Probability in the Engineering and Informational Sciences, **30**, 2016, 212–223.
doi:10.1017/S026996481500039X

THE RUNNING MAXIMUM OF A LEVEL-DEPENDENT QUASI-BIRTH–DEATH PROCESS

MICHEL MANDJES
Korteweg-de Vries Institute for Mathematics,
University of Amsterdam,
Science Park 904, 1098 XH Amsterdam,
the Netherlands

CWI, P.O. Box 94079,
1090 GB Amsterdam,
the Netherlands

Eurandom, Eindhoven University of Technology,
Eindhoven,
the Netherlands

IBIS, Faculty of Economics and Business,
University of Amsterdam,
Amsterdam,
the Netherlands
E-mail: m.r.h.mandjes@uva.nl

PETER TAYLOR
School of Mathematics and Statistics,
University of Melbourne,
Victoria 3010,
Australia
E-mail: p.taylor@ms.unimelb.edu.au

The objective of this note is to study the distribution of the running maximum of the level in a level-dependent quasi-birth–death process. By considering this running maximum at an exponentially distributed "killing epoch" T, we devise a technique to accomplish this, relying on elementary arguments only; importantly, it yields the distribution of the running maximum jointly with the level and phase at the killing epoch. We also point out how our procedure can be adapted to facilitate the computation of the distribution of the running maximum at a deterministic (rather than an exponential) epoch.

1. INTRODUCTION

A *birth–death process* $X(t)$ is a continuous-time Markov chain defined on the non-negative integers $\{0,1,\ldots\}$ which has the special property that it can jump only one level up or down at a time – one typically writes λ_i for the upward rate from state $i \in \{0,1,\ldots\}$, and μ_i for the downward rate from state $i \in \{1,2,\ldots\}$. A *quasi-birth–death process* (usually abbreviated to QBD) is a generalization of the birth–death process, in which the upward

 0269-9648/15 $25.00

and downward transition rates depend on the state of an underlying Markov chain, usually known as the *phase*; see, for example, [12,13] and [5, Section XI.3c]. In the standard variant, the phase process $J(t)$ is modeled as an irreducible, finite-state, continuous-time Markov chain, attaining values in $\{1, \ldots, d\}$ for some $d \in \mathbb{N}$, but more general variants have been proposed in the literature. One such generalization, discussed in, for example, [7,14], is the *level-dependent* QBD in which arrival rates and departure rates depend on the current level, and in which the phase process is also level-specific, defined on the state space $\{1, \ldots, d_i\}$ when the level is i, with $d_i \in \mathbb{N}$.

There are many examples of level-dependent QBD models of interest to researchers. In Section 5, we shall analyze a Markov-modulated $M/M/\infty$ queue, in which the service rates and arrival rates of an $M/M/\infty$ queue depend on the state of an underlying Markov chain, and a generalization of a cable network model of Ellens et al. [9].

Another class of level-dependent QBD models that has attracted considerable interest in the literature is the class of *retrial queues*; see, for example, the survey by Artalejo and Gomez-Corral [4]. In a retrial queue, customers who cannot gain immediate admittance to a finite-capacity queueing system move to an *orbit*, from which they periodically retry to enter the queue. Such queues can be modeled via a QBD state description $(C(t), N(t))$, where $C(t)$ gives the state of the finite-capacity part of the system and $N(t) \in \mathbb{Z}_+$ gives the number of customers in the orbit. The retrial rate is usually a linear function of $N(t)$, which is the reason that the models are level dependent.

Much research on QBDs has concentrated on the (time-dependent and stationary) distribution of the level, typically in combination with the phase. While it is a key object of study in Lévy fluctuation theory [11], considerably less attention has been paid to the distributional properties of the *running maximum* of a QBD process, defined by

$$\bar{X}(t) := \sup_{s \in [0,t]} X(s). \tag{1}$$

The exceptions that the authors are aware of have occurred in the retrial queue literature. There [3,10] provided a computational method to derive the distribution of the maximum queue length attained, either in a busy period or a fixed time, for an $M/M/c$ retrial queue and [2] gave a similar analysis for a more general model of a call center queue, with a specific application to a retrial queue.

The objective of the present paper is to determine the distribution of the running maximum $\bar{X}(t)$ attained by a level-dependent QBD. We first devise a procedure to evaluate the distribution of $\bar{X}(T)$, where T is *exponentially distributed* (say with mean τ^{-1}), conditional on $X(0)$ and $J(0)$. Importantly, we obtain the distribution of $\bar{X}(T)$ *jointly with* the level $X(T)$ and phase $J(T)$ at the exponential "killing epoch" T. The fact that we are able to compute this joint distribution facilitates the evaluation of $\bar{X}(t)$ at a deterministic time $t \geqslant 0$, relying on the concept of "Erlangization", as an efficient and easily implementable alternative to Laplace inversion.

This note is organized as follows. The model is introduced in Section 2, while Section 3 presents the analysis and Section 4 contains a discussion of the computational aspects of implementing our expressions. Section 5 includes a discussion of two examples and Section 6 a short conclusion.

2. MODEL

Consider a level-dependent QBD, which is a bivariate process comprising *levels* and *phases*. The level process, in the sequel denoted by $X(\cdot)$, attains values in $\{0, 1, \ldots\}$. When $X(t) = i$, the phase $J(t)$ attains values in $\{1, \ldots, d_i\}$, for some $d_i \in \mathbb{N}$.

