivalently, $\boldsymbol{P}\boldsymbol{E}_m \le \tilde{\boldsymbol{P}}\boldsymbol{E}_m$.

Remark 2.1 It is easy to show that $\boldsymbol{a} \le_{m-st} \boldsymbol{b}$ if and only if $\sum_{k=0}^{v} \boldsymbol{a}_k \ge_{el} \sum_{k=0}^{v} \boldsymbol{b}_k$ for $v \in \mathbb{N}_0$, i.e., $\boldsymbol{a}\boldsymbol{E}_m^T \ge_{el} \ \boldsymbol{b}\boldsymbol{E}_m^T$. $\boldsymbol{P} \le_{m-st} \tilde{\boldsymbol{P}}$ if and only if $A_{i,\cdot} \le_{m-st} \tilde{A}_{i,\cdot}$, where $A_{i,\cdot}$ and $\tilde{A}_{i,\cdot}$ denote the i row of matrices $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ respectively.

The following example shows that the block-increasing ordering of probability vectors can be obtained by changing their elements with block size m.

Example 2.1 Let $\boldsymbol{a} = (\boldsymbol{a}_0, \boldsymbol{a}_1, \boldsymbol{a}_2, \ldots)$ be a probability vector with block size m. Then for $k = 0, 1, 2, \ldots$, we have

(i) $\boldsymbol{b} \le_{m-st} \boldsymbol{a}$, where $\boldsymbol{b} = (\sum_{j=0}^{k} \boldsymbol{a}_j, \boldsymbol{a}_{k+1}, \boldsymbol{a}_{k+2}, \ldots)$.

(ii) $\boldsymbol{b} \le_{m-st} \boldsymbol{a}$, where $\boldsymbol{b} = (\boldsymbol{a}_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_k, \sum_{j=k+1}^{\infty} \boldsymbol{a}_j, \boldsymbol{0}, \boldsymbol{0}, \ldots)$.

(iii) $\boldsymbol{a}^{[k]} \le_{m-st} \boldsymbol{a}^{[k+1]}$, where $\boldsymbol{a}^{[k]} = (\boldsymbol{0}, \boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{a}_1, \boldsymbol{a}_2, \ldots)$ be a block probability vectors of size m which there is k zero vector before $\boldsymbol{a}_1$.

Probability vectors and stochastic matrices under the block-increasing ordering can be characterized by corresponding block-monotone functions introduced by Li and Zhao (2000): A real function $\boldsymbol{f}(\cdot)$ defined on $\mathbb{N}_0$ can be written as a row vector with block size m, that is, $\boldsymbol{f} = (\boldsymbol{f}_0, \boldsymbol{f}_1, \ldots)$, where $\boldsymbol{f}_k$ is a row vector of dimension m for $k \in \mathbb{N}_0$. The function $\boldsymbol{f}$ is said to be block-increasing with block size m if

$\boldsymbol{f}_{k+1} - \boldsymbol{f}_k \geq_{el} \mathbf{0}$ for all $k \in \mathbb{N}_0$, equivalently, $\boldsymbol{E}_m^{-1}\boldsymbol{f}^T \geq_{el} \mathbf{0}^T$. Let $\mathcal{F}_{m-st}$ denote the set of all block-increasing functions with block size m.

It is obvious that every increasing function is block-increasing with any block size m, but $\boldsymbol{f}$ in Definition 2.1 (i) is not necessary an increasing function. For example, let $\boldsymbol{a} = (2, 1, 1)$ and $\boldsymbol{b} = (2, 3, 3)$ be two sub-vectors of size 3, then $\boldsymbol{f} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{b}, \ldots) = (2, 1, 1, 2, 3, 3, 2, 3, 3, \ldots)$ is a block-increasing function with block size 3. Note that $\boldsymbol{f}$ is not an increasing function.

Proposition 2.1 Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two probability vectors of block size m defined on $\mathcal{P}(\mathbb{N}_0)$. Then the following statements are equivalent:

(i) $\boldsymbol{a} \leq_{m-st} \boldsymbol{b}$.

(ii) $\boldsymbol{a}\boldsymbol{f}^T \leq \boldsymbol{b}\boldsymbol{f}^T$ for all $\boldsymbol{f} \in \mathcal{F}_{m-st}$.

Proof. *(i)* $\Leftarrow$ *(ii)*: Taking $\boldsymbol{f} = \boldsymbol{t}_j^{[0]}, \boldsymbol{t}_j^{[1]}, \boldsymbol{t}_j^{[2]}, \ldots$ in part (ii), for $j = 1, 2, \ldots, m$, where $\boldsymbol{t}_j^{[k]} = (\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{e}_j^T, \boldsymbol{e}_j^T, \ldots)$ is a probability vectors with block size m which there is k zero vectors before the first $\boldsymbol{e}_j^T$, $\boldsymbol{e}_j^T = (0, \ldots, 0, 1, 0, \ldots)$ is a m-dimension vector which its jth element is equal to 1, and otherwise are equal to zero for $j = 1, 2, \ldots, m$ and $k = 0, 1, 2, \ldots$. Combining Definition 2.1 (i) and Eq.(1), which leads to the desired result.

(i) $\Rightarrow$ *(ii)*: By Definition 2.1(i), $\boldsymbol{a} \leq_{m-st} \boldsymbol{b}$ implies that $\boldsymbol{a}\ {\boldsymbol{t}_j^{[k]}}^T \leq \boldsymbol{b}\ {\boldsymbol{t}_j^{[k]}}^T$ follows for all $j = 1, 2, \ldots, m$ and $k = 0, 1, , 2, \ldots$, where $\boldsymbol{t}_j^{[k]}$ are given by above for $j = 1, 2, \ldots, m$. By using the fact that any block-increasing function $\boldsymbol{f}$ can be expressed as combination of functions $\boldsymbol{t}_j^{[k]}$ for $j = 1, 2, \ldots, m$ and $k = 0, 1, , 2, \ldots$, which leads to the desired result. □

Proposition 2.2(Masuyama (2015)Proposition 2.2) Suppose that $\boldsymbol{a}$ and $\boldsymbol{b}$ be two block probability vectors of size m in $\mathcal{P}(\mathbb{N}_0)$, $\boldsymbol{P}$ is a stochastic matrix with block size m. Then the following statements are equivalent:

(i) $\boldsymbol{P}$ is block-monotone with respect to ordering $\leq_{m-st}$.

(ii) $\boldsymbol{a} \leq_{m-st} \boldsymbol{b}$ implies $\boldsymbol{a}\boldsymbol{P} \leq_{m-st} \boldsymbol{b}\boldsymbol{P}$.

(iii) $\boldsymbol{P}\boldsymbol{f}^T \in \mathcal{F}_{m-st}$ hold for all $\boldsymbol{f} \in \mathcal{F}_{m-st}$.

Remark 2.2 Block-increasing ordering of probability vectors and stochastic matrices may be defined by Proposition 2.1 (ii) and Proposition 2.2 (ii) respectively see e.g., Definition 2.2 and Definition 2.5 in Li and Zhao (2000).

Remark 2.3 In particular, when $m = 1$, a stochastic matrix $A = (a_{k,l})$, $k, l \in \mathbb{N}_0$ satisfying $E_1^{-1} A E_1 \geq 0$ is called stochastically monotone in the sense of usual stochastic ordering $\leq_{st}$(see, e.g., Keilson J, Kester A. 1977 pp.233 Eq.(1.2)).

Stochastic matrices in sense of stochastically block-increasing ordering have the following basic properties(see, Proposition 2.10 and Proposition 2.11 in Li and Zhao (2000)).

Lemma 2.1 Let $\mathcal{M}_{m-st}$ denote the set of the block-monotone matrix with respect to $\leq_{m-st}$, $\boldsymbol{P} \in \mathcal{M}_{m-st}$ means $\boldsymbol{P}$ is block-monotone with respect to $\leq_{m-st}$.

(i) If $\boldsymbol{A} \in \mathcal{M}_{m-st}$ and $\boldsymbol{B} \in \mathcal{M}_{m-st}$, then
(a) $\alpha\boldsymbol{A} + (1-\alpha)\boldsymbol{B} \in \mathcal{M}_{m-st}$ for $0 \leq \alpha \leq 1$.
(b) $\boldsymbol{A}\boldsymbol{B} \in \mathcal{M}_{m-st}$ and $\boldsymbol{A}^n \in \mathcal{M}_{m-st}$ for $n = 0, 1, 2, \ldots$.

(ii) Suppose that $\boldsymbol{A}$ and $\boldsymbol{B}$ are stochastic matrices with $\boldsymbol{A} \leq_{m-st} \boldsymbol{B}$. If either $\boldsymbol{A} \in \mathcal{M}_{m-st}$ or $\boldsymbol{B} \in \mathcal{M}_{m-st}$, then $\boldsymbol{A}^n \leq_{m-st} \boldsymbol{B}^n$ for $n = 1, 2, \ldots$.

Block-monotone matrices with respect to block-increasing ordering have the following monotonicity and comparability, it can be proved by using similar approach as Proposition 2.1, we omit its proofs.

Theorem 2.1 Let $\mathcal{F}_{m-st}$ denote the set of the block-increasing functions, $\mathcal{M}_{m-st}$ denote the set of the block-monotone matrix with respect to $\leq_{m-st}$, and let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two block probability vectors of size m in $\mathcal{P}(\mathbb{N}_0)$. Assume that $\boldsymbol{P}$ is block-monotone with respect to $\leq_{m-st}$.

(i) If $\boldsymbol{a}\ \leq_{m-st}\ \boldsymbol{b}$, then $\boldsymbol{a}\ \boldsymbol{P}^n \leq_{m-st}\ \boldsymbol{b}\ \boldsymbol{P}^n$ for all $n \in \mathbb{N}_0$.

(ii) If $\boldsymbol{a}\ \leq_{m-st}\ \boldsymbol{a}\boldsymbol{P}$, then $\boldsymbol{a}\ \leq_{m-st}\ \boldsymbol{a}\boldsymbol{P}^n$ for all $n \in \mathbb{N}_0$; If $\boldsymbol{a}\boldsymbol{P}\ \leq_{m-st}\ \boldsymbol{a}$, then $\boldsymbol{a}\boldsymbol{P}^n\ \leq_{m-st}\ \boldsymbol{a}$ for all $n \in \mathbb{N}_0$.

(iii) Suppose that $\boldsymbol{\pi}$ is the steady-state distribution of $\boldsymbol{P}$ (i,e,. $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}$ and $\boldsymbol{\pi}\boldsymbol{e}^T = 1$). If $\boldsymbol{a}\ \leq_{m-st}\ \boldsymbol{a}\boldsymbol{P}$, then $\boldsymbol{a}\ \leq_{m-st}\ \boldsymbol{\pi}$; if $\boldsymbol{a}\boldsymbol{P}\ \leq_{m-st}\ \boldsymbol{a}$, then $\boldsymbol{\pi}\ \leq_{m-st}\ \boldsymbol{a}$.

(iv) Let $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ are two block-increasing stochastic matrices of block size m, i.e., $\boldsymbol{P} \in \mathcal{M}_{m-st}$ and $\tilde{\boldsymbol{P}} \in \mathcal{M}_{m-st}$. Then (a) $\boldsymbol{P}\tilde{\boldsymbol{P}} \in \mathcal{M}_{m-st}$; (b) $\boldsymbol{P}^n \in \mathcal{M}_{m-st}$; (c) $\sum_{k=0}^{\infty} w_k \boldsymbol{P}^k \in \mathcal{M}_{m-st}$ for $0 \leq w_k \leq 1$, $\sum_{k=0}^{\infty} w_k = 1$. In particularly, $\sum_{k=0}^{\infty} w_k \boldsymbol{P}^k \in \mathcal{M}_{m-st}$ holds for $w_k = e^{-\theta}\theta^k/k!$, $\theta > 0$.

Proof. Part (i) holds by Proposition 2.2 (ii). Proof of part (ii) is obtained by part (i). Part (iii) holds by part (ii).

Part (iv)(a): by Proposition 2.2 (ii) and under the assumption, we have $\tilde{\boldsymbol{P}}\boldsymbol{f}^T \in \mathcal{F}_{m-st}$ hold for all $\boldsymbol{f} \in \mathcal{F}_{m-st}$. By Proposition 2.2 (ii), $\boldsymbol{P}\tilde{\boldsymbol{P}}\boldsymbol{f}^T \in \mathcal{F}_{m-st}$ hold for all $\boldsymbol{f} \in \mathcal{F}_{m-st}$, that is $\boldsymbol{P}\tilde{\boldsymbol{P}} \in \mathcal{M}_{m-st}$.

Parts (iv)(b) and (c): by Lemma 2.1 and under the assumption, which leads to the desired result. □

Secondly, we introduce the definitions associated with the block monotonicity and block-wise dominance relation for probability vectors and stochastic matrices in the sense of block-increasing convex ordering.

Definition 2.2 Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two probability vectors of block size m defined on $\mathcal{P}(\mathbb{N}_0)$; $\boldsymbol{P} = (A_{k,l})$ is a stochastic matrix of block size m. $\boldsymbol{E}_m$ and $\boldsymbol{E}_m^{-1}$ are given in Eq.(1) and (2) respectively. $\boldsymbol{O}$ be a zero matrix with block size m.

(i) $\boldsymbol{a}$ is said to be less than $\boldsymbol{b}$ in the sense of block-increasing convex ordering with block size m (written as $\boldsymbol{a} \leq_{m-icx} \boldsymbol{b}$) if $\sum_{l=v}^{\infty}\sum_{k=l}^{\infty} \boldsymbol{a}_k \leq_{el} \sum_{l=v}^{\infty}\sum_{k=l}^{\infty} \boldsymbol{b}_k$ for all $v \in \mathbb{N}_0$, or equivalently, $\boldsymbol{a}\boldsymbol{E}_m^2 \leq_{el}\ \boldsymbol{b}\boldsymbol{E}_m^2$.

(ii) $\boldsymbol{P}$ is said to be block-monotone with respect to block-increasing convex ordering $\leq_{m-icx}$ (denoted by $\boldsymbol{P} \in \mathcal{M}_{m-icx}$) if $\sum_{s=v}^{\infty}\sum_{l=s}^{\infty} A_{k,l} \leq_{el} \sum_{s=v}^{\infty}\sum_{l=s}^{\infty} A_{k+1,l}$ for all $k, v \in \mathbb{N}_0$, or equivalently, $\boldsymbol{E}_m^{-1}\ \boldsymbol{P}\boldsymbol{E}_m^2 \geq \boldsymbol{O}$.

(iii) Let $\boldsymbol{P} = (A_{k,l})$ and $\tilde{\boldsymbol{P}} = (\tilde{A}_{k,l})$ are two block-increasing convex stochastic matrices of block size m, i.e., $\boldsymbol{P} \in \mathcal{M}_{m-icx}$ and $\tilde{\boldsymbol{P}} \in \mathcal{M}_{m-icx}$. $\boldsymbol{P}$ is said to be less than $\tilde{\boldsymbol{P}}$ with respect to block-increasing convex ordering $\leq_{m-st}$ (denoted by $\boldsymbol{P} \leq_{m-icx} \tilde{\boldsymbol{P}}$) if $\sum_{s=v}^{\infty}\sum_{l=s}^{\infty} A_{k,l} \leq_{el} \sum_{s=v}^{\infty}\sum_{l=s}^{\infty} \tilde{A}_{k+1,l}$ for all $k, l \in \mathbb{N}_0$, or equivalently, $\boldsymbol{P}\boldsymbol{E}_m^2 \leq \tilde{\boldsymbol{P}}\boldsymbol{E}_m^2$.

Now we define a class of functions called block-increasing convex function which will be used to compare two probability vectors with block structure, and characterize the block-monotonicity of stochastic matrices.

Definition 2.3 A real function $\boldsymbol{f}(\cdot)$ defined on $\mathbb{N}_0$ can be written as a row vector with block size m, that is, $\boldsymbol{f} = (\boldsymbol{f}_0, \boldsymbol{f}_1, \ldots)$, where $\boldsymbol{f}_k$ is a row vector of dimension m for $k \in \mathbb{N}_0$. The function $\boldsymbol{f}$ is said to be block-convex with block size m if $\boldsymbol{f}_{k+2} + \boldsymbol{f}_k - 2\boldsymbol{f}_{k+1} \geq_{el} \boldsymbol{0}$ for all $k \in \mathbb{N}_0$, equivalently, $\boldsymbol{E}_m^{-2}\boldsymbol{f}^T \geq_{el} \boldsymbol{0}^T$.

