Construction of algorithms for discrete-time quasi-birth-and-death processes through physical interpretation

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

Małgorzata M. O’Reilly∗†
School of Natural Sciences
University of Tasmania
TAS 7001, Australia
malgorzata.oreilly@utas.edu.au

Nigel G. Bean∗
School of Mathematical Sciences
University of Adelaide
SA 5005, Australia
nigel.bean@adelaide.edu.au

ABSTRACT

We apply physical interpretations to construct algorithms for the key matrix \( G \) in discrete-time quasi-birth-and-death (dtQBD) and its \( z \)-transform \( G(z) \), motivated by the work on stochastic fluid models (SFMs) in [13]. In this methodology, we first write a summation expression for \( G(z) \) by considering a physical interpretation similar to that of an algorithm in [13]. Next, we construct the corresponding iterative scheme, and prove its convergence to \( G(z) \).

In particular, here we consider the physical interpretation of Algorithm 1 for \( \Psi(s) \) in [13], and use a similar physical interpretation for \( G(z) \) partitioned into three sections, each expressed in terms of matrices analogous to block matrices in the fluid generator \( Q(s) \) in stochastic fluid models.

1. INTRODUCTION

We consider stochastic fluid models (SFMs) and quasi-birth-and-death processes (QBDs), which are key processes in the literature of matrix-analytic methods (MAMs). We exploit the similarities between them in order to apply analogous physical interpretations to construct new expressions and algorithms for the key matrix \( G \) in discrete-time QBDs (dtQBDs) and its \( z \)-transform \( G(z) \).

Many expressions in the theory of SFMs, including those for the matrix \( \Psi(s) \), have underlying physical interpretations, which are obtained by deconstructing sample paths into various components, and then writing corresponding expressions in terms of fluid generator \( Q(s) \) [9, 11, 12, 13].

Here, we apply a physical interpretation and conditioning similar to that used in the construction of [13, Algorithm 1] for matrix \( \Psi(s) \) in SFMs to derive an expression and algorithm for the matrix \( G(z) \).

The main algorithms for QBDs were generated with physical interpretations where the set of included sample paths were partitioned, according to the iteration count, by the maximum level reached. The linearly-convergent algorithms then had a linearly increasing maximum level, while the quadratically convergent algorithms had a geometrically increasing maximum level.

Such a partitioning was not entirely appropriate when algorithms were being developed for SFMs. Instead, in [11, Sections 3.1-3.5], sample paths included in the matrix \( \Psi(s) \) were partitioned according to a key level, \( y \), and the behaviours on the required three sample path components were carefully controlled: starting at level 0 and reaching level \( y \), leaving level \( y \) before returning to level \( y \), and starting at level \( y \) until reaching level 0 for the first time. Each of these components can be expressed in terms of the fluid generator \( Q(s) \).

To construct an analogous algorithm for \( G(z) \), we use a similar partitioning principle and three sample path components to construct the iterations for \( G(z) \). We express each component in terms of the matrix \( M(z) \), the dtQBD-equivalent of the fluid generator \( Q(s) \).

2. PRELIMINARIES

2.1 Discrete-time QBDS

Consider a dtQBD, denoted \( \{X_t : t = 0, 1, 2, \ldots\} \), on a two-dimensional state space \( \{(n, i) : n \geq 0, 0 \leq i \leq m\} \), with level variable \( n \) and phase variable \( i \), and the one-step transition probability matrix

\[
P = \begin{bmatrix}
B & A_- & 0 & 0 & \cdots \\
A_+ & A_0 & A_+ & 0 & \cdots \\
0 & A_- & A_0 & A_+ & \cdots \\
0 & 0 & A_- & A_0 & \cdots \\
& \vdots & \vdots & \vdots & \ddots
\end{bmatrix},
\]

where matrices \( B, A_+, A_- \), and \( A_0 \) are square matrices of order \( m \) such that, for all \( i, j \in \{0 \leq i \leq m\} \),

\[
[B]_{ij} = P(X_{t+1} = (0, j) \mid X_t = (0, i)),
\]

\[
[A_+]_{ij} = P(X_{t+1} = (n + 1, j) \mid X_t = (n, i)),
\]

\[
[A_-]_{ij} = P(X_{t+1} = (n - 1, j) \mid X_t = (n, i)),
\]

\[
[A_0]_{ij} = P(X_{t+1} = (n, j) \mid X_t = (n, i)).
\]