The birth–death nature of a level-dependent QBD is reflected by the fact that the level can increase or decrease by at most 1. Its transition structure is defined below.

- $Q^{(i)}$ is a $d_i \times d_i$ transition rate matrix corresponding to a continuous-time Markov chain, living on state space $\{1,\ldots,d_i\}$, with elements $q^{(i)}_{k\ell} \geqslant 0$. At level i, a jump from phase k to phase ℓ that leaves the level unchanged occurs with rate $q^{(i)}_{k\ell}$, for $k \neq \ell$. In addition, we define the diagonal elements to be such that

$$q^{(i)}_{kk} := -q^{(i)}_{k} = -\sum_{\ell \neq k} q^{(i)}_{k\ell};$$

  where the sum on the right-hand side should be understood to be over all $\ell \in \{1,\cdots,d_i\}$ such that $\ell \neq k$.

- The matrix $\Lambda^{(i)}$ has dimension $d_i \times d_{i+1}$. Its (k,ℓ)th element contains the rate $\lambda^{(i)}_{k\ell} \geqslant 0$ that the level increases by 1 while the phase jumps from k to ℓ; note that $k = \ell$ is now allowed (if $k \leqslant d_{i+1}$). We use the compact notation

$$\lambda^{(i)}_{k} := \sum_{\ell=1}^{d_{i+1}} \lambda^{(i)}_{k\ell},$$

  to denote the total rate corresponding to an increase in level.

- Finally, the (k,ℓ)th element of the matrix $\mathcal{M}^{(i)}$, which has dimension $d_i \times d_{i-1}$, contains the rate $\mu^{(i)}_{k\ell} \geqslant 0$ that the level decreases by 1 while the phase jumps from k to ℓ; again, $k = \ell$ is allowed (if $k \leqslant d_{i-1}$). In the sequel we write

$$\mu^{(i)}_{k} := \sum_{\ell=1}^{d_{i-1}} \mu^{(i)}_{k\ell}$$

  for the total rate of a decrease in level.

We assume that the matrices $Q^{(i)}$ (for $i = 0,1,\ldots$), $\Lambda^{(i)}$ (for $i = 0,1,\ldots$), and $\mathcal{M}^{(i)}$ (for $i = 1,2,\ldots$) are such that the process $(X(t),J(t))$ is irreducible.

As mentioned above, $X(t)$ denotes the level of the level-dependent QBD at time $t \geqslant 0$. The objective of this note is to find, for given $t \geqslant 0$, the distribution of the running maximum $\bar{X}(t)$ defined in (1). In our analysis, we first identify the distribution of $\bar{X}(T)$, jointly with the level $X(T)$ and phase $J(T)$, conditional on having started in $X(0) = i \in \{0,1,\ldots\}$ and $J(0) = k \in \{1,\ldots,d_i\}$. Either by Laplace inversion or by relying on the concept of "Erlangization", this facilitates the numerical computation of the distribution of $\bar{X}(t)$ at a deterministic time t, as will be explained in Section 4.

3. ANALYSIS

As mentioned above, let T be an exponentially distributed random variable with mean τ^{-1}, sampled independently of the level-dependent QBD that we introduced in the previous section. Our primary goal is to derive a computationally efficient method to evaluate the probabilities

$$s_{i,j,m}[k,\ell] := \mathbb{P}(\bar{X}(T) = m, X(T) = j, J(T) = \ell \,|\, X(0) = i, J(0) = k). \tag{2}$$

One can obviously assume that $\max\{i,j\} \leqslant m$, with $k \in \{1,\ldots,d_i\}$ and $\ell \in \{1,\ldots,d_j\}$.

To identify the $s_{i,j,m}[k,\ell]$, observe that the event described on the right-hand side of (2) can be decomposed into events over two time intervals:

- first, the level increases from i to m, and then
- it goes down to j while remaining consistently below $m+1$.

This decomposition is reflected in the two steps discussed below. Letting $T_j \geqslant 0$ represent the first entrance time to level j, the argument essentially involves judicious use of "first step" decomposition to derive equations for matrices $G_{i,j}$ of hitting probabilities whose (k,ℓ)th elements are of the form

$$g_{i,j}[k,\ell] := \mathbb{P}(T_j < T, J(T_j) = \ell \,|\, X(0) = i, J(0) = k), \tag{3}$$

with the level transition matrices constructed appropriately to take account of the exponential killing and, in the second time interval, the fact that level $m+1$ is taboo.

Path from i to m.

We first concentrate on the first time period in which the level gradually increases from i to m. To this end, we introduce the matrix $P_i \equiv P_i(\tau)$ of probabilities that the running maximum is at least $i+1$, starting at level i, jointly with the phases. This matrix has entries

$$p_i[k,\ell] := \mathbb{P}(\bar{X}(T) \geqslant i+1, J(T_{i+1}) = \ell \,|\, X(0) = i, J(0) = k).$$

Since the event $\bar{X}(T) \geqslant i+1$ is the same as the event $T_{i+1} < T$ for a process that starts in level i, the matrix P_i coincides with the matrix $G_{i,i+1}$.

