*IMA Journal of Numerical Analysis* (2021) Page 1 of 23
doi:10.1093/imanum/drn000

# A family of fast fixed point iterations for M/G/1-type Markov chains†

DARIO A. BINI‡
*Dipartimento di Matematica, Università di Pisa, Largo B. Pontecorvo, 5, 56127 Pisa, Italy*

GUY LATOUCHE§
*Université Libre de Bruxelles, Département d'Informatique, CP212, 1050 Bruxelles, Belgium*

BEATRICE MEINI¶
*Dipartimento di Matematica, Università di Pisa, Largo B. Pontecorvo, 5, 56127 Pisa, Italy*

[Received on 21 January 2021]

We consider the problem of computing the minimal nonnegative solution $G$ of the nonlinear matrix equation $X = \sum_{i=-1}^{\infty} A_i X^{i+1}$ where $A_i$, for $i \geqslant -1$, are nonnegative square matrices such that $\sum_{i=-1}^{\infty} A_i$ is stochastic. This equation is fundamental in the analysis of M/G/1-type Markov chains, since the matrix $G$ provides probabilistic measures of interest. A new family of fixed point iterations for the numerical computation of $G$, that includes the classical iterations, is introduced. A detailed convergence analysis proves that the iterations in the new class converge faster than the classical iterations. Numerical experiments confirm the effectiveness of our extension.

*Keywords*: Nonlinear matrix equations; fixed point iterations; nonnegative matrices; convergence analysis; M/G/1-type Markov chains.

## 1. Introduction

M/G/1-type Markov chains (see Neuts, 1981) are characterised by a transition matrix in block Hessenberg form of the kind

$$P = \begin{bmatrix} B_0 & B_1 & B_2 & \dots \\ A_{-1} & A_0 & A_1 & \dots \\ & A_{-1} & A_0 & \ddots \\ & & \ddots & \ddots \end{bmatrix} \tag{1.1}$$

where $B_i, A_{i-1}$, $i \geqslant 0$ are nonnegative square matrices of order $m$, such that $\sum_{i=0}^{\infty} B_i$ and $\sum_{i=-1}^{\infty} A_i$ are stochastic matrices. The matrix $P$ is associated with the nonlinear matrix equation

$$X = A_{-1} + A_0 X + A_1 X^2 + A_2 X^3 + \cdots \tag{1.2}$$

where the unknown $X$ is an $m \times m$ matrix. We may write that equation as $X = A(X)$, where $A(z) = A_{-1} + A_0 z + A_1 z^2 + A_2 z^3 + \cdots$. It is well known that (1.2) has a component-wise minimal nonnegative solution $G$, which, besides having a relevant probabilistic interpretation, is fundamental in providing an

†Research partially supported by INdAM-GNCS
‡Email: dario.bini@unipi.it
§Email: latouche@ulb.ac.be
¶Corresponding author. Email: beatrice.meini@unipi.it

explicit representation of the invariant probability measure of the Markov chain, see Bini *et al.* (2005), and Neuts (1981). Motivated by the important role of the matrix $G$, numerous algorithms for the numerical computation of this matrix have been designed and analyzed in the literature. A review of the earliest methods for this problem, specifically in the context of Markov chains, is given in Ramaswami (1988), while the more recent analysis in Higham & Kim (2000), Higham & Kim (2001) deals with matrix quadratic equations in a general framework. Fixed point iterations for Markov chains applications are treated in Chen *et al.* (2019), Guo (1999), Guo (2003), Rhee (2010), and Newton's iteration is the focus in Latouche (1994), Pérez *et al.* (2012), Seo & Kim (2014) and Seo *et al.* (2018). A vast literature is devoted to methods based on cyclic reduction and on doubling algorithms; we refer the reader to the survey paper Bini & Meini (2009) and to the literature cited therein, to the paper Latouche & Ramaswami (1993) for the logarithmic reduction algorithm and to Chiang *et al.* (2009) for a convergence analysis of SDA-based iterations. A conditioning analysis is performed in Meng *et al.* (2018).

A general overview of the stochastic processes theory behind the problem, and of the probabilistic aspects of the early algorithms, is found in the classical books Neuts (1981) and Latouche & Ramaswami (1999), while the more numerically oriented book Bini *et al.* (2005) provides a detailed algorithmic analysis of the problem. In particular, the following three classical fixed point iterations are analysed in (Bini *et al.*, 2005, Chapter 6); as we shall see, these are special cases of the family introduced and analysed in the present paper, and they are known as

$$\text{Natural} \qquad X_{k+1} = \sum_{i=0}^{\infty} A_{i-1} X_k^i, \tag{1.3}$$

$$\text{Traditional} \qquad (I - A_0) X_{k+1} = A_{-1} + \sum_{i=2}^{\infty} A_{i-1} X_k^i, \tag{1.4}$$

$$U\text{-based} \qquad \left( I - \sum_{i=0}^{\infty} A_i X_k^i \right) X_{k+1} = A_{-1}, \tag{1.5}$$

with $k \geqslant 0$, starting from an initial approximation $X_0$. It is shown that, if $X_0 = 0$, the $U$-based iteration converges to $G$ faster than the Traditional iteration which in turn converges faster than the Natural iteration. Observe that in all three cases, the matrix function $A(z)$ goes through a decomposition of the form $A(z) = A_{-1}(z) + A_0(z)z$ where $A_{-1}(z)$ and $A_0(z)$ depend on the specific iteration, and (1.2) is replaced by the equivalent equation $X = A_{-1}(X) + A_0(X)X$ solved by the fixed point iteration $X_{k+1} = A_{-1}(X_k) + A_0(X_k)X_{k+1}$, for $k = 0, 1, \ldots$. For the Natural iteration we have $A_{-1}(z) = \sum_{i=0}^{\infty} A_{i-1} z^i$, $A_0(z) = 0$; here, the nonlinear part of the equation, the linear factor $A_0$ and the constant $A_{-1}$ are embedded in the constant term. For the Traditional iteration we have $A_{-1}(z) = A_{-1} + \sum_{i=2}^{\infty} A_{i-1} z^i$, $A_0(z) = A_0$, i.e., the nonlinear part only of (1.2) is embedded in the function $A_{-1}(z)$. Finally, for the $U$-based iteration we have

$$A_{-1}(z) = A_{-1}, \quad A_0(z) = \sum_{i=0}^{\infty} A_i z^i, \tag{1.6}$$

so that the nonlinear part is embedded in $A_0(z)$, together with $A_0$.

In summary, the computation of $X_{k+1}$ given $X_k$ is reduced to solving a *linear* matrix equation; the original matrix power series equation is reduced to a linear equation by embedding the nonlinear part either in the constant or in the linear coefficient of equation (1.2).

The general strategy suggested by this new interpretation of well-known algorithms is to reduce the

original equation (1.2) to a polynomial matrix equation of the kind

$$X = A_{-1}(X) + A_0(X)X + A_1(X)X^2 + \cdots + A_q(X)X^{q+1}, \tag{1.7}$$

where $q \geqslant -1$, and to embed the terms of degree higher than $q+1$ in the matrix coefficients $A_i(X)$, for $i = -1, \ldots, q$. Indeed, we easily verify that (1.2) is equivalent to (1.7) if the matrix power series $A_\ell(z) = \sum_{i=0}^{\infty} A_{\ell,i} z^i$, $\ell = -1, 0, 1, \ldots, q$, satisfy the only conditions $A_{\ell,i} \geqslant 0$ and $\sum_{\ell=-1}^{q} A_\ell(z) z^{\ell+1} = \sum_{i=0}^{\infty} A_{i-1} z^i$.

Equation (1.7) is solved by the fixed point iteration

$$X_{k+1} = \sum_{\ell=-1}^{q} A_\ell(X_k) X_{k+1}^{\ell+1}, \quad k = 0, 1, \ldots, \tag{1.8}$$

for a given initial approximation $X_0$, where the matrix $X_{k+1}$ is defined as the minimal nonnegative solution of the matrix equation (1.8).

As already pointed out, the classical fixed point iterations can be viewed as specific instances of the new class of iterations (1.8): the Natural iteration is obtained by choosing $q = -1$ and embedding all the terms in $A_{-1}(z)$, $q = 0$ for the Traditional iteration and all the terms of degree higher than 2 are embedded in $A_{-1}(z)$ while for the $U$-based iteration, $q = 0$ also but all the terms of degree higher than 2 are embedded in $A_0(z)$.

Observe that, if $q \geqslant 1$, then the polynomial equation (3.4) has degree at least 2 and its solution requires an iterative technique. Therefore, the algorithms in the new class may be viewed as formed by an outer iteration, which generates $X_{k+1}$ given $X_k$, and an inner iteration needed to approximate $X_{k+1}$ by solving the polynomial equation of degree $q+1$.

We present in this paper our analysis of this new class of fixed point iterations. We rely on the properties of nonnegative matrices and of regular splittings of M-matrices, and show that the sequence $\{X_k\}_k$ given by (3.4) is well defined and monotonically convergent to the solution $G$ if $X_0 = 0$, and convergent if $X_0$ is any stochastic matrix. Moreover, we show that for a given degree $q+1$, the highest speed of convergence is obtained by choosing $A_\ell(z) = A_\ell$, $\ell = -1, 0, \ldots, q-1$, $\quad A_q(z) = \sum_{i=0}^{\infty} A_{i+q} z^i$, that is, by embedding all $A_i$s, $i \geqslant q$, in $A_q(z)$. For this particular family of equations, we show that the convergence speed is higher for larger values of $q$, and so, for $q \geqslant 1$, all the iterations in this class outperform the traditional iterations (1.3 – 1.5).

A further analysis shows that the polynomial matrix equation of degree $q+1$ is better conditioned and easier to solve for smaller values of $q$. Thus, the outer iteration is faster the larger $q$ is, whereas the inner iteration is faster the smaller $q$ is. Apparently, this provides a trade-off in the choice of $q$. However, as we will see from the numerical experiments, the number of inner iterations can be substantially reduced by a suitable choice of the starting approximation, and it turns out that in practice, the number of inner iterations is a nonincreasing function of $q$. This fact makes the new class of iterations not only faster in terms of convergence speed, but also more effective in terms of CPU time. In fact, our numerical experiments show that the speed-up in the CPU time, with respect to the available iterations, is generally larger than 2 and in many cases reaches values larger than 20.

The paper is organized as follows. In Section 2 we recall the main tools used in our analysis. In Section 3 we analyse the convergence of the new family of iterations in the case where $X_0 = 0$. Different embedding strategies are discussed in Section 4, with the aim to determine an optimal value of $q$. In Section 5, we show convergence properties in the case where $X_0$ is a stochastic matrix. A computational cost and conditioning analysis is performed in Section 6, and we report in Section 7 on our numerical experiments.

## 2. Preliminary results

In this section we recall results on nonnegative matrices and M/G/1-type Markov chains that will be used in the rest of the paper.

### 2.1 *Nonnegative matrices*

A real matrix $A$ is called nonnegative if $A \geqslant 0$, where the inequality is meant entry-wise. If $A$ and $B$ are real nonnegative matrices of the same size, the inequality $A \geqslant B$ means $A - B \geqslant 0$. Given a complex matrix $B$, we denote by $|B|$ the matrix whose entries are the moduli of the entries of $B$, while $\rho(B)$ denotes the spectral radius of $B$, i.e., the maximum modulus of its eigenvalues.

Given a matrix power series $A(z) = \sum_{i=0}^{\infty} A_i z^i$ with real matrix coefficients, we write $A(z) \geqslant 0$ if $A_i \geqslant 0$ for any $i \geqslant 0$; moreover, if $B(z)$ is a matrix power series, we write $A(z) \geqslant B(z)$ if $A(z) - B(z) \geqslant 0$.

A matrix of the kind $B = sI - A$, where $A \geqslant 0$ and $s \geqslant \rho(A)$, is called an M-matrix. It is nonsingular if $s > \rho(A)$, otherwise it is a singular M-matrix. An additive splitting $A = M - N$ of the real square matrix $A$ is called a regular splitting if $\det M \neq 0$, $M^{-1} \geqslant 0$ and $N \geqslant 0$.

The following result synthesizes some properties from the Perron-Frobenius theory of nonnegative matrices (Berman & Plemmons, 1994).