Block monotone probability vectors and stochastic matrices under the block-increasing convex ordering can be characterized by corresponding block-increasing convex functions. Let $\mathcal{F}_{m-icx}$ denote the set of the block-increasing convex functions. The following Proposition 2.3 and Proposition 2.4 provide a necessary and sufficient condition to compare two probability vectors or two stochastic matrices. Proposition 2.3 and Proposition 2.4 can be proved by using similar approach as Proposition 2.1, we omit its proofs.

Proposition 2.3 Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two probability vectors of block size m defined on $\mathcal{P}(\mathbb{N}_0)$. Then the following statements are equivalent:

(i) $\boldsymbol{a} \leq_{m-icx} \boldsymbol{b}$.

(ii) $\boldsymbol{a}\boldsymbol{f}^T \leq \boldsymbol{b}\boldsymbol{f}^T$ for all $\boldsymbol{f} \in \mathcal{F}_{m-icx}$.

Proposition 2.4 Suppose that $\boldsymbol{a}$ and $\boldsymbol{b}$ be two block probability vectors of size m in $\mathcal{P}(\mathbb{N}_0)$, $\boldsymbol{P}$ is a stochastic matrix with block size m. The following statements are equivalent:

(i) $\boldsymbol{P}$ is block-monotone with respect to order $\leq_{m-icx}$.

(ii) $\boldsymbol{a} \leq_{m-icx} \boldsymbol{b}$ implies $\boldsymbol{a}\boldsymbol{P} \leq_{m-icx} \boldsymbol{b}\boldsymbol{P}$.

(iii) $\boldsymbol{P}\boldsymbol{f}^T \in \mathcal{F}_{m-icx}$ hold for all $\boldsymbol{f} \in \mathcal{F}_{m-icx}$.

Block-monotone matrices with respect to block-increasing ordering have the following monotonicity and comparability, its can be proved by using similar approach as Theorem 2.1, we omit its proofs.

Theorem 2.2 Let $\mathcal{F}_{m-icx}$ denote the set of the block-increasing functions, $\mathcal{M}_{m-icx}$ denote the set of the block-monotone matrix with respect to $\leq_{m-icx}$, and let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two block probability vectors of size m in $\mathcal{P}(\mathbb{N}_0)$. Assume that $\boldsymbol{P}$ is block-monotone with respect to $\leq_{m-icx}$.

(i) If $\boldsymbol{a} \leq_{m-icx} \boldsymbol{b}$, then $\boldsymbol{a}\,\boldsymbol{P}^n \leq_{m-icx} \boldsymbol{b}\,\boldsymbol{P}^n$ for all $n \in \mathbb{N}_0$.

(ii) If $\boldsymbol{a} \leq_{m-icx} \boldsymbol{a}\boldsymbol{P}$, then $\boldsymbol{a} \leq_{m-icx} \boldsymbol{a}\boldsymbol{P}^n$ for all $n \in \mathbb{N}_0$; If $\boldsymbol{a}\boldsymbol{P} \leq_{m-st} \boldsymbol{a}$, then $\boldsymbol{a}\boldsymbol{P}^n \leq_{m-st} \boldsymbol{a}$ for all $n \in \mathbb{N}_0$.

(iii) Suppose that $\boldsymbol{\pi}$ is the steady-state distribution of $\boldsymbol{P}$ (i,e,. $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}$ and $\boldsymbol{\pi}\boldsymbol{e}^T = 1$). If $\boldsymbol{a} \leq_{m-icx} \boldsymbol{a}\boldsymbol{P}$, then $\boldsymbol{a} \leq_{m-icx} \boldsymbol{\pi}$; if $\boldsymbol{a}\boldsymbol{P} \leq_{m-icx} \boldsymbol{a}$, then $\boldsymbol{\pi} \leq_{m-icx} \boldsymbol{a}$.

(iv) Let $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ are two block-increasing stochastic matrices of block size m, i.e., $\boldsymbol{P} \in \mathcal{M}_{m-icx}$ and $\tilde{\boldsymbol{P}} \in \mathcal{M}_{m-icx}$. Then (a) $\boldsymbol{P}\tilde{\boldsymbol{P}} \in \mathcal{M}_{m-icx}$; (b) $\boldsymbol{P}^n \in \mathcal{M}_{m-icx}$; (c) $\sum_{k=0}^{\infty} w_k \boldsymbol{P}^k \in \mathcal{M}_{m-icx}$ for $0 \leq w_k \leq 1$, $\sum_{k=0}^{\infty} w_k = 1$. In particularly, $\sum_{k=0}^{\infty} w_k \boldsymbol{P}^k \in \mathcal{M}_{m-icx}$ holds for $w_k = e^{-\theta}\theta^k/k!$, $\theta > 0$.

Example 2.2 Consider a discrete-time Markov chain $\boldsymbol{Z} = \{Z_n,\ n \in \mathbb{N}_0\}$, it is a countable state Markov chain whose transition matrix has the following block structure: For $k = 0, 1, 2, \ldots$,

$$\boldsymbol{P} = \begin{pmatrix} \sum_{i=0}^{k} A_i & A_{k+1} & A_{k+2} & A_{k+3} & A_{k+4} & \cdots \\ A_0 & A_1 & A_2 & A_3 & A_4 & \cdots \\ O & A_0 & A_1 & A_2 & A_3 & \cdots \\ O & O & A_0 & A_1 & A_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}, \quad (4)$$

where A_l, $l = 1, 2, 3$, are all $m \times m$ non-negative matrices with $\sum_{j=1}^{3} A_j \boldsymbol{e}^T = \boldsymbol{e}^T$, $\boldsymbol{e}$ is a m row vector which consists of ones, O is a zero matrix of order $m \times m$. Let B_j denote the jth block row of $\boldsymbol{P}$, which is a matrix of size $m \times \infty$. Using Example 2.1(i), we have $B_1\boldsymbol{f}^T \leq B_2\boldsymbol{f}^T$ for all block-increasing function $\boldsymbol{f} \in \mathcal{F}_{m-st}$. From Example 2.1(iii), it follows that for any $j \geq 2$, $B_j\boldsymbol{f}^T \leq B_{j+1}\boldsymbol{f}^T$ for all block-increasing function $\boldsymbol{f} \in \mathcal{F}_{m-st}$. Therefore, $\boldsymbol{P}\boldsymbol{f}^T$ is block-monotone with respect to order $\leq_{m-st}$, by Proposition 2.2, $\boldsymbol{P}$ is block-monotone with respect to order $\leq_{m-st}$.

In particular, taking $k = 1$, Li and Zhao (2000 Example 2.8) showed that $\boldsymbol{P}$ in Eq.(4) is block-monotone with respect to order $\leq_{m-st}$.

Example 2.3 Consider a discrete-time Markov chain $\boldsymbol{Z} = \{Z_n,\ n \in \mathbb{N}_0\}$, it is a countable state Markov chain whose transition matrix has the following block structure:

$$\boldsymbol{P} = \begin{pmatrix} I_m - B_0 & B_0 & O & O & O & \cdots \\ I_m - \sum_{i=0}^{1} B_i & B_1 & B_0 & O & O & \cdots \\ I_m - \sum_{i=0}^{2} B_i & B_2 & B_1 & B_0 & O & \cdots \\ I_m - \sum_{i=0}^{3} B_i & B_3 & B_2 & B_1 & B_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}, \quad (5)$$

where B_k, $k = 1, 2, \ldots$, are all $m \times m$ non-negative matrices, O is a zero matrix of order $m \times m$. By Definition 2.1 and 2.2, $\boldsymbol{P}$ in (5) is block monotone with respect to order $\leq_{m-st}$ and is block-monotone with respect to order $\leq_{m-icx}$.

Example 2.4 Consider a discrete-time Markov chain $\boldsymbol{Z} = \{Z_n,\ n \in \mathbb{N}_0\}$, it is a countable state Markov chain whose transition matrix has the following block structure:

$$\boldsymbol{P} = \begin{pmatrix} B & C & O & O & O & \cdots \\ D & A_1 & A_0 & O & O & \cdots \\ O & A_2 & A_1 & A_0 & O & \cdots \\ O & O & A_2 & A_1 & A_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}, \quad (6)$$

where B, C, D and A_k, $k = 1, 2, 3$, are all $m \times m$ non-negative matrices with $\sum_{k=1}^{3} A_k \boldsymbol{e}^T = \mathbf{e}^T$, $\boldsymbol{e}$ is a m row vector which consists of ones, O is a zero matrix of order $m \times m$. By Definition 2.1 and 2.2, we have

(i) If $B \geq_{el} D$, $B + C \geq_{el} D + A_1$ and $D + A_1 \geq_{el} A_2$, then $\boldsymbol{P}$ in (6) is block-monotone with respect to order $\leq_{m-st}$.

(ii) If $A_1 + 2A_0 \geq_{el} C$, $B + C \leq_{el} D + A_1 + A_2 \leq_{el} 2(A_0 + A_1 + A_2)$, then $\boldsymbol{P}$ in (6) is block-monotone with respect to order $\leq_{m-icx}$.

2.2 Block-Monotone Discrete-Time Markov Chains

Consider a discrete-time Markov chain $\boldsymbol{Z} = \{Z_n,\ n \in \mathbb{N}_0\}$ with countable state space $\mathcal{E} = \{(k, i),\ k \in \mathbb{N}_0,\ i = 1, 2, \ldots, m\}$, whose transition matrix $\boldsymbol{P}$ has the block structure of block size m, i.e., $\boldsymbol{P} = (A_{k,l})$, $k, l \in \mathbb{N}_0$, where $A_{k,l}$ are all $m \times m$ matrices. Let $\boldsymbol{v}$ be the initial probability distribution of the Markov chain $\boldsymbol{Z}$.

In general, to ensure stochastically monotonicity of Markov chain $\boldsymbol{Z}$, certain conditions must be imposed on the transition matrix $\boldsymbol{P}$ and the initial distribution $\boldsymbol{v}$.

$$\begin{array}{ll} \text{Condition I}_{m-st}: & \boldsymbol{P} \in \mathcal{M}_{m-st}, \\ \text{Condition I}_{m-icx}: & \boldsymbol{P} \in \mathcal{M}_{m-icx}, \\ \text{Condition II}_{m-st}: & \boldsymbol{v} \leq_{m-st} \boldsymbol{vP}, \\ \text{Condition II}_{m-icx}: & \boldsymbol{v} \leq_{m-icx} \boldsymbol{vP}. \end{array} \tag{7}$$

Under the above conditions, the following Theorem 3.1 shows the stochastically monotonicity of block-monotone Markov chain, Theorem 3.2 provide stochastic comparison results for two block-monotone Markov chains, Theorem 3.3 give sufficient conditions to obtain monotonicity and convexity of function associated block-monotone Markov chain. The three theorems also explains why the above conditions are utilized in this paper. Theorem 3.1 and 3.2 can be obtained by Theorem 2.1 and 2.2, we omit their proofs.

Theorem 3.1 Consider a discrete-time Markov chain $\boldsymbol{Z}$ with transition matrix $\boldsymbol{P}$ having the block structure. Let $\boldsymbol{v}$ be the initial distribution of $\boldsymbol{Z}$.

(i) If Conditions I_{m-st} and II_{m-st} in Eq.(7) hold, then $Z_n \leq_{m-st} Z_{n+1}$ for all $n \in \mathbb{N}_0$.

(ii) If Conditions I_{m-icx} and II_{m-icx} in Eq.(7) hold, then $Z_n \leq_{m-icx} Z_{n+1}$ for all $n \in \mathbb{N}_0$.

(iii) Suppose that $\boldsymbol{\pi}$ is steady-state distributions of $\boldsymbol{P}$. If Conditions I_{m-st} and II_{m-st} in Eq.(7) hold, then $\boldsymbol{v} \leq_{m-st} \boldsymbol{\pi}$.

(iv) Suppose that $\boldsymbol{\pi}$ is steady-state distributions of $\boldsymbol{P}$. If Conditions I_{m-icx} and II_{m-icx} in Eq.(7) hold, then $\boldsymbol{v} \leq_{m-icx} \boldsymbol{\pi}$.

Theorem 3.2 Consider two discrete-time Markov chain $\boldsymbol{Z} = \{Z_n,\ n \in \mathbb{N}_0\}$ and $\tilde{\boldsymbol{Z}} = \{\tilde{Z}_n,\ n \in \mathbb{N}_0\}$ with the same state space $\mathcal{E} = \{(k, i),\ k \in \mathbb{N}_0,\ i = 1, 2, \ldots, m\}$, their transition matrices $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ have the block structure. Let $\boldsymbol{v}$ and $\tilde{\boldsymbol{v}}$ be the initial distribution of $\boldsymbol{Z}$ and $\boldsymbol{Z}$ respectively.

(i) If $\boldsymbol{v} \leq_{m-st} \tilde{\boldsymbol{v}}$, $\boldsymbol{P} \leq_{m-st} \tilde{\boldsymbol{P}}$ and one of $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ is block monotone with respect to order $\leq_{m-st}$, then $Z_n \leq_{m-st} \tilde{Z}_n$ for all $n \in \mathbb{N}_0$.

(ii) If $\boldsymbol{v} \leq_{m-icx} \tilde{\boldsymbol{v}}$, $\boldsymbol{P} \leq_{m-icx} \tilde{\boldsymbol{P}}$ and one of $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ is block monotone with respect to order $\leq_{m-icx}$, then $Z_n \leq_{m-icx} \tilde{Z}_n$ for all $n \in \mathbb{N}_0$.

(iii) Suppose that $\boldsymbol{\pi}$ and $\tilde{\boldsymbol{\pi}}$ are steady-state distributions of $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ respectively. If conditions in part (i) hold, then $\boldsymbol{\pi} \leq_{m-st} \tilde{\boldsymbol{\pi}}$.

(iv) Suppose that $\boldsymbol{\pi}$ and $\tilde{\boldsymbol{\pi}}$ are steady-state distributions of $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ respectively. If conditions in part (ii) hold, then $\boldsymbol{\pi} \leq_{m-icx} \tilde{\boldsymbol{\pi}}$.

Define

$$\boldsymbol{h_f} = \boldsymbol{P f}^T - \boldsymbol{f}, \tag{8}$$

for function $\boldsymbol{f}(\cdot)$ defined on $\mathbb{N}_0$ and have a block form with block size m. We state another conditions as follows.

$$\begin{array}{ll} \text{Condition II}'_{m-st}: & \boldsymbol{vP} \leq_{m-st} \boldsymbol{v}, \\ \text{Condition III}_{m-st}: & \boldsymbol{h_f} \text{ in Eq. (8) is block-decreasing,} \\ \text{Condition III}'_{m-st}: & \boldsymbol{h_f} \text{ in Eq. (8) is block-increasing.} \end{array} \tag{9}$$

The following theorem shows the monotonicity, convexity, and concavity of the function $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(Z_n)] = \boldsymbol{v P}^n \boldsymbol{f}^T$ under the above conditions.

Theorem 3.3 Consider a discrete-time Markov chain $\boldsymbol{Z} = \{Z_n,\ n \in \mathbb{N}_0\}$ with state space $\mathcal{E} = \{(k, j),\ k \in \mathbb{Z}_+,\ j = 1, 2, \ldots, m\}$, its transition matrix $\boldsymbol{P}$ has the block structure. We assume that $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(Z_n)]$ is finite for $n \geq 0$.

(i) If Conditions I_{m-st}, II_{m-st} in Eq.(7) and III_{m-st} in Eq.(9) hold and the function $\boldsymbol{f}(\cdot)$ is block-increasing, then the function $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(Z_n)]$ is increasing concave in n.

(ii) If Conditions I_{m-st} in Eq.(7) and II'_{m-st} and III_{m-st} in Eq.(9) hold and the function $\boldsymbol{f}(\cdot)$ is block-increasing, then the function $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(Z_n)]$ is decreasing convex in n.

Proof. Define $\varphi_{\boldsymbol{v}}(n) = \mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(Z_n)] = \boldsymbol{v P}^n \boldsymbol{f}^T$ for $n = 0, 1, 2, \ldots$.

Part (i): under the assumption, we have $\varphi_{\boldsymbol{v}}(n+1) - \varphi_{\boldsymbol{v}}(n) = (\boldsymbol{vP} - \boldsymbol{v})\boldsymbol{P}^n \boldsymbol{f}^T$ for $n = 1, 2, \ldots$. By combining Proposition 2.2 and the assumption, we have $\varphi_{\boldsymbol{v}}(n)$ is increasing in n for $n = 1, 2, \ldots$. Furthermore, we have $\varphi_{\boldsymbol{v}}(n+2) + \varphi_{\boldsymbol{v}}(n) - 2\varphi_{\boldsymbol{v}}(n+1) = (\boldsymbol{vP}^2 + \boldsymbol{v} - 2\boldsymbol{vP})\boldsymbol{P}^n \boldsymbol{f}^T$ for $n = 1, 2, \ldots$. By Proposition 2.2 and the assumption, which leads to the desired result.