By reversing the order of the levels and applying [14, Eq. (6)] to the discrete-time *jump chain* obtained by observing the QBD only at transition points, we see that the matrices P_i satisfy

$$P_i = \left(\bar{Q}_i + \bar{\Lambda}_i + \bar{\mathcal{M}}_i\, 1_{\{i>0\}} + \tau I^{(i)}\right)^{-1} \left[\Lambda^{(i)} + (Q^{(i)} + \bar{Q}_i)P_i + \mathcal{M}^{(i)} P_{i-1} P_i\, 1_{\{i>0\}}\right], \tag{4}$$

where $\bar{Q}_i := \operatorname{diag}\{q_1^{(i)}, \ldots, q_{d_i}^{(i)}\}$, $\bar{\Lambda}_i := \operatorname{diag}\{\lambda_1^{(i)}, \ldots, \lambda_{d_i}^{(i)}\}$ and $\bar{\mathcal{M}}_i := \operatorname{diag}\{\mu_1^{(i)}, \ldots, \mu_{d_i}^{(i)}\}$ are matrices of dimension $d_i \times d_i$ and $I^{(i)}$ is the d_i-dimensional identity matrix.

The (k,ℓ)th entry of the matrices on both sides of Eq. (4) contains the probability that the QBD enters level $i+1$ in phase ℓ before the killing time, conditional on it starting in phase k of level i. On the right-hand side, this probability is decomposed according to the first transition that the QBD undertakes: The first term contains the probability that the QBD moves from level i to level $i+1$ on the first transition, the second the probability that it remains at level i and then reaches level $i+1$ before the killing time, and the third the probability that the first transition takes the QBD to level $i-1$ from where it reaches level $i+1$ via level i before the killing time.

Multiplying Eq. (4) by $\left(\bar{Q}_i + \bar{\Lambda}_i + \bar{\mathcal{M}}_i\, 1_{\{i>0\}} + \tau I^{(i)}\right)$ and rearranging, we obtain the equation

$$\Lambda^{(i)} + (Q^{(i)} - \bar{\Lambda}^{(i)} - \bar{\mathcal{M}}^{(i)}\, 1_{\{i>0\}} - \tau I^{(i)})P_i + \mathcal{M}^{(i)} P_{i-1} P_i\, 1_{\{i>0\}} = 0. \tag{5}$$

Now observe that the $d_i \times d_i$ matrix $Q^{(i)} - \bar{\Lambda}^{(i)} - \bar{\mathcal{M}}^{(i)}\, 1_{\{i>0\}} + \mathcal{M}^{(i)} P_{i-1} - \tau I^{(i)}$ is the transition matrix of the censored continuous-time Markov chain whose sample paths trace the progression through the phases when the Markov chain is in level i before the process either moves to level $i+1$ (corresponding to the random variable T_{i+1}) or the exponential killing lifetime (corresponding to T) expires. Since, with probability one, one of these events

will occur in finite time whatever the starting state, this censored Markov chain is transient and therefore, with U_{i+1} denoting the amount of time spent in level i before time T_{i+1} and U denoting the amount of time spent in level i before the killing time T, we know that

$$\begin{aligned}
&\int_0^\infty \mathbb{P}(U_{i+1} > u, U > u, J(u) = \ell \,|\, X(0) = i, J(0) = k)du \\
&\quad = \int_0^\infty \exp\left(\left(Q^{(i)} - \bar{\Lambda}^{(i)} - \bar{\mathcal{M}}^{(i)} 1_{\{i>0\}} + \mathcal{M}^{(i)} P_{i-1} - \tau I^{(i)}\right) u\right)_{k,\ell} du \\
&\quad = \left(Q^{(i)} - \bar{\Lambda}^{(i)} - \bar{\mathcal{M}}^{(i)} 1_{\{i>0\}} + \mathcal{M}^{(i)} P_{i-1} - \tau I^{(i)}\right)^{-1}_{k,\ell}
\end{aligned}$$

is finite for all $k \in \{1, \ldots, d_i\}$ and $\ell \in \{1, \ldots, d_{i+1}\}$. We conclude that, for $i = 1, \ldots, m$, the matrix

$$K^{(i)}(P_{i-1}) := (Q^{(i)} - \bar{\Lambda}^{(i)} - \bar{\mathcal{M}}^{(i)} 1_{\{i>0\}} + \mathcal{M}^{(i)} P_{i-1} - \tau I^{(i)})^{-1} \tag{6}$$

exists. This allows us to rewrite Eq. (5) as

$$P_i = -K^{(i)}(P_{i-1})\Lambda^{(i)} 1_{\{i>0\}}, \tag{7}$$

which serves as a recursion to calculate P_i in terms of P_{i-1}. The recursion can be initiated by observing that

$$P_0 = -(Q^{(0)} - \bar{\Lambda}^{(0)} - \tau I^{(0)})^{-1}\Lambda^{(0)}. \tag{8}$$

Now, with

$$\begin{aligned}
[S_{i,m}]_{k,\ell} &:= \mathbb{P}(\bar{X}(T) \geqslant m, J(T_m) = \ell \,|\, X(0) = i, J(0) = k), \\
&= \mathbb{P}(T_m < T, J(T_m) = \ell \,|\, X(0) = i, J(0) = k),
\end{aligned}$$

by decomposing the event that the process reaches level m from level i into the successive events that it first reaches level $i+1$, then level $i+2$, up to level m, we see that, for $i = 0, \ldots, m-1$,

$$\begin{aligned}
S_{i,m} &= P_i P_{i+1} \cdots P_{m-1} \\
&= \left(-K^{(i)}(P_{i-1})\Lambda^{(i)}\right)\left(-K^{(i+1)}(P_i)\Lambda^{(i+1)}\right)\cdots\left(-K^{(m-1)}(P_{m-2})\Lambda^{(m-1)}\right),
\end{aligned} \tag{9}$$

with $S_{m,m} = I$.