THEOREM 2.1 Let $A \geqslant 0$ be a square matrix. The following properties hold:

- $\rho(A)$ is an eigenvalue; if, in addition, $A$ is irreducible, then $\rho(A) > 0$;
- if $B \geqslant A$ then $\rho(B) \geqslant \rho(A)$; if $A$ is irreducible and $B \neq A$, then $\rho(B) > \rho(A)$;
- if $B$ is a complex matrix such that $|B| \leqslant A$, then $\rho(B) \leqslant \rho(A)$.

We provide some properties of M-matrices and their regular splittings (Varga, 2000).

THEOREM 2.2 Assume that $B$ is a nonsingular M-matrix. Then:

- $B^{-1} \geqslant 0$;
- if $B = M - N$ is a regular splitting, then $\rho(M^{-1}N) = \rho(B^{-1}N)/(1 + \rho(B^{-1}N)) < 1$;
- if $B = M_1 - N_1 = M_2 - N_2$ are two regular splittings and $N_1 \leqslant N_2$, then $\rho(M_1^{-1}N_1) \leqslant \rho(M_2^{-1}N_2)$; if $M_1^{-1}N_1 \neq M_2^{-1}N_2$ and $M_1^{-1}N_1$ is irreducible then the inequality is strict.

The following result on matrix polynomials is a consequence of Theorem 2.1.

LEMMA 2.1 Let $B(z) = z^n I - \sum_{i=0}^{n-1} B_i z^i$ and $\tilde{B}(z) = z^n I - \sum_{i=0}^{n} \tilde{B}_i z^i$ be monic matrix polynomials such that $0 \leqslant \tilde{B}_i \leqslant B_i$ for $i = 0, \ldots, n-1$. Then the polynomials $\det B(z)$ and $\det \tilde{B}(z)$ have a nonnegative real root $\lambda$ and $\tilde{\lambda}$, respectively, which is the root of largest modulus. Moreover, $\tilde{\lambda} \leqslant \lambda$.

*Proof.* The roots of $\det B(z)$ and of $\det \tilde{B}(z)$ are the eigenvalues of the block companion matrices $C$ and $\tilde{C}$ associated with the monic matrix polynomials $B(z)$ and $\tilde{B}(z)$, respectively (Gohberg *et al.*, 2009). Since $0 \leqslant \tilde{C} \leqslant C$, from Theorem 2.1 we deduce that $\lambda = \rho(C)$ and $\tilde{\lambda} = \rho(\tilde{C})$, moreover $\rho(\tilde{C}) \leqslant \rho(C)$. □

### 2.2 *M/G/1-type Markov chains*

Assume the M/G/1-type Markov chain with transition matrix (1.1) is irreducible and aperiodic, and that $A = \sum_{i=-1}^{\infty} A_i$ is irreducible and stochastic. Assume also that the series $\sum_{i=-1}^{\infty} iA_i$ is convergent and

define the vector $\boldsymbol{a} = \sum_{i=-1}^{\infty} iA_i\mathbf{1}$, where $\mathbf{1}$ is the vector with all entries equal to 1. Let $\boldsymbol{\alpha}$ be the vector such that $\boldsymbol{\alpha}^{\mathrm{T}}A = \boldsymbol{\alpha}^{\mathrm{T}}$, $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} = 1$; the drift of the Markov chain is defined as (Bini *et al.*, 2005; Neuts, 1981)

$$\mu = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{a} \tag{2.1}$$

and $\mu \leqslant 0$ if and only if the Markov chain is recurrent, $\mu > 0$ if and only if it is transient.

The minimal nonnegative solution of the matrix equation (1.2) is characterised as follows (Bini *et al.*, 2005).

THEOREM 2.3 Let $A_i$, $i = -1, 0, 1, \ldots$, be nonnegative square matrices such that $(\sum_{i=-1}^{\infty} A_i)\mathbf{1} \leqslant \mathbf{1}$. The matrix equation (1.2) has a unique minimal nonnegative solution $G$, i.e., if $Y$ is any other nonnegative solution, then $G \leqslant Y$. Moreover, $G\mathbf{1} \leqslant \mathbf{1}$ and $G$ is the limit of the sequence $X_{k+1} = A_{-1} + A_0X_k + A_1X_k^2 + \cdots$, $k = 0, 1, \ldots$, with $X_0 = 0$.

The next result provides a comparison between the minimal nonnegative solutions of two matrix equations.

THEOREM 2.4 Let $A_i$ and $\widetilde{A}_i$, $i = -1, 0, 1, \ldots$, be nonnegative square matrices such that $\widetilde{A}_i \leqslant A_i$ for any $i \geqslant -1$ and $(\sum_{i=-1}^{\infty} A_i)\mathbf{1} \leqslant \mathbf{1}$. Let $G$ and $H$ be the minimal nonnegative solutions of the matrix equations (1.2) and $X = \sum_{i=0}^{\infty} \widetilde{A}_{i-1}X^i$, respectively. Then $H \leqslant G$.

*Proof.* We have $G = \lim_{k\to\infty} X_k$, where $X_k$ is defined in Theorem 2.3. Similarly, $H = \lim_{k\to\infty} Y_k$, where $Y_k$ is defined by $Y_{k+1} = \sum_{i=0}^{\infty} \widetilde{A}_{i-1}Y_k^i$, with $Y_0 = 0$. By induction on $k$, one has $Y_k \leqslant X_k$, for $k \geqslant 0$, therefore the inequality holds also in the limit. □

For the convergence analysis of fixed point iterations, it is useful to introduce the matrices

$$V = \sum_{j=0}^{\infty} A_j^*, \qquad A_i^* = \sum_{j=i}^{\infty} A_j G^{j-i}, \ \ i \geqslant 0. \tag{2.2}$$

If $\mu < 0$ then $\rho(V) < 1$, so that $H = I - V$ is a nonsingular M-matrix (Bini *et al.*, 2005).

The following result (see Bini *et al.*, 2005), concerns the convergence of fixed point iterations when the starting approximation $X_0$ is stochastic.

THEOREM 2.5 Let $A_i$, $i = -1, 0, 1, \ldots$, be nonnegative square matrices such that $(\sum_{i=-1}^{\infty} A_i)\mathbf{1} = \mathbf{1}$. The matrix equation (1.2) has a unique stochastic solution $G_{\mathrm{sto}}$. It is the limit of the sequence $\{X_k\}$ defined in Theorem 2.3 with $X_0$ stochastic, and $G \leqslant G_{\mathrm{sto}}$, where $G$ is the minimal nonnegative solution of (1.2). Finally, if $\mu \leqslant 0$ then $G_{\mathrm{sto}} = G$.

## 3. A new family of fixed point iterations

In this section we introduce the new family of fixed point iterations, we prove the well-posedness, analyze their convergence properties and, for a given $q$, determine the optimal choice of the matrix power series $A_\ell(z)$ that maximizes the convergence rate. Moreover, we give a physical interpretation of the new family in terms of Markov chain properties.

Let $q \geqslant -1$ and let

$$A_\ell(z) = \sum_{i=0}^{\infty} A_{\ell,i}z^i, \ \ \ell = -1, 0, 1, \ldots, q, \tag{3.1}$$

be matrix power series such that $A_{\ell,i} \geqslant 0$ and

$$\sum_{\ell=-1}^{q} A_\ell(z) z^{\ell+1} = \sum_{i=-1}^{\infty} A_i z^{i+1}. \tag{3.2}$$

Equating the coefficients of $z^i$ on both sides of (3.2) yields

$$A_i = A_{-1,i+1} + A_{0,i} + A_{1,i-1} + \cdots + A_{q,i-q}, \quad i \geqslant -1, \tag{3.3}$$

where, for $\ell = -1, 0, \ldots, q$, we set $A_{\ell,i} = 0$ if $i < 0$. This means that the probability to jump $i$ levels, represented by the matrix $A_i$, is spread into the probability to jump $\ell$ levels, where $\ell$ ranges from $-1$ to $q$.

By replacing the variable $z$ in (3.2) with the matrix $X$, we easily check that the original equation (1.2) is equivalent to the matrix equation $X = A_{-1}(X) + A_0(X)X + A_1(X)X^2 + \cdots + A_q(X)X^{q+1}$. If we interpret $A_i(X)$, for $i = -1, \ldots, q$, as matrix coefficients, then the latter matrix equation can be seen as a matrix polynomial equation of degree $q+1$. The new family of fixed point iterations consists in solving the above equation by means of the fixed point iteration

$$X_{k+1} = \sum_{\ell=-1}^{q} A_\ell(X_k) X_{k+1}^{\ell+1}, \quad k = 0, 1, \ldots, \tag{3.4}$$

for a given initial approximation $X_0$, where the matrix $X_{k+1}$ is defined as the minimal nonnegative solution of the matrix equation (3.4). We prove with the next theorem that, if $X_0 = 0$, the sequence $\{X_k\}_k$ generated by (3.4) is well defined and converges monotonically to the solution $G$ of (1.2).

THEOREM 3.1 Set $X_0 = 0$ and, for $k \geqslant 0$, define $X_{k+1}$ as the minimal nonnegative solution of (3.4). Then the sequence $\{X_k\}_k$ is well defined, $0 \leqslant X_k \leqslant X_{k+1}$ and $X_{k+1}\mathbf{1} \leqslant \mathbf{1}$ for $k = 0, 1, \ldots$. Moreover, the sequence $\{X_k\}_k$ converges monotonically to the minimal nonnegative solution $G$ of (1.2).

*Proof.* We first observe that, since $A_{\ell,i} \geqslant 0$ for any $\ell$ and $i$, then $0 \leqslant A_\ell(X) \leqslant A_\ell(Y)$ whenever $0 \leqslant X \leqslant Y$. We prove the theorem by induction on $k$. If $k = 0$ then $A_\ell(0) \geqslant 0$, for $\ell = -1, 0, \ldots, q$, and $\sum_{\ell=-1}^{q} A_\ell(0)\mathbf{1} \leqslant \mathbf{1}$, therefore the minimal nonnegative solution $X_1$ exists and $X_1\mathbf{1} \leqslant \mathbf{1}$ by Theorem 2.3. Assume that $0 \leqslant X_k \leqslant X_{k+1}$ and $X_{k+1}\mathbf{1} \leqslant \mathbf{1}$. Then $0 \leqslant A_\ell(X_k) \leqslant A_\ell(X_{k+1})$. Moreover, since $X_{k+1}\mathbf{1} \leqslant \mathbf{1}$, then $\sum_{\ell=-1}^{q} A_\ell(X_{k+1})\mathbf{1} \leqslant \mathbf{1}$. Therefore, according to Theorems 2.3 and 2.4, the minimal nonnegative solution $X_{k+2}$ exists, $0 \leqslant X_{k+1} \leqslant X_{k+2}$ and $X_{k+2}\mathbf{1} \leqslant \mathbf{1}$. Since the sequence $\{X_k\}_k$ is monotonic non-decreasing and bounded from above, it is convergent to a limit $G$, which solves equation (1.2) by continuity. Such limit is the minimal nonnegative solution since, if $Y \geqslant 0$ is any other nonnegative solution, then we easily prove by induction on $k$ that $X_k \leqslant Y$ for any $k$. Therefore, taking the limit yields $G \leqslant Y$. □

To further analyse the convergence, we need the following technical lemma, which can be easily proved by induction.

LEMMA 3.1 If $X$ and $Y$ are square matrices of the same size, and $n \geqslant 0$ is an integer, then $X^n - Y^n = \sum_{j=0}^{n-1} X^j (X-Y) Y^{n-j-1}$.

The following result provides a relation between the error $E_k = G - X_k$ at two subsequent steps.