Discrete-time QBDS have been used to analyse a variety of real-life situations including modelling embedded queues [16] and maintenance [14]. The analytical expressions for the
key quantities in the transient and stationary analysis of this model have led to powerful algorithms. The derivations of both the analytic expressions and subsequent algorithms appear in [1, 6, 16]. We now define a key building block used in the construction of the algorithm in later sections.

**Definition 1.** For complex number \(z\) inside the unit disk, let the matrices \(M_+(z) = [M_{i,j}(z)]_{1 \leq i,j \leq m}\) and \(M_-(z) = [M_{-i,j}(z)]_{1 \leq i,j \leq m}\) be given by

\[
M_+(z) = \sum_{n=0}^{\infty} (A_0 z)^n A_+ z = (I - A_0 z)^{-1} A_+ z,
\]

\[
M_-(z) = \sum_{n=0}^{\infty} (A_0 z)^n A_- z = (I - A_0 z)^{-1} A_- z.
\]

For an irreducible discrete-time QBD, \((I - A_0 z)^{-1}\) exists [17].

The entry \([M_+(z)]_{ij}\) is the probability generating function (PGF) of the time for the process to reach level \((n+1)\) for the first time and do so in phase \(j\), whilst avoiding level \((n-1)\), given the process starts in level \(n > 0\) in phase \(i\) at time zero. The entry \([M_-(z)]_{ij}\) is the PGF of the time taken for the process to reach level \(n-1\) for the first time and do so in phase \(j\), whilst avoiding level \((n+1)\), given the process starts in level \(n > 0\) in phase \(i\) at time zero.

The particular quantity we consider in this paper is the matrix \(G(z)\). Before defining \(G(z)\), we define \(r\) as the time taken to first reach level \((n-1)\). Then the \((i,j)\)-th entry of the matrix \(G(z)\) is defined

\[
[G(z)]_{ij} = E[z^r \mathbb{I}[r < \infty, X_r = (n-1, j)] | X_0 = (n, i)],
\]

where \([G(z)]_{ij}\) records the PGF of the time taken for the process to reach level \(n-1\) for the first time and do so in phase \(j\), given the process starts in level \(n > 0\) at phase \(i\). Note \(J[-]\) denotes the indicator function throughout the paper.

For \(0 \leq z \leq 1\), the matrix \(G(z)\) is the minimal non-negative solution [17] to

\[
G(z) = A_+ z + A_0 z G(z) + A_- z (G(z))^2.
\]

Transforming this into a fixed-point equation by using the recommended iteration by Neuts [17], and then representing in terms of \(M_+(z)\) and \(M_-(z)\), we get

\[
G(z) = M_-(z) + M_+(z) (G(z))^2
\]

or

\[
G(z) = (I - M_+(z) G(z))^{-1} M_-(z).
\]

### 2.2 SFMs

Consider a SFM, denoted \(\{(\varphi(t), X(t)) : t \geq 0\}\), with phase variable \(\varphi(t) \in \mathcal{S} = \{1, \ldots, n\}\) and level variable \(X(t)\) with lower bound at zero, such that:

- the phase process \(\{\varphi(t) : t \geq 0\}\) is an irreducible continuous-time Markov chain (CTMC) with state space \(\mathcal{S}\) and generator \(\mathbf{Q}\);
- the level variable \(X(t)\) changes at rate \(dX(t)/dt = c_{\varphi(t)}\) at time \(t\) whenever \(X(t) > 0\), and at rate \(\max\{c_{\varphi(t)}, 0\}\) whenever \(X(t) = 0\).

SFMs have been used in the analysis of a variety of real-life situations, including telecommunications systems [19], risk assessment [7], power generation systems [10] and congestion control [18].