Part (ii) can be obtained by similar approach as part (i). □

2.3 Block-Monotone Continuous-Time Markov Chains

Let $\boldsymbol{X} = \{X(t),\ t \in \mathbb{R}_+\}$ be a homogeneous and continuous-time Markov chain with state space $\mathcal{E} = \{(k, i),\ k \in \mathbb{N}_0,\ i \in \mathbb{D}_m\}$, $\mathbb{D}_m = \{1, 2, \ldots, m\}$, its infinitesimal generator $\boldsymbol{Q}$ has the block structure, $\boldsymbol{Q} = (B_{k,l})$, $B_{k,l} = (B_{k,l}(i,j))$ are $m \times m$ squire matrices, $m \geq 1, k, l \in \mathbb{N}_0$, $i, j \in \mathbb{D}_m$. If $\sup_{k \in \mathbb{N}_0, i \in \mathbb{D}_m}(-B_{k,k}(i,i)) < \infty$, then Markov process $\boldsymbol{X}$ is said to be uniformizable ((see Dijk (1990)). Let ξ be equal or greater than the biggest absolute value of all diagonal elements of $\boldsymbol{Q}$, that is, $\xi \geq \sup_{k \in \mathbb{N}_0, i \in \mathbb{D}_m}(-B_{k,k}(i,i))$, we define

$$\boldsymbol{P} = I + \boldsymbol{Q}/\xi, \tag{10}$$

where I is the identity matrix. Note that $\boldsymbol{P}$ defined by Eq.(10) is a stochastic matrix and it has the block structure, i.e., $\boldsymbol{P} = (A_{k,l})$, $A_{k,l}$ are $m \times m$ squire matrices, $m \geq 1, k, l \in \mathbb{N}_0$, $A_{k,l} = I + B_{k,l}/\xi$ for $k, l \in \mathbb{N}_0$. We have:

$$exp\{\boldsymbol{Q}t\} = e^{-\xi t} \sum_{k=0}^{\infty} \frac{(\xi t)^k}{k!} \boldsymbol{P}^k. \tag{11}$$

Counterparts of conditions in Eq.(7) and (9) are given as follows.

$$\begin{array}{ll} \text{Condition I}_{cont}: & I + \boldsymbol{Q}/\xi \text{ is } \leq_{m-st} \text{-block-monotone,} \\ \text{Condition II}_{cont}: & \boldsymbol{v} \leq_{m-st} \boldsymbol{v}(I + \boldsymbol{Q}/\xi), \\ \text{Condition II}'_{cont}: & \boldsymbol{v}(I + \boldsymbol{Q}/\xi) \leq_{m-st} \boldsymbol{v}, \\ \text{Condition III}_{cont}: & \boldsymbol{Q f}^T \text{ is block-decreasing,} \\ \text{Condition III}'_{cont}: & \boldsymbol{Q f}^T \text{ is block-increasing.} \end{array} \tag{12}$$

We state the monotonicity, convexity, and concavity of the function $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(X(t))]$ in the following theorem, it can be proved by combining Theorem 3.3 and Eq.(10), we omit its proof.

Theorem 3.4 Suppose that Markov chain $\boldsymbol{X}$ is uniformizable, and Markov chain $\boldsymbol{Y}$ with transition matrix $\boldsymbol{P}$ given by Eq.(10). Suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same initial distribution $\boldsymbol{v}$. We assume that $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(X(t))]$ is finite for $t \geq 0$.

(i) If Conditions I_{cont} and II_{cont} in Eq.(12) hold and if the function $\boldsymbol{f}$ defined on $\mathbb{N}_0$ is block-increasing (block-decreasing), then the function $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(X(t))]$ is increasing (decreasing) in t.

(ii) If Conditions I_{cont} and II'_{cont} in Eq.(12) hold and if the function $\boldsymbol{f}$ defined on $\mathbb{N}_0$ is block-increasing (block-decreasing), then the function $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(X(t))]$ is decreasing (increasing) in t.

(iii) If Conditions I_{cont}, II_{cont} and III_{cont} in Eq.(12) hold, then the function $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(X(t))]$ is concave in t. If Conditions I_{cont}, II_{cont} and III'_{cont} in Eq.(12) hold, then the function $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(X(t))]$ is convex in t.

(iv) If Conditions I_{cont}, II'_{cont} and III_{cont} in Eq.(12) hold, then the function $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(X(t))]$ is convex in t. If Conditions I_{cont}, II'_{cont} and III'_{cont} in Eq.(12) hold, then the function $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(X(t))]$ is concave in t. □

3. APPLICATIONS TO QUEUEING SYSTEMS

General type distribution, other than the phase types, can be used to describe both inter-arrive and service times. In continuous times it is well know that general distributions encountered in queueing systems can be approximated by continuous time PH distributions. This is also true for the discrete distributions. However, discrete distributions have an added advantage in that if the distribution has a finite support then it can be represented exactly by discrete PH. Alfa (2016) show that this is true by using a general inter-event time X with finite support and a general distribution given as $Pr\{X = j\} = g_j, j = 1, 2, \ldots, m_t$, $m_t < \infty$.

In this section, we show how the block-monotonicity and convexity can help us gain insight into the queue length processes for queueing systems with general arrival or general service time, such as the discrete-time $GI/Geo/1$ queue (see, e.g., Alfa (2016)pp.156-159) and $PH/M/c$ queue (see, e.g., Li and Zhao (2000)Example 4.1).

3.1 The $GI/Geo/1$ queue

Consider a single server $GI/Geo/1$ queue, suppose that the arrivals are of the general independent type with inter-arrival times τ having a probability mass function of $g_j = Pr\{\tau = j\}, j = 1, 2, \ldots, m$, $m < \infty$ whose mean is λ^{-1} and the service times is of the geometric distribution with parameter μ and mean service time μ^{-1}, $\bar{\mu} = 1 - \mu$, $0 < \mu < 1$. According to Alfa (2016) pp.90, the inter-arrival times τ follow a discrete PH distribution with remaining time representation (β, S), where $\beta = (g_1, g_2, \ldots, g_m)$, and the $m \times m$ matrix S has the same form as B in Eq.(14).

Let X_n and J_n be the number of packets in the system and the remaining inter-arrival time at time $n \geq 0$. It is immediately that $\{(X_n, J_n), n \geq 0\}$ is a bivariate discrete time Markov chain with transition matrix $\boldsymbol{P}$ written as

$$\boldsymbol{P} = \begin{pmatrix} B & C & O & O & O & \cdots \\ A_2 & A_1 & A_0 & O & O & \cdots \\ O & A_2 & A_1 & A_0 & O & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{13}$$

where

$$B = \begin{pmatrix} 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \ddots & \end{pmatrix}, C = \begin{pmatrix} g_1 & g_2 & g_3 & \cdots \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \ddots & \end{pmatrix} \tag{14}$$

$$A_2 = \begin{pmatrix} 0 & 0 & 0 & \cdots \\ \mu & 0 & 0 & \cdots \\ 0 & \mu & 0 & \cdots \\ \vdots & \vdots & \ddots & \end{pmatrix}, A_1 = \begin{pmatrix} \mu g_1 & \mu g_2 & \mu g_3 & \cdots \\ \bar{\mu} & 0 & 0 & \cdots \\ 0 & \bar{\mu} & 0 & \cdots \\ \vdots & \vdots & \ddots & \end{pmatrix} \tag{15}$$

$$A_0 = \begin{pmatrix} \bar{\mu} g_1 & \bar{\mu} g_2 & \bar{\mu} g_3 & \cdots \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \ddots & \end{pmatrix}. \tag{16}$$

It is easily to show that if $0 < \mu < 1$ then B, C and A_k, $k = 0, 1, 2$ satisfy the conditions in Example 2.4 (i) and (ii), so $\boldsymbol{P}$ in Eq.(13) is block-monotone with respect to order $\leq_{m-st}$ and is block-monotone with respect to order $\leq_{m-icx}$.

For $GI/Geo/1$ queue, conditions in Eq.(7) and Eq.(9) can be further simplified to the following:

(a) Condition I_{m-st} always holds for $0 < \mu < 1$.

(b) Condition II_{m-st}: For $k = 0, 1, 2, \ldots$,

$$\sum_{j=0}^{k} v_j(i) - [\sum_{j=0}^{k} v_j(i+1) + \mu v_{k+1}(i+1)]$$
$$-g_i[(1+\mu)\sum_{j=1}^{k} v_j(i) + v_0(i) - v_k(i)] \geq 0,$$
$$\text{for } i = 1, 2, \ldots, m-1, \tag{17}$$
$$\sum_{j=0}^{k} v_j(m) - g_m[(1+\mu)\sum_{j=1}^{k} v_j(m)$$
$$+v_0(m) - v_k(m)] \geq 0. \tag{18}$$

(c) Condition II'_{m-st}: For $k = 0, 1, 2, \ldots$,

$$\sum_{j=0}^{k} v_j(i) - [\sum_{j=0}^{k} v_j(i+1) + \mu v_{k+1}(i+1)]$$
$$-g_i[(1+\mu)\sum_{j=1}^{k} v_j(i) + v_0(i) - v_k(i)] \leq 0,$$
$$\text{for } i = 1, 2, \ldots, m-1, \tag{19}$$
$$\sum_{j=0}^{k} v_j(m) - g_m[(1+\mu)\sum_{j=1}^{k} v_j(m)$$
$$+v_0(m) - v_k(m)] \leq 0. \tag{20}$$

(d) Condition III_{m-st}: For $k = 0, 1, 2, \ldots$,

$$\begin{aligned}\Delta f_{k+1}(1) \geq \mu \sum_{j=1}^{m} g_j[\Delta f_{k+1}(j) - \Delta f_{k+2}(j)] \\ + \sum_{j=1}^{m} g_j \Delta f_{k+2}(j), \qquad (21)\end{aligned}$$

$$\begin{aligned}\Delta f_{k+1}(i+1) - \Delta f_{k+1}(i) \geq \mu(\Delta f_k(i) - \Delta f_{k+1}(i)), \\ \text{for } i = 1, 2, \ldots, m-1. \qquad (22)\end{aligned}$$

where $\Delta f_{k+1}(j) = f_{k+1}(j) - f_k(j)$, for $k = 0, 1, 2, \ldots$, and $j = 1, 2, \ldots, m$.

First, it is clear that the matrix $\boldsymbol{P}$ in Eq.(13) is always block-monotone with respect to order $\leq_{m-st}$, that is, Condition Condition I_{m-st} always holds for $0 < \mu < 1$. Second, by routine calculations, it is easy to verify that the initial distribution $\boldsymbol{v} = (\boldsymbol{e}_1, \boldsymbol{0}, \boldsymbol{0}, \ldots)$ satisfy Condition II_{m-st} in (17), (18), where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)$ and $\boldsymbol{0}$ is a zero vector with m-dimension; the initial distribution $\boldsymbol{v} = (\boldsymbol{v}_0, \boldsymbol{0}, \boldsymbol{0}, \ldots)$ satisfy Condition II_{m-st} in (17), (18), where $\boldsymbol{v}_0 = (\theta, 1-\theta, \ldots, 0)$, $1/2 < \theta < 1$. It can also be verified that the initial distribution $\boldsymbol{v} = (\boldsymbol{v}_0, \boldsymbol{0}, \boldsymbol{0}, \ldots)$ satisfy Condition II'_{m-st} in (21) and (22), where $\boldsymbol{v}_0 = (0, 0, r^{m-1}, r^{m-1}, \ldots, r^2, r)$, $0 < r < 1$ is unique solution of equation $r^{m-2} - 2r + 1 = 0$,$m \geq 4$. For Condition III_{m-st} in (21) and (22), it always holds for $\boldsymbol{f} = (\boldsymbol{f}_0, \boldsymbol{f}_1, \boldsymbol{f}_2, \ldots)$,where $\boldsymbol{f}_k(j) = k$ for $k = 0, 1, 2, \ldots$ and $j = 1, 2, \ldots, m$.

By Theorem 3.3, a) if Eq.(17), (18), (21) and (22) satisfy, then the function $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(Z_n)]$ is increasing concave in n. b) if Eq.(19), (20), (21) and (22) satisfy, then the function $\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(Z_n)]$ is decreasing convex in n. By Theorem 2.1, under above conditions, for any initial distribution $\boldsymbol{v}$ we have $\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{v}}[\boldsymbol{f}(Z_n)] = \boldsymbol{\pi}\boldsymbol{f}^T$, where $\boldsymbol{\pi}$ is the steady-state distribution of $\boldsymbol{P}$ (i,e,. $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}$ and $\boldsymbol{\pi}\boldsymbol{e}^T = 1$).

Remark 4.1 Alfa(2016) studied the $GI/Geo/1$ queue by applying the Matrix-Analytic Methods, and showed that the rate matrix R satisfying $R = A_0 + RA_1 + R^2A_2$ has the following properties:

$$R = \begin{pmatrix} r_1 & r_2 & r_3 & \cdots \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \qquad (23)$$

where r_k satisfy: $r_k = \bar{\mu} g_k + r_1 \mu g_k + \mu r_1 r_{k+1}$ for $k = 1, 2, \ldots$. $r_k, k = 1, 2, \ldots$

$$\begin{aligned} r_m &= \bar{\mu} g_m + r_1 \mu g_m, \\ r_{m-1} &= \bar{\mu} g_{m-1} + r_1 \mu g_{m-1} + r_1 \mu \bar{\mu} g_m + \mu^2 (r_1)^2 g_m, \\ &\quad \cdots\cdots \\ r_1 &= \xi_0 + \xi_1 r_1 + \xi_2 (r_1)^2 + \ldots + \xi_m (r_1)^m, \\ &\quad \text{where } \xi_j \ j = 1, 2, \ldots, m, \text{ are known constants.} \end{aligned} \qquad (24)$$

Remark 4.2 Comparison the matrix-analytic results for the $GI/Geo/1$ queue, e.g., Alfa(2016), our approach can provide those conditions for some functions of the queue length and phase of arrival to be monotone, convex, or concave. Those conditions only are related to the initial distributions and transition matrix of Markov chain consist of the queue length and phase of arrival.

3.2 The $PH/M/c$ queue

Consider the $PH/M/c$ ($c \geq 1$) queue in which the arrivals form a PH-renewal process with interarrival time distribution of phase type with representation (α, T) of order m (Page 88, Neuts (1981)). The service rate for each of c servers is denoted by μ. The $PH/M/c$ queue is a QBD process on the state space $E = \{(i, j), i \geq 0, 1 \leq j \leq m\}$ where i denotes the number of customers in the system and j denotes the phase of the arrival process. The generator $\boldsymbol{Q}$ of the Markov chain is given by

$$\boldsymbol{Q} = \begin{pmatrix} B_0 & A_0 & & & & & \\ B_1 & A_{11} & A_0 & & & & \\ & B_2 & A_{12} & A_0 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & A_2 & A_1 & A_0 & \\ & & & & A_2 & A_1 & A_0 \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix}, \quad (25)$$

where $B_0 = T$, $A_0 = T^0\alpha$, $A_1 = T - c\mu I$, $A_2 = c\mu I$, $B_k = k\mu I$, $A_{1k} = T - k\mu I$, $k = 1, 2, \ldots, c-1$, I is the $m \times m$ identity matrix. By using uniformization $P = I + Q/\xi$ with ξ equal or greater than the biggest absolute value of all diagonal elements of Q, one can easily convert the generator $\boldsymbol{Q}$ in Eq.(25) into the transition matrix $\boldsymbol{P}$ of a discrete time Markov chain, where $\boldsymbol{P}$ is written as

$$\boldsymbol{P} = \begin{pmatrix} \tilde{B}_0 & \tilde{A}_0 & & & & & \\ \tilde{B}_1 & \tilde{A}_{11} & \tilde{A}_0 & & & & \\ & \tilde{B}_2 & \tilde{A}_{12} & \tilde{A}_0 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \tilde{A}_2 & \tilde{A}_1 & \tilde{A}_0 & \\ & & & & \tilde{A}_2 & \tilde{A}_1 & \tilde{A}_0 \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix}, \quad (26)$$

where $\tilde{B}_0 = I + T/\xi$, $\eta = \mu/\xi$, $\tilde{A}_0 = T^0\alpha/\xi$, $\tilde{A}_1 = I + (T - c\mu I)/\xi$, $\tilde{A}_2 = c\mu I/\xi$, $\tilde{B}_k = k\mu I/\xi$, $\tilde{A}_{1k} = I + (T - k\mu I)/\xi$, $k = 1, 2, \ldots, c-1$.