THEOREM 3.2 Let $E_k = G - X_k$, $k \geqslant 0$, where the sequence $\{X_k\}$ is defined by means of (3.4) for any

given $X_0$. One has

$$\begin{aligned} E_{k+1} &= \sum_{\ell=0}^{q} A_\ell(G) \sum_{j=0}^{\ell} G^j E_{k+1} X_{k+1}^{\ell-j} + S(E_k), \\ S(E_k) &= \sum_{\ell=-1}^{q} \sum_{i=1}^{\infty} A_{\ell,i} \sum_{j=0}^{i-1} G^j E_k X_k^{i-j-1} X_{k+1}^{\ell+1}. \end{aligned} \tag{3.5}$$

*Proof.* By subtracting (3.4) from the equation $G = \sum_{\ell=-1}^{q} A_\ell(G) G^{\ell+1}$, we obtain

$$E_{k+1} = \sum_{\ell=-1}^{q} \left( A_\ell(G) \left( G^{\ell+1} - X_{k+1}^{\ell+1} \right) + (A_\ell(G) - A_\ell(X_k)) X_{k+1}^{\ell+1} \right).$$

From (3.1) and Lemma 3.1, we find that $G^{\ell+1} - X_{k+1}^{\ell+1} = \sum_{j=0}^{\ell} G^j E_{k+1} X_{k+1}^{\ell-j}$, and $A_\ell(G) - A_\ell(X_k) = \sum_{i=1}^{\infty} A_{\ell,i} \sum_{j=0}^{i-1} G^j E_k X_k^{i-j-1}$, $\ell = -1, 0, 1, \ldots, q$, hence we arrive at (3.5). □

If $X_0 = 0$ then, in view of Theorem 3.1, $E_k \geqslant 0$ so that $\|E_k\|_\infty = \|E_k \mathbf{1}\|_\infty$. Therefore, we may estimate the convergence rate of the sequence $\{X_k\}_k$ by analyzing the convergence of the sequence $\boldsymbol{\varepsilon}_k = E_k \mathbf{1} \geqslant 0$. The following theorem provides information in this regard.

THEOREM 3.3 If $X_0 = 0$, then $M\boldsymbol{\varepsilon}_{k+1} \leqslant N\boldsymbol{\varepsilon}_k$, where

$$M = I - F, \quad N = \sum_{\ell=-1}^{q} \sum_{i=1}^{\infty} A_{\ell,i} \sum_{j=0}^{i-1} G^j, \quad F = \sum_{\ell=0}^{q} A_\ell(G) \sum_{j=0}^{\ell} G^j. \tag{3.6}$$

Moreover, $M - N = H$, where $H = I - V$ and $V$ is defined in (2.2). If the drift $\mu$ of (2.1) is negative, then $\rho(F) < 1$, $M$ is a nonsingular M-matrix and $H = M - N$ is a regular splitting, therefore $\rho(M^{-1}N) = \rho(H^{-1}N)/(1 + \rho(H^{-1}N)) < 1$.

*Proof.* Since $X_0 = 0$, it results from Theorem 3.1 that $0 \leqslant X_k \leqslant G$ and $X_k \mathbf{1} \leqslant \mathbf{1}$ and, multiplying (3.5) on the right by $\mathbf{1}$, we obtain $\boldsymbol{\varepsilon}_{k+1} \leqslant \sum_{\ell=0}^{q} A_\ell(G) \sum_{j=0}^{\ell} G^j \boldsymbol{\varepsilon}_{k+1} + S(E_k)\mathbf{1}$. On the other hand $S(E_k)\mathbf{1} \leqslant \sum_{\ell=-1}^{q} \sum_{i=1}^{\infty} A_{\ell,i} \sum_{j=0}^{i-1} G^j \boldsymbol{\varepsilon}_k$ whence $M\boldsymbol{\varepsilon}_{k+1} \leqslant N\boldsymbol{\varepsilon}_k$. By using equation (3.3), we find that

$$I - (M - N) = \sum_{\ell=-1}^{q} A_\ell(G) \sum_{j=0}^{\ell} G^j + \sum_{\ell=-1}^{q} \sum_{i=1}^{\infty} A_{\ell,i} \sum_{j=0}^{i-1} G^j = \sum_{\ell=-1}^{q} \left( \sum_{i=0}^{\infty} A_{\ell,i} G^i \sum_{j=0}^{\ell} G^j + \sum_{i=1}^{\infty} A_{\ell,i} \sum_{j=0}^{i-1} G^j \right),$$

where in the first equality we used the fact that $\sum_{j=0}^{\ell} G^j = 0$ if $\ell = -1$, whence we get $I - (M - N) = \sum_{\ell=-1}^{q} \sum_{i=0}^{\infty} A_{\ell,i} \sum_{j=0}^{i+\ell} G^j = \sum_{\ell=-1}^{q} \sum_{k=\ell}^{\infty} A_{\ell,k-\ell} \sum_{j=0}^{k} G^j$. By using the convention that $A_{\ell,i} = 0$ if $i < 0$ and the property that all the series are absolutely convergent, we may exchange the order of the summations in the last term of the above equation, and, by using (3.3) again, arrive at $I - (M - N) = \sum_{k=-1}^{\infty} \sum_{\ell=-1}^{q} A_{\ell,k-\ell} \sum_{j=0}^{k} G^j = \sum_{k=-1}^{\infty} A_k \sum_{j=0}^{k} G^j = V$. To show that $H = M - N$ is a regular splitting, we need to prove that $\det M \neq 0$ and $M^{-1} \geqslant 0$. Since $M = I - F$, where $0 \leqslant F \leqslant V$ and $\rho(V) < 1$ if $\mu < 0$ (see Bini *et al.*, 2005, Theorem 4.14), we have $\rho(F) \leqslant \rho(V) < 1$ by Theorem 2.1, so that $M$ is a nonsingular M-matrix and $M^{-1} \geqslant 0$. As $N$ is nonnegative, $H = M - N$ is a regular splitting, and we obtain the expression for $\rho(M^{-1}N)$ from Theorem 2.2. □

The spectral radius of $M^{-1}N$ provides an estimate of the convergence rate of the sequence $\{X_k\}_k$, when $X_0 = 0$. Indeed, from Theorem 3.3, $0 \leqslant \boldsymbol{\varepsilon}_k \leqslant (M^{-1}N)^k \boldsymbol{\varepsilon}_0$, therefore $\|\boldsymbol{\varepsilon}_k\|_\infty \leqslant \|(M^{-1}N)^k\|_\infty \|\boldsymbol{\varepsilon}_0\|_\infty$.

The ratio $\|\boldsymbol{\varepsilon}_k\|_\infty/\|\boldsymbol{\varepsilon}_{k-1}\|_\infty$ represents the norm reduction of the error at step $k$, while the geometric mean of the reduction of the errors in the first $k$ steps, i.e.,

$$r_k := \left(\frac{\|\boldsymbol{\varepsilon}_1\|}{\|\boldsymbol{\varepsilon}_0\|}\frac{\|\boldsymbol{\varepsilon}_2\|}{\|\boldsymbol{\varepsilon}_1\|}\cdots\frac{\|\boldsymbol{\varepsilon}_k\|}{\|\boldsymbol{\varepsilon}_{k-1}\|}\right)^{\frac{1}{k}} = \left(\frac{\|\boldsymbol{\varepsilon}_k\|}{\|\boldsymbol{\varepsilon}_0\|}\right)^{\frac{1}{k}},$$

represents the average reduction of the errors per step after $k$ steps. Observe that, $r_k \leqslant \|(M^{-1}N)^k\|_\infty^{1/k}$. By following Guo (1999), Meini (1997), we define the asymptotic rate of convergence $r = \limsup_k r_k$. From the latter inequality and the property $\lim_k \|(M^{-1}N)^k\|_\infty^{1/k} = \rho(M^{-1}N)$, we find that $r \leqslant \rho(M^{-1}N)$.

In consequence of Theorem 3.3, we next compare the speed of convergence of different iterations by comparing the corresponding matrix $F$ of (3.6).

THEOREM 3.4 Let $\{X_k^{(h)}\}_k$, for $h = 1,2$, be two sequences generated by (3.4), with $X_0^{(1)} = X_0^{(2)} = 0$, defined by $A_\ell^{(h)}(z) = \sum_{i=0}^\infty A_{\ell,i}^{(h)} z^i$, $\ell = -1,\ldots,q_h$, for $h = 1,2$. Let $r^{(h)}$, $h = 1,2$, be their asymptotic rates of convergence. If

$$\sum_{\ell=0}^{q_1} A_\ell^{(1)}(G)\sum_{j=0}^{\ell} G^j \geqslant \sum_{\ell=0}^{q_2} A_\ell^{(2)}(G)\sum_{j=0}^{\ell} G^j, \tag{3.7}$$

then $r^{(1)} \leqslant r^{(2)}$, i.e., the sequence $\{X_k^{(1)}\}_k$ converges faster than the sequence $\{X_k^{(2)}\}_k$.

*Proof.* By Theorem 3.3, $H = M^{(h)} - N^{(h)}$, $h = 1,2$, are two regular splittings. Therefore, if $N^{(1)} \leqslant N^{(2)}$, then $\rho({M^{(1)}}^{-1}N^{(1)}) \leqslant \rho({M^{(2)}}^{-1}N^{(2)})$. On the other hand, $N^{(1)} \leqslant N^{(2)}$ is equivalent to $I - M^{(1)} \geqslant I - M^{(2)}$, i.e., equivalent to (3.7). □

As a corollary of Theorem 3.4, we obtain the results shown in Meini (1997), whereby the $U$-based iteration (1.5) is faster than the traditional iteration (1.4), which is in turn faster than the natural iteration (1.3). Indeed, for these three iterations $q \leqslant 0$ and, denoting by $F^{(N)}$, $F^{(T)}$ and $F^{(U)}$ the matrix $F$ in (3.6) for the three iterations, we find that $F^{(N)} = 0 \leqslant F^{(T)} = A_0 \leqslant F^{(U)} = \sum_{i=0}^\infty A_i G^i$.

For the next theorem, we assume that $A_{-1}(z) = A_{-1}$: this means that the transition probability matrices $A_0, A_1, \ldots$ may be variously embedded in the coefficients of order 0 to $q$, but none in $A_{-1}(G)$. We show below that any iteration of the kind (3.4), with $q \geqslant 1$, which satisfies this constraint converges faster than the $U$-based iteration; we discuss at the end of Section 3.1 the physical significance of this assumption.

If $A_{-1}(z) = A_{-1}$, Equation (3.2) may be rewritten as

$$\sum_{\ell=0}^{q} A_\ell(z) z^\ell = \sum_{i=0}^{\infty} A_i z^i, \tag{3.8}$$

so that, by replacing $z$ with $G$, we deduce that

$$\sum_{\ell=0}^{q} A_\ell(G) G^\ell = \sum_{i=0}^{\infty} A_i G^i. \tag{3.9}$$

THEOREM 3.5 Let $q \geqslant 1$, assume that $X_0 = 0$, and that $A_{-1}(z) = A_{-1}$. The sequence $\{X_k\}_k$ generated by (3.4) converges faster than the sequence (1.5) generated by the $U$-based iteration.

*Proof.* The proof is a consequence of Theorem 3.4, where $\{X_k^{(1)}\}_k$ is the sequence defined by (3.4), and $\{X_k^{(2)}\}_k$ is the sequence (1.5). From (1.6), we have $A_0^{(2)}(G) = \sum_{i=0}^\infty A_i G^i$ and $A_\ell^{(2)}(G) = 0$ for $\ell \geqslant 1$,

therefore the inequality (3.7) to be verified is equivalent to $\sum_{\ell=0}^{q} A_\ell(G) \sum_{j=0}^{\ell} G^j \geqslant \sum_{i=0}^{\infty} A_i G^i$, this clearly holds by (3.9). □

### 3.1 *Physical interpretation*

We may give an interpretation of Theorems 3.4 and 3.5, based on the physical significance of the matrices $V = (v_{ij})_{ij}$ and $F = (f_{i,j})_{ij}$. Assume that the Markov chain with transition matrix (1.1) starts at time 0 in an arbitrary but fixed level $n \geqslant 1$. Define $\tau_{-1}$, $\tau_0$, $\tau_1$, ... to be the epochs when the Markov chain makes a first transition to level $n+i$, for $i \geqslant -1$, and define $N_j = \sum_{\nu \geqslant 0} \mathbb{I}[\tau_\nu < \tau_{-1}, \varphi(\tau_\nu) = j]$, for $j = 1, \ldots, m$, where $\mathbb{I}[\cdot]$ is the indicator function and $\varphi(t)$ represents the phase occupied by the Markov chain at time $t$; that is, $N_j$ is the total number of times the Markov chain visits phase $j$, at the epochs of first visit to a new level, under taboo of the levels below level $n$. It is easy to verify that $v_{ij}$ is the conditional expected number of such visits, given that the initial phase is $i$. Similarly, $f_{ij}$ is the expected number of such visits for the Markov chain where the transition blocks $A_k$ are replaced by $A_k(G)$, for $k \geqslant -1$, and we consider it as providing an approximation of $v_{ij}$.

From Theorem 3.3, we know that $F \leqslant V$, the inequality (3.7) may be rewritten as $V \geqslant F^{(1)} \geqslant F^{(2)}$, and so Theorem 3.4 states that $\{X_k^{(1)}\}_k$ converges faster than $\{X_k^{(2)}\}_k$ if the approximation $f_{ij}^{(1)}$ is uniformly better than $f_{ij}^{(2)}$ for all $i$ and $j$.