The stationary and transient analysis of SFMs and powerful algorithms for the numerical evaluations of various performance measures can be found in [2, 3, 4, 5, 11, 13, 20]. We now define the key building blocks in SFMs used in the construction of the algorithms in [10, 11, 12, 13, 21]. Let \(\mathcal{S}_+= \{i \in \mathcal{S} : c_i > 0\}\), \(\mathcal{S}_- = \{i \in \mathcal{S} : c_i < 0\}\), \(\mathcal{S}_0 = \{i \in \mathcal{S} : c_i = 0\}\). Block matrices \(Q_{++}(s)\) and \(Q_{--}(s)\) in the fluid generator \(Q(s)\) introduced in [12] are analogous to the building blocks \(M_+(z)\) and \(M_-(z)\) in QBDs, respectively.

**Definition 2.** For \(s \in \mathbb{C}\) with \(\Re(s) \geq 0\), let

\[
Q(s) = \begin{bmatrix} Q_{++}(s) & Q_{+-}(s) \\ Q_{-+}(s) & Q_{--}(s) \end{bmatrix},
\]

where

\[
Q_{++}(s) = C_+^{-1} \mathbb{E}[\sum_{t=0}^{\tau} -s I - T_{s0}(T_{00} - s I)^{-1} T_{01}],
\]

\[
Q_{--}(s) = C_-^{-1} \mathbb{E}[\sum_{t=0}^{\tau} -s I - T_{00}(T_{00} - s I)^{-1} T_{01}],
\]

\[
Q_{+-}(s) = C_+^{-1} \mathbb{E}[\sum_{t=0}^{\tau} -s I - T_{s0}(T_{00} - s I)^{-1} T_{01}],
\]

\[
Q_{-+}(s) = C_-^{-1} \mathbb{E}[\sum_{t=0}^{\tau} -s I - T_{00}(T_{00} - s I)^{-1} T_{01}].
\]

The physical interpretation is reliant on the following definitions of the in-out fluid \(h(t)\) and the first time for \(h(t)\) to hit some \(y > 0\). For any \(t \geq 0\), define the random variable

\[
h(t) = \int_{0}^{t} |\varphi(u)| du,
\]

interpreted as the total amount of fluid that has entered or exited the buffer \(X(t)\) during the time interval \([0, t]\), and referred to as the in-out fluid [12] of the process \(X(t)\). Also, for any \(y > 0\), define the random variable

\[
\omega(y) = \inf\{t > 0 : h(t) = y\},
\]

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

Subsequently, the physical interpretation of \([e^{Q_{++}(s)}]_{ij}\) as shown in [12] and extended in [21] is the Laplace-Stieltjes Transform (LST) of the distribution of time of the in-out fluid to reach \(y\) for the first time and do so when the process is in phase \(j \in \mathcal{S}_+\) whilst avoiding phases in \(\mathcal{S}_-\), given that the in-out fluid starts at \(0\) and the process starts in phase \(i \in \mathcal{S}_+\). Further, let \(\theta(x) = \inf\{t > 0 : X(t) = x\}\) be the first passage time to level \(x\). For \(i \in \mathcal{S}_+\), \(j \in \mathcal{S}_-\), and \(s \in \mathbb{C}\), where \(\Re(s) \geq 0\), \([\Psi(s)]_{ij}\) is given by the conditional expectation

\[
[\Psi(s)]_{ij} = \mathbb{E}[e^{s\theta(x)} \mathbb{I}[\theta(x) < \infty, \varphi(\theta(x)) = j] | X(0) = x, \varphi(0) = i].
\]

The physical interpretation of \([\Psi(s)]_{ij}\) is the LST of the time taken for the process to hit level \(x\) for the first time and does so in phase \(j\), given the process starts from level \(x\) whilst avoiding levels below \(x\).

### 3. Lowest-Trough Algorithm

We construct a lowest-trough algorithm (LT) for the matrix \(G(z)\) by considering the physical interpretation of the sample path corresponding to Algorithm 1 for matrix \(\Psi(s)\) in [13]. Algorithm 1 partitions all relevant sample paths according to the lowest-trough observed in any sample path, corresponding to \(\Psi(s)\), occurring at level \(y\).