Path from m to j, remaining below $m+1$.

Now that we have derived an expression for the probability of reaching level m starting from level i, we now concentrate on the second part of the path, moving from level m to level $j \leqslant m$ while consistently remaining below level $m+1$.

For $j \leqslant m$, if we were interested in determining the probabilities

$$\hat{s}_{m,j}[k,\ell] := \mathbb{P}(\bar{X}(T) = m, J(T_j) = \ell \,|\, X(0) = m, J(0) = k), \tag{10}$$

we could do so using an analysis similar to that above, applied to a level-dependent QBD with $\Lambda^{(m)}$ set to zero. So, again applying a first transition decomposition, this time to the

jump chain of the QBD with with $\Lambda^{(m)}$ set to zero, we see that, for $i = j+1, \dots m-1$, the matrices $\hat{P}_i$ whose entries are

$$\hat{p}_{i,i-1}[k,\ell] := \mathbb{P}(\bar{X}(T) = m, J(T_{i-1}) = \ell \,|\, X(0) = i, J(0) = k),$$

satisfy the system of equations

$$\mathcal{M}^{(i)} + (Q^{(i)} - \bar{\Lambda}^{(i)} - \bar{\mathcal{M}}^{(i)} - \tau I^{(i)})\hat{P}_i + \Lambda^{(i)}\hat{P}_{i+1}\hat{P}_i = 0, \tag{11}$$

with

$$\mathcal{M}^{(m)} + (Q^{(i)} - \bar{\mathcal{M}}^{(i)} - \tau I^{(i)})\hat{P}_m = 0. \tag{12}$$

These can be solved in a similar manner to Eqs. (5) and (8). We then have

$$\hat{S}_{m,j} \quad = \quad \hat{P}_m \hat{P}_{m-1} \cdots \hat{P}_{j+1}. \tag{13}$$

However, it is not quite the probabilities (10) that we are interested in. We want to record both the level and the phase when the exponential lifetime T expires, not the phase when the QBD first hits level j on its downward path. We thus want to derive

$$\bar{p}_{m,j}[k,\ell] := \mathbb{P}(\bar{X}(T) = m, X(T) = j, J(T) = \ell \,|\, X(0) = m, J(0) = k). \tag{14}$$

To this end, concentrate on $j = m$. Again relying on standard Markovian reasoning, for $m \geqslant 1$, a first step decomposition yields the relation

$$\bar{p}_{m,m}[k,\ell] = \sum_{k'=1}^{d_{m-1}} \frac{\mu^{(m)}_{kk'}}{\nu^{(m)}_k} \sum_{k''=1}^{d_m} p_{m-1}[k',k'']\,\bar{p}_{m,m}[k'',\ell] + \sum_{k' \neq k}^{d_m} \frac{q^{(m)}_{kk'}}{\nu^{(m)}_k} \bar{p}_{m,m}[k',\ell] + \frac{\tau}{\nu^{(m)}_k} 1_{\{k=\ell\}}, \tag{15}$$

where, for $i = 0, \dots, m$, $\nu^{(i)}_k = \left(\bar{Q}_i + \bar{\Lambda}_i + \bar{\mathcal{M}}_i\, 1_{\{i>0\}} + \tau I^{(i)}\right)_{kk}$.

With $\bar{P}_{m,m}$ the $d_m \times d_m$ dimensional matrix consisting of the entries $\bar{p}_{m,m}[k,\ell]$, this equation can be written in matrix form as

$$-(Q^{(m)} - \bar{\Lambda}^{(m)} - \bar{\mathcal{M}}^{(m)} - \tau I^{(m)})\bar{P}_{m,m} = \mathcal{M}^{(m)} P_{m-1} \bar{P}_{m,m} + \tau I^{(m)}. \tag{16}$$

This equation directly implies that

$$\bar{P}_{m,m} = -\tau\, K^{(m)}(P_{m-1}). \tag{17}$$

We can compute this quantity; it contains the matrix P_{m-1}, and above we have already presented a recursive procedure to determine this matrix.

Now consider the case $j = m-1$. Via similar reasoning we can show

$$-(Q^{(m)} - \bar{\Lambda}^{(m)} - \bar{\mathcal{M}}^{(m)} - \tau I^{(m)})\bar{P}_{m,m-1} = \mathcal{M}^{(m)} P_{m-1} \bar{P}_{m,m-1} + \mathcal{M}^{(m)} \bar{P}_{m-1,m-1}. \tag{18}$$

Combining this with (17), this implies

$$\begin{aligned}
\bar{P}_{m,m-1} &= -(Q^{(m)} - \bar{\Lambda}^{(m)} - \bar{\mathcal{M}}^{(m)} + \mathcal{M}^{(m)} P_{m-1} - \tau I^{(m)})^{-1} \mathcal{M}^{(m)} \bar{P}_{m-1,m-1} \\
&= \left(-K^{(m)}(P_{m-1})\mathcal{M}^{(m)}\right)\left(-\tau K^{(m-1)}(P_{m-2})\right) \\
&= \tau \left(-K^{(m)}(P_{m-1})\mathcal{M}^{(m)}\right)\left(-K^{(m-1)}(P_{m-2})\right).
\end{aligned}$$