Define

$$N'_j = \mathbb{I}[\tau_0 < \tau_{-1}, \varphi(\tau_0) = j] \leqslant N_j. \tag{3.10}$$

The expected value of $N'_j$ is the probability of returning to the initial level $n$ in phase $j$, under taboo of level $n-1$, and (3.9) shows that this probability is the same for all embedding $A_{\ell,k}$, $\ell \geqslant 0$, if $A_{-1}(G) = A_{-1}$, that is, provided that the probability of transiting immediately to the lower level remains unchanged.

Now, we have for the $U$-based iteration $A_{-1}^{(U)} = A_{-1}$, $A_0^{(U)} = \sum_{k \geqslant 0} A_k G^k$, $A_\ell^{(U)} = 0$, for $\ell \geqslant 2$, by (1.6), and we readily verify that $F^{(U)}$ may also be interpreted as the matrix of expected values of $N'$ and (3.10), together with Theorem 3.4 provides a physical justification for Theorem 3.5.

## 4. Optimal embedding

In this section we examine the role of the integer $q$ which determines the degree of the matrix equation (3.4) to be solved at each step. In particular, the goal is to give properties to determine an optimal value of $q$, in terms of speed of convergence of the sequence $\{X_k\}_k$ and in terms of numerical properties of the matrix equation (4.8).

### 4.1 *Comparisons*

To embed the tail of the series only in the coefficients of the terms of degrees 0 to $q+1$, for a given integer $q \geqslant 1$, according to (3.3), the matrices $A_{\ell,k}$ satisfy

$$A_\nu = A_{-1,\nu+1} + A_{0,\nu} + \cdots + A_{\nu,0}, \qquad \text{for } -1 \leqslant \nu \leqslant q, \tag{4.1}$$

$$A_\nu = A_{-1,\nu+1} + A_{0,\nu} + \cdots + A_{q,\nu-q}, \qquad \text{for } \nu \geqslant q+1, \tag{4.2}$$

$$A_{\ell,i} = 0, \qquad \text{for } \ell \geqslant q+1,\ i \geqslant 0. \tag{4.3}$$

THEOREM 4.1 Given an integer $q \geqslant 1$, the parameters that maximize the matrix $F$ of (3.6) are given by

$$A_{\ell,0} = A_\ell, \qquad \text{for } -1 \leqslant \ell \leqslant q, \tag{4.4}$$

$$A_{q,\ell-q} = A_\ell, \qquad \text{for } \ell \geqslant q+1, \tag{4.5}$$

$$A_{\ell,i} = 0, \qquad \text{for } -1 \leqslant \ell \leqslant q-1,\ i \geqslant 1. \tag{4.6}$$

In other words,

$$A_\ell(z) = A_\ell, \quad \text{for } -1 \leqslant \ell \leqslant q-1, \qquad A_q(z) = \sum_{i=q}^{\infty} A_i z^{i-q}. \tag{4.7}$$

and the tail of the series is embedded in the coefficient of degree $q+1$ only.

*Proof.* The sum $F$ of (3.6) may be re-written as

$$F = \sum_{0\leqslant k\leqslant q} \sum_{i\geqslant 0} A_{k,i} \sum_{i\leqslant \nu\leqslant k+i} G^\nu = \sum_{\ell\geqslant 0} \sum_{0\leqslant k\leqslant \min(\ell,q)} A_{k,\ell-k} \sum_{\ell-k\leqslant \nu\leqslant \ell} G^\nu$$

which coincides with $\sum_{\ell\geqslant 0} \sum_{0\leqslant k\leqslant \min(\ell,q)} A_{k,\ell-k} B_{k,\ell}$, where $B_{k,\ell} = \sum_{\ell-k\leqslant \nu\leqslant \ell} G^\nu$ is increasing with $k$, for any given $\ell$. Thus, it suffices to use the matrices in (4.4 – 4.6) to maximise $F$ under the constraints (4.1, 4.2). Finally, if we equate the coefficients of $z^0$ in (3.2), we find that $A_{-1,0} = A_{-1}$, and this completes the proof. □

### 4.2 *Spectral properties*

In this analysis we restrict the attention to the case where $A_\ell(z) \geqslant A_\ell$, for $\ell = 0, 1, \ldots, q$, where we recall that the inequality involving matrix power series is meant coefficient-wise. We show some spectral properties which clarify the role of $q$ in the numerical properties of the matrix equation

$$X = A_{-1}(G) + A_0(G)X + A_1(G)X^2 + \cdots + A_{q-1}(G)X^q + A_q(G)X^{q+1}, \tag{4.8}$$

in terms of conditioning and speed of convergence of fixed point iterations.

Define the matrix Laurent power series $S(z) = I - \sum_{i=-1}^{\infty} A_i z^i$, and the polynomial $S_q(z) = I - \sum_{i=-1}^{q} A_i(G) z^i$, associated with the matrix equations (1.2) and (4.8) respectively. If the drift $\mu$ is negative, it is well known (see Bini *et al.*, 2005, Theorem 4.12) that there exists $\xi > 1$ such that $\det S(\xi) = 0$ and $\xi = \min\{|z| : z \in \mathbb{C}, |z| > 1, \det S(z) = 0\}$. The closeness of $\xi$ to 1 governs the convergence of numerical methods for solving the matrix equation (1.2), as well as the conditioning of the problem: the closer is $\xi$ to 1, the slower is the convergence of numerical methods and the worse is the conditioning (Bini *et al.*, 2005, Chapter 7).

We show in the theorem below that the smallest root of $\det S_q(z)$ outside the closed unit disk is larger than $\xi$, and so the matrix equation (4.8) has better numerical properties than the original equation (1.2), if $A_\ell(z) \geqslant A_\ell$, $0 \leqslant \ell \leqslant q$, and $A_{-1}(z) = A_{-1}$. In view of Theorem 4.1, this is not a very restrictive assumption.

THEOREM 4.2 Assume $q \geqslant 1$. If $A_\ell(z) \geqslant A_\ell$, for $\ell = 0, 1, \ldots, q$, and $A_{-1}(z) = A_{-1}$, then the matrix functions $S(z)$ and $S_q(z)$ may be factorized as

$$S(z) = U(z)(I - z^{-1}G), \quad S_q(z) = U_q(z)(I - z^{-1}G),$$

where $U(z) = I - \sum_{i=0}^{\infty} A_i^* z^i$, $U_q(z) = I - \sum_{i=0}^{q} B_i^* z^i$, the matrices $A_i^*$, $i \geqslant 0$, are defined in (2.2) and $B_i^* = \sum_{j=i}^{q} A_j(G) G^{j-i}$, $i = 0, \ldots, q$.

Moreover, $B_0^* = A_0^*$, $B_i^* \leqslant A_i^*$ for $i = 1, \ldots, q$, and, if $\sum_{i=-1}^{\infty} A_i z^{i+1}$ is a matrix polynomial and if the drift $\mu$ is negative, there exists $\xi_q \geqslant \xi > 1$ such that $\det S_q(\xi_q) = 0$ and $\xi_q = \min\{|z| : z \in \mathbb{C}, |z| > 1, \det S_q(z) = 0\}$.

*Proof.* The factorization of $S(z)$ is known (see Bini *et al.*, 2005, Theorem 4.13). To verify that $S_q(z) = (I - \sum_{i=0}^{q} B_i^* z^i)(I - z^{-1}G)$, we use the fact that $G$ is a solution of the matrix equation (4.8) and verify that the coefficients of equal powers of $z$ are equal.

To verify that $B_0^* = A_0^*$, we replace $z$ with $G$ in (3.8). To verify that $B_i^* \leqslant A_i^*$, $i = 1, \ldots, q$, we define the matrix power series $B_i^*(z) = \sum_{j=i}^{q} A_j(z) z^{j-i}$ and $A_i^*(z) = \sum_{j=i}^{\infty} A_j z^{j-i}$ and show that $B_i^*(z) \leqslant A_i^*(z)$. Since the coefficients of these power series and $G$ are nonnegative matrices, by replacing $z$ with $G$, that inequality implies $B_i^* \leqslant A_i^*$, $i = 1, \ldots, q$. The inequality $B_i^*(z) \leqslant A_i^*(z)$ is equivalent to $z^i B_i^*(z) \leqslant z^i A_i^*(z)$ and we obtain from (3.8) that $B_i^*(z) z^i - A_i^*(z) z^i = \sum_{j=0}^{i-1} A_j z^j - \sum_{j=0}^{i-1} A_j(z) z^j \leqslant 0$, under the assumption $A_\ell(z) \geqslant A_\ell$, $\ell = 0, \ldots, q$. By denoting $u(z) = \det U(z)$, if $\mu < 0$ then the roots of $u(z)$ lie outside the closed unit disk, $u(\xi) = 0$ and $\xi = \min\{|z| : z \in \mathbb{C}, |z| > 1, u(z) = 0\}$. Moreover, since $\det(I - A_0^*) \neq 0$, then $u(z) = 0$ if and only if $\tilde{u}(z) = 0$, where $\tilde{u}(z) = \det(I - \sum_{i=1}^{d} C_i z^i)$, with $C_i = (I - A_0^*)^{-1} A_i^*$ and $d$ is such that $A_i = 0$ for $i > d$. Similarly, by denoting $u_q(z) = \det U_q(z)$, since $B_0^* = A_0^*$, we have $u_q(z) = 0$ if and only if $\tilde{u}_q(z) = 0$, where $\tilde{u}_q(z) = \det(I - \sum_{i=1}^{q} \tilde{C}_i z^i)$, with $\tilde{C}_i = (I - A_0^*)^{-1} B_i^*$, $i = 1, \ldots, q$. The inequality $\xi_q \geqslant \xi$ follows by applying Lemma 2.1 to the reversed matrix polynomials $z^d (I - \sum_{i=1}^{d} C_i z^{-i})$ and $z^d (I - \sum_{i=1}^{q} \tilde{C}_i z^{-i})$. □

### 4.3 *Embedding the mass in the largest degree coefficient*

It follows from Theorem 4.1 that the fastest convergence of the sequence $\{X_k\}_k$ is obtained by embedding the tail of the series into the coefficient of largest degree. Moreover, embedding the mass into a coefficient of index $q_2 > q_1$, gives a sequence having a faster convergence rate. Hence, the larger $q$, the faster convergence of the sequence $\{X_k\}_k$. In the limit case where $A(z)$ is a polynomial of degree $d$, one iteration is sufficient to obtain $G$ if one sets $q = d - 1$. However, in this latter case, the new algorithm does not provide any advantage, since the equation to be solved coincides with the original one. If the coefficients are defined as in (4.7), one easily checks from Theorem 4.2 that $U_q(z)$ is obtained by truncating the series $U(z)$ at a polynomial of degree $q$, i.e., $U_q(z) = I - \sum_{i=0}^{q} A_i^* z^i$. Furthermore, it follows from Lemma 2.1, that $\xi_{q_1} \geqslant \xi_{q_2}$ if $q_1 < q_2$. Therefore, the numerical properties of the matrix equation (4.8) with coefficients defined in (4.7) are better for smaller values of $q$. Hence, there is an optimal value $q$ which results from a trade-off between the good convergence properties of the sequence $\{X_k\}_k$ and the good numerical properties of the matrix equation (4.8) to be solved at each step $k$. From a theoretical point of view, it is difficult to determine the optimal value of $q$. We will discuss this issue in Section 6.

## 5. The case of stochastic initial approximation

In this section we study the convergence of the sequence (3.4), in the case where the starting approximation $X_0$ is a stochastic matrix, and we prove that it is formed of stochastic matrices and converges to the stochastic solution of (1.2).

THEOREM 5.1 Assume that the drift $\mu$ is nonpositive and let $X_0$ be a stochastic matrix. Then, for any $k \geqslant 0$, the matrix equation (3.4) has a unique stochastic solution $X_{k+1}$, so that the sequence $\{X_k\}_k$ is well defined. Moreover, the sequence $\{X_k\}_k$ converges to the minimal nonnegative solution $G$ of (1.2), which is stochastic.

