# SOLVING QUADRATIC MATRIX EQUATIONS ARISING IN RANDOM WALKS IN THE QUARTER PLANE*

DARIO A. BINI†, BEATRICE MEINI‡, AND JIE MENG§

**Abstract.** Quadratic matrix equations of the kind $A_1X^2 + A_0X + A_{-1} = X$ are encountered in the analysis of Quasi–Birth-Death stochastic processes where the solution of interest is the minimal nonnegative solution $G$. In many queueing models, described by random walks in the quarter plane, the coefficients $A_1, A_0, A_{-1}$ are infinite tridiagonal matrices with an almost Toeplitz structure. Here, we analyze some fixed point iterations, including Newton's iteration, for the computation of $G$ and introduce effective algorithms and acceleration strategies which fully exploit the Toeplitz structure of the matrix coefficients and of the current approximation. Moreover, we provide a structured perturbation analysis for the solution $G$. The results of some numerical experiments which demonstrate the effectiveness of our approach are reported.

**Keywords:** Matrix equations, random walks, Markov chains, Toeplitz matrices, infinite matrices, fixed point iteration, Newton iteration.

**MSC:** 65F30, 15A24, 60J22, 15B05

**1. Introduction.** Random walks in the quarter plane describe a wide variety of two-queue models with various service policies such as nonpreemptive priority, $K$-limited service, server vacation and server setup [26]. Models of this kind concern, for instance, bi-lingual call centers [30], generalized two-node Jackson networks [27], two-demand models [10], two-stage inventory queues [11], and more.

A theoretical analysis of stability, of tail decay rates and of other asymptotic properties has been carried out by several authors, in particular in [14], [22], [23], [26], and in the book [9], in which the invariant measure and the transient behavior are investigated by means of analytic and functional tools.

A different approach is based on representing a random walk in the quarter plane as a 2-dimensional Quasi–Birth-Death (QBD) stochastic process. This latter framework, based on the matrix analytic approach of [25], allows expressing the invariant probability measure, and other quantities of interest for the stochastic model, in terms of a solution of suitable quadratic matrix equations. This provides a further tool for the theoretical analysis [17], [19] and paves the way for the design of effective algorithms based on the numerical solution of quadratic matrix equations.

In fact, relying on the matrix analytic theory of [25], the problem of computing the invariant probability measure of a QBD process is reduced to computing the minimal nonnegative solutions $G$ and $R$ of the two matrix equations

$$A_1X^2 + A_0X + A_{-1} = X, \tag{1.1}$$

$$A_1 + XA_0 + X^2A_{-1} = X, \tag{1.2}$$

respectively, where the coefficients $A_{-1}, A_0, A_1$ are nonnegative matrices such that $A_{-1} + A_0 + A_1$ is row-stochastic and $X$ is the unknown. We say that a matrix $X$ is nonnegative, and we write $X \geq 0$, if its entries are nonnegative. Moreover we say that a solution $X$ of a matrix equation is minimal nonnegative if $X \geq 0$ and for any other nonnegative solution $Y$ it holds $Y - X \geq 0$. For more details in this regard, we refer the reader to the books [2], [19], and [25].

*Research partially supported by INdAM-GNCS

†Dipartimento di Matematica, Università di Pisa, Italy, (dario.bini@unipi.it)

‡Dipartimento di Matematica, Università di Pisa, Italy, (beatrice.meini@unipi.it)

§Finance·Fishery·Manufacture Industrial Mathematics Center on Big Data, Pusan National University, Busan, South Korea (mengjie@pusan.ac.kr)

In the case where the coefficients are finite dimensional, several algorithms have been introduced to compute $G$ and $R$. They include fixed point iterations and doubling algorithms like Logarithmic Reduction and Cyclic Reduction (CR) [7], [2], [19].