This argument can be iterated, to obtain, for any $j \in \{0, \ldots, m-1\}$,

$$\bar{P}_{m,j} = \tau \left(-K^{(m)}(P_{m-1})\mathcal{M}^{(m)}\right)\left(-K^{(m-1)}(P_{m-2})\mathcal{M}^{(m-1)}\right) \cdots \left(-K^{(j+1)}(P_j)\mathcal{M}^{(j+1)}\right)\left(-K^{(j)}(P_{j-1})\right), \tag{19}$$

and so we have expressed, for any $j \in \{1, \ldots, m\}$, the matrices $\bar{P}_{m,j}$ in terms of the matrices P_{j-1} up to (and including) P_{m-1}.

Putting together our decomposition of the time interval with expressions (9) and (19), we have proved the following.

THEOREM 3.1: *Let $S_{i,j,m}$ be the $d_i \times d_j$ matrix with entries $s_{i,j,m}[k,\ell]$. For any i, j,*

$$S_{i,j,m} = \tau \left(-K^{(i)}(P_{i-1})\Lambda^{(i)}\right) \cdots \left(-K^{(m-1)}(P_{m-2})\Lambda^{(m-1)}\right) \times \left(-K^{(m)}(P_{m-1})\mathcal{M}^{(m)}\right) \cdots \left(-K^{(j+1)}(P_j)\mathcal{M}^{(j+1)}\right)\left(-K^{(j)}(P_{j-1})\right).$$

4. COMPUTATIONAL ASPECTS

In this section, we consider the problem of computing the distribution of the running maximum at a deterministic time rather than an exponential time. In addition, we describe various ramifications.

Suppose that, for some $t \geqslant 0$, we wish to evaluate the probabilities

$$r_{i,j,m}[k,\ell]_t := \mathbb{P}(\bar{X}(t) \leqslant m, X(t) = j, J(t) = \ell \,|\, X(0) = i, J(0) = k). \tag{20}$$

We could do this by deriving

$$s_{i,j,m}[k,\ell]_t := \mathbb{P}(\bar{X}(t) = m, X(t) = j, J(t) = \ell \,|\, X(0) = i, J(0) = k),$$

using the observation that $s_{i,j,m}[k,\ell]$ defined in Eq. (2) is the Laplace transform

$$\frac{s_{i,j,m}[k,\ell]}{\tau} = \int_0^\infty e^{-\tau t} s_{i,j,m}[k,\ell]_t \, dt$$

of $s_{i,j,m}[k,\ell]_t$. Thus, we can derive the functions $s_{i,j,m}[k,\ell]_t$ by performing numerical Laplace inversion on $s_{i,j,m}[k,\ell]/\tau$ derived from from Theorem 3.1 (with respect to τ). The numbers $r_{i,j,m}[k,\ell]_t$ then follow from the trivial relation

$$r_{i,j,m}[k,\ell]_t = \sum_{m'=\max\{i,j\}}^{m} s_{i,j,m'}[k,\ell]_t.$$

There are fast and reliable techniques to perform Laplace transform inversion; see, for example, the classical reference [1], and [8] for a more-recent sophisticated variant.

We now advocate an alternative technique, which has (for reasons that will become obvious) been coined "Erlangization" [6,15]. It uses the results of Section 3, which focus on the distribution of $\bar{X}(T)$ after an exponentially distributed time, in order to evaluate the distribution of $\bar{X}(t)$ after a deterministic time $t \geqslant 0$. The key fact that makes this possible is that not only do we have the distribution of the running maximum over an interval of length T, but also its distribution *jointly with* the level and phase at the killing epoch T.

More specifically, consider the $d_i \times d_j$ matrix

$$R_{i,j,m} = \sum_{m'=\max\{i,j\}}^{m} S_{i,j,m'}, \tag{21}$$

and, with $D := \sum_{i=0}^{m} d_i$, the $D \times D$ matrix

$$R_m := \begin{pmatrix} R_{0,0,m} & \cdots & R_{0,m,m} \\ \vdots & \ddots & \vdots \\ R_{m,0,m} & \cdots & R_{m,m,m} \end{pmatrix}. \tag{22}$$

Denote by $R^{(N)}_{i,j,m}$ the (i,j)th block matrix in $(R_m)^N$ (which has dimension $d_i \times d_j$), and let $r^{(N)}_{i,j,m}[k,\ell]$ denote the (k,ℓ)th entry in this matrix. Now observe that the quantity $r^{(N)}_{i,j,m}[k,\ell]$ equals the probability

$$r_{i,j,m}[k,\ell] := \mathbb{P}(\bar{X}(T) \leqslant m, X(T) = j, J(T) = \ell \,|\, X(0) = i, J(0) = k), \tag{23}$$

but where T no longer has an exponential distribution with parameter τ, it has an Erlang(N,τ) distribution. Replacing τ by τN and letting N become large, by virtue of the law of large numbers, we can approximate the counterpart of (20) after a *deterministic* time τ^{-1}. This technique has been extensively tested in [9].