*Proof.* We prove by induction that $X_k$ is stochastic. For $k=0$, $X_k$ is stochastic. Assume that, for a $k \geqslant 0$, the matrix $X_k$ is stochastic. Observe that $A_i(X_k)\mathbf{1} = A_i(1)\mathbf{1}$ so that from (3.2) it follows that $\sum_{i=-1}^{q} A_i(1)\mathbf{1} = \mathbf{1}$. Thus, applying Theorem 2.5, we obtain that (3.4) has a unique stochastic solution $X_{k+1}$. Since $\mu \leqslant 0$ then $G$ is stochastic. We prove that $\lim_k X_k = G$. Observe that stochastic matrices form a compact set so that the sequence $\{X_k\}_k$ has a converging subsequence $\{X_{k_i}\}_i$ which converges to a stochastic matrix $S$. Consider the sequence defined by recursion (3.4), obtained by starting with the null matrix, and denote, within this proof, such sequence by $\{Y_k\}_k$. We may easily show by induction that $Y_k \leqslant X_k$ for any $k \geqslant 0$. Since $\lim_{k\to\infty} Y_k = G$, then $G \leqslant S$. Since both $G$ and $S$ are stochastic, then $G = S$. Therefore, any converging subsequence of $\{X_k\}_k$ converges to the same limit $G$, therefore the sequence $\{X_k\}_k$ is convergent and converges to $G$. □

Now we will show that, if $\mu \leqslant 0$, so that $G$ is stochastic, then the sequence obtained with $X_0$ stochastic converges faster than the sequence obtained with $X_0 = 0$. To this aim, we need to rewrite (3.5) in a slightly different way. We subtract (3.4) from the equation $G = \sum_{\ell=-1}^{q} A_\ell(G)G^{\ell+1}$ and obtain $E_{k+1} = \sum_{\ell=-1}^{q}(A_\ell(X_k)G^{\ell+1} - X_{k+1}^{\ell+1}) + (A_\ell(G) - A_\ell(X_k))G^{\ell+1})$, where $E_k = G - X_k$. We use Lemma 3.1 and find that

$$\begin{aligned} E_{k+1} &= \sum_{\ell=0}^{q} A_\ell(X_k) \sum_{j=0}^{\ell} X_{k+1}^j E_{k+1} G^{\ell-j} + \widehat{S}(E_k), \\ \widehat{S}(E_k) &= \sum_{\ell=-1}^{q} \sum_{i=1}^{\infty} A_{\ell,i} \sum_{j=0}^{i-1} X_k^j E_k G^{i-j+\ell}. \end{aligned} \tag{5.1}$$

We write the matrix product $Y = AXB$ as $\mathrm{vec}(Y) = (B^{\mathrm{T}} \otimes A)\mathrm{vec}(X)$, where $\otimes$ is the Kronecker product and $\mathrm{vec}(C)$ is the vector obtained by stacking the columns of the matrix $C$. Setting $\eta_k = \mathrm{vec}(E_k)$, we rewrite (5.1) as

$$(I - Q(X_k, X_{k+1}))\eta_{k+1} = P(X_k)\eta_k, \quad k = 0, 1, \ldots, \tag{5.2}$$

where

$$\begin{aligned} Q(X,Y) &= \sum_{\ell=0}^{q} \sum_{s=0}^{\ell} \left( (G^s)^{\mathrm{T}} \otimes A_\ell(X) Y^{\ell-s} \right), \\ P(X) &= \sum_{\ell=-1}^{q} \sum_{i=1}^{\infty} \sum_{s=\ell+1}^{\ell+i} (G^s)^{\mathrm{T}} \otimes (A_{\ell,i} X^{i+\ell-s}). \end{aligned} \tag{5.3}$$

As we are interested in asymptotic convergence results, we analyse in the next lemma the spectral properties of the matrices $Q(G,G)$ and $P(G)$.

LEMMA 5.1 Let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $G$. The set of eigenvalues of the matrix $Q(G,G)$, defined in (5.3), is the union of the sets of eigenvalues of the matrices $\sum_{\ell=0}^{q}\sum_{s=0}^{\ell} \lambda_i^s A_\ell(G)G^{\ell-s}$ for $i = 1, \ldots, m$. In particular, $\rho(Q(G,G)) \leqslant \rho(\sum_{\ell=0}^{q}\sum_{s=0}^{\ell} A_\ell(G)G^{\ell-s})$ and, if the drift $\mu$ of (2.1) is negative, then $\rho(Q(G,G)) < 1$, so that $I - Q(G,G)$ is invertible.

*Proof.* Let $T = SG^{\mathrm{T}}S^*$ be the Schur form of the matrix $G^{\mathrm{T}}$, where $T$ is upper triangular with diagonal entries $\lambda_1, \ldots, \lambda_m$, $S$ is a unitary matrix and the symbol $*$ denotes conjugate transposition. The matrix $(S \otimes I)Q(G,G)(S^* \otimes I)$ is block upper triangular with diagonal blocks $\sum_{\ell=0}^{q}\sum_{s=0}^{\ell} \lambda_i^s A_\ell(G)G^{\ell-s}$ for $i = 1, \ldots, m$. Since the set of eigenvalues of a block triangular matrix is the union of the sets of eigenvalues of the diagonal blocks, then the first claim follows. Moreover, as $|\lambda_i| \leqslant 1$ for any $i$, we have $|\sum_{\ell=0}^{q}\sum_{s=0}^{\ell} \lambda_i^s A_\ell(G)G^{\ell-s}| \leqslant \sum_{\ell=0}^{q}\sum_{s=0}^{\ell} |\lambda_i|^s A_\ell(G)G^{\ell-s} \leqslant \sum_{\ell=0}^{q}\sum_{s=0}^{\ell} A_\ell(G)G^{\ell-s}$. If $\mu < 0$, the rightmost matrix in the inequality above has spectral radius less than 1 in view of Theorem 3.3. Therefore, $\rho(Q(G,G)) < 1$ by Theorem 2.1. □

If the drift $\mu$ is negative, it follows from the invertibility of $I - Q(G,G)$ that, if the sequence $\{X_k\}_k$ converges to $G$, there exists $k_0 > 0$ such that for any $k \geqslant k_0$ the matrix $I - Q(X_k, X_{k+1})$ is invertible and, from (5.2), we may write

$$\eta_{k+1} = (I - Q(X_k, X_{k+1}))^{-1} P(X_k)\eta_k. \tag{5.4}$$

If $X_0 = 0$, since the sequence $\{X_k\}_k$ converges monotonically to $G$, then $\eta_k \geqslant 0$ for any $k$ and $(I - Q(X_k, X_{k+1}))^{-1} P(X_k) \leqslant (I - Q(G,G))^{-1} P(G)$. Therefore, $\eta_k \leqslant W^k \eta_0$, where $W = (I - Q(G,G))^{-1} P(G)$, and $\|\eta_k\| \leqslant \|W^k\| \|\eta_0\|$ for any operator norm $\|\cdot\|$, so that the asymptotic rate of convergence is

$$r^{(0)} = \limsup_k \left( \frac{\|\eta_k\|}{\|\eta_0\|} \right)^{\frac{1}{k}} \leqslant \lim_k \|W^k\|^{\frac{1}{k}} = \rho(W). \tag{5.5}$$

To study the spectral properties of the matrix $W$, we follow an argument similar to the one used in the proof of Lemma 5.1. Since $\mathbf{1}^{\mathrm{T}} G^{\mathrm{T}} = \mathbf{1}^{\mathrm{T}}$, we may find a unitary matrix $S$, having as first row $\frac{1}{\sqrt{m}} \mathbf{1}^{\mathrm{T}}$, such that $T = S G^{\mathrm{T}} S^*$ is a Schur form of $G^{\mathrm{T}}$. With this choice, the diagonal entries of $T$ are $1, \lambda_2, \ldots, \lambda_m$.

Define $\mathscr{P} = S \otimes I$. We may verify that the matrix $\mathscr{P} W \mathscr{P}^*$ is a block upper triangular matrix of the form

$$\mathscr{P} W \mathscr{P}^* = \begin{bmatrix} T_1 & * \\ 0 & T_2 \end{bmatrix}, \tag{5.6}$$

where $T_1 = M^{-1} N$, with $M$ and $N$ defined in (3.6), and $T_2$ has size $(m^2 - m) \times (m^2 - m)$. By following the same arguments used in the proof of Lemma 5.1, we may show that $\rho(T_2) \leqslant \rho(T_1)$, so that $\rho(W) = \rho(M^{-1}N)$. In particular equation (5.5) provides the same bound $r \leqslant \rho(M^{-1}N)$ obtained in Section 3.

If $X_0$ is a stochastic matrix, then $\{X_k\}_k$ is a sequence of stochastic matrices that converges to $G$. Therefore $E_k \mathbf{1} = 0$ and $\eta_k = \mathrm{vec}(E_k)$ belongs to the subspace orthogonal to the vectors of the form $\mathbf{1} \otimes v$ for any $v \in \mathbb{R}^m$, and the vector $s_k = \mathscr{P} \eta_k$ has its first $m$ entries equal to zero, i.e., $s_k^{\mathrm{T}} = [0, \ldots, 0, \hat{s}_k^{\mathrm{T}}]$, where $\hat{s}_k$ has size $m^2 - m$. Since $X_k$ is stochastic, the first column of $S X_k S^*$ is the first column of the identity matrix. Therefore, defining $W_k = (I - Q(X_k, X_{k+1}))^{-1} P(X_k)$, we have

$$\mathscr{P} W_k \mathscr{P}^* = \begin{bmatrix} T_{1,k} & * \\ 0 & T_{2,k} \end{bmatrix}, \tag{5.7}$$

where $T_{1,k}$ is an $m \times m$ matrix and $T_{2,k}$ is $(m^2 - m) \times (m^2 - m)$. From (5.4) and (5.7), we conclude that $\hat{s}_{k+1} = T_{2,k} \hat{s}_k$.

Since the asymptotic rate of convergence is independent of the norm, we may choose the norm $\|x\|' := \|\mathscr{P} x\|_\infty$. Therefore $\|\eta_k\|' = \|\hat{s}_k\|_\infty$ and the asymptotic rate of convergence in the stochastic case is $r^{(\mathrm{sto})} = \limsup_k \|\hat{s}_k\|_\infty^{1/k} = \limsup_k \sigma_k$, $\sigma_k = \|T_{2,k-1} T_{2,k-2} \cdots T_{2,k_0} \hat{s}_{k_0}\|_\infty^{1/k}$, where $k_0$ is such that $\det(I - Q(X_k, X_{k+1})) \neq 0$ for any $k \geqslant k_0$. Since $\sigma_k \leqslant \|T_{2,k-1} T_{2,k-2} \cdots T_{2,k_0}\|_\infty \|\hat{s}_{k_0}\|_\infty$ and $\lim_{k \to \infty} T_{2,k} = T_2$, we have $r^{(\mathrm{sto})} \leqslant \lim_{k \to \infty} \|T_{2,k}\|_\infty^{1/k} = \rho(T_2)$, where the latter implication follows by the same arguments as used in Meini (1997). We may conclude with the following theorem.

THEOREM 5.2 Let $r^{(0)}$ and $r^{(\mathrm{sto})}$ be the asymptotic rates of convergence of the sequences (3.4) with $X_0 = 0$, and with $X_0$ equal to a stochastic matrix, respectively. Then $r^{(0)} \leqslant \rho(W)$. If the drift $\mu$ of (2.1) is negative, then $r^{(\mathrm{sto})} \leqslant \rho(T_2)$, where $T_2$ is the matrix in (5.6). If $W$ is irreducible and aperiodic, then $\rho(T_2) < \rho(W)$, otherwise the weak inequality holds.

## 6. Computational cost and stability analysis

In the analysis of the computational cost, we assume that the power series $A(z)$ is a matrix polynomial of degree $d$. We look at the proposed method as a two-level iterative method, where the *outer iteration* is the iteration defined by (3.4), while the *inner iteration* is the iteration applied to solve the matrix equation of degree $q+1$ at each step $k$ of the outer iteration. At each step of the outer iteration, we have to compute the coefficients of the matrix equation (3.4). In this analysis we restrict the attention to the case where the mass is embedded in the coefficient $A_q$ of the term of degree $q+1$, i.e., the coefficients are defined by (4.7). From (3.4) and (4.7), at each step of the outer iteration, we have to compute $A_q(X_k) = \sum_{i=0}^{d-q-1} A_{q+i} X_k^i$. By using Horner's rule, the cost of the computation of $A_q(X_k)$ is $2m^3(d-q-1)$ arithmetic operations, where we neglect the $O(m^2)$ terms.