In the case of 2-dimensional QBDs the coefficients $A_{-1}, A_0, A_1$ are semi-infinite and have a special structure, more precisely,

$$A_i = \begin{bmatrix} b_{i,0} & b_{i,1} & & & \\ a_{i,-1} & a_{i,0} & a_{i,1} & & \\ & a_{i,-1} & a_{i,0} & a_{i,1} & \\ & & \ddots & \ddots & \ddots \end{bmatrix}, \quad i = -1, 0, 1, \tag{1.3}$$

where $a_{i,j} \geq 0$, $b_{i,j} \geq 0$ and $\sum_{i,j=-1}^{1} a_{i,j} = 1$, $\sum_{i=-1}^{1} \sum_{j=0}^{1} b_{i,j} = 1$. These blocks belong to the class of matrices representable in the form $A = T(s)+E$ where $T(s)$ is the Toeplitz matrix associated with the symbol $s(z) = \sum_{i\in\mathbb{Z}} s_i z^i$, that is, $(T(s))_{i,j} = s_{j-i}$, and $E = (e_{i,j})$ is such that $v_i = \sum_j |e_{i,j}|$ is finite and $\lim_i v_i = 0$. The matrix $T(s)$ is called *Toeplitz part* while $E$ is called *correction*. Here $s(z)$ is a function belonging to the Wiener class $\mathcal{W} = \{f(z) = \sum_{i\in\mathbb{Z}} f_i z^i,\ \|f\|_w := \sum_{i\in\mathbb{Z}} |f_i| < \infty\}$. In particular, for the matrix in (1.3) it is easy to check that $A_i = T(a_i)+E_i$ where $a_i(z) = \sum_{j=-1}^{1} a_{i,j} z^j$ and $E_i$ has zero entries except for the first row which is equal to $[b_{i,0}-a_{i,0}, b_{i,1}-a_{i,1}, 0, \ldots]$. Matrices of this kind are called *Quasi-Toeplitz (QT)* in [3].

The case of QBD with infinite blocks has been initially investigated in [17], [18], and [20], by reducing the problem to finite size relying on truncation and augmentation of the blocks. However, this approach does not lead to reliable computational techniques since the result of the numerical computation is strongly dependent on the way the infinite matrices have been truncated. In fact, the models obtained by truncating the infinite dimensional problem may have asymptotic properties, like the decay rate, which are not consistent with the original problem [15], [29].

More recently, conditions under which the solution $G$ of (1.1) can be represented as the sum $G = T(g) + E_g$ are given in [5], so that, despite the solution $G$ has infinitely many entries, it can be represented up to any arbitrary approximation error by using a finite number of parameters. Moreover, in [3] and [4], by using the structure properties of QT matrices, the algorithm of Cyclic Reduction has been extended to the case of infinite matrices. This algorithm still keeps a fast convergence speed in terms of number of iterative steps. However, in certain cases the cost of each step becomes extremely large due to the cost of certain operations with QT matrices, like matrix inversion and the compression of the correction part. Another drawback of CR is that this iteration is not self-correcting.

In this paper, we restrict the attention to the computation of the matrix $G$ since the matrix $R$ can be explicitly expressed in function of $G$ [19, Theorem 6.2.9]. We propose and analyze some fixed point iterations which have a low cost per step and, unlike CR, are self-correcting and allow keeping separated the computation of the Toeplitz part $T(g)$ and the correction part $E_g$.

In fact, we show that the symbol $g(z)$ defining the Toeplitz part satisfies the functional equation

$$a_1(z)g(z)^2 + a_0(z)g(z) + a_{-1}(z) = g(z), \quad |z| = 1.$$

We use this property to design an algorithm based on evaluation and interpolation at the roots of 1 for approximating the coefficients of $g(z)$, which is extremely fast and allows for an automatic control of the number of interpolation points, according to the desired approximation error.

The correction part $E_g$ is obtained by simply applying fixed point iterations. We consider three iterations of the kind $X_{k+1} = F(X_k)$, $k = 0, 1, \ldots$, defined by suitable functions $F_1$, $F_2$, $F_3$, where $F_1$ requires no matrix inversion, $F_2$ requires to compute an inverse matrix once for all, while $F_3$ requires one inversion per step. These iterations are well known in the case of QBD with finitely many phases [2], [19], [21], and are here extended to coefficients with the QT structure (1.3).