Note that, in the case where T is exponential, it is "cheap" to find the matrices $S_{i,j,m}$, as this essentially involves inversions of $d_i \times d_i$ matrices and a recursion consisting of m iterations. Also observe that finding the probabilities after a *deterministic* time can be accomplished efficiently as well: The evaluation of the matrix $R_m^{(2^N)}$ can be acccomplished by repeatedly *squaring* R_m (that is, N times); recall that R_m has dimension $D \times D$.

5. EXAMPLES

Example 5.1: Suppose that we monitor a QBD by recording its values at time epochs $0, \Delta, 2\Delta, \ldots$, for some $\Delta > 0$. A natural question that arises is: What is the distribution of the maximum level attained between two subsequent measurements? In the terminology of this note, we wish to evaluate

$$\mathbb{P}(\bar{X}(\Delta) \leqslant m \,|\, X(0) = i, J(0) = k, X(\Delta) = j, J(\Delta) = \ell),$$

which can be rewritten as

$$\frac{r_{i,j,m}[k,\ell]_\Delta}{\sum_{m=0}^{\infty} s_{i,j,m}[k,\ell]_\Delta}.$$

To illustrate the type of numerical analysis that can be performed, we report some results for a simple Markov-modulated $M/M/\infty$ queueing model with two phases. In this example, we concentrate on a model in which only the arrival rate is Markov modulated; we take, for all i,

$$\Lambda^{(i)} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad \mathcal{M}^{(i)} = \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix}, \quad Q^{(i)} = \begin{pmatrix} -1 & 1 \\ 2 & -2 \end{pmatrix}.$$

Our goal is to evaluate $r_{i,j,m}[k,\ell]_t$; in the sequel we take $m = 5$ and $t = 1$. To this end, we compute $(R_m)^{2^N}$ for values of N varying from 0 to 7, where τ in the Nth experiment

is taken to be 2^{-N}. We thus obtain the probability of remaining below level $m+1$ after an Erlang-distributed time E_N consisting 2^N exponential phases each having mean 2^{-N}, such that $\mathbb{E}[E_N]=1$ and $\mathbb{V}\mathrm{ar}[E_N]=2^{-N}$.

With $D=12$, the full matrices R_m and $(R_m)^{2^N}$ are of dimension 12×12. Below we present eight matrices of dimension 6×6, each of them consisting of nine submatrices of dimension 2×2, corresponding to the top-left corner of $(R_m)^{2^N}$; the rest of the matrix shows the same rate of convergence. The nth of these eight matrices corresponds to $N=n-1$.

$$\begin{pmatrix} 0.4704 & 0.1103 & 0.1905 & 0.0813 & 0.0644 & 0.0384\\ 0.2378 & 0.2752 & 0.1630 & 0.1382 & 0.0689 & 0.0579\\ 0.1733 & 0.0558 & 0.4084 & 0.1057 & 0.1243 & 0.0572\\ 0.1241 & 0.0777 & 0.2170 & 0.2642 & 0.1078 & 0.1065\\ 0.0988 & 0.0353 & 0.2259 & 0.0774 & 0.3250 & 0.0841\\ 0.0802 & 0.0396 & 0.1613 & 0.1178 & 0.1647 & 0.2332 \end{pmatrix}, \begin{pmatrix} 0.4102 & 0.1204 & 0.2119 & 0.0944 & 0.0723 & 0.0439\\ 0.2601 & 0.1989 & 0.1884 & 0.1433 & 0.0789 & 0.0657\\ 0.1926 & 0.0648 & 0.3434 & 0.1141 & 0.1348 & 0.0656\\ 0.1430 & 0.0813 & 0.2345 & 0.1927 & 0.1239 & 0.1076\\ 0.1124 & 0.0403 & 0.2442 & 0.0888 & 0.2560 & 0.0899\\ 0.0904 & 0.0442 & 0.1851 & 0.1203 & 0.1764 & 0.1639 \end{pmatrix},$$

$$\begin{pmatrix} 0.3747 & 0.1242 & 0.2239 & 0.1034 & 0.0778 & 0.0481\\ 0.2693 & 0.1585 & 0.2056 & 0.1412 & 0.0863 & 0.0700\\ 0.2031 & 0.0710 & 0.3074 & 0.1172 & 0.1391 & 0.0709\\ 0.1560 & 0.0810 & 0.2411 & 0.1574 & 0.1341 & 0.1038\\ 0.1219 & 0.0440 & 0.2507 & 0.0960 & 0.2202 & 0.0916\\ 0.0978 & 0.0469 & 0.2004 & 0.1175 & 0.1801 & 0.1316 \end{pmatrix}, \begin{pmatrix} 0.3557 & 0.1251 & 0.2300 & 0.1087 & 0.0811 & 0.0508\\ 0.2720 & 0.1395 & 0.2157 & 0.1378 & 0.0912 & 0.0720\\ 0.2081 & 0.0747 & 0.2896 & 0.1180 & 0.1403 & 0.0738\\ 0.1637 & 0.0798 & 0.2429 & 0.1418 & 0.1396 & 0.1000\\ 0.1277 & 0.0465 & 0.2518 & 0.0999 & 0.2035 & 0.0919\\ 0.1026 & 0.0482 & 0.2088 & 0.1144 & 0.1808 & 0.1180 \end{pmatrix},$$