The computational cost of the inner iterations depends on the numerical method used to solve the matrix equation of degree $q+1$. By applying the $U$-based functional iteration, we generate the sequence

$$Z_{\nu+1} = \left( I - \sum_{i=0}^{q-1} A_i Z_\nu^i - A_q(X_k) Z_\nu^q \right)^{-1} A_{-1}, \qquad \nu = 0, 1, \ldots. \tag{6.1}$$

If the matrix inversion is performed by computing the $LU$ factorization and by solving the linear systems, the computational cost per step is $2m^3(q+4/3)$ arithmetic operations, where we neglect the $O(m^2)$ terms.

Therefore, by denoting $N_{\text{out}}$ and $N_{\text{in}}$ the number of outer iterations and the overall number of inner iterations, the computational cost is $2m^3(N_{\text{out}}(d-q-1) + N_{\text{in}}(q+4/3))$ arithmetic operations. This estimate should be compared with the cost of the $U$-based iteration, which is $2m^3 N_{U\text{-based}}(d+1/3)$, where $N_{U\text{-based}}$ is the number of iterations.

As shown in Section 4, the number of outer iterations decreases as $q$ increases. Indeed, if $A_i = 0$ for $i \geqslant d$, in the limit case of $q = d-1$, one outer iteration is enough to compute the solution $G$. On the other hand, as pointed in Section 4.3, smaller values of $q$ provide larger values of $\xi_q$. This properties implies that, with smaller values of $q$ the conditioning of the matrix equation (4.8) is better, and the number of inner iterations for (4.8) is lower.

Hence, there is a trade-off between the good convergence properties of $\{X_k\}_k$ and good numerical properties of the matrix equation (4.8). It is difficult to determine the optimal value of $q$. However, this is an asymptotic analysis and, in practice, the number of inner iterations strongly depends on the starting approximation and on the stop condition.

In the numerical experiments presented in Section 7 we have halted the outer iteration if the residual error in the infinity norm, that is $\delta_k = \frac{1}{m}\|X_k - \sum_{i=0}^{d} A_{i-1} X_k^i\|_\infty$, is less than $\varepsilon = 10^{-15}$ or if $\delta_k$ is significantly larger than the error at the previous step, i.e., $\delta_k > \delta_{k-1}(1+10^{-3})$. At the $k$-th outer step, we choose $Z_0 = X_k$ as starting approximation of the inner iteration (6.1) for computing $X_{k+1}$. The inner iteration is stopped if the residual error $\tilde{\delta}_\nu = \frac{1}{m}\|Z_\nu - \sum_{i=0}^{q} A_{i-1} Z_\nu^i - A_q(X_k) Z_\nu^{q+1}\|_\infty$ satisfies the condition $\tilde{\delta}_\nu < \max\left\{\frac{1}{10}\delta_k, 4u, \frac{1}{4}\varepsilon\right\}$, where $u$ is the machine precision, or if it is significantly larger than the error at the previous step, i.e, $\tilde{\delta}_\nu > \tilde{\delta}_{\nu-1}(1+10^{-3})$. As a consequence of this choice, the number of inner iterations does not grow as $q$ grows (see Figure 2). In all the numerical experiments, the optimal value of $q$, in terms of overall CPU time, is generally much smaller than $d$.

## 7. Numerical experiments

In this section we report some numerical experiments which validate the theoretical results obtained in the previous sections and show the improvement of the computational efficiency of the new fixed point iterations with respect to classical iterations. All the algorithms have been implemented in Matlab and tested on a Laptop i3-7100 CPU 3.90GHz×4.

### 7.1 *Test problems*

We have considered two kinds of test problems. The tests of the first kind are generated synthetically in such a way that all the eigenvalues of the matrix $G$ have modulus close to 1. The tests of the second kind are PH/PH/1 queues (see He & Neuts, 1998; Latouche & Ramaswami, 1997). We provide below a description of these two classes of problems.

**Synthetic examples** We have generated an M/G/1-type Markov chain associated with the matrix polynomial $A(z) = \sum_{i=0}^{d} A_{i-1} z^i$ of degree $d$, where the matrix coefficients have size $m \times m$. The matrix coefficients have been constructed in such a way that the drift $\mu$ of the Markov chain (2.1) is close to a given negative value.

Let $C = (c_{i,j})$ be the $m \times m$ circulant matrix such that $c_{i,j} = 1$ if $j - i \equiv 1 \bmod m$, $c_{i,j} = 0$ elsewhere. Let $v_i$, $i = -1, \ldots, d-1$, be nonnegative real numbers such that $\sum_{i=0}^{d} v_{i-1} = 1$. Since $CC^{\mathrm{T}} = I$, if $A_i = v_i C^i$ for $i = -1, 0, \ldots, d-1$, then $G = C^{\mathrm{T}}$ solves the equation $G = \sum_{i=0}^{d} A_{i-1} G^i$. Moreover, since $\mathbf{1}^{\mathrm{T}} C^i = \mathbf{1}^{\mathrm{T}}$, and $\sum_{i=0}^{d} v_{i-1} = 1$ then $\mathbf{1}^{\mathrm{T}} \sum_{i=0}^{d} A_{i-1} = \mathbf{1}^{\mathrm{T}}$ so that the drift (2.1) is $\mu = -v_{-1} + \sum_{i=1}^{d-1} i v_i$.

We use the above properties to generate matrix coefficients in such a way that the drift is close to an assigned negative value. More specifically, given $\mu < 0$, $0 < s_1, s_2 < 1$, and a small positive number $\sigma$, we define $\tilde{A}_i = v_i C^i + \sigma s_2^{m(i+1)} R_i \Delta$, $i = -1, \ldots, d-1$, where $R_i$ is a random $m \times m$ matrix with entries uniformly distributed between 0 and 1, $\Delta$ is the diagonal matrix with diagonal entries $1, s_2, \ldots, s_2^{m-1}$, and $v_{-1} = \frac{1 - s_1^{d-1}}{1 - s_1} - \mu$, $v_0 = 1 - v_{-1} - \sum_{i=1}^{d-1} \frac{s_1^{i-1}}{i}$, $v_i = \frac{s_1^{i-1}}{i}$ for $i = 1, \ldots, d-1$. The basis $s_1$ cannot be chosen too close to 1, otherwise $v_0$ is negative. We may easily check that $\sum_{i=-1}^{d-1} v_i = 1$ and that $-v_{-1} + \sum_{i=1}^{d-1} i v_i = \mu$. Therefore, if $\sigma = 0$, then the drift is exactly $\mu$ and $G = C^{\mathrm{T}}$ is the minimal nonnegative solution of (1.2). If $\sigma > 0$ the matrix $\sum_{i=-1}^{d-1} \tilde{A}_i$ is not stochastic, therefore we define $A_i = D^{-1} \tilde{A}_i$, $i = -1, \ldots, d-1$, where $D$ is the diagonal matrix with diagonal entries equal to the components of the vector $\sum_{i=-1}^{d-1} \tilde{A}_i \mathbf{1}$. If $\sigma > 0$ is a small number, then the drift is close to the given value $\mu$ and the minimal nonnegative solution $G$ is a small perturbation of $C^{\mathrm{T}}$. Since the eigenvalues of $C$ are the $m$-th roots of 1, then all the eigenvalues of $G$, except the eigenvalue equal to 1, have modulus close to 1. This latter property increases the difficulty of the computation of $G$ (see Bini *et al.*, 2005).

In our experiments, we have chosen size $m = 20$, degree $d = 1500$, and drift $\mu \in \{-0.1, -0.05, -0.01, -0.005\}$. The two bases $s_1$ and $s_2$ of the exponential decay of the coefficients have been chosen as $s_1 = 0.6$ and $s_2 = 0.9995$, while the parameter $\sigma$ for the random perturbation has been chosen as $\sigma = 10^{-11}$. We recall that $\mu = 0$ means that the Markov chain is null recurrent, in this case the problem is more difficult from the computational point of view, in fact, the convergence of the fixed point iterations slows down and the problem is more ill-conditioned.

**PH/PH/1 queues** We briefly recall the definition of PH/PH/1 queues. For a detailed description we refer the reader to He & Neuts (1998); Latouche & Ramaswami (1997). Consider two sequences $\{X_h\}_h$ and $\{Y_h\}_h$ of independent continuous random variables with PH($\boldsymbol{\tau}, T$) and PH($\boldsymbol{\beta}, S$) distributions, respectively. Here $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$ are probability vectors of length $n_1$ and $n_2$, respectively, while $T$ and $S$ are subgenerators of size $n_1 \times n_1$ and $n_2 \times n_2$, respectively, i.e., $-T$ and $-S$ are nonsingular M-matrices. In

the queueing applications, $\{X_h\}_h$ represents the intervals between successive arrivals and $\{Y_h\}_h$ represents the service durations. Assume points are marked on a time axis at the epochs $X_1+X_2+\cdots+X_h$ and at the epochs $Y_1+Y_2+\cdots+Y_h$, $h\geqslant 1$.

Let $(A_{h-1})_{ij}$ be the probability that $h$ points of type $X$ occur in an interval of type $Y$, and the phase of the last interval of type $X$ is $j$, given the phase of the first interval is $i$; the first phase of the interval $Y$ has distribution $\boldsymbol{\beta}$. In queueing applications, these would be the probabilities that the queue increases by $h$ units during a service interval. Define $M_1=-(T\otimes I_{n_1}+I_{n_2}\otimes S)^{-1}(\boldsymbol{t}\cdot\boldsymbol{\tau}^{\mathrm{T}}\otimes I_{n_2})$, $M_0=-(T\otimes I_{n_1}+I_{n_2}\otimes S)^{-1}(I_{n_1}\otimes\boldsymbol{s}\cdot\boldsymbol{\beta}^{\mathrm{T}})$, where $\boldsymbol{t}=-T\mathbf{1}$ and $\boldsymbol{s}=-S\mathbf{1}$. We have

$$A_h=(I_{n_1}\otimes\boldsymbol{\beta}^{\mathrm{T}})M_1^{h+1}M_0(I_{n_1}\otimes\mathbf{1})\qquad\text{for all } h\geqslant -1. \tag{7.1}$$

The above matrices are nonnegative and their sum is stochastic if $-\boldsymbol{\beta}^{\mathrm{T}}S^{-1}\mathbf{1}<-\boldsymbol{\tau}^{\mathrm{T}}T^{-1}\mathbf{1}$. In our experiments, we have chosen as PH$(\boldsymbol{\beta},S)$ an Erlang distribution, see Ramaswami & Latouche (1989). We start from an Er$(n_2,\lambda)$ distribution with $n_2=10$ phases and $\lambda=10$, that is, $\boldsymbol{\beta}^{\mathrm{T}}=\begin{bmatrix}1 & 0 & \dots & 0\end{bmatrix}$, $S=(s_{i,j}), s_{i,i}=-\lambda$, $s_{i,i+1}=\lambda$, $s_{i,j}=0$ elsewhere. This matrix is such that $-\boldsymbol{\beta}^{\mathrm{T}}S^{-1}\mathbf{1}=1$.

For the PH$(\boldsymbol{\tau},T)$ distribution, we have taken a pseudo heavy-tailed distribution, which is used in Deelstra *et al.* (2020) and borrowed from Robert & Le Boudec (1997). Define the transition matrix $Q=(q_{i,j})$ such that $q_{1,1}=-(c+s_a)$, $q_{i,1}=q_{i,i}=-(b/a)^{i-1}$ for $i=2,\dots,n_2$, $q_{1,i}=(1/a)^{i-1}$, $i=2,\dots,n_2$, $q_{i,j}=0$ elsewhere, where $s_a=(1/a)+(1/a)^2+(1/a)^3+\cdots+(1/a)^{n_2-1}$. The parameters must satisfy the conditions $a>1$, $a>b>0$, $c>0$. The initial probability vector is $\boldsymbol{\tau}^{\mathrm{T}}=\begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}$ and the matrix $T_0$, defined as $T_0=(-\boldsymbol{\tau}^{\mathrm{T}}Q^{-1}\mathbf{1})Q$, is such that $-\boldsymbol{\tau}^{\mathrm{T}}T_0^{-1}\mathbf{1}=1$. We have chosen the values $a=2$, $b=1$, $c=1.5$ and $n_1=10$. In order to have the expected interval between arrivals equal to $1/\rho$, we take $T=\rho T_0$. In summary, the two distributions are normalized in such a way that $-\boldsymbol{\beta}^{\mathrm{T}}S^{-1}\mathbf{1}<-\boldsymbol{\tau}^{\mathrm{T}}T^{-1}\mathbf{1}=1/\rho$ so that the queue is stable if $\rho<1$. Moreover, we may verify that the drift of the M/G/1-type Markov chain defined by the matrices (7.1) is $\mu=1-\rho$. The value of $\rho$ has been taken to be $\rho=0.85$.