We show that, under mild assumptions, starting with $X_0 = 0$, the sequences generated by $F_1, F_2, F_3$ converge monotonically and linearly to $G$ in the infinity norm. Moreover, we prove that the rate of convergence of the sequence generated by $F_3$ is better than the rate of the sequence generated by $F_2$, which in turn is better than that generated by $F_1$. We prove that if $X_0$ is row-stochastic then all the matrices $X_k$ are row-stochastic and the rate of convergence of each of the three iterations is better than that obtained with $X_0 = 0$. Numerical experiments show also the evidence that for $X_0 = T(g)$ the rate of convergence of the three sequences is even better.

Then we adapt Newton's iteration, in the form given by [16], to the case of QT coefficients. In order to solve the Sylvester equation arising at each step of Newton's iteration we rely on the solver introduced in [28]. Under mild conditions, we prove that, for $X_0 = 0$, convergence holds in the infinity norm, is monotonic and quadratic.

In order to evaluate an *a posteriori* bound on the approximation error in the computation of $G$, we also perform the analysis of the structured condition number. More specifically, we provide perturbation results related to perturbations of the Toeplitz part and of the correction part in the matrix coefficients $A_i$, $i = -1, 0, 1$, and we estimate the consequent variation of the solution $G$ and of its Toeplitz part $T(g)$.

Numerical experiments are reported which show the effectiveness of our approach and the reliability of our algorithms with respect to the algorithm CR. In particular we show that in certain cases the combination of Newton's iteration and cyclic reduction provides a substantial acceleration of the convergence.

The paper is organized as follows: in Section 2 we recall some preliminary properties and concepts useful for the analysis of the problem; in Section 3 we present an algorithm for computing the Toeplitz part $T(g)$ of the solution; Section 4 deals with the analysis of three fixed-point iterations applied to infinite QT matrices, while Section 5 concerns the algorithmic analysis of Newton iteration; in Section 6 we carry out the analysis of the conditioning by providing some perturbation results, while in Section 7 we present and discuss some numerical experiments which show the effectiveness of our approach.

**2. Preliminaries.** Let $\ell^\infty$ be the set of sequences $x = (x_i)_{i\in\mathbb{N}}$ such that $\|x\|_\infty := \sup_i |x_i|$ is finite. Consider the set of matrices $A = (a_{i,j})$ such that the application $x \to y = Ax$, where $y_j = \sum_{j=1}^\infty a_{i,j}x_j$, defines a linear operator from $\ell^\infty$ to $\ell^\infty$. Denote this set by $\mathcal{L}_\infty$ and define the induced norm $\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty$. It can be verified that $\|A\|_\infty = \sup_i \sum_{j=1}^\infty |a_{i,j}|$. Recall that $\mathcal{L}_\infty$ is a Banach algebra, that is, it is closed under the row-by-column product, the norm satisfies $\|AB\|_\infty \le \|A\|_\infty \cdot \|B\|_\infty$ for any $A, B \in \mathcal{L}_\infty$, and the normed space is complete.

We introduce the following notation

$$\begin{aligned} a_i(z) &= a_{i,-1}z^{-1} + a_{i,0} + a_{i,1}z \\ b_i(z) &= b_{i,0} + b_{i,1}z, \end{aligned} \tag{2.1}$$

so that we may write $A_i = T(a_i) + E_i$, for $i = -1, 0, 1$, where $E_i$ has null entries except for those in the first row which are equal to $[b_{i,0} - a_{i,0}, b_{i,1} - a_{i,1}, 0, \ldots]$. We assume that the entries of $A_i$ are nonnegative and $(A_{-1} + A_0 + A_1)\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$

is the vector of all ones of appropriate dimension. It is known [19], [31] that under these conditions, there exist the minimal nonnegative solutions $R$ and $G$ of (1.1) and (1.2), respectively, and the Laurent matrix polynomial $\varphi(z) = z^{-1}A_{-1} + A_0 - I + zA_1$ admits the factorization

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

where

$$\begin{aligned} &A_1 = RW, \quad A_{-1} = WG, \quad A_0 = I - W - RWG, \\ &W = I - A_0 - A_1G = I - A_0 - RA_{-1}, \\ &G\mathbf{1} \leq \mathbf{1}. \end{aligned} \tag{2.2}$$

From the above equations it follows that, once $G$ has been computed, the matrix $R$ can be obtained explicitly by means of $R = A_1(I - A_0 - A_1G)^{-1}$. Observe that if $a_{-1}(1) = 0$, i.e., $A_{-1} = e_1w^T$, $w^T = (b_{-1,0}, b_{-1,1}, 0, \ldots) \neq 0$, then the minimal nonnegative solution $G$ of equation (1.1) can be expressed in the form $G = \mathbf{1}v^T$, where $v = \frac{1}{b_{-1}(1)}w$, [19, Theorem 8.5.1]. Therefore, without loss of generality we may assume that $a_{-1}(1) > 0$.