Li and Zhao (2000 Example 4.1) proved that $\boldsymbol{P}$ in Eq.(26) is stochastically block-monotone with respect to the order $\leq_{m-st}$. Hence, Li and Zhao (2000) showed the stationary probability vector can be obtained by approximation from a stochastic block augmentation of $\boldsymbol{P}$, and the last block-column augmentation provides the best approximation in the sense Eq.(3.6) in Li and Zhao (2000).

For $PH/M/c$ queue, conditions in Eq.(7) and Eq.(9) can be further simplified to the following:

(a) Condition II_{m-st}:

$$\begin{aligned} &\boldsymbol{0} \geq_{el} \boldsymbol{v}_0 T + \boldsymbol{v}_1 \mu I, \\ &\boldsymbol{0} \geq_{el} \sum_{i=0}^{k} \boldsymbol{v}_i Q^* + \boldsymbol{v}_{k+1} T + \boldsymbol{v}_{k+2}(k+2)\mu I, \ 0 \leq k \leq c-2. \\ &\boldsymbol{0} \geq_{el} \sum_{i=0}^{k} \boldsymbol{v}_i Q^* + \boldsymbol{v}_{k+1} T + \boldsymbol{v}_{k+2} c\mu I, \ k \geq c-1. \end{aligned} \qquad (27)$$

(b) Condition II'_{m-st}:

$$\begin{aligned} &\boldsymbol{0} \leq_{el} \boldsymbol{v}_0 T + \boldsymbol{v}_1 \mu I, \\ &\boldsymbol{0} \leq_{el} \sum_{i=0}^{k} \boldsymbol{v}_i Q^* + \boldsymbol{v}_{k+1} T + \boldsymbol{v}_{k+2}(k+2)\mu I, \ 0 \leq k \leq c-2. \\ &\boldsymbol{0} \leq_{el} \sum_{i=0}^{k} \boldsymbol{v}_i Q^* + \boldsymbol{v}_{k+1} T + \boldsymbol{v}_{k+2} c\mu I, \ k \geq c-1. \end{aligned} \qquad (28)$$

(c) Condition III$_{m-st}$:

$$\begin{aligned}
&(\mu I - T)\boldsymbol{f}_0 + (T - \mu I - T^0\alpha)\boldsymbol{f}_1 + T\alpha \boldsymbol{f}_2 \leq_{el} \mathbf{0},\\
&-(i+1)\mu I \boldsymbol{f}_i + [(2i+3)\mu I - T]\boldsymbol{f}_{i+1}\\
&+[T - (i+2)\mu I - T^0\alpha]\boldsymbol{f}_{i+2} + T^0\alpha \boldsymbol{f}_{i+3} \leq_{el} \mathbf{0},\\
&\text{for } i = 0, 1, \ldots, c-2,\\
&-c\mu I \boldsymbol{f}_i + [2c\mu I - T]\boldsymbol{f}_{i+1}\\
&+[T - c\mu I - T^0\alpha]\boldsymbol{f}_{i+2} + T^0\alpha \boldsymbol{f}_{i+3} \leq_{el} \mathbf{0},\\
&\text{for } i = c-1, c, \ldots. \qquad (29)
\end{aligned}$$

It is clear that the matrix $\boldsymbol{P}$ in Eq.(26) is always block-monotone with respect to order $\leq_{m-st}$ according to Li and Zhao (2000), that is, Condition Condition I$_{m-st}$ always holds. Condition II$_{m-st}$ in (27), Condition II$'_{m-st}$ in (28) and Condition III$_{m-st}$ in (29) can be checked by using similar method as $GI/Geo/1$ queue above subsection.

Remark 4.3 Similar the discussion in Remark 4.2, comparison the results for the $PH/M/c$ queue, e.g., Li and Zhao (2000), our approach can provide those conditions for some functions of the queue length and phase of arrival to be monotone, convex, or concave. Those conditions are only related to the initial distributions and transition matrix of Markov chain consist of the queue length and phase of arrival.

4. CONCLUSIONS

This paper generalizes Yu, He and Zhang (2006)'s work to bivariate Markov chains with a general block transition matrices, here the important condition is the stochastically block-monotonicity of matrices. We provide the sufficient conditions under which the monotonicity, convexity and comparability of some functions associated with discrete-time or continuous-time denumerable Markov chains with a general block transition matrices are obtained.

5. REFERENCES

[1] Alfa A S. 2016. Applied Discrete-Time Queue[M]. New York: Springer.

[2] Daley D J. 1968. Stochastically monotone Markov chains[J]. Z. Wahrscheinlichkeitstheorie verw. Geb. 10: 305-317.

[3] Dijk N M V. 1990. On a simple proof of uniformization for continuous and discrete-state continuous-time Markov chains[J]. Advances Applied Probability, 22: 749-750.

[4] Gail H R, Hantler S L, Taylor B A. 1996. Spectral analysis of M/G/1 and GI/M/1 type Markov chains[J]. Advances in Applied Probability, 28, 114-165

[5] Gail H R, Hantler S L, Taylor B A. 1997. M/G/1 and GI/M/1 type Markov chains with multiple boundary levels[J]. Advances in Applied Probability, 29, 773-758

[6] Gibson, D. and E. Seneta. 1987. Monotone infinite stochastic matrices and their augmented truncations[J]. Stochastic Processes and Their Applications, 24, 287-292.

[7] He Q-M. 2014. Fundamentals of Matrix-Analytic Methods, Springer, New York

[8] Lucantoni D M. 1991. New results on the single server queue with a batch Markovian arrival process[J]. Stochastic Models. 7: 1-46.

[9] Li H, Shaked M. 1994. Stochastic convexity and concavity of Markov processes[J]. Mathematics of Operations Research, 19:477-493.

[10] Li H, Zhao Y Q. 2000. Stochastic block-monotonicity in the approximation of the stationary distribution of infinite Markov chains[J]. Stochastic Models, 16(2):313-333.

[11] Liu Y. 2010. Augmented truncation approximations of discrete-time Markov chains[J]. Operations Research Letter, 38: 218-222.

[12] Masuyama H. 2015. Error bounds for augmented truncations of discrete-time block-monotone Markov chains under geometric drift conditions[J]. Advances in Applied Probability, 47(1):83-105.

[13] Masuyama H. 2016. Continuous-time block-monotone Markov chains and their block-augmental truncations[J]. Submitted for publication in Linear Algebra and its Applications (Submitted on 15 Nov 2015 (v1), revised 18 Nov 2015 (v2), latest version 29 Jan 2016 (v3))

[14] Neuts M F. 1981. Matrix-Geometric Solutions in Stochastic Models: an algorithmic approach[M]. John Hopkins University Press.

[15] Neuts M F. 1989. Stuctured Stochastic Matrices of M/G/1 Type and their applications[M]. Marcel Dekker, Inc., New York and Basel.

[16] Neuts M F. 1998. Some promising directions in algorithmic probability, in Advances in Matrix Analysis Methods for Stochastic Models[M]. Alfa, A.S. and Chakravarthy, S.R. Eds, Notable Publications, New Jersey, 429-443.

[17] Tweedie R L. 1998. Truncation approximations of invariant measures for Markov chains[J]. Journal of Applied Probability, 35: 517-536.

[18] Yu H-B, He Q M, Zhang H. 2006. Convexity of some functions associated with denumerable-state-space Markov chains and applications to queuing systems [J]. Probability in the Engineering and Information Sciences, 20: 67-86.

[19] Zhao Y Q, Li W, Braun W J. 1998. Infinite block-structured transition matrices and their properties[J]. Advances in Applied Probability, 30(2):365-384.

[20] Zhao Y Q, Li W, Alfa A S. 2000. Duality results for block-structured transition matrices, Infinite block-structured transition matrices and their properties[J]. Journal of Applied Probability, 2000, 36(4):1045-1057.

A matrix geometric approach for random walks in the quadrant

[Full Paper]

Stella Kapodistria
Department of Mathematics and Computer Science, Eindhoven University of Technology
Eindhoven, The Netherlands
s.kapodistria@tue.nl

Zbigniew Palmowski
Mathematical Institute
University of Wroclaw
Wroclaw, Poland
zbigniew.palmowski@gmail.com

ABSTRACT

In this manuscript, the authors consider a sub-class of the two-dimensional homogeneous nearest neighbor (simple) random walk restricted on the lattice. In particular, the subclass of random walks with equilibrium distributions given as series of product-forms is considered and the derivations for the calculations of the terms involved in the equilibrium distribution representation, as well as the eigenvalues and the corresponding eigenvectors of the matrix $\boldsymbol{R}$ are presented. The above results are obtained by connecting three existing approaches available for such an analysis: the matrix geometric approach, the compensation approach and the boundary value problem method.

Keywords

Random walks; Equilibrium distribution; Matrix geometric approach; Compensation approach; Boundary value problem method.

1. INTRODUCTION

The objective of this work is to demonstrate how to obtain the equilibrium distribution of the state of a two-dimensional homogeneous nearest neighbor (simple) random walk restricted on the lattice using the matrix geometric approach. This type of random walk can be modeled as a Quasi-Birth-Death (QBD) process with the characteristic that both the levels and the phases are countably infinite. Then, based on the matrix geometric approach, if $\boldsymbol{\pi}_n = (\pi_{n,0} \quad \pi_{n,1} \quad \cdots)$ denotes the vector of the equilibrium distribution at level n, $n = 0, 1, \ldots$, it is known that $\boldsymbol{\pi}_{n+1} = \boldsymbol{\pi}_n \boldsymbol{R}$. This is a very well known result, but the complexity of the solution lies in the calculation of the infinite dimension matrix $\boldsymbol{R}$. In this manuscript, the authors develop a new methodological approach for the calculation of the eigenvalues and eigenvectors of matrix $\boldsymbol{R}$. Moreover, this approach can be numerically used for the approximation of the matrix $\boldsymbol{R}$.

As a first step and for illustration purposes, as well as for reasons of simplicity, we restrict our analysis to random walks whose equilibrium distribution away from the origin $(0, 0)$ can be written as a series (finite or infinite) of product-forms. In particular, under the following sufficient conditions, referred to as *conditions for meromorphicity*, cf. [1, 4],

- Step size: Only transitions to neighboring states are allowed;
- Forbidden steps: No transitions from interior states to the North, North-East, and East are allowed;
- Homogeneity: All transitions in the same direction occur according to the same rate;

the equilibrium distribution of the simple random walk restricted on the lattice can be written as a series of product-forms for all states $n, m > 0$, say

$$\pi_{n,m} = \sum_{k=0}^{\infty} \tilde{c}_k \tilde{\alpha}_k^n \tilde{\beta}_k^m, \ n, m > 0, \tag{1}$$

cf. [1]. Note that the above conditions do not necessarily imply that the transitions on the boundaries are identical to the transitions in the interior, thus we allow for $\pi_{0,m}$ and $\pi_{n,0}$ to exhibit a slightly different structural pattern, i.e.

$$\pi_{n,0} = \sum_{k=0}^{\infty} \tilde{e}_k \tilde{\alpha}_k^n, \ n > 0, \tag{2}$$

$$\pi_{0,m} = \sum_{k=0}^{\infty} \tilde{d}_k \tilde{\beta}_k^m, \ m > 0, \tag{3}$$

while $\pi_{0,0}$ can be obtained as a function of (1)-(3) by solving the system of equilibrium equations at the origin.

Some examples of queueing systems that satisfy the above conditions are the 2×2 switch and the join the shortest queue, see e.g., [2]. Also, we would like to remark that the conditions for meromorphicity are not necessary for a random walk to have an equilibrium distribution in the form of a series of product-forms, e.g. two-node Jackson networks although violate the above conditions have equilibrium distributions with a product-form representation.

In this manuscript, the authors illustrate the connection between the form of the equilibrium distribution depicted in Equation (1) and the derivation of the eigenvalues and eigenvectors of the infinite matrix $\boldsymbol{R}$. Moreover, this work sets

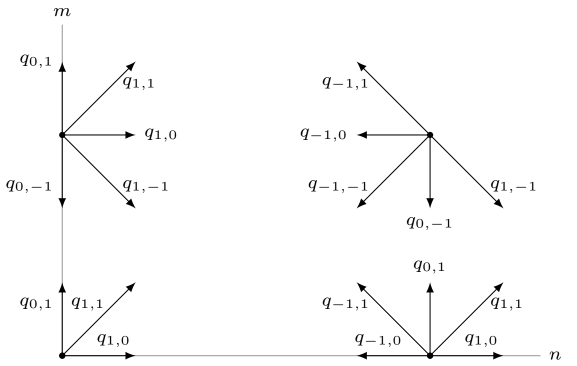


Figure 1: Transition rate diagram of the homogeneous simple random walk on the state space (n, m) with no transitions in the interior to the North, North-East, and East. Only the transitions at a few selected states are depicted as an indication.

the groundwork for the probabilistic interpretation of the terms α and β appearing in the series of product-forms.

The paper is organized as follows: in Section 2 the model is described and in Section 3 the three relevant methods are sketched; more concretely, the matrix geometric approach is presented in Section 3.1, the compensation approach in Section 3.2 and the boundary value problem method in Section 3.3. In Section 4 the derivations for the calculations of the terms involved in the equilibrium distribution representation are presented and in Section 4.1 the eigenvalues and the eigenvectors of matrix $\boldsymbol{R}$ are derived. Finally, in Section 5 conclusions and future work is discussed.

2. MODEL DESCRIPTION

To investigate the scope of applicability of the method described in this manuscript we study a class of Markov processes on the lattice in the non-negative quadrant of $\mathbb{R}^2$. We consider random walks (processes) for which the transition rates are constant, i.e. do not depend on the state, and we further assume that transitions are restricted to neighboring states. The transition rates are depicted in Figure 1.

3. RELATED WORK

3.1 QBD processes and matrix geometric approach

For the model described in the section above, we define n to be the *level* and m the *phase*. Thus, the generator of the random walk can be written as follows

$$\boldsymbol{G} = \begin{bmatrix} \tilde{\boldsymbol{A}}_0 & \tilde{\boldsymbol{A}}_1 & 0 & 0 & \cdots \\ \boldsymbol{A}_{-1} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & 0 & \cdots \\ 0 & \boldsymbol{A}_{-1} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \cdots \\ 0 & 0 & \boldsymbol{A}_{-1} & \boldsymbol{A}_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

with

$$\boldsymbol{A}_{-1} = \begin{bmatrix} q_{-1,0} & q_{-1,1} & 0 & 0 & \cdots \\ q_{-1,-1} & q_{-1,0} & q_{-1,1} & 0 & \cdots \\ 0 & q_{-1,-1} & q_{-1,0} & q_{-1,1} & \cdots \\ 0 & 0 & q_{-1,-1} & q_{-1,0} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$\boldsymbol{A}_0 = \begin{bmatrix} -\tilde{q} & q_{0,1} & 0 & 0 & \cdots \\ q_{0,-1} & -q & 0 & 0 & \cdots \\ 0 & q_{0,-1} & -q & 0 & \cdots \\ 0 & 0 & q_{0,-1} & -q & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$\boldsymbol{A}_1 = \begin{bmatrix} q_{1,0} & q_{1,1} & 0 & 0 & \cdots \\ q_{1,-1} & 0 & 0 & 0 & \cdots \\ 0 & q_{1,-1} & 0 & 0 & \cdots \\ 0 & 0 & q_{1,-1} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$\tilde{\boldsymbol{A}}_1 = \begin{bmatrix} q_{1,0} & q_{1,1} & 0 & 0 & \cdots \\ q_{1,-1} & q_{1,0} & q_{1,1} & 0 & \cdots \\ 0 & q_{1,-1} & q_{1,0} & q_{1,1} & \cdots \\ 0 & 0 & q_{1,-1} & q_{1,0} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$\tilde{\boldsymbol{A}}_0 = \begin{bmatrix} -\tilde{q} & q_{1,0} & 0 & 0 & \cdots \\ q_{-1,0} & -q & q_{1,0} & 0 & \cdots \\ 0 & q_{-1,0} & -q-q_{1,0} & q_{1,0} & \cdots \\ 0 & 0 & q_{-1,0} & -q-q_{1,0} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where $\tilde{q} = q_{1,0} + q_{1,1} + q_{0,1} + q_{-1,1} + q_{-1,0}$ and $q = q_{1,-1} + q_{0,-1} + q_{-1,-1} + q_{-1,0} + q_{-1,1}$.