The matrix power series obtained this way has blocks of size $10\times 10$, and has been truncated to a matrix polynomial of degree $d=61$ so that the infinity norm of the remainder is less than $10^{-16}$.

### 7.2 *Numerical results*

We have performed different kinds of tests, where the sequences generated by all the fixed point iterations have been started either with $X_0=0$ or with $X_0=I$. In its wider generality, we generate the sequence (3.4), where at each step $k$ we solve a matrix equation of degree $q+1$. This way we obtain a two-level iterative method, where the *outer iteration* is the iteration defined by (3.4), while the *inner iteration* is the iteration applied to solve the matrix equation of degree $q$ at each step $k$ of the outer iteration.

**The synthetic examples.** For the synthetic case, as starting approximation for the tested fixed point iterations, we have always chosen $X_0=I$. Indeed, since all the eigenvalues of the matrix $G$ are close to one, the performances of the fixed point iterations are not much different if we start with the null matrix or with a stochastic matrix.

The first test aims to compare the convergence speed of the $U$-based iteration and of the three iterations obtained by setting $q=1$ and embedding the tail of the matrix polynomial into the constant, the linear and the quadratic coefficient, respectively.

In Figure 1 we report the semi-logarithmic plot of the residual error in the infinity norm $e_k=\frac{1}{m}\|X_k-\sum_{j=0}^{d}A_{j-1}X_k^j\|_\infty$, where $m$ is the matrix size, for the three iterations together with the residual error of

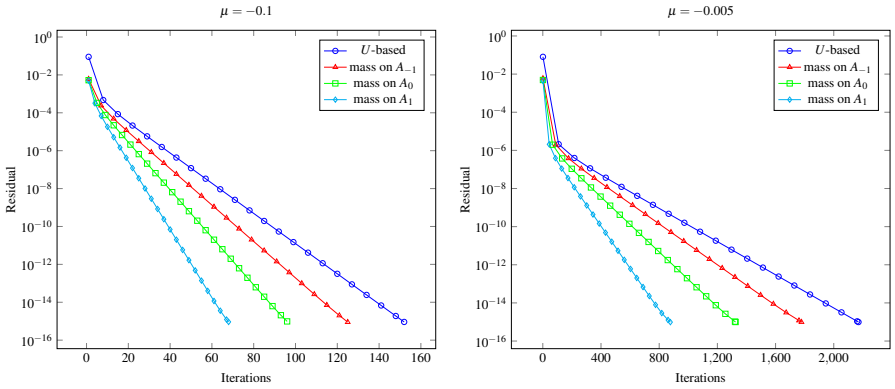


FIG. 1. Residual errors, for two values of the drift $\mu$, of the three iterations obtained by embedding the tail of the series into the constant, linear and quadratic term, respectively, and of the $U$-based iteration.

the $U$-based iteration; to the left the case with drift $\mu = -0.1$, to the right the case where $\mu = -0.005$. As we can see from this plot, the four graphs have different slopes, in accordance with Theorems 3.4 and 3.5. In particular, for $\mu = -0.005$, the number of steps needed to have a residual error less than $\varepsilon = 10^{-15}$ for the $U$-based iteration and for the three iterations relying on the solution of the quadratic equation is 2170, 1778, 1325, 877, respectively. This shows a substantial improvement of our approach in terms of convergence speed.

The second test aims to compare the effective gain that we have in terms of CPU time by using the $U$-based iteration or the iteration obtained by embedding the tail in the constant, linear, and quadratic term. In this case, we may have different possibilities according to the way in which the quadratic equation is solved at each step (inner iteration). We have considered two different implementations for this resolution which differ by the way the quadratic equation is solved. The first implementation simply applies the $U$-based iteration as inner iteration, the second implementation applies Cyclic Reduction (CR) as inner iteration.

In Table 1 we report the speed–up of the CPU time with respect to the $U$-based iteration. For different values of $\mu$, we report in the first three columns the values of the speed-up obtained by embedding the mass in $A_{-1}$, $A_0$ and $A_1$, respectively, where the quadratic equation is solved by means of inner $U$-based iterations. The last column corresponds to the case where the mass is embedded in $A_1$ and the quadratic equation is solved by means of the CR algorithm. We see that the acceleration in terms of CPU time obtained by the combination of our algorithm with CR is by a factor greater than 2. A substantially larger speed–up can be obtained by embedding the mass in the coefficients of higher degree terms. This computational analysis is performed in the next test.

In the third test, we implemented the algorithm where at each step (outer iteration) a matrix equation of degree $q+1$ is solved, and the tail of the matrix polynomial is embedded into the coefficient $A_q$ of the term of degree $q+1$. The matrix equation at each outer iteration is solved by means of the $U$-based algorithm as inner iteration, where the starting approximation is the current approximation of the outer iteration. For any value of $q+1$ in the range $[2,30]$ we computed the CPU time needed to arrive at a residual error less than $10^{-15}$ together with the number of outer iterations and the overall number of inner iterations. The graphs with these values are reported in Figure 2. As we can see from this figure,

| $\mu \setminus$ Alg | $A_{-1}$ | $A_0$ | $A_1$ | $A_1$-CR |
|---|---|---|---|---|
| $-0.1$ | 1.2 | 1.4 | 2.1 | 2.3 |
| $-0.05$ | 1.2 | 1.4 | 2.0 | 2.3 |
| $-0.01$ | 1.2 | 1.5 | 2.2 | 2.7 |
| $-0.005$ | 1.2 | 1.5 | 2.3 | 2.8 |
| $-0.001$ | 1.2 | 1.5 | 2.5 | 2.8 |
| $-0.0005$ | 1.2 | 1.5 | 2.3 | 2.9 |
| $-0.0001$ | 1.1 | 1.5 | 2.2 | 3.0 |

Table 1. Speed–up, in terms of the CPU time, with respect to the $U$-based iteration, of the algorithms obtained by embedding the mass in the coefficient $A_{-1}$, $A_0$, and $A_1$, respectively, where the quadratic equation is solved by means of the $U$-based iteration. In the last column the quadratic equation is solved by means of Cyclic Reduction with the mass embedded in $A_1$.

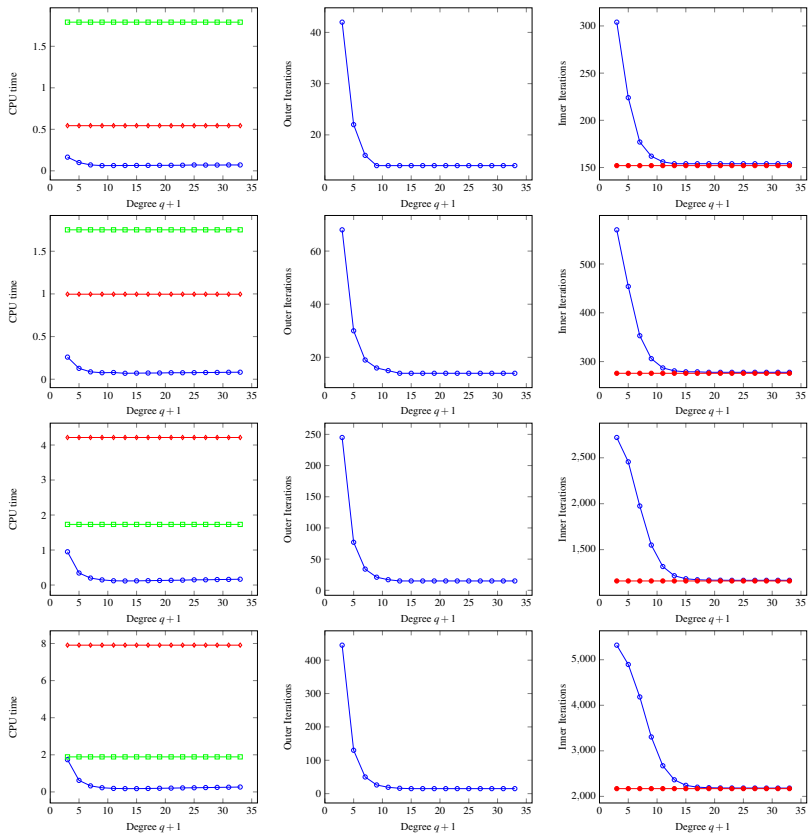


FIG. 2. CPU time, in seconds, and number of inner / outer iterations required to compute $G$ within a residual error less than `1.e-15` by embedding the tail into the coefficient of degree $q+1$. The equation of degree $q+1$ is solved by means of the $U$-based iteration. The two lines denoted by a green square and a red diamond, mark the CPU time needed by PWCR (Bini *et al.*, 2005) and by the $U$-based iteration, respectively. The line denoted by a red star, in the third column, denotes the number of iterations needed by the $U$-based method. From top to bottom the problems with drift $-0.1$, $-0.05$, $-0.01$ and $-0.005$, are considered.

| $\mu$ | -1e-1 | -5e-2 | -1e-2 | -5e-3 | -1e-3 | -5e-4 |
|---|---|---|---|---|---|---|
| CR | 1.7e-14 | 1.7e-14 | 1.7e-14 | 1.7e-14 | 1.8e-14 | 1.8e-14 |
| New | 8.5e-16 | 9.8e-16 | 9.8e-16 | 9.9e-16 | 9.8e-16 | 9.9e-16 |

Table 2. Residual errors for different values of the drift in the approximations provided by CR and by the new iteration with error bound $10^{-15}$.

| $\mu\backslash q$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| -1e-1 | 3.2 | 4.6 | 5.9 | 7.0 | 7.4 | 8.1 | 8.6 | 8.5 | 8.5 |
| -5e-2 | 3.3 | 5.2 | 7.8 | 9.6 | 11.3 | 12.4 | 13.0 | 13.5 | 13.4 |
| -1e-2 | 4.4 | 7.7 | 12.0 | 16.9 | 21.4 | 25.6 | 29.3 | 32.2 | 33.3 |
| -5e-3 | 4.5 | 8.1 | 12.9 | 18.3 | 24.1 | 29.3 | 33.6 | 37.8 | 40.0 |
| -1e-3 | 4.6 | 8.5 | 13.7 | 20.2 | 26.3 | 31.3 | 35.3 | 38.4 | 41.9 |
| -5e-4 | 4.7 | 8.6 | 14.1 | 20.5 | 26.6 | 31.3 | 34.6 | 37.6 | 39.8 |

Table 3. Speed–up of the CPU time of the iteration obtained by embedding the tail in the coefficient of degree $q+1$, with respect to the $U$-based iteration, obtained for some values of the drift $\mu$.

by increasing the value of $q$, we obtain a rapid decrease of the CPU time. After reaching the minimum value, the time slowly increases. It is also interesting to observe that in this case, the optimal value of $q$ is less than 20. This value is rather small with respect to the degree $d$ of the matrix polynomial. Another interesting observation is that the number of outer iterations rapidly decreases and stabilizes for values of $q$ greater than 14. A similar behaviour has the number of inner iterations which tend to stabilize for $q \geqslant 14$ as well. This explains why the graph of the CPU time has an almost linear growth for $q \geqslant 14$. In fact, with the number of inner and outer iterations being almost constant, the time spent for the inner iterations is proportional to $q$ and so is the time spent for the outer iterations. It is also interesting to observe that the overall number of inner iterations, reported in the graphs at the third column of Figure 1, tends to stabilize on the value of the overall number of iterations required by the $U$-based method (line marked by red stars). This explains the higher efficiency of the new iteration with respect to the $U$-based method, since the cost of one step of the $U$-based method is proportional to the degree $d$ of the matrix polynomial while the cost of performing an inner iteration is proportional to the degree $q+1$ of the matrix equation which must be solved at each step. It is important to point out that, while CR is not self-correcting, the methods based on fixed point iterations are self-correcting. In fact, as shown in Table 2, unlike CR, the new iteration allows to obtain approximation to $G$ with a smaller residual error.