The following result is valid if $a_{-1}(1) > 0$ and $b_{-1}(1) > 0$, that is, $A_{-1}\mathbf{1} > 0$.

LEMMA 2.1. *Assume $A_{-1}\mathbf{1} > 0$ and define*

$$\theta = \min\{a_{-1}(1), b_{-1}(1)\}, \quad \gamma = \max\left\{\frac{a_1(1)}{a_{-1}(1)}, \frac{b_1(1)}{b_{-1}(1)}\right\}. \tag{2.3}$$

*Then the matrix $W = I - A_0 - A_1G$ is invertible in $\mathcal{L}_\infty$, has nonnegative inverse, and $\|W^{-1}\|_\infty \leq \frac{1}{1-\|(A_0+A_1)\mathbf{1}\|_\infty} = \frac{1}{\theta}$. Moreover, $\|W^{-1}RW\|_\infty \leq \gamma$. If $A_{-1}\mathbf{1} > A_1\mathbf{1}$ then $\gamma < 1$.*

*Proof.* Observe that $\|A_0 + A_1G\|_\infty = \|(A_0 + A_1G)\mathbf{1}\|_\infty \leq \|(A_0 + A_1)\mathbf{1}\|_\infty = \|(I - A_{-1})\mathbf{1}\|_\infty = 1 - \theta < 1$, since $(A_0 + A_1)\mathbf{1} = (I - A_{-1})\mathbf{1}$, $A_{-1}\mathbf{1} > 0$ and $A_{-1}\mathbf{1} = (b_{-1}(1), a_{-1}(1), a_{-1}(1), \ldots)^T$. This implies that $W^{-1} \in \mathcal{L}_\infty$, since in general, if $I - B \in \mathcal{L}_\infty$ is such that $\|B\|_\infty < 1$ then the series $\sum_{i=0}^{\infty} B^i$ has norm bounded by $1/(1 - \|B\|_\infty)$ and coincides with $(I - B)^{-1}$. Moreover, $W^{-1}$ is nonnegative, being $W^{-1} = \sum_{i=0}^{\infty}(A_0 + A_1G)^i$ and

$$\|W^{-1}\|_\infty = \|\sum_{k=0}^{\infty}(A_0 + A_1G)^k\mathbf{1}\|_\infty \leq \frac{1}{1 - \|A_0 + A_1\|_\infty} = \frac{1}{\theta}.$$

Now we show that $\|W^{-1}RW\|_\infty \leq \gamma$. Since $W^{-1}RW = W^{-1}A_1$ by (2.2), it is sufficient to consider $\|W^{-1}A_1\|_\infty = \|W^{-1}A_1\mathbf{1}\|_\infty$. By definition of $\gamma$ we have $A_1\mathbf{1} \leq \gamma A_{-1}\mathbf{1}$, therefore

$$W^{-1}A_1\mathbf{1} \leq \gamma W^{-1}A_{-1}\mathbf{1} = \gamma G\mathbf{1} \leq \gamma\mathbf{1},$$

where we used the fact that $W^{-1}A_{-1} = G$. Thus we have $\|W^{-1}RW\|_\infty \leq \gamma$. Since $A_1\mathbf{1} = (b_1(1), a_1(1), a_1(1), \ldots)^T$ and $A_{-1}\mathbf{1} = (b_{-1}(1), a_{-1}(1), a_{-1}(1), \ldots)^T$ then $A_{-1}\mathbf{1} > A_1\mathbf{1}$ implies $\gamma < 1$. ∎

Define $\mathcal{W} = \{f(z) = \sum_{i\in\mathbb{Z}} f_i z^i : \quad \|f\|_w := \sum_{i\in\mathbb{Z}} |f_i| < \infty\}$. Consider the following class

$$\mathcal{QT} := \{A = T(f) + E\}$$

where $f(z) \in \mathcal{W}$, the matrix $E = (e_{i,j}) \in \mathcal{L}_\infty$ is such that $\lim_i v_i = 0$, where $v_i = \sum_{j=1}^{\infty} |e_{i,j}|$.