In this section, we first summarise Algorithm 1 in [13] by stating the iterative scheme, the integral equation equivalent...
to the iterative expression, and the physical interpretation of the expression. Next, we apply a similar physical interpretation to derive a summation equation for $G(z)$. Further, we construct an iterative scheme resulting from that equation, denoting by $G^{LT}_n(z)$ the matrix calculated in the $n$-th iteration of the scheme. Finally, we let the corresponding algorithm be called the LT Algorithm, and prove its convergence to $G(z)$.

3.1 Algorithm 1 for $\Psi(s)$ in SFMs.

Algorithm 1 in [13] is based upon the following iterative scheme, which continues until an appropriate stopping criterion is met.

1. Let $\Psi_0(s) = 0$.

2. For $n = 0, 1, 2, \ldots$, let $\Psi_{n+1}(s)$ be the unique solution to

$$Q_{++}(s)\Psi_{n+1}(s) + \Psi_{n+1}(s)Q_{--}(s) = -Q_{+-} - \Psi_n(s)Q_{-+}(s)\Psi_n(s).$$

For $s \geq 0$, the above iterative scheme converges to the minimal non-negative solution of the Riccati equation for $\Psi(s)$ [11, 13].

As described in [13], equation (14) is equivalent to

$$\Psi_{n+1}(s) = \int_{y=0}^{\infty} e^{Q_{++}(s)y}(Q_{--}(s)$$

$$+ \Psi_n(s)Q_{-+}(s)\Psi_n(s))e^{Q_{-+}(s)y}dy,$$

which has the following physical interpretation, after defining $\Phi_n$ as the set of sample paths contributing to $\Psi_n$.

Each sample path contributing to $[\Psi_{(n+1)}(s)]_{ij}$ has three stages outlined below and depicted in Figure 1.

1. Given the process starts at level 0 in phase $i \in S_+$, the process remains in some phases in $S_+ \cup S_0$ until the process reaches level $y$ in phase $i_1 \in S_+$ whilst avoiding any transitions into $S_-$. The LST corresponding to this stage is $[e^{Q_{++}(s)y}]_{i_1}$. 

2. Given the process starts at level $y$ in phase $i_1 \in S_+$, the process:
   - Either makes a transition from phase $i_1$ to $i_2 \in S_-$ with instantaneous LST rate $[Q_{++}(s)]_{i_1i_2}$.
   - Or, the process leaves level $y$ in phase $i_1 \in S_+$, before returning to level $y$ in some phase $i_3 \in S_+$ along a path in $\Phi_n$, with LST $[\Psi_n(s)]_{i_1i_3}$. Then the process makes a transition from phase $i_3$ to phase $i_4 \in S_+$ with instantaneous LST rate $[Q_{++}(s)]_{i_3i_4}$. The process then leaves level $y$ in phase $i_4 \in S_+$, before returning to level $y$ in some phase $i_2$ along a path in $\Phi_n$ with LST $[\Psi_n(s)]_{i_4i_2}$. The LST rate corresponding to this stage is $([Q_{--}(s)] + \Psi_n(s)Q_{-+}(s)\Psi_n(s)]_{i_2i_1}$.

3. Given the process starts from level $y$ in phase $i_2 \in S_-$, the process remains in some phases in $S_- \cup S_0$ until the process drains to level 0 and does so in phase $j$, whilst avoiding any transitions into $S_+$. The LST of the time taken to complete this stage is $[e^{Q_{--}(s)y}]_{i_2}$.

By integrating over all possible $y$, all possible sample paths are captured in (15).

3.2 Lowest-trough algorithm for $G(s)$ in dQBDs

We now construct the LT Algorithm using a similar physical interpretation to that of Algorithm 1 in [13], as described in Section 3.1 above. Let $G_0(z) = 0$ and define $\Omega^L_n$, $n \geq 0$, as the set of sample paths contributing to $G_n(z)$.