3.1.1 Stability condition

Let $\boldsymbol{A} = \boldsymbol{A}_{-1} + \boldsymbol{A}_0 + \boldsymbol{A}_1$ and $\boldsymbol{x}$ be the unique solution to

$$\boldsymbol{x}\boldsymbol{A} = 0$$

such that $x\mathbb{1} = 1$, with $\mathbb{1}$ a column vector of ones. For the random walk at hand $\boldsymbol{x}$ corresponds to the vector of the equilibrium distribution of a Birth-Death process with birth rates $\lambda_0 = q_{-1,1} + q_{0,1} + q_{1,1}$, $\lambda_n = q_{-1,1}$, $n \geq 1$, and death rates $\mu_n = q_{-1,-1} + q_{0,-1} + q_{1,-1}$, $n \geq 1$. Then, it was shown in Theorem 1.7.1 [13] that the QBD is positive recurrent if and only if

$$\boldsymbol{x}\boldsymbol{A}_{-1}\mathbb{1} < \boldsymbol{x}\boldsymbol{A}_1\mathbb{1}.$$

3.1.2 Structure of the QBD solution

Let $\pi_{n,m},\ n, m \geq 0$ denote the equilibrium distribution of the QBD. Then, if $\boldsymbol{\pi}_n = \begin{pmatrix} \pi_{n,0} & \pi_{n,1} & \cdots \end{pmatrix}$ denotes the equilibrium vector of level n, $n = 0, 1, \ldots$, it is known that

$$\boldsymbol{\pi}_{n+1} = \boldsymbol{\pi}_n \boldsymbol{R}, \tag{4}$$

where the infinite dimensional matrix $\boldsymbol{R}$ is obtained as the minimal non-negative solution to the matrix quadratic equation

$$\boldsymbol{A}_1 + \boldsymbol{R}\boldsymbol{A}_0 + \boldsymbol{R}^2\boldsymbol{A}_{-1} = 0,$$

cf. [13]. Solving this last equation in terms of $\boldsymbol{R}$ we obtain recursively the equilibrium vector $\boldsymbol{\pi}_{n+1}$, $n \geq 0$, in terms of the matrix $\boldsymbol{R}$ and the vector of the equilibrium distribution corresponding to level 1. However, the structure of the random walk is overly generic and thus does not permit the calculation of the infinite matrix $\boldsymbol{R}$. This will be achieved by combining the two other approaches used in the analysis of random walks on the lattice: the compensation approach and the boundary value problem method.

3.2 Compensation approach

The compensation approach is developed by Adan et al. in a series of papers [1, 2, 3] and aims at a direct solution for the sub-class of two-dimensional random walks on the lattice of the first quadrant that obey the conditions for meromorphicity. The compensation approach can also be effectively used in cases that the random walk at hand does not satisfy the aforementioned conditions, but the equilibrium distribution still can be written in the form of series of product-forms. This is due to the fact that this approach exploits the structure of the equilibrium equations in the interior of the quarter plane by imposing that linear (finite or infinite) combinations of product-forms satisfy them. This leads to a kernel equation for the terms appearing in the product-forms. Then, it is required that these linear combinations satisfy the equilibrium equations on the boundaries as well. As it turns out, this can be done by alternatingly compensating for the errors on the two boundaries, which eventually leads to a (potentially) infinite series of product-forms.
For the model described in Section 2 one can easily show, cf. [1], that

Step 1: $\pi_{n,m} = \tilde{\alpha}^n\tilde{\beta}^m$, $m, n > 0$, is a solution to the equilibrium equations in the interior if and only if $\tilde{\alpha}$ and $\tilde{\beta}$ satisfy

$$\begin{aligned}&\tilde{\alpha}\tilde{\beta}(q_{-1,1} + q_{1,-1} + q_{0,-1} + q_{-1,-1} + q_{-1,0})\\&= \tilde{\alpha}^2 q_{-1,1} + \tilde{\beta}^2 q_{1,-1} + \tilde{\alpha}\tilde{\beta}^2 q_{0,-1}\\&\quad+\tilde{\alpha}^2\tilde{\beta}^2 q_{-1,-1} + \tilde{\alpha}^2\tilde{\beta} q_{-1,0}.\end{aligned} \tag{5}$$

Step 2: Consider a product-form $c_0\tilde{\alpha}_0^n\tilde{\beta}_0^m$ that satisfies the kernel equation (5) and also satisfies the equilibrium equations of the horizontal boundary. Without loss of generality we can assume that $c_0 = 1$. If the product-form $c_0\tilde{\alpha}_0^n\tilde{\beta}_0^m$ also satisfies the equilibrium equations of the vertical boundary then this constitutes the solution of the equilibrium equations up to a multiplicative constant that can be obtained using the normalizing equation. Otherwise, consider a linear combination of two product-forms, say $\tilde{\alpha}_0^n\tilde{\beta}_0^m + c_1\tilde{\alpha}_1^n\tilde{\beta}_1^m$, $m, n > 0$, such that this combination satisfies now the equilibrium equations of the vertical boundary. For this to happen it must be that $\tilde{\beta}_1 = \tilde{\beta}_0$ and that $\tilde{\alpha}_1$ satisfies the kernel equation (5) for $\tilde{\beta} = \tilde{\beta}_0$.

Step 3: Finally, as long as our expression of linear combinations of product-forms violates one of the two equilibrium equations on the boundary, we continue by adding new product-form terms satisfying the kernel equation (5). This will eventually lead to Equation (1). Of course, one still needs to show that the series expression of Equation (1) converges for all $n, m > 0$.

3.3 Boundary value problem method

The boundary value problem method is an analytic method which is applicable to some two-dimensional random walks restricted to the first quadrant. The bivariate probability generating function (PGF), say

$$\Pi(x,y) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\pi_{n,m}x^n y^m,\ |x|,|y| \leq 1,$$

of the position of a homogeneous nearest neighbor random walk satisfies a functional equation of the form

$$\begin{aligned}&K(x,y)\Pi(x,y) + A(x,y)\Pi(x,0) + B(x,y)\Pi(0,y)\\&+ C(x,y)\Pi(0,0) = 0,\end{aligned} \tag{6}$$

with $K(x,y)$, $A(x,y)$, $B(x,y)$ and $C(x,y)$ known bivariate polynomials in x and y, depending only on the parameters of the random walk. In particular,

$$\begin{aligned}K(x,y) &= y^2 q_{-1,1} + x^2 q_{1,-1} + x q_{0,-1}\\&\quad+ q_{-1,-1} + y q_{-1,0}\\&\quad- xy(q_{-1,1} + q_{1,-1} + q_{0,-1} + q_{-1,-1} + q_{-1,0})\end{aligned}$$

and hence $K(1/\tilde{\alpha}, 1/\tilde{\beta}) = 0$ reduces to exactly Equation (5).

The boundary value problem method consists of the following steps:

i) First, define the zero tuples (x,y) such that $K(x,y) = 0$, $|x|,|y| < 1$.

ii) Then, along the curve $K(x,y) = 0$ (and provided that $\Pi(x,y)$ is defined on this curve), Equation (6) reads

$$\begin{aligned}&A(x,y)\Pi(x,0) + B(x,y)\Pi(0,y)\\&+ C(x,y)\Pi(0,0) = 0.\end{aligned} \tag{7}$$

iii) Finally, in same instances Equation (7) can be solved as a Riemann (Hilbert) boundary value problem.

Malyshev pioneered this approach of transforming the functional equation to a boundary value problem in the 1970's. The idea to reduce the functional equation for the generating function to a standard Riemann-Hilbert boundary value problem stems from the work of Fayolle and Iasnogorodski [7] on two parallel M/M/1 queues with coupled processors (the service speed of a server depends on whether or not the other server is busy). Extensive treatments of the boundary value technique for functional equations can be found in Cohen and Boxma [6, Part II] and Fayolle, Iasnogorodski and Malyshev [8]. The model depicted in Figure 1 can be analyzed by the approach developed by Fayolle and Iasnogorodski [7, 8] and Cohen and Boxma [6], however this approach does not lead to the direct determination of the equilibrium distribution, since it requires inverting the PGF, and the existing numerical approaches for this method are oftentimes tedious and case specific.

4. ANALYSIS

In this paper, we connect for the first time the three approaches: the matrix geometric approach, the compensation approach and the boundary value problem method. We will demonstrate now how to easily compute recursively these $\tilde{\alpha}$'s and $\tilde{\beta}$'s.

$$\begin{tikzcd}(\tilde\alpha_0,\tilde\beta_0) \arrow[r] & (\tilde\alpha_1,\tilde\beta_1=\tilde\beta_0) \arrow[r] & (\tilde\alpha_2=\tilde\alpha_1,\tilde\beta_2) \arrow[r] & (\tilde\alpha_3,\tilde\beta_3=\tilde\beta_2) & \cdots\end{tikzcd}$$

Figure 2: The recursive structure of the product-form terms.

$$\begin{tikzcd}\alpha_0 \arrow[r] \arrow[r, bend left, "y_+"] & \beta_0 \arrow[l, bend left, "x_-"] \arrow[r] \arrow[r, bend left, "x_+"] & \alpha_1 \arrow[l, bend left, "y_-"] \arrow[r] \arrow[r, bend left, "y_+"] & \beta_1 \arrow[l, bend left, "x_-"] & \cdots\end{tikzcd}$$

Figure 3: Evolutions of α's and β's.

Step 1: Rewrite the PGF as

$$\begin{aligned}\Pi(x,y) &= \sum_{n=0}^{\infty} x^n\,\pi_n\,(\;1\quad y\quad y^2\quad\cdots\;)^T\\ &= \pi_0(\;1\quad y\quad y^2\quad\cdots\;)^T\\ &\quad+\sum_{n=1}^{\infty} x^n\,\pi_1\,\boldsymbol{R}^{n-1}(\;1\quad y\quad y^2\quad\cdots\;)^T\\ &= \pi_0(\;1\quad y\quad y^2\quad\cdots\;)^T\\ &\quad+\pi_1\,(x^{-1}\boldsymbol{I}-\boldsymbol{R})^{-1}(\;1\quad y\quad y^2\quad\cdots\;)^T.\end{aligned}\tag{8}$$

In the last equation the term $(x^{-1}\boldsymbol{I}-\boldsymbol{R})^{-1}$ should be interpreted as an operator instead of the inverse of a matrix. Then, Equation (6) reduces to

$$\begin{aligned}&\pi_1\,(x^{-1}\boldsymbol{I}-\boldsymbol{R})^{-1}\Big(K(x,y)(\;1\quad y\quad y^2\quad\cdots\;)^T\\ &\quad+A(x,y)(\;1\quad 0\quad 0\quad\cdots\;)^T\Big)\\ &=-\,\pi_0\left(\,(K(x,y)+B(x,y))\,(\;1\quad y\quad y^2\quad\cdots\;)^T\right.\\ &\quad\left.+\,(A(x,y)+C(x,y))\,(\;1\quad 0\quad 0\quad\cdots\;)^T\right).\end{aligned}\tag{9}$$

Note that due to (1) we can meromorphically continue the PGF on the entire complex domain, i.e. the PGF is holomorphic on the entire complex domain except for a set of isolated points (the poles of the function) $x^{-1}=\tilde\alpha_k$ and $y^{-1}=\tilde\beta_k$, $k\ge 0$. More concretely, it can be shown that these PGFs are meromorphic functions, i.e., they have a finite number of poles in every finite domain, cf. [5].

Step 2: For the isolated point $x^{-1}=\tilde\alpha_0$ the right hand side of (9) is well defined, which implies that

$$\begin{aligned}&K(x,y)(\;1\quad y\quad y^2\quad\cdots\;)^T\\ &+A(x,y)(\;1\quad 0\quad 0\quad\cdots\;)^T=0,\end{aligned}$$

or equivalently that $K(x,y)=0$ and $A(x,y)=0$. This reveals the starting solution $\tilde\alpha_0$, with $|\tilde\alpha_0|<1$ for the iterative calculation of the sequences $\{\tilde\alpha_k\}_{k\ge0}$ and $\{\tilde\beta_k\}_{k\ge0}$.

For the starting solution $x^{-1}=\tilde\alpha_0$ we calculate recursively $y^{-1}=\tilde\beta_0$ by the kernel equation $K(x,y)=0$. This will produce a single $\tilde\beta$ with $|\tilde\beta|\le|\tilde\alpha|$. We can proceed in an analogous manner and construct the entire set of product-form terms. Moreover, one can easily show, cf. [1], that

$$\tilde\beta_{2k+1}=\tilde\beta_{2k},\;\;k\ge0$$

and $\tilde\alpha_{2k+1}$, $k\ge0$, is obtained as the root of

$$K(1/\tilde\alpha_{2k+1},1/\tilde\beta_{2k+1})=0$$

with $|\tilde\alpha_{2k+1}|<|\tilde\beta_{2k+1}|$. Also,

$$\tilde\alpha_{2k}=\tilde\alpha_{2k-1},\;\;k\ge1,$$

and $\tilde\beta_{2k}$, $k\ge1$, is obtained as the root of

$$K(1/\tilde\alpha_{2k},1/\tilde\beta_{2k})=0$$

with $|\tilde\beta_{2k}|<|\tilde\alpha_{2k}|$. Figure 2 displays the way in which the product-form terms are generated.

Step 3: We now consider a new representation of $\pi_{n,m}$ so as to avoid repetitions in the family of roots of the kernel. To this purpose, for $k\ge0$ we denote:

$$\alpha_k=\tilde\alpha_{2k},\qquad\beta_k=\tilde\beta_{2k}.$$

Then, Equations (1), (2) and (3) are re-written, for $n,\;m>0$

$$\pi_{n,m}=c_0\alpha_0^n\beta_0^m+\sum_{k=1}^{\infty}c_k\alpha_k^n(\beta_{k-1}^m+f_k\beta_k^m),\tag{10}$$

and

$$\pi_{n,0}=\sum_{k=0}^{\infty}e_k\alpha_k^n,\qquad\pi_{0,m}=\sum_{k=0}^{\infty}d_k\beta_k^m,\tag{11}$$

with $c_0=\tilde c_0$, $e_0=\tilde e_0$ and $d_0=\tilde d_0$, and, for $k\ge1$, $c_k=\tilde c_{2k-1}$, $f_k=\tilde c_{2k}/\tilde c_{2k-1}$, $e_k=\tilde e_{2k}+\tilde e_{2k-1}$ and $d_k=\tilde d_{2k}+\tilde d_{2k-1}$. Equivalently, we can also consider the following representation, for $n,m>0$,

$$\pi_{n,m}=\sum_{k=0}^{\infty}c_k^*\beta_k^m(\alpha_k^n+f_{k+1}^*\alpha_{k+1}^n),\tag{12}$$

with $c_0^*=c_0$, $c_k^*=c_kf_k$ for $k>0$ and $f_{k+1}^*=c_{k+1}/c_k^*$. Now we have the sequence of the zero-tuples of the kernel equation, cf. Figure 3, with 'forward' operators y_+ and x_+ and 'backward' operators y_- and x_- constructed as the solutions of the kernel equation $K(x,y)=0$ with respect to y and x. More concretely, by re-writing the kernel equation $K(x,y) = 0$ as a quadratic function in either y or x

$$K(x,y) = a_x(y - y_+(x))(y - y_-(x)),$$
$$K(x,y) = \tilde{a}_y(x - x_+(y))(x - x_-(y))$$

for some functions a_x and $\tilde{a}_y$, yields

$$y_+\left(\frac{1}{\alpha_k}\right) = \frac{1}{\beta_k}, \qquad x_+\left(\frac{1}{\beta_k}\right) = \frac{1}{\alpha_{k+1}}, \tag{13}$$
$$y_-\left(\frac{1}{\alpha_k}\right) = \frac{1}{\beta_{k-1}}, \qquad x_-\left(\frac{1}{\beta_k}\right) = \frac{1}{\alpha_k}. \tag{14}$$

Step 4: It remains to show how to calculate the coefficients of the product-form terms. Observe that the representations (10)-(11) for the $\pi_{n,m}$ (or equivalently (12)-(11)) yield

$$\begin{aligned}\Pi(x,y) =&\Pi(x,0) + (\Pi(0,y) - \pi_{0,0}) \\ &+ c_0\frac{\alpha_0\beta_0 xy}{(1-\alpha_0 x)(1-\beta_0 y)} \\ &+\sum_{k=1}^{\infty} c_k\frac{\alpha_k x}{1-\alpha_k x}\left(\frac{\beta_{k-1}y}{1-\beta_{k-1}y} + f_k\frac{\beta_k y}{1-\beta_k y}\right),\end{aligned} \tag{15}$$

with

$$\Pi(x,0) = \pi_{0,0} + \sum_{k=1}^{\infty} e_k\frac{\alpha_k x}{1-\alpha_k x}, \tag{16}$$

$$\Pi(0,y) = \pi_{0,0} + \sum_{k=1}^{\infty} d_k\frac{\beta_k y}{1-\beta_k y}. \tag{17}$$

We first identify the sequences $\{e_k\}_{k\geq 0}$ and $\{d_k\}_{k\geq 0}$ appearing in (11). Next, we set $y = y_+(x)$ in (9) (for which $K(x, y_+(x)) = 0$) and substitute the representations for $\Pi(x,0)$ and $\Pi(0,y)$. We multiply the resulting equation with $1 - \alpha_i x$ and take the limit as $x \to 1/\alpha_i$. This yields

$$\begin{aligned}0 = &A\left(\frac{1}{\alpha_i}, y_+\left(\frac{1}{\alpha_i}\right)\right)e_i \\ &+ B\left(\frac{1}{\alpha_i}, y_+\left(\frac{1}{\alpha_i}\right)\right)d_i \lim_{x\to\frac{1}{\alpha_i}}\frac{1-\alpha_i x}{1-\beta_i y_+(x)}.\end{aligned} \tag{18}$$

Similarly, repeating the above procedure for $y = y_-(x)$ produces

$$\begin{aligned}0 = &A\left(\frac{1}{\alpha_i}, y_-\left(\frac{1}{\alpha_i}\right)\right)e_i \\ &+ B\left(\frac{1}{\alpha_i}, y_-\left(\frac{1}{\alpha_i}\right)\right)d_{i-1} \lim_{x\to\frac{1}{\alpha_i}}\frac{1-\alpha_i x}{1-\beta_{i-1} y_-(x)}.\end{aligned} \tag{19}$$

Similarly, we could have chosen $x = x_\pm(y)$, but the resulting equations would be identical to the ones derived above.