Observe that $A_i \in \mathcal{QT}$ for $i = -1, 0, 1$, moreover, in [5] it is shown that $\mathcal{QT}$ is an algebra with the infinity norm, and the matrices $W, G$ and $R$ in (2.2) belong to $\mathcal{QT}$ if $A_{-1}\mathbf{1} > A_1\mathbf{1}$ or if $A_{-1}\mathbf{1} \geq A_1\mathbf{1} > 0$. More precisely we have the following result [5, Theorem 9].

THEOREM 2.2. *The minimal nonnegative solution $G$ of the matrix equation (1.1) can be written as $G = T(g) + E_g$ where $E_g \in \mathcal{L}_\infty$, $\|E_g\|_\infty \leq 1 + g(1)$, $|E_g|\mathbf{1} \leq 2 \cdot \mathbf{1}$, and $g(z) = \sum_{i\in\mathbb{Z}} g_i z^i \in \mathcal{W}$ is such that $g_i \geq 0$ and $\|g\|_w = g(1) \leq 1$. Moreover, for any $z$ such that $|z| = 1$, $g(z)$ is a solution of minimum modulus of the quadratic equation*

$$a_{-1}(z) + a_0(z)\lambda + a_1(z)\lambda^2 = \lambda. \tag{2.4}$$

*This solution is unique if there exists $j$ such that $a_{i,j} \neq 0$ for at least two different values of $i$. If*

$$A_{-1}\mathbf{1} > A_1\mathbf{1}, \quad \text{or} \quad A_{-1}\mathbf{1} \geq A_1\mathbf{1} > 0,$$

*then $G \in \mathcal{QT}$, $G\mathbf{1} = \mathbf{1}$ and $g(1) = 1$. Conversely, if $G\mathbf{1} = \mathbf{1}$ and $G \in \mathcal{QT}$ then $a_{-1}(1) \geq a_1(1)$ and $g(1) = 1$.*

Observe that the condition

$$A_{-1}\mathbf{1} > A_1\mathbf{1} \tag{2.5}$$

is equivalent to $a_{-1}(1) > a_1(1)$ and $b_{-1}(1) > b_1(1)$. In the following we assume that (2.5) holds and that there exists $i$ such that $a_{i,j} \neq 0$ for at least two values of $j$. The latter condition is very mild.

**3. Computing the symbol $g(z)$.** Relying on Theorem 2.2, we provide an algorithm, based on the evaluation/interpolation at the roots of 1, for computing an approximation $\hat{g}_i$, $i = -n+1, \ldots, n$, to the coefficients $g_i$ of $g(z)$, where $n$ is such that $|\hat{g}_i - g_i| \leq \epsilon/(2n)$, for any $i$ and for a given tolerance $\epsilon > 0$. In this analysis we may relax the assumption $a_{-1}(1) > a_1(1)$ so that the result holds in general.

Let $n > 0$ be an integer and set $m = 2n$. Define $\omega_m = \cos\frac{2\pi}{m} + \mathbf{i}\sin\frac{2\pi}{m}$ a principal $m$th root of 1, where $\mathbf{i}$ is the imaginary unit such that $\mathbf{i}^2 = -1$. Rewrite $g(z)$ as

$$g(z) = \sum_{j=-n+1}^{n} g_j z^j + \sum_{j=-n+1}^{n} \sum_{k\geq 1} (z^{mk+j} g_{mk+j} + z^{-mk-j+1} g_{-mk-j+1}). \tag{3.1}$$

Since $\omega_m^{km} = 1$, from (3.1) we have

$$g(\omega_m^i) = \sum_{j=-n+1}^{n} g_j \omega_m^{ij} + \sum_{j=-n+1}^{n} (\omega_m^{ij} \sum_{k\geq 1} g_{mk+j} + \omega_m^{-i(j-1)} \sum_{k\geq 1} g_{-mk-j+1}).$$