$$\begin{pmatrix} 0.3461 & 0.1250 & 0.2330 & 0.1115 & 0.0829 & 0.0525\\ 0.2725 & 0.1309 & 0.2210 & 0.1353 & 0.0942 & 0.0727\\ 0.2105 & 0.0767 & 0.2810 & 0.1182 & 0.1405 & 0.0751\\ 0.1678 & 0.0789 & 0.2433 & 0.1349 & 0.1423 & 0.0976\\ 0.1309 & 0.0480 & 0.2516 & 0.1018 & 0.1959 & 0.0918\\ 0.1055 & 0.0487 & 0.2130 & 0.1123 & 0.1808 & 0.1122 \end{pmatrix}, \begin{pmatrix} 0.3413 & 0.1249 & 0.2344 & 0.1130 & 0.0839 & 0.0534\\ 0.2725 & 0.1268 & 0.2238 & 0.1339 & 0.0958 & 0.0730\\ 0.2116 & 0.0777 & 0.2769 & 0.1182 & 0.1405 & 0.0758\\ 0.1699 & 0.0784 & 0.2434 & 0.1318 & 0.1436 & 0.0963\\ 0.1326 & 0.0489 & 0.2512 & 0.1028 & 0.1924 & 0.0918\\ 0.1071 & 0.0489 & 0.2150 & 0.1112 & 0.1808 & 0.1096 \end{pmatrix},$$

$$\begin{pmatrix} 0.3389 & 0.1247 & 0.2351 & 0.1137 & 0.0844 & 0.0539\\ 0.2724 & 0.1249 & 0.2251 & 0.1331 & 0.0966 & 0.0731\\ 0.2121 & 0.0782 & 0.2748 & 0.1182 & 0.1405 & 0.0761\\ 0.1710 & 0.0781 & 0.2434 & 0.1303 & 0.1442 & 0.0956\\ 0.1335 & 0.0493 & 0.2510 & 0.1032 & 0.1907 & 0.0917\\ 0.1079 & 0.0490 & 0.2160 & 0.1106 & 0.1807 & 0.1083 \end{pmatrix}, \begin{pmatrix} 0.3377 & 0.1246 & 0.2355 & 0.1140 & 0.0847 & 0.0541\\ 0.2724 & 0.1239 & 0.2258 & 0.1327 & 0.0970 & 0.0732\\ 0.2124 & 0.0785 & 0.2738 & 0.1182 & 0.1405 & 0.0763\\ 0.1716 & 0.0779 & 0.2433 & 0.1296 & 0.1445 & 0.0952\\ 0.1340 & 0.0495 & 0.2508 & 0.1034 & 0.1898 & 0.0917\\ 0.1084 & 0.0491 & 0.2165 & 0.1103 & 0.1807 & 0.1077 \end{pmatrix}.$$

We can observe that the matrices rapidly converge; the matrix corresponding to $N=7$ just marginally differs from the one corresponding to $N=6$.

The above example shows how our approach offers an effective procedure to evaluate the probability that a level-dependent QBD $X(s)$ does not exceed level m for $s\in[0,t]$, for given (deterministic) t. For instance, the entry $(5,4)$ in the matrix corresponding to $N=7$, that is, 0.1034, is an approximation for $r_{2,1,5}[1,2]_1$, which is the probability of being at level 1 and in phase 2 at time $t=1$ and having remained below level $m+1=6$ during the interval $[0,t]=[0,1]$, having started at level 2, in phase 1, at time 0. In addition, the probability of remaining below $m+1$ for a given initial level and phase can be computed as well. It is given by the sum of the entries of the fifth row of the entire matrix $(R_m)^{2^7}$. This is 0.9937 (compare with the sum of the first three block matrices shown above, which 0.8192), corresponding to the probability of remaining below level 6 in the interval $[0,1]$, starting from level 2 and phase 1 at time 0.

Example 5.2: The authors of [9] presented a numerical analysis of a model that can be used to determine whether a service level agreement in a cable network is being adhered to. In their model, jobs arrive according to a Poisson process with rate λ, file sizes are exponential with mean $1/f$, and there is a maximum per-user transmission rate $R_{\max}$. Users receive this maximum rate as long as it can be accommodated, otherwise they receive an equal share of

the available capacity. This results in a birth and death process model with transition rates

$$q_{i,i+1} = \lambda, \quad \text{and} \quad q_{i,i-1} = \min\{C, R_{\max} i\}/f. \tag{24}$$

Among the performance measures of interest to the authors of [9] was the probability that the number of customers exceeds a given level m during a fixed time interval, given the numbers at the beginning and end of the interval. For $C = 800$, $R_{\max} = 80$, $\lambda = 100$, $f = 6$ and a period length $t = 2$, Figure 1(a) of [9] provided a plot of the probability that the maximum number of customers exceeds $m = 15$ as a function of the numbers at the beginning and ending of the interval.

In this example, we carried out a similar analysis for a Markov-modulated version of the model in [9]. Specifically we allowed C and $R_{\max}$ to vary according to the state of an underlying Markov chain on five phases whose transition rates

$$Q = \begin{pmatrix} -25.6921 & 9.9979 & 0.4890 & 5.5337 & 9.6715 \\ 10.5165 & -31.4895 & 8.4180 & 6.9109 & 5.6441 \\ 9.9951 & 1.9202 & -29.5037 & 14.7246 & 2.8639 \\ 8.0869 & 14.9862 & 10.0376 & -39.5345 & 6.4238 \\ 10.4716 & 2.5668 & 2.8565 & 12.8328 & -28.7277 \end{pmatrix} \tag{25}$$

were generated randomly, and with $\lambda = 100$, $f = 6$ and a period length $t = 1$. The values of C and $R_{\max}$ corresponding to each of the phases are given in Table 1.