Suppose that the process starts from level 1 in phase $i$ and first reaches level 0 in phase $j$. Then for each sample path contributing to $G_n(z)$ the following three stages must occur, also depicted in Figure 2.

1. Given the process starts at level 1 in phase $i$, the process reaches level $k$ in some phase $i_2$ without making a downwards transition. The PGF corresponding to this stage is $[M_+(z)^{k-1}]_{i_2i_1}$.

2. Given the process starts from level $k$ in phase $i_1$, the process:
   - Either makes a transition to level $(k-1)$ in phase $i_2$, without reaching leve $(k+1)$, with PGF $[M_-(z)]_{i_1i_2}$.
   - Or, the process first reaches level $(k+1)$ for the first time whilst avoiding level $(k-1)$, and does so in some phase $i_3$. The corresponding PGF is $[M_+(z)]_{i_1i_2}$.

Next, given the process starts from level $(k+1)$ in phase $i_2$, it reaches level $k$ for the first time and does so in some phase $i_4$ along a path $\Omega^L_n$. The corresponding PGF is $[G_n(z)]_{i_2i_4}$.

The process must repeat the above at least one more time ending at level $k$ in some phase $i_5$ under the same restrictions. The corresponding PGF is $\sum_{k=2}^{\infty}([M_+(z)G_n(z)])^k [M_-(z)]_{i_1i_2}$.

3. Finally, given the process starts from level $k-1$ in phase $i_2$, the process first drains to level 0 in phase $j$ without making an upwards transition. The PGF corresponding to this stage is $[M_-(z)^{k-1}]_{i_2j}$.

By stages 1 – 3 above and partitioning on $k \geq 1$, we obtain the following expression,

$$G_{n+1}(z) = \sum_{k=1}^{\infty} M_+(z)^{k-1} \left( I + \sum_{t=2}^{\infty} [M_+(z)G_n(z)]^t \right) M_-(z)^k$$

$$= \sum_{k=1}^{\infty} M_+(z)^{k-1} (I - M_+(z)G_n(z))^{-1} - M_+(z)G_n(z)M_-(z)^k.$$  (16)
We use [15, Theorem 3] below, where \( \rho(\cdot) \) represents the spectral radius of a given matrix.

**Lemma 1.** Equation

\[
X = AXB + C,
\]

for appropriately sized matrices \( A, B \) and \( C \), has the unique solution given by

\[
X = \sum_{k=0}^{\infty} A^kCB^k
\]

if and only if \( \rho(A)\rho(B) < 1 \).

After applying Lemma 1, we express equation (16) as the iterative scheme

\[
GLT_{n+1}(z) - M(z)GLT_n(z)
\]

(19)

\[
GLT_{n+1}(z) = (I - M(z)GLT_n(z))^{-1} - M(z)GLT_n(z)
\]

(20)

to construct the LT Algorithm and then show its convergence to \( G(z) \).

**Lemma 2.** \( GLT_n(z) \) converges to \( G(z) \) as \( n \to \infty \).

**Proof:** Firstly, let \( \Omega \) be the set of sample paths that contribute to \( G(z) \) and recall that \( \OmegaLT_n \) is the set of sample paths that contribute to the \( n \)-th iteration of \( GLT(z) \). The physical interpretation of \( GLT_n(z) \) is the PGF of the time taken to traverse paths in \( \OmegaLT_n \subseteq \Omega \), and so \( 0 \leq GLT_n(z) \leq G(z) \).

Now, from the physical interpretation, for all \( n \), the sample paths contributing to \( GLT_n(z) \), \( \OmegaLT_n \) must also contribute to \( G(z) \), that is, \( \OmegaLT_n \subseteq \Omega \). Now, consider an arbitrary sample path in \( \Omega \). That path must either have a single peak at some level \( k \geq 1 \) or a minimum trough at some level \( k \). As such that path is counted by the \( n \)-th iteration of \( \OmegaLT_n \) for all \( n \) sufficiently large to allow for the necessary sample path components. Since the sample path was chosen arbitrarily, then all sample paths in \( \Omega \) are contained within \( \OmegaLT_n \) for some \( n \). The result follows.