Now starting from e_0 all the coefficients $\{e_k\}_{k\geq 0}$ and $\{d_k\}_{k\geq 0}$ are obtained recursively as follows: for a given e_k using Equation (18) one can derive d_k, next Equation (19) produces e_{k+1}.

Step 5: Having $\{e_k\}_{k\geq 0}$ and $\{d_k\}_{k\geq 0}$ recursively identified in terms of e_0, we show in this paragraph how to obtain the sequence $\{c_k\}_{k\geq 0}$. Multiplying Equation (9) with $1 - \alpha_i x$, then substituting (15)-(17) therein, and afterwards taking the limit as $x \to 1/\alpha_i$ and setting $y = 1$ gives

$$\begin{aligned}0 = &\left(A\left(\frac{1}{\alpha_i}, 1\right) + K\left(\frac{1}{\alpha_i}, 1\right)\right)e_i \\ &+ K\left(\frac{1}{\alpha_i}, 1\right)c_0\frac{\beta_0}{1-\beta_0}\end{aligned} \tag{20}$$

and

$$\begin{aligned}0 = &\left(A\left(\frac{1}{\alpha_i}, 1\right) + K\left(\frac{1}{\alpha_i}, 1\right)\right)e_i \\ &+ K\left(\frac{1}{\alpha_i}, 1\right)c_i\left(\frac{\beta_{i-1}}{1-\beta_{i-1}} + f_i\frac{\beta_i}{1-\beta_i}\right),\end{aligned} \tag{21}$$

for $i > 0$. Equivalently, using representation (12) yields

$$\begin{aligned}\Pi(x,y) =&\Pi(x,0) + (\Pi(0,y) - \pi_{0,0}) \\ &+ c_0\frac{\alpha_0\beta_0 xy}{(1-\alpha_0 x)(1-\beta_0 y)} \\ &+\sum_{k=0}^{\infty} c_k^*\frac{\beta_k y}{1-\beta_k y}\left(\frac{\alpha_k x}{1-\alpha_k x} + f_{k+1}^*\frac{\alpha_{k+1} x}{1-\alpha_{k+1} x}\right).\end{aligned} \tag{22}$$

Now using (22), multiplying Equation (9) by $1 - \beta_i y$, and afterwards taking the limit as $y \to 1/\beta_i$ and setting $x = 1$ yields, for $i \geq 0$,

$$\begin{aligned}0 = &\left(B\left(1, \frac{1}{\beta_i}\right) + K\left(1, \frac{1}{\beta_i}\right)\right)d_i \\ &+ K\left(1, \frac{1}{\beta_i}\right)c_i^*\left(\frac{\alpha_i}{1-\alpha_i} + f_{i+1}^*\frac{\alpha_{i+1}}{1-\alpha_{i+1}}\right)\end{aligned} \tag{23}$$

which is equivalent to

$$\begin{aligned}0 = &\left(B\left(1, \frac{1}{\beta_0}\right) + K\left(1, \frac{1}{\beta_0}\right)\right)d_0 \\ &+ K\left(1, \frac{1}{\beta_0}\right)\left(c_0\frac{\alpha_0}{1-\alpha_0} + c_1\frac{\alpha_1}{1-\alpha_1}\right)\end{aligned} \tag{24}$$

and

$$\begin{aligned}0 = &\left(B\left(1, \frac{1}{\beta_i}\right) + K\left(1, \frac{1}{\beta_i}\right)\right)d_i \\ &+ K\left(1, \frac{1}{\beta_i}\right)\left(c_i f_i\frac{\alpha_i}{1-\alpha_i} + c_{i+1}\frac{\alpha_{i+1}}{1-\alpha_{i+1}}\right),\end{aligned} \tag{25}$$

for $i \geq 1$. Now the iterative procedure is as follows: Starting from c_0, we derive e_0 from Equation (20) (and hence from Step 4 all coefficients $\{e_k\}_{k\geq 0}$ and $\{d_k\}_{k\geq 0}$ are produced in terms of c_0). Then, from Equation (24) we calculate c_1. Having c_1 and using (21) we identify f_1, which allows us to derive c_2 from Equation (25). Continuing this procedure permits the identification of the sequences $\{c_k\}_{k\geq 0}$ and $\{f_k\}_{k\geq 1}$. The starting constant c_0 is uniquely identified by the normalization equation.

Note that from the construction of Equation (6) it follows that we can take any $|y| \leq 1$ instead of $y = 1$ (respectively any $|x| \leq 1$) since the rhs of the above equation does not in practice depend on y.

4.1 The matrix $\boldsymbol{R}$

We are now in position to present the main result of the manuscript, that connects the derivation of the matrix $\boldsymbol{R}$ with Equations (10) and (11), and hence the boundary value problem with the matrix geometric approach.

Theorem 1. *The terms* $\{\alpha_k\}_{k\geq 0}$ *constitute the different eigenvalues of the matrix* $\boldsymbol{R}$. *For eigenvalue* α_k *the corresponding eigenvector of the matrix* $\boldsymbol{R}$ *is* $\boldsymbol{h}_k = (h_{k,0}, h_{k,1}, k_{k,2} \ldots)$, *with* $h_{k,0} = e_k$ *and* $h_{k,m} = c_k(\beta_{k-1}^m + f_k\beta_k^m)$ $(m = 1, 2, \ldots)$, *if and only if* $c_k \neq 0$.

Proof. From (10) and (11) note that, for $n > 0$,

$$\begin{aligned}\pi_n &= \begin{pmatrix}\pi_{n,0} & \pi_{n,1} & \pi_{n,2} & \cdots\end{pmatrix}\\ &= \sum_{k=0}^{\infty}\alpha_k^n\boldsymbol{h}_k.\end{aligned}$$

Plugging this last result into (4) and after straightforward manipulations yields

$$\sum_{k=0}^{\infty}\alpha_k^n\boldsymbol{h}_k(\alpha_k\boldsymbol{I} - \boldsymbol{R}) = 0, \ \forall n > 0.$$

From this last equation it is needed that

$$\boldsymbol{h}_k(\alpha_k\boldsymbol{I} - \boldsymbol{R}) = 0, \ \forall k \geq 0,$$

which implies the statement of the Theorem, cf. [11]. □

Remark 1. *Note that one could use in the above proof the representations of Equations* (1) *and* (2), *instead of* (10) *and* (11). *Such a choice, would reveal that the sequence* $\{\tilde{\alpha}_k\}_{k\geq 0}$ *constitutes the eigenvalues of matrix* $\boldsymbol{R}$, *with eigenvalues* $\tilde{\alpha}_k$, $k \geq 1$, *having an algebraic multiplicity of 2, since* $\tilde{\alpha}_{2k} = \tilde{\alpha}_{2k-1}$, $k \geq 1$, *and geometric multiplicity equal to 2, since for eigenvalue* $\tilde{\alpha}_{2k}$ *there are two eigenvectors*

$$\begin{pmatrix}\tilde{e}_{2k-1} & \tilde{c}_{2k-1}\tilde{\beta}_{2k-1} & \tilde{c}_{2k-1}\tilde{\beta}_{2k-1}^2 & \cdots\end{pmatrix}$$
$$\begin{pmatrix}\tilde{e}_{2k} & \tilde{c}_{2k}\tilde{\beta}_{2k} & \tilde{c}_{2k}\tilde{\beta}_{2k}^2 & \cdots\end{pmatrix}$$

which if added together and taking into account the connections between the various representations produce exactly the eigenvector $\boldsymbol{h}_k$, $k \geq 1$, *appearing in Theorem 1.*

4.1.1 Numerical evaluation of matrix $\boldsymbol{R}$

It is known, see [1, 4], that the sequences $\{\alpha_k\}_{k\geq 0}$ and $\{\beta_k\}_{k\geq 0}$ decrease exponentially fast to 0. Based on this fact, we suggest to truncate the dimension of the matrix $\boldsymbol{R}$, say at phase N, and obtain its approximation, say $\boldsymbol{R}_N$, as

$$\boldsymbol{R}_N = \boldsymbol{H}_N^{-1}\boldsymbol{D}_N\boldsymbol{H}_N,$$

with $\boldsymbol{D}_N = \text{diag}(\alpha_0, \alpha_1, \ldots, \alpha_{N-1})$ and the matrix $\boldsymbol{H}_N = (\boldsymbol{h}_0^{(N)}, \ldots \boldsymbol{h}_{N-1}^{(N)})$, where $\boldsymbol{h}_k^{(N)} = (h_{k,0}, \ldots, h_{k,N-1})$. Then, as $N \to \infty$ the matrix $\boldsymbol{R}_N = \boldsymbol{H}_N\boldsymbol{D}_N\boldsymbol{H}_N^{-1}$ converges to the infinite matrix $\boldsymbol{R}$, cf. [11]. Furthermore, closely inspecting the structure of the matrix $\boldsymbol{H}_N = (h_{k,m})_{0\leq k,m\leq N-1}$ we observe that it can be written as a generalized Vandermonde matrix for which the inverse can be easily calculated.

5. CONCLUSIONS AND FUTURE WORK

In this manuscript, the authors present a methodological approach that on the one hand permits a straightforward derivation of the equilibrium distribution of random walks in the quadrant satisfying the conditions for meromorphicity and on the other hand connects three existing techniques: the matrix geometric approach, the compensation approach, and the boundary value problem method. Furthermore, the derivation of the eigenvalues and eigenvectors of matrix $\boldsymbol{R}$ sets the groundwork for the probabilistic interpretation of the terms α and β appearing in the expression of the equilibrium distribution. This work can be easily extended to cover a wider spectrum of random walks with meromorphic probability generating functions.

6. REFERENCES

[1] Adan, I.J.B.F. (1991). *A Compensation Approach for Queueing Problems.* PhD dissertation, Eindhoven University of Technology, Eindhoven.

[2] Adan, I.J.B.F., Boxma, O.J. and Resing J.A.C. (2001). Queueing models with multiple waiting lines. *Queueing Systems*, **37**(1-3) 65–98.

[3] Adan, I.J.B.F., Wessels, J. and Zijm, W.H.M. (1993). A compensation approach for two-dimensional Markov processes. *Advances in Applied Probability*, **25**(4) 783–817.

[4] Cohen, J.W. (1994). On a Class of Two-Dimensional Nearest-Neighbour Random Walks *Journal of Applied Probability*, **31** 207-237

[5] Cohen, J.W. (1998). Analysis of the asymmetrical shortest two-server queueing model. *Journal of Applied Mathematics and Stochastic Analysis*, **11**(2) 115–162.

[6] Cohen, J.W. and Boxma, O.J. (1983). *Boundary Value Problems in Queueing System Analysis*, North-Holland, Amsterdam.

[7] Fayolle, G. and Iasnogorodski, R. (1979). Two coupled processors: the reduction to a Riemann-Hilbert problem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **47** 325–351.

[8] Fayolle, G., Iasnogorodski, R. and Malyshev, V. (1999). *Random Walks in the Quarter Plane*, Springer-Verlag, New York.

[9] Gertsbakh, I. (1984). The shorter queue problem: A numerical study using the matrix-geometric solution. *European Journal of Operational Research*, **15**(3) 374–381.

[10] Kingman, J. F. (1961). Two similar queues in parallel. *The Annals of Mathematical Statistics*, **32**(4) 1314–1323.

[11] Shivakumar, P.N. and Sivakumar, K.C. (2009). A review of infinite matrices and their applications. *Linear Algebra and its Applications*, **430**(4) 976–998.

[12] Kroese, D. P., Scheinhardt, W. R. W., and Taylor, P. G. (2004). Spectral properties of the tandem Jackson network, seen as a quasi-birth-and-death process. *Annals of Applied Probability*, **14**(4) 2057–2089.

[13] Neuts, M.F. (1981). *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore.

[14] Li, H., Miyazawa, M., and Zhao, Y. Q. (2007). Geometric decay in a QBD process with countable background states with applications to a join-the-shortest-queue model. *Stochastic Models*, **23**(3) 413–438.

SIR epidemics with stages of infection

[Extended Abstract]

Claude Lefèvre
Université Libre de Bruxelles
Campus de la Plaine C.P. 210
B-1050 Bruxelles, Belgium
clefevre@ulb.ac.be

Matthieu Simon
Université Libre de Bruxelles
Campus de la Plaine C.P. 210
B-1050 Bruxelles, Belgium
matsimon@ulb.ac.be

ABSTRACT

This work is concerned with a stochastic model for the spread of an epidemic in a closed homogeneously mixing population. We assume that any infective can go through several stages of infection before being removed. The transitions between stages are governed by either a Markov process or a semi-Markov process. An infective of any stage makes contacts amongst the population at the epochs of a Poisson process. Our main purpose is to derive the distribution of the final epidemic size and severity by using simple matrix analytic methods and martingale arguments.

CCS Concepts

•Mathematics of computing → Markov processes; Renewal theory;

Keywords

Epidemic models, final size and severity, Markovian or semi-Markovian infection process, matrix analytic methods.

1. INTRODUCTION

SIR epidemic models describe the spread of an infectious disease in a closed homogeneously mixing population subdivided into three classes of individuals: the susceptibles, the infectives and the removed cases. In short, an infective remains infectious during a random period of time. While infected, it can contact all the susceptibles present, independently of the other infectives. At the end of the infectious period, it becomes a removed case and has no further part in the infection process. The epidemic ceases as soon as there are no more infectives in the population.

The final size of the epidemic is defined to be the number of initial susceptibles who ultimately become infected. Its distribution and various approximations have received considerable attention in the literature, as well as some related statistics. Much on the epidemic theory can be found in the books by Daley and Gani [2] and Andersson and Britton [1].

We represent an infectious period as a set of different stages that the infective can go through before being removed. The transitions between stages are ruled by either a Markov process or a more general semi-Markov process. In each stage, an infective makes contacts at the epochs of a Poisson process with a rate specific to each stage. Full details are presented in Lefèvre and Simon [3].

We determine the exact final epidemic outcome for this class of epidemic models with phases. The evolution through infection phases is modelled by a Markovian process in Section 2 and by a semi-Markovian process in Section 3. The analysis exploits matrix analytic methods and simple martingale arguments. This method has the merit to provide us with closed expressions for the final size and severity distributions in terms of parameters that are explicitly calculable. It also points out how the only study of the contaminations made per a single infective amongst a fixed number of susceptibles enables us to determine the final behaviour of the epidemic.

2. MARKOVIAN INFECTION PROCESS

2.1 Model

Any infective stays infectious during a random period of time before being removed. During this period, it can go through L different stages of infection (corresponding to the degree of infectiousness, light or strong for example); these stages are labelled $j = 1, \dots, L$. At the end of the period, the infective is removed in one of the R possible removal states (corresponding to the form of removal, death or immunization for example); these states are denoted $\odot_r$, $r = 1, \dots R$. The evolution through infection phases is modelled by a Markov process $\{\varphi(t),\, t \geq 0\}$ with state space $\{\odot_1, \dots, \odot_R, 1, \dots, L\}$ and whose generator is of the form

$$Q = \left[\begin{array}{ccc|c} \multicolumn{3}{c|}{0} & 0 \\ \hline \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_R & A \end{array}\right],$$

where A is a $L \times L$ matrix that contains the transition rates between the L phases of infection and $\boldsymbol{a}_r$ is the column vector containing the removal rates in state $\odot_r$ from stage r.