Therefore, the Laurent polynomial defined by

$$\begin{aligned} \widehat{g}(z) &= \sum_{j=-n+1}^{n} g_j z^j + \sum_{j=-n+1}^{n} (z^j \hat{g}_j^+ + z^{-j+1} \hat{g}_{-j+1}^-), \\ \hat{g}_j^+ &= \sum_{k\geq 1} g_{mk+j}, \quad \hat{g}_{-j+1}^- = \sum_{k\geq 1} g_{-mk-j+1}, \quad j = -n+1, \ldots, n, \end{aligned} \tag{3.2}$$

is such that $g(\omega_m^i) = \hat{g}(\omega_m^i)$, that is, it interpolates $g(z)$ at the $m$-th roots of 1.

The following lemma provides a bound to the tail of the Laurent series $g(z)$ and extends to the case of Laurent series a similar property proved in [8] valid for power series.

LEMMA 3.1. *Let $g(z)$ be the solution of minimum modulus of equation (2.4). Let $\hat{g}(z) = \sum_{j=-n+1}^{n} \hat{g}_j z^j$ be the Laurent polynomial interpolating $g(z)$ at the $m$-th roots of 1, i.e., such that $g(\omega_m^i) = \hat{g}(\omega_m^i)$, $i = -n+1, \ldots, n$, where $m = 2n$. If $g''(x) \in \mathcal{W}$, then $g''(1) \geq 0$ and*

$$g''(1) - \hat{g}''(1) \geq 2n \left( \sum_{j<-n+1} g_j + \sum_{j>n} g_j \right), \tag{3.3}$$

*moreover $0 \leq \hat{g}_j - g_j \leq \frac{1}{2n}(g''(1) - \hat{g}''(1))$, for $j = -n+1, \ldots, n$.*

*Proof.* Since the coefficients $g_i$ are nonnegative then also $g''(z)$ has nonnegative coefficients, moreover, since $g''(z) \in \mathcal{W}$, then the series $g''(1)$ is absolutely convergent and $g''(1) \geq 0$. Thus, from the representation (3.1), in view of (3.2), we deduce that

$$g''(1) - \hat{g}''(1) = \sum_{j=-n+1}^{n} \sum_{k \geq 1} \left( g_{mk+j} \alpha_{j,k} + g_{-mk-j+1} \alpha_{j,k} \right),$$

where $\alpha_{j,k} = (mk+j)(mk+j-1) - j(j-1)$. The inequality (3.3) follows from the nonnegativity of the coefficients and from the property $\alpha_{j,k} \geq m$, valid for $k \geq 1$, $j = -n+1, \ldots, n$ which can be verified by a direct inspection. The bound on $\hat{g}_j - g_j$ follows from (3.3) since $\hat{g}_j = g_j + \hat{g}_j^+ + \hat{g}_j^-$ and $\hat{g}_j^+ + \hat{g}_j^- \leq \sum_{j<-n+1} g_j + \sum_{j>n} g_j$ in view of (3.2). ∎

Observe that $\hat{g}''(1)$ is computable once the coefficients of the polynomial $\hat{g}(z)$ have been computed. Moreover, the value of $g''(1)$ is computable even though $g(z)$ is not known. In fact, by taking the second derivative in the equation obtained by replacing $\lambda$ with $g(z)$ in (2.4), i.e.,

$$a_1(z)g(z)^2 + (a_0(z) - 1)g(z) + a_{-1}(z) = 0,$$

for $z = 1$, the value of $g''(1)$ can be easily expressed in terms of $a_i(1)$, $a_i'(1)$ and $a_i''(1)$. More precisely, by taking the first derivative we obtain

$$a_1'(z)g(z)^2 + 2g(z)g'(z)a_1(z) + (a_0(z) - 1)g'(z) + a_0'(z)g(z) + a_{-1}'(z) = 0,$$

which yields

$$g'(1) = \frac{a_1'(1)g(1)^2 + a_0'(1)g(1) + a_{-1}'(1)}{1 - 2a_1(1)g(1) - a_0(1)}, \quad g(1) = \min(1, a_{-1}(1)/a_1(1)). \tag{3