TABLE 1. Values of C and $R_{\max}$.

	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
C	700	750	800	850	900
$R_{\max}$	90	85	80	75	70

TABLE 2. Probability that the number of customers exceeds 15 in the interval $[0, 1]$, as a function of initial number of customers and phase.

$X(0)$	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
0	0.4238	0.4239	0.4242	0.4243	0.4250
1	0.4276	0.4278	0.4281	0.4283	0.4292
2	0.4319	0.4321	0.4326	0.4329	0.4340
3	0.4368	0.4371	0.4376	0.4381	0.4394
4	0.4425	0.4428	0.4435	0.4441	0.4457
5	0.4494	0.4496	0.4504	0.4511	0.4531
6	0.4578	0.4578	0.4586	0.4597	0.4620
7	0.4685	0.4681	0.4688	0.4701	0.4728
8	0.4830	0.4815	0.4817	0.4833	0.4863
9	0.5032	0.4995	0.4986	0.5003	0.5034
10	0.5308	0.5250	0.5213	0.5227	0.5255
11	0.5678	0.5599	0.5536	0.5529	0.5544
12	0.6168	0.6071	0.5983	0.5951	0.5931
13	0.6812	0.6701	0.6593	0.6538	0.6470
14	0.7646	0.7532	0.7416	0.7347	0.7248
15	0.8703	0.8614	0.8522	0.8459	0.8367

For each value of the initial level $i \leqslant 15$ and initial phase k, Table 2 contains the probabilities that the number of customers reaches 16 or greater during the interval $[0, 1]$. We calculated this by summing the entries of the array $r_{i,j,15}[k, \ell]_1$ over j and ℓ and subtracting the result from 1. As we would expect, the exceedance probabilities are increasing with the level i. There is not much variation in these probabilities with the initial phase, but it is interesting to see that the initial phase that has the lowest exceedance probability varies with the level. For the highest levels, phase 5 has the lowest probability of exceedance, but then as the level becomes lower, the minimum is achieved at phases 4, 3 and 2, respectively.

6. CONCLUSION

In this short note, we have presented a simple recursion to calculate the distribution of the running maximum of a level-dependent QBD, conditional on the initial and final levels and phases. We accomplished this first by calculating the probabilities over an exponentially distributed random time T and then, via Erlangization, at a fixed time t. In Section 5, we have applied this to two different classes of examples and demonstrated the numerical efficiency of the method.

Acknowledgements

The authors wish to thank an anonymous referee for some insightful comments and for pointing out references [2,3,10]. Michel Mandjes' research is partly funded by the NWO Gravitation Project NETWORKS—grant number 024.002.003. Peter Taylor's research is supported by the Australian Research Council (ARC) Laureate Fellowship FL130100039 and the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).

References

1. Abate, J. & Whitt, W. (1995). Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing* 7: 36–43.
2. Artalejo, J.R., Economou A. & Gómez-Corral A. (2007). Applications of maximum queue lengths to call center management. *Computers and Operations Research* 34: 983–996.
3. Artalejo, J.R., Economou A. & Lopez-Herrero M.J. (2007). Algorithmic analysis of the maximum queue length in a busy period for the M/M/c retrial queue. *INFORMS Journal on Computing* 19: 121–126.
4. Artalejo, J.R. & Gómez-Corral, A. (2008). *Retrial queueing systems. a computational approach*, Berlin: Springer-Verlag, 2008.
5. Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edn., New York, NY, USA: Springer.
6. Asmussen, S., Avram, F. & Usabel, M. (2002). The Erlang approximation of finite time ruin probabilities. *ASTIN Bulletin* 32: 267–281.
7. Bright, L.W. & Taylor, P.G. (1995). Calculating the equilibrium distribution in level dependent Quasi-Birth-and-Death processes. *Stochastic Models* 11: 497–526.
8. den Iseger, P. (2006). Numerical transform inversion using Gaussian quadrature. *Probability in the Engineering and Informational Sciences* 20: 1–44.
9. Ellens, W., Mandjes, M., van den Berg, H., Worm, D. & Błaszczuk, S. (2015). Performance evaluation using periodic system-state measurements. *Performance Evaluation* 93: 27–46.
10. Gómez-Corral, A. & García, M.L. (2014). Maximum queue lengths during a fixed time interval in the $M/M/c$ retrial queue. *Applied Mathematics and Computation* 235: 124–136.
11. Kyprianou, A. (2006). *Introductory lectures on fluctuations of Lévy processes with applications,* Berlin, Germany: Springer.
12. Latouche, G. & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modelling.* ASA/SIAM Series on Statistics and Applied Probability. Philadelphia PA, USA.
13. Neuts, M.F. (1981). *Matrix-geometric solutions in stochastic models*, Baltimore, MD, USA: Johns Hopkins University Press.

14. Ramaswami, V. & Taylor, P.G. (1996). Some properties of the rate matrices in level dependent Quasi-Birth-and-Death processes with a countable number of phases. *Stochastic Models* 12: 143–164.
15. Ramaswami, V., Woolford, D. & Stanford, D. (2008). The Erlangization method for Markovian fluid flows. *Annals of Operations Research* 160: 215–225.