Now, while in any phase j of infection, the infective makes contact at the epochs of a Poisson process of rate β_j. Each contact is with an individual chosen at random from the n

initial susceptibles. Thus, if there are s susceptibles present, one of them is contacted according to a Poisson process of rate $s\beta_j/n$. Let B be the diagonal matrix whose non-null entries are the L rates β_j. When a susceptible is contacted, it becomes infected and begins an infection process in one of the L stages according to the probability row vector $\boldsymbol{\alpha}$. We further assume that all the infectives behave independently.

2.2 Final epidemic outcome

Consider a population that initially counts n susceptibles and m_j infectives in stage j of infection. For the new infectives, the infectious periods are i.i.d. and distributed as a random variable $D =_d PH(\boldsymbol{\alpha}, A)$. For the infectives initially in stage j, the infectious periods are i.i.d. and distributed as $D_j =_d PH(\boldsymbol{e}_j, A)$, where $\boldsymbol{e}_j$ is the j-th vector of the usual basis of $\mathbb{R}^L$. The epidemic terminates at time T when all the infectives are removed.

At the end of the epidemic, there are in the population S_T susceptibles and $R_T^{(r)}$ removed cases, $r = 1, \ldots, R$. A measure of the virulence of the disease is given by the final epidemic size $n - S_T$. A complementary measure is the severity A_T defined as the cumulative total duration of infection, which is also the area under the trajectory of the infectives. Our aim is to determine the joint distribution of the statistics S_T, A_T and $R_T^{(r)}$. To achieve it, we are going to make an artificial time change as e.g. in Lefèvre and Utev [4]. Specifically, we follow the successive occurrences of removals one after the other. This gives a representation of the epidemic using a discrete-time scale $t = 0, 1, 2, \ldots$. Let S_t, I_t and $R_t^{(r)}$ be the numbers of susceptibles, infectives in stage j and removal cases of type r at time t. By construction, $t + S_t + I_t = n + m$ for all t, with $m = m_1 + \ldots + m_L$. Thus, the epidemic terminates at time $\tilde{T}$ when $\tilde{T} = \inf\{t : t + S_t = n + m\}$. Of course, the time change modifies the true order of events, but it is easily seen that the final outcome is not affected, i.e. $S_{\tilde{T}}$, $A_{\tilde{T}}$ and $R_{\tilde{T}}^{(r)}$ are distributed as in the real course of time.

Suppose that the t-th removal is the removal of an infective that started in stage j. Denote by D_j its infectious period, by $1_{j,r}$ the indicator that it becomes removed with type r and by $1_j(k)$ the indicator that a group of k fixed susceptibles escapes infection from the infective. The artificial time provides then the following relations:

$$\begin{aligned}
\binom{S_t}{k} &= \sum_{u=1}^{\binom{S_{t-1}}{k}} 1_j(k;u), \text{ with } S_0 = n,\\
A_t &= A_{t-1} + D_j, \text{ with } A_0 = 0,\\
R_t^{(r)} &= R_{t-1}^{(r)} + 1_{j,r}, \text{ with } R_0^{(r)} = 0,
\end{aligned}$$

where the $1_j(k;u)$ are i.i.d. and distributed as $1_j(k)$. Let $\mathcal{F}_t$ be the filtration $\sigma\{S_\tau, A_\tau, R_\tau^{(1)}, \ldots, R_\tau^{(R)}, 0 \leq \tau \leq t\}$. The preceding relations can be used to show that, for each $k = 0, 1, ..., n$, $\theta \geq 0$ and $\boldsymbol{z} \in \mathbb{R}^R$, the process

$$\left\{\binom{S_t}{k} \frac{e^{-\theta A_t}}{q(k,\theta,\boldsymbol{z})^t} \prod_{r=1}^{R} {z_r}^{R_t^{(r)}} \quad t \geq m\right\}$$

is a $\mathcal{F}_t$-martingale if the parameters $q(k,\theta,\boldsymbol{z})$ and $q_j(k,\theta,\boldsymbol{z})$ are defined as

$$\begin{aligned}
q_j(k,\theta,\boldsymbol{z}) &= E\left[1_j(k) e^{-\theta D_j} \prod_{r=1}^{R} {z_r}^{1_{j,r}}\right], \qquad (1)\\
q(k,\theta,\boldsymbol{z}) &= \sum_{j=1}^{L} \alpha_j q_j(k,\theta,\boldsymbol{z}), \qquad (2)
\end{aligned}$$

for $0 \leq k \leq n$, $1 \leq j \leq L$. Applying the optional stopping theorem on this martingale yields a transform of the joint distribution of S_T, A_T and $R_T^{(r)}$, after having considered the effect of the m initial infectives:

Proposition 1. *For $0 \leq k \leq n$,*

$$\begin{aligned}
&E\left[\binom{S_T}{k} e^{-\theta A_T} q(k,\theta,\boldsymbol{z})^{S_T} \prod_{r=1}^{R} {z_r}^{R_T^{(r)}}\right]\\
&\qquad = \binom{n}{k} q(k,\theta,\boldsymbol{z})^n \prod_{j=1}^{L} q_j(k,\theta,\boldsymbol{z})^{m_j}. \qquad (3)
\end{aligned}$$

In particular, Equation (3) provides a triangular system of $n+1$ linear equations for the final susceptible state probabilities $P(S_T = s)$:

$$\begin{cases}
\sum_{s=k}^{n} \binom{s}{k} q(k)^s \, P(S_T = s) = \binom{n}{k} q(k)^n \prod_{j=1}^{L} q_j(k)^{m_j}\\
\sum_{s=0}^{n} P(S_T = s) = 1
\end{cases}$$

where $q_j(k) = q(k, \mathbf{0}, \mathbf{1})$. The moments of S_T, A_T and $R_T^{(r)}$ can also be obtained from (3).

2.3 Contagion per infective

Equation (3) shows that the final epidemic outcome only depends on the parameters $q_j(k,\theta,\boldsymbol{z})$, that is, we only need to analyse the behaviour of a unique infected facing a group of k susceptibles to determine the final state of the population at time T. Consider an infective who begins its infectious period in a stage given by the probability vector $\boldsymbol{\gamma}$. This infective faces a group of k susceptibles and we want to determine the total number $N_{\boldsymbol{\gamma}}(k)$ of infections generated by this single infective. The new infected cases will be here supposed to be directly removed. We can model the behaviour of the infected by the Markov process $\{[N_{\boldsymbol{\gamma}}(k;t), \varphi(t)], \, t \in \mathbb{R}^+\}$ where $N_{\boldsymbol{\gamma}}(k;t)$ is the number of susceptibles contacted up to time t and $\varphi(t)$ is the stage of infection at time t; the removal of the infective corresponds to the absorption in any state $\odot_r$. Thus, the state space of the process is

$$\{\odot_1, ..., \odot_R, [(0,1), ..., (0,L)], \ldots, [(k,1), ..., (k,L)]\},$$

and the associated generator is given by

$$\left[\begin{array}{c|ccccc}
0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\
\hline
\boldsymbol{a}_1 \cdots \boldsymbol{a}_R & A_0(k) & A_1(k) & 0 & \cdots & 0\\
\boldsymbol{a}_1 \cdots \boldsymbol{a}_R & 0 & A_0(k-1) & A_1(k-1) & \cdots & 0\\
\boldsymbol{a}_1 \cdots \boldsymbol{a}_R & 0 & 0 & A_0(k-2) & \cdots & 0\\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\
\boldsymbol{a}_1 \cdots \boldsymbol{a}_R & 0 & 0 & 0 & \cdots & A_1(1)\\
\boldsymbol{a}_1 \cdots \boldsymbol{a}_R & 0 & 0 & 0 & \cdots & A
\end{array}\right],$$

where $A_1(h) = hB/n = A - A_0(h)$ for $1 \leq h \leq k$. In particular, the parameters $q_j(k,\theta,\boldsymbol{z})$ can be derived from this Markov process, by using probabilistic arguments:

PROPOSITION 2. *For $1 \leq j \leq L$,*

$$q_j(k,\theta,\boldsymbol{z}) = \boldsymbol{e}_j \,[\theta I - A_0(k)]^{-1} \sum_{r=1}^{R} z_r \boldsymbol{a}_r.$$

The same formula holds for $q(k,\theta,\boldsymbol{z})$ except that $\boldsymbol{\alpha}$ is substituted for $\boldsymbol{e}_j$.

3. SEMI-MARKOV INFECTION PROCESS

In this Section, we adapt the above results to the case where the evolution of an infective through stages of infection is described by a semi-Markov process $\{\varphi(t),\, t \geq 0\}$. The state space of the contagion process is the same as in Section 2.1 and the semi-Markov kernel is of the form

$$Q(t) = \left[\begin{array}{ccc|c} & I & & 0 \\ \hline \boldsymbol{a}_1(t) & \ldots & \boldsymbol{a}_R(t) & A(t) \end{array}\right],$$

where, if τ denotes the first renewal time in the associated Markov renewal process,

$$\begin{aligned} A_{j,v}(t) &= P[\tau \leq t, \varphi(\tau) = v \mid \varphi(0) = j], \\ (\boldsymbol{a}_r)_j(t) &= P[\tau \leq t, \varphi(\tau) = \odot_r \mid \varphi(0) = j], \end{aligned}$$

for $1 \leq r \leq R, 1 \leq j, v \leq L$. We assume that the time before removal is finite a.s. As before, an infective in phase j can infect any given susceptible according to a Poisson process of rate β_j/n. If infected, a susceptible begins in one of the L stages according to the row vector $\boldsymbol{\alpha}$.

By comparison with Section 2, only the infection process has been modified here. Thus, Proposition 1 still holds but, of course, the expression for $q_j(k,\theta,\boldsymbol{z})$ needs to be adapted. To that aim, consider as before a single infective who is facing k susceptibles and begins its infectious period in a stage given by $\boldsymbol{\gamma}$. As in Section 2.3, we want to evaluate the total number $N_{\boldsymbol{\gamma}}(k)$ of susceptibles that will be contacted by this single infective. For that, we construct the semi-Markov process $\{[N_{\boldsymbol{\gamma}}(k;t), \varphi(t)],\, t \geq 0\}$ where $\varphi(t)$ is the stage of infection at time t and $N_{\boldsymbol{\gamma}}(k;t)$ is the number of susceptibles contacted up to time t but only updated at the successive renewal instants of $\varphi(.)$. For instance, if the first renewal occurs at time τ and the infective contacts three susceptibles during the period $(0,\tau)$, then $N_{\boldsymbol{\gamma}}(k;t) = 0$ for $t < \tau$ and $N_{\boldsymbol{\gamma}}(k;\tau) = 3$. Although such a clock does not count the infections in real time, we observe that it does not modify the total number of contacts made by the infective. The removal of the infective leads to the absorption in a state $\odot$. So, the state space of the process is

$$\{(0,\odot), \cdots, (k,\odot), [(0,1), ..., (0,L)], \ldots, [(k,1), ..., (k,L)]\},$$

and the associated semi-Markov kernel is given by

$$\left[\begin{array}{ccc|ccc} & I & & & 0 & \\ \hline \boldsymbol{u}_{kk}(t) & \cdots & \boldsymbol{u}_{k\,0}(t) & \mathcal{U}_{kk}(t) & \cdots & \mathcal{U}_{k\,0}(t) \\ \boldsymbol{0} & \cdots & \boldsymbol{u}_{k-1\,0}(t) & 0 & \cdots & \mathcal{U}_{k-1\,0}(t) \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{u}_{00}(t) & 0 & \cdots & \mathcal{U}_{00}(t) \end{array}\right],$$

where, if $Y(t)$ denotes the number of susceptibles present at time t, the matrices $\mathcal{U}_{hl}(t)$ and the vectors $\boldsymbol{u}_{hl}(t)$, $0 \leq h \leq s$, $0 \leq l \leq h$, are defined as

$$\begin{aligned} (\mathcal{U}_{hl})_{j,v}(t) &= P[\tau \leq t, Y(\tau) = l, \varphi(\tau) = v], \\ (\boldsymbol{u}_{hl})_j(t) &= P[\tau \leq t, Y(\tau) = l, \varphi(\tau) = \odot], \end{aligned}$$

conditionally to $Y(0) = h, \varphi(0) = j$, and for $1 \leq j, v \leq L$. As for the Markovian case, the calculation of the coefficients can be done by using the structure of this last kernel, as well as some probabilistic arguments. This leads to the following final expression:

PROPOSITION 3. *For $1 \leq j \leq L$,*

$$q_j(k,\theta,\boldsymbol{z}) = \boldsymbol{e}_j \,[I - C_k(\theta)]^{-1} \sum_{r=1}^{R} z_r \boldsymbol{c}_{k,r}(\theta),$$

where for $0 \leq k \leq n$,

$$\begin{aligned} (C_k)_{j,v}(\theta) &= \widehat{A}_{j,v}(\theta + k\beta_j/n), \quad 1 \leq v \leq L, \\ (\boldsymbol{c}_{k,r})_j(\theta) &= (\widehat{\boldsymbol{a}}_r)_j(\theta + k\beta_j/n), \quad 1 \leq r \leq R, \end{aligned}$$

where $\widehat{A}_{j,v}$ and $(\widehat{\boldsymbol{a}}_r)_j$ are the Laplace transforms of the functions $A_{j,v}$ and $(\boldsymbol{a}_r)_j$. The same formula holds for $q(k,\theta,\boldsymbol{z})$ except that $\boldsymbol{\alpha}$ is substituted for $\boldsymbol{e}_j$.

4. REFERENCES

[1] H. Andersson and T. Britton. *Stochastic Epidemic Models and their Statistical Analysis.* Lecture Notes in Statistics 151, New York, 2000.

[2] D. Daley and J. Gani. *Epidemic Modelling: an Introduction.* Cambridge University Press, Cambridge, 1999.

[3] C. Lefèvre and M. Simon. SIR epidemics with stages of infection. *To appear in Advances in Applied Probability*, 48, 2016.

[4] C. Lefèvre and S. Utev. Branching approximation for the collective epidemic model. *Methodology and Computing in Applied Probability*, 1:211–228, 1999.

Modelling Mortality and Discharge of Hospitalized Stroke Patients using a Phase-Type Recovery Model

Bruce Jones
Department of Statistical and Actuarial Sciences
Western University
London, Ontario, Canada N6A 5B7
jones@stats.uwo.ca

Sally McClean
School of Computing and Information Engineering
Ulster University
Coleraine, Northern Ireland BT52 1SA, UK
si.mcclean@ulster.ac.uk

David Stanford
Department of Statistical and Actuarial Sciences
Western University
London, Ontario, Canada N6A 5B7
stanford@stats.uwo.ca

ABSTRACT

We model the length of in-patient hospital stays due to stroke and the mode of discharge using a phase-type stroke recovery model. The model allows for three different types of stroke: haemorrhagic (the most severe, caused by ruptured blood vessels that cause brain bleeding), cerebral infarction (less severe, caused by blood clots) and transient ischemic attack or TIA (the least severe, a mini-stroke caused by a temporary blood clot). A four-phase recovery process is used, where the initial phase depends on the type of stroke, and transition from one phase to the next depends on the age of the patient. There are three differing modes of absorption: from a typical recovery phase, a patient may die (mode 1), be transferred to a nursing home (mode 2) or be discharged to the individual's usual residence (mode 3).

The first recovery phase is characterized by a very high rate of mortality and very low rates of discharge by the other two modes. The next two recovery phases have progressively lower mortality rates and higher mode 2 and 3 discharge rates. The fourth recovery phase is visited only by those who experience a very mild TIA, and they are discharged to home after a short stay.

The model has practical value as it yields length of stay distributions by age and type of stroke, which are useful in resource planning. Also, inclusion of the three modes of discharge permits analyses of outcomes.

1. BACKGROUND

Due to the debilitating nature of a stroke and complex makeup of the disease there is an urgent need for stochastic models that can be used for bed occupancy analysis, capacity planning, performance modeling and prediction, with a view to decreasing patient delays, better use of resources, and improved adherence to targets.

Modeling length-of-stay (LOS) in hospital is an important aspect of characterising patient stay in hospital and outcomes in the form of discharge destinations. We focus on using easily accessible administrative data routinely collected at discharge. Such data, which include information on patient date of birth, date of admission, diagnosis and discharge date, are not appropriate for patient prognostication but can rather been aimed towards supporting planning, service organization, and allocation of resources, e.g. Shahani et al. (2008), Faddy and McClean (2000), Marshall and McClean (2003). In such cases we are interested in the behavior of future patient populations rather than individuals, with a focus on system wide planning.