In Tables 3 and 4 we report the speed–up factor of the CPU time obtained with the different values of $q > 1$ with respect to the time needed by the $U$-based iteration and by the CR algorithm, respectively. It is interesting to observe that the optimal speed up for each problem ranges from 8.5 to 48.4 if compared to the $U$-based method. This value increases as the drift gets close to 0. The speed-up with respect to CR takes large values only for problems which are far from being null recurrent. This happens since CR, unlike functional iteration, is not much depending on the drift of the stochastic process and has an almost constant CPU time. However, it must be said that CR cannot provide the highest accuracy in the approximation as we have already pointed out. The acceleration in the CPU time can be further increased if we implement the algorithm in a recursive fashion where, instead of the $U$-based iteration to solve the equation of degree $q+1$, we use the same approach by embedding the mass of the matrix polynomial of degree $q+1$ into the leading coefficient of a matrix polynomial of lower degree.

In Figure 3 we report the residual errors per step, for the methods obtained by embedding the tail into the coefficient of degree $q+1$ for a few values of $q$. This graph extends to higher values of $q$ the graph in Figure 1. We may observe that, the larger $q$, the steeper is the slope of the curve. Moreover, the

| $\mu\backslash q$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| -1e-1 | 10.9 | 15.1 | 18.0 | 21.4 | 25.2 | 26.8 | 28.5 | 28.3 | 28.3 |
| -5e-2 | 6.8 | 10.2 | 13.7 | 17.1 | 20.3 | 22.0 | 23.0 | 24.2 | 22.2 |
| -1e-2 | 1.8 | 3.2 | 5.0 | 7.0 | 8.7 | 10.0 | 11.8 | 12.6 | 14.2 |
| -5e-3 | 1.1 | 1.9 | 3.0 | 4.4 | 5.8 | 7.1 | 8.4 | 9.2 | 10.0 |
| -1e-3 | 0.26 | 0.47 | 0.78 | 1.1 | 1.5 | 1.8 | 2.1 | 2.3 | 2.5 |
| -5e-4 | 0.14 | 0.27 | 0.44 | 0.64 | 0.85 | 1.0 | 1.1 | 1.2 | 1.3 |

Table 4. Speed–up of the CPU time of the iteration obtained by embedding the tail in the coefficient of degree $q+1$, with respect to Cyclic Reduction, obtained for some values of the drift $\mu$.

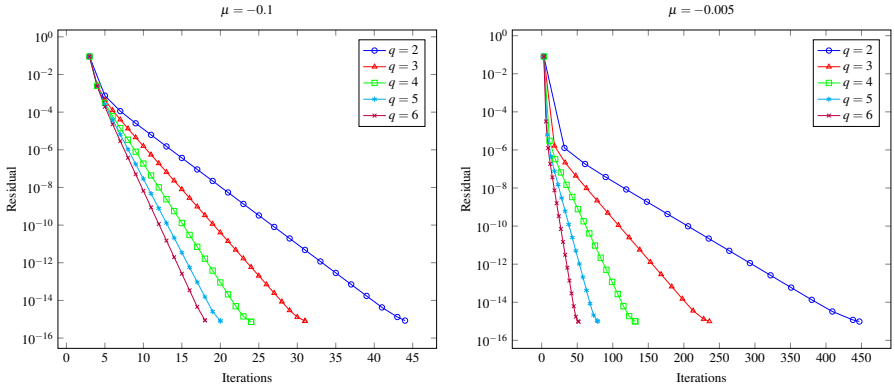


FIG. 3. Residual errors per step, for the methods obtained by embedding the tail into the coefficient of degree $q+1$. At each step, an equation of degree $q+1$ is solved. The test problem is the same as the one of Figure 1.

convergence turns out to be linear with a factor which is smaller for larger $q$.

**PH/PH/1 queues.** In this case, we compare the performances of the fixed point iterations, when the starting approximation is the null matrix or a stochastic matrix.

The $U$-based iteration and the new iterations are applied with $X_0 = 0$ and $X_0 = I$. In the first row of Figure 4 we report the case $X_0 = 0$, in the second row the case $X_0 = I$. We may observe that the same behaviour reported in Figure 2 is maintained. For $X_0 = I$ the speed-up, in terms of the CPU time, of the new iteration with respect to the $U$-based method and to CR is 2.6 and 3.8, respectively. For the new iteration, the speed-up obtained for $X_0 = I$ with respect to $X_0 = 0$ is 2.0.

In Tables 5 and 6, for each value of the degree $q+1$, we have reported the number of outer and inner iterations, respectively, obtained with the two different initial approximations. In the first column, we report the number of iterations required by the $U$-based method. In the last column, in boldface, we report the minimum number of outer and inner iterations, respectively, together with the degree $q+1$ of the associated embedding, where the minimum is taken for $q+1$ in the range $[3, d-1]$, being $d$ the degree of the matrix polynomial. Also in this test, the number of inner iterations gets closer to the number of iterations required by the $U$-based method.

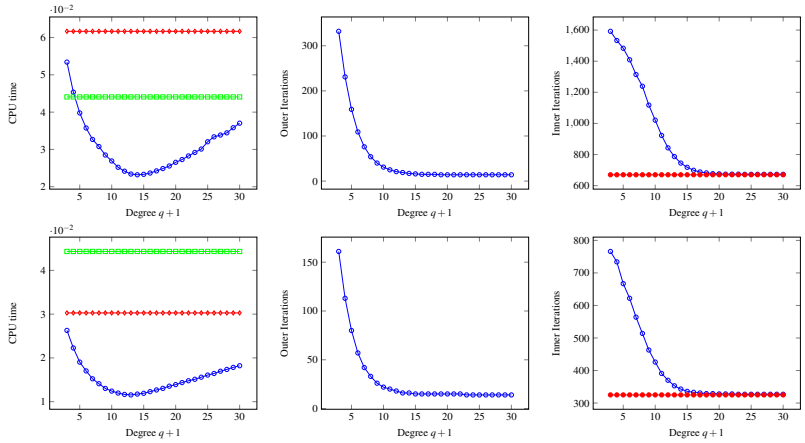


FIG. 4. CPU time, in seconds, number of outer and inner iterations required to compute $G$ within a residual error less than `1.e-15` for the PH/PH/1 problem. The two lines denoted by a green square and a red diamond, mark the CPU time needed by PWCR and by the $U$-based iteration, respectively. The line denoted by a red star, in the third column, denotes the number of iterations needed by the $U$-based method. The graphs in the first row concern the case where $X_0 = 0$, the graphs in the second row the case where $X_0 = I$.

| | | $q+1$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $U$-based | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
| $X_0 = 0$ | 670 | 231 | 159 | 109 | 76 | 54 | 40 | 31 | **14/19** |
| $X_0 = I$ | 325 | 113 | 80 | 57 | 42 | 33 | 26 | 22 | **14/23** |

Table 5. For different choices of $X_0$, number of $U$-based iteration (first column) and number of outer iterations of the method obtained by embedding the mass in the term of degree $q+1$. In the last column, in bold, the minimum number of outer iterations together with the degree $q+1$ of the corresponding embedding.

| | | $q+1$ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $U$-based | 3 | 4 | 5 | 6 | 7 | 8 | |
| $X_0 = 0$ | 670 | 1592 | 1533 | 1483 | 1409 | 1314 | 1239 | **673/25** |
| $X_0 = I$ | 325 | 766 | 734 | 667 | 622 | 564 | 514 | **327/23** |

Table 6. For different choices of $X_0$, number of $U$-based iteration (first column) and number of inner iterations of the method obtained by embedding the mass in the term of degree $q+1$. In the last column, in bold, the minimum number of inner iterations together with the degree $q+1$ of the corresponding embedding.

**Acknowledgements**

The authors thank the anonymous referees for providing very useful comments to improve the presentation of this paper.

**References**

BERMAN, A. & PLEMMONS, R. J. (1994) *Nonnegative matrices in the mathematical sciences*. Classics in Applied Mathematics, vol. 9. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, pp. xx+340. Revised reprint of the 1979 original.

BINI, D. A., LATOUCHE, G. & MEINI, B. (2005) *Numerical Methods for Structured Markov Chains*. Numerical Mathematics and Scientific Computation. New York: Oxford University Press, pp. xii+327. Oxford Science Publications.

BINI, D. A. & MEINI, B. (2009) The cyclic reduction algorithm: from Poisson equation to stochastic processes and beyond. *Numer. Algorithms*, **51**, 23–60.

CHEN, C., LI, R.-C. & MA, C. (2019) Highly accurate doubling algorithm for quadratic matrix equation from quasi-birth-and-death process. *Linear Algebra Appl.*, **583**, 1–45.

CHIANG, C.-Y., CHU, E. K.-W., GUO, C.-H., HUANG, T.-M., LIN, W.-W. & XU, S.-F. (2009) Convergence analysis of the doubling algorithm for several nonlinear matrix equations in the critical case. *SIAM J. Matrix Anal. Appl.*, **31**, 227–247.

DEELSTRA, G., LATOUCHE, G. & SIMON, M. (2020) On barrier option pricing by Erlangization in a regime-switching model with jumps. *J. Comput. Appl. Math.*, **371**, 112606, 16.

GOHBERG, I., LANCASTER, P. & RODMAN, L. (2009) *Matrix Polynomials*, vol. 58. SIAM.

GUO, C.-H. (1999) On the numerical solution of a nonlinear matrix equation in Markov chains. *Linear Algebra Appl.*, **288**, 175–186.

GUO, C.-H. (2003) On a quadratic matrix equation associated with an $M$-matrix. *IMA J. Numer. Anal.*, **23**, 11–27.

HE, Q.-M. & NEUTS, M. F. (1998) Markov chains with marked transitions. *Stochastic Process. Appl.*, **74**, 37–52.

HIGHAM, N. J. & KIM, H.-M. (2000) Numerical analysis of a quadratic matrix equation. *IMA J. Numer. Anal.*, **20**, 499–519.

HIGHAM, N. J. & KIM, H.-M. (2001) Solving a quadratric matrix equation by Newton's method with exact line searches. *SIAM J. Matrix Anal. Appl.*, **23**, 303–316.

LATOUCHE, G. (1994) Newton's iteration for non-linear equations in Markov chains. *IMA J. Numer. Anal.*, **14**, 583–598.

LATOUCHE, G. & RAMASWAMI, V. (1993) A logarithmic reduction algorithm for quasi-birth-death processes. *J. Appl. Probab.*, **30**, 650–674.

LATOUCHE, G. & RAMASWAMI, V. (1997) The $PH/PH/1$ queue at epochs of queue size change. *Queueing Systems Theory Appl.*, **25**, 97–114.

LATOUCHE, G. & RAMASWAMI, V. (1999) *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Philadelphia PA: SIAM.

MEINI, B. (1997) New convergence results on functional iteration techniques for the numerical solution of $M/G/1$ type Markov chains. *Numer. Math.*, **78**, 39–58.

MENG, J., SEO, S.-H. & KIM, H.-M. (2018) Condition numbers and backward error of a matrix polynomial equation arising in stochastic models. *J. Sci. Comput.*, **76**, 759–776.

NEUTS, M. F. (1981) *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Baltimore, MD: The Johns Hopkins University Press.

PÉREZ, J. F., TELEK, M. & VAN HOUDT, B. (2012) A fast Newton's iteration for $M/G/1$-type and $GI/M/1$-type Markov chains. *Stoch. Models*, **28**, 557–583.

RAMASWAMI, V. (1988) Nonlinear matrix equations in applied probability—solution techniques and open problems. *SIAM Rev.*, **30**, 256–263.

RAMASWAMI, V. & LATOUCHE, G. (1989) An experimental evaluation of the matrix-geometric method for the $GI/PH/1$ queue. *Comm. Statist. Stochastic Models*, **5**, 629–667.

RHEE, N. H. (2010) Note on functional iteration technique for $M/G/1$ type Markov chains. *Linear Algebra Appl.*, **432**, 1042–1048.

ROBERT, S. & LE BOUDEC, J.-Y. (1997) New models for pseudo self-similar traffic. *Performance Evaluation*, **30**, 56–68.

SEO, J.-H. & KIM, H.-M. (2014) Convergence of pure and relaxed Newton methods for solving a matrix polynomial equation arising in stochastic models. *Linear Algebra Appl.*, **440**, 34–49.

SEO, S.-H., SEO, J.-H. & KIM, H.-M. (2018) A modified Newton method for a matrix polynomial equation arising in stochastic problem. *Electron. J. Linear Algebra*, **34**, 500–513.

VARGA, R. S. (2000) *Matrix Iterative Analysis*. Springer Series in Computational Mathematics, vol. 27, expanded edn. Springer-Verlag, Berlin, pp. x+358.