**4. NUMERICAL EXAMPLE**

Consider the six-phase dtQBD version of Example 1 in [8], with

\[
A_+ = \begin{bmatrix}
0.0151 & 0.3021 & 0 & 0 & 0 & 0 \\
0 & 0.0151 & 0.3021 & 0 & 0 & 0 \\
0 & 0 & 0.0151 & 0.3021 & 0 & 0 \\
0 & 0 & 0 & 0.0151 & 0.3021 & 0 \\
0 & 0 & 0 & 0 & 0.0151 & 0.3021 \\
0 & 0 & 0 & 0 & 0 & 0.0151
\end{bmatrix}
\]

(21)

**Algorithm 1** LT Algorithm for calculating \( G(z) \)

**Input:** \( A_-, A_0, A_+ \)

Set a real \( \epsilon > 0 \), \( z \in \mathbb{R}^c > 0 \).

**Set:**

\( M_+ (z) = (I - A_+ z)^{-1} A_+ z \), \( M_- (z) = (I - A_- z)^{-1} A_- z \), and \( GLT_0 (z) = 0 \).

while \( \| GLT_n (z) - GLT_{n+1} (z) \|_\infty > \epsilon \) do

**Compute:**

\[
C = ((I - M_+ (z) GLT_n (z))^{-1} - M_+ (z) GLT_n (z)) M_- (z)
\]

**Solve:**

\[
X - M_+ (z) XM_- (z) = C
\]

**Set:**

\( GLT_n (z) = X \)

end while

**Output:** \( G(z) \approx GLT_n (z) \)

**Work Count:**

One matrix inversion, two matrix products, and solving the Sylvester equation, for approximately \( 61 m^3 \) floating-point operations per iteration.

\[
\begin{bmatrix}
0.6344 & 0.0302 & 0 & 0 & 0 & 0 \\
0.0302 & 0.6042 & 0.0302 & 0 & 0 & 0 \\
0 & 0.0302 & 0.6042 & 0.0302 & 0 & 0 \\
0 & 0 & 0.0302 & 0 & 0.0302 & 0 \\
0 & 0 & 0 & 0 & 0.0302 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.0302
\end{bmatrix}
\]

(22)

and

\[
A_0 = \begin{bmatrix}
0.0181 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.0181 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.0181 & 0.9063 & 0 & 0 \\
0 & 0 & 0 & 0.0181 & 0.9063 & 0 \\
0.9063 & 0 & 0 & 0 & 0.0181 & 0.9063 \\
0.9063 & 0 & 0 & 0 & 0 & 0.0181
\end{bmatrix}
\]

(23)

We set the stopping criterion to \( \epsilon = 10^{-12} \).

The algorithm achieved the desired precision in the production of the matrix

\[
\begin{bmatrix}
0.7831 & 0.0149 & 0.0016 & 0.1084 & 0.0015 & 0.0905 \\
0.6538 & 0.0492 & 0.0030 & 0.1889 & 0.0018 & 0.1033 \\
0.0533 & 0.0016 & 0.0183 & 0.9180 & 0.0002 & 0.0087 \\
0.7426 & 0.0015 & 0.0016 & 0.1270 & 0.0022 & 0.1252 \\
0.0650 & 0.0001 & 0.0000 & 0.0040 & 0.0182 & 0.9126 \\
0.9489 & 0.0002 & 0.0000 & 0.0017 & 0.0006 & 0.0485
\end{bmatrix}
\]

(24)

LT algorithm converges in 60 iterations with an average time of 0.015 seconds on a Dell OptiPlex 7450 AIO.

**5. CONCLUSION**

We constructed a linearly-convergent lowest-trough algorithm for \( G(z) \) by applying physical interpretation analogous to that of Algorithm 1 for \( \Psi(s) \) in [13].

Future work includes using similar methodology to construct further algorithms through their physical interpretation. That is, we partition sample paths according to some key level \( k \), derive the corresponding iterative schemes, and numerically compare them with the existing algorithms.

**6. REFERENCES**

56