Heterogeneity of patient pathways and LOS characteristics have been investigated by a number of authors. Such heterogeneity arises from a number of sources, for example, method of admission, diagnosis, severity of illness, age, gender, and treatment e.g. Faddy and McClean 2005, Marshall and McClean 2004, Harper et al. 2012. Such covariates have previously been incorporated into phase-type models via conditional phase-type models Marshall and McClean (2003), a Coxian proportional hazards approach (Faddy and McClean, 2000) and classification trees (Harper et al., 2012).

2. STROKE IN-PATIENT DATA

We here focus on incorporating age and diagnosis into a model of stroke patients in hospital. The modelling was based on five yearsâĂŹ retrospective data for patients admitted to the Belfast City Hospital with a diagnosis of stroke (cerebral hemorrhage, bleed in the brain; cerebral infarction, clot in the brain; transient ischemic attack, minor stroke; and unspecified or undetermined type of stroke). Data were obtained from the Patient Administration System, PAS (a computerized system that records patient activity relating to inpatients, outpatients, and waiting lists, A&E and case note tracking). Data retrieved from the Belfast City Hospital PAS included age, diagnosis, LOS, and discharge destination. Diagnosis and age were previously shown to be highly significant with regard to LOS (McClean et al., 2011) Our approach then models the patient journey through hospital as a phase-type model incorporating diagnosis and age.

The large number of deaths after a relatively short average length of stay confirms the high mortality rate of hemorrhagic stroke patients. Also, the large number of discharges to home after a very short average stay indicates the very mild nature of the TIA type of stroke. Summary information is shown in Table 1.

Table 1: Summary by Type of Stroke and Mode of Discharge

Discharge Counts			
Mode of Discharge	Hemorrhagic	Cerebral Infarction	TIA
Death	65	125	13
Nursing Home	5	59	8
Usual Residence	69	432	389
Average Lengths of Stay (days)			
Mode of Discharge	Hemorrhagic	Cerebral Infarction	TIA
Death	18.3	34.6	37.5
Nursing Home	85.5	83.7	25.8
Usual Residence	51.3	31.9	8.2

3. MODEL DETAILS

The model which we have decided upon strikes us as the best compromise between allowing for sufficient distinction of the various types of stroke to be considered, while maintaining a reasonable level of parsimony for parameter estimation. In fact, when we tried to make the model smaller as a check, there was a statistically significant reduction in the loglikelihood that was more than would be justified by the reduction in the number of parameters. Also, both our goodness of fit tests and the work of McClean et al. (2011) show that phase-type models with this level of simplicity tend to fit this hospital length of stay data well. Furthermore, we wish to be able to distinguish paths based upon type of stroke incurred, and ultimate disposition upon absorption. The model's state transition diagram is shown in Figure 1, although initially we made provision for transitions from all transient recovery states to all modes of discharge.

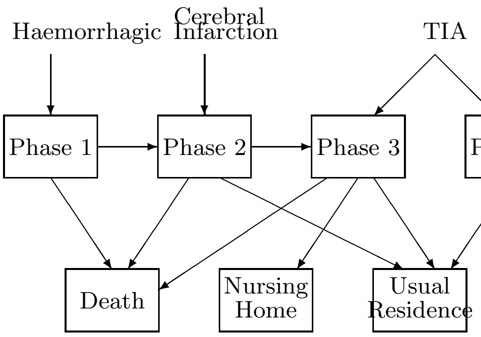


Figure 1: State Transition Diagram

Since haemorrhagic strokes are generally the most debilitating, we anticipate three recovery phases for such patients. Those suffering a cerebral infarction, typically less severe in its debilitation, involve the latter two of the recovery phases that haemorrhagic patients encounter. TIAs are the least severe of all, but our study of the data revealed that they appear to comprise two distinction groups in terms of the duration of their recovery period. TIA patients with longer recoveries involve the last recovery phase of haemorrhagic & infraction patients, with a distinct state for very short TIA durations.

Transition rates for the model comprise some that depend upon the age and stroke type of the patient, and others independent of age. Both for reasons of parsimony and for simplicity, we have chosen exponential forms for selected transition intensities and for the mixing probability for TIA patients; this ensures positive rates for the former and results in the interval (0,1) for the latter. For $i = 1, 2$, let $\lambda_i(x)$ be the transition intensity from phase i to phase $i+1$ for a patient who is age x, which we formulate as taking the form $\lambda_i(x) = \exp(\gamma_i + \beta_i x)$. Also, let $p(x)$ represent the probability that a TIA stroke patient age x is in recovery phase 4 upon admission (representing the less severe TIAs); the other TIA patients start in phase 3. We assume that $p(x) = \exp\{-\exp(\theta_0 + \theta_1 x)\}$. As regards the parameters which are independent of age, for $i = 1, \ldots, 4$, μ_i denotes the mortality rate from phase i, ν_i the rate of discharge to a nursing home from phase i, and ρ_i the rate of discharge to the usual residence from phase i. As indicated in Figure 1, it is assumed that $\mu_4 = 0$ and $\nu_4 = 0$.

Let $\mathbf{T} = (t_{ij})$ be a 4×4 matrix containing the transition rates among the transient states and $\mathbf{T}_A = (t_{ij});\ i = 1, 2, 3, 4;\ j = 5, 6, 7$ is a 4×3 matrix containing the absorption rates for the various discharge modes (death, nursing home, and usual residence, respectively). Then the usual densities and distributions for the stroke length of stay X are readily obtained in the usual way; for instance, given an initial distribution of recovery phases $\boldsymbol{\alpha}$, one finds

$$f_X(x\,|\boldsymbol{\alpha}, \mathbf{T}, \mathbf{T}_A) = \boldsymbol{\alpha}' \exp(\mathbf{T}x)\, \mathbf{T}_A \mathbf{1}_3\,,\ x \geq 0\,. \qquad (1)$$

where $\mathbf{1}_3$ denotes a column matrix of ones of length three.

The 4×3 matrix $\mathbf{P} = (-\mathbf{T})^{-1}\mathbf{T}_A$ is of interest in its own right, and can be interpreted as the probability of being absorbed into the various discharge modes (death, nursing home, or regular residence), for each of the recovery phases. This is equivalent to similar expressions previously given in McClean et al. (2011). We are also interested in length of stay by discharge mode, which can be obtained from the conditional distributions by dividing by the ultimate probability of absorption into the appropriate states.

4. ESTIMATION

The total number of parameters to be estimated is 16. We use maximum likelihood estimation. The likelihood function is given by the product over all 1,234 stroke patients of the likelihood contribution for that patient. The likelihood contribution for a given patient is (proportional to) the model probability or probability density associated with our observation for that patient.

Let L_j represent the likelihood contribution for patient j. Space limitations prevent us from presenting expressions for L_j herein; the interested reader can obtain the fuller form of the paper from the authors.

The results of the initial estimation process showed that three of the parameters were not meaningfully different from zero (namely, ν_1, ν_2 and ρ_1), as the associated p-values were all in excess of 90 percent. We therefore revised the original state transition diagrams to eliminate these transitions. These eliminations also make sense, in that Phase 1 pertains to seriously ill patients, for whom any sort of discharge other than by death is unrealistic, while there would be no reason to transfer patients from Phase 2 to a nursing home without availing of the normal amount of recovery time provided by Phase 3. The resulting diagram appears in Figure 1.

We then obtained the revised parameter estimates shown in Table 2. An asymptotic covariance matrix is obtained as the inverse of the observed information matrix evaluated at the maximum likelihood estimates. The latter matrix is found as a byproduct of the numerical method used to maximize the log-likelihood function. The standard error estimates shown in Table 2 are the square roots of the diagonal elements of the asymptotic covariance matrix. The Z-statistics are the parameter estimates divided by the standard errors. Each can be used to test the hypothesis that the corresponding parameter equals 0. The p-values, based on asymptotic normality of the parameter estimators, indicate rather strong evidence against the hypothesis in each case.

Table 2: Parameter Estimates

Parameter	Estimate	Std Error	Z-Stat	p-value
γ_1	6.63570	1.21893	5.44388	0.00000
β_1	-0.03652	0.01631	-2.23902	0.02515
γ_2	-3.06931	1.22697	-2.50153	0.01237
β_2	0.07153	0.01667	4.29057	0.00002
θ_0	-8.66118	1.48644	-5.82680	0.00000
θ_1	0.08801	0.01828	4.81391	0.00000
μ_1	22.10156	4.95434	4.46105	0.00001
μ_2	2.48820	0.37993	6.54912	0.00000
μ_3	1.56162	0.20294	7.69509	0.00000
ν_3	1.27849	0.17391	7.35165	0.00000
ρ_2	11.76860	0.99634	11.81180	0.00000
ρ_3	3.41989	0.38393	8.90762	0.00000
ρ_4	63.92514	4.11394	15.53865	0.00000

In order to check the fit of our model, we considered comparisons of nonparametric estimates of the cumulative intensity functions for the different modes of discharge with estimates of the cumulative intensity function based on our fitted model. Since the latter estimates depend on age at admission, we examined nonparametric estimates for three age groups. Specifically, we plotted the well-known Nelson-Åalen estimates of the cumulative intensity function for each of the age intervals $[60, 70)$, $[70, 80)$ and $[80, 90)$ for each type of stroke and each mode of discharge for which we have a meaningful number of discharges. (Due to limited space in this volume, interested readers are invited to contact the authors for a longer version which includes these graphs.)

5. RESULTS

At this point, we present several examples of what the model can be used for. The first such measure we present is the "destination" matrix P for an individual aged 65, 75, and 85, respectively, indicating the likelihoods of the possible destinations upon discharge for each initial recovery phase.

The table shows the notable dependence upon age of the likelihoods of death and discharge to the usual residence for an individual who suffers a haemorrhagic stroke. The likelihood of death increases with age, and there is a slight increase as well in the number of discharges to a nursing home. For cerebral infractions, there is a slight increase in the chance of death with age, but the more notable increase is in the chance of discharge to a nursing home. Qualitatively, these relative likelihoods are in keeping with what one might anticipate from the relative severity of these two types of stroke.

Table 3: Ultimate Destination Percentage by Age and Type of Stroke

	Age 65		
	Death	Nursing Home	Usual Residence
Haemorrhagic	38.5	4.0	57.5
Cerebral Infarction	19.4	5.2	75.5
TIA complex	24.9	20.4	54.6
TIA simple	0	0	100.0
	Age 75		
	Death	Nursing Home	Usual Residence
Haemorrhagic	45.1	5.8	49.1
Cerebral Infarction	20.5	8.4	71.1
TIA complex	24.9	20.4	54.6
TIA simple	0	0	100.0
	Age 85		
	Death	Nursing Home	Usual Residence
Haemorrhagic	52.5	7.3	40.1
Cerebral Infarction	21.9	12.0	66.1
TIA complex	24.9	20.4	54.6
TIA simple	0	0	100.0

Figure 2 presents the cumulative probability of discharge as a function of time by the type of stroke for each of the modes of discharge: death, nursing home, and usual residence (i.e. home). For haemorrhagic strokes (top panel), we see that the deaths that occur tend to happen quickly, with most of them happening within the first 10 days since onset of the stroke. The discharges to the usual residence take much longer, as an extended period is needed to pass through the corresponding recovery phases before being discharged home. With few cases on record of discharge to a nursing home, little can be inferred, other than the fact that it, too, tends to take a lot of time.

Figure 2's middle panel reveals that the chance of death is markedly reduced in the case of cerebral infractions, and that those deaths that do occur tend to happen over the course of the stay. The likelihood of discharge to a nursing home is also greater than either of the other types of stroke; nenetheless, death is about twice as likely as discharge to nursing home. The situation for TIAs (bottom panel) is rather straightforward, with in excess of ninety percent of patients being discharged to their usual residence.

6. CONCLUSIONS AND FURTHER WORK

We have developed a phase-type modelling approach with particular applicability to stroke patient care. The model includes a number of absorbing states to account for the possible outcomes for patients (discharge to normal residence, nursing home, or death). Among these, of special note are discharges to private nursing homes, which may be responsible for bottlenecks, and resulting delayed discharge, which can have a significant effect on expected LOS in hospital (a key performance metric).

Based on data for stroke patients from the Belfast City Hospital, various scenarios have been explored with a focus on modelling phases which represent different recovery

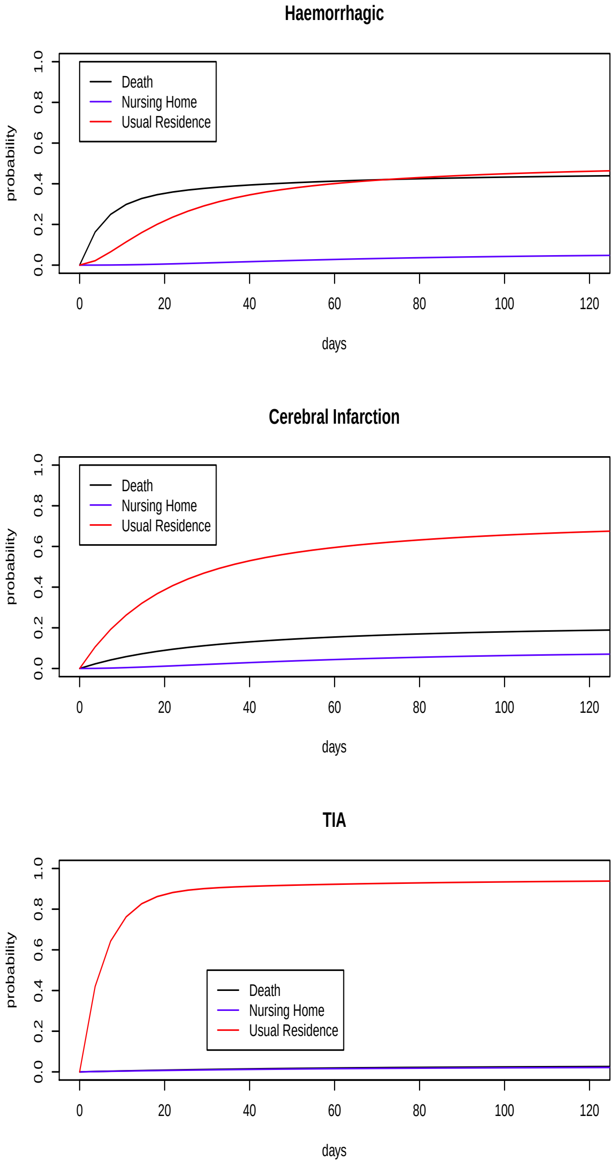


Figure 2: Cumulative probability of discharge by type of stroke and destination

phases whose transition rates are functions of important covariates, in this case age. The admission phase is characterised by the type of stroke, as different types of strokes have corresponding severity of illness and outcome. The results demonstrate the relationship between phase of discharge and expected total LOS, with its inherent impact on bed occupancy. By exploring such scenarios, the key mechanisms for delay can thus be explored and their impact assessed.

Our current analytic model has the advantage that the results are based on routinely available discharge data. Our current framework represents initial work towards developing integrated models for stroke services, including both hospital and community care, with the aim of supporting integrated planning. However, we believe that this approach also has considerable potential to include more detailed and explicit models of stroke services that allow us to assess complex scenarios involving interactions between services. Another important aspect of extending our current framework is to attach costs to various options within the model. For example, we would like to be able to answer questions such as: should additional resources be put into thrombolysis for patients immediately after they have suffered a stroke, or is it better to focus on rehabilitative services in the community? Stroke is an excellent paradigm example enabling modelling of a whole health and social care system. The experience gained and techniques learned are likely to be relevant to the health and care of older persons in general. Phase-type models have an important role in this work.

7. REFERENCES

[1] Faddy, M. J. and McClean, S. I. (2000). Analysing data on lengths of stay of hospital patients using phase-type distributions. Appl Stoch. Models and Data Anal. 15, 311-317.

[2] Faddy, M. J. and McClean S. I. (2005). Markov chain modeling for geriatric patient care. Meth. Inform. Med. 44, 369-373.

[3] Harper, P.R., Knight, V.A., Marshall, A.H. (2012). Discrete Conditional Phase-type models utilising classification trees: Application to modelling health service capacities. Eur. J. Oper. Res. 219(3), 522-530.

[4] Marshall, A. H. and McClean, S. I. (2003). Conditional phase-type distributions for modeling patient LOS in hospital. Intern. Trans. Oper. Res. 10, 6, 565-576.

[5] Marshall, A. H. and McClean, S. I. (2004). Using Coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. Health Care Manag. Sci. 7, 285-289.

[6] McClean S. I., M. Barton, L. Garg, and K. Fullerton (2011). Combining Analytical and Simulation approaches to model Patient Flows, ACM Transactions on Modelling and Computer Simulation (TOMACS), 21(4): 25.

[7] Shahani, A. K., Ridley, S. A., and Nielsen, M. S. (2008). Modelling patient flows as an aid to decision-making for critical care capacities and organisation. Anaesthesia. 63, 10, 1074-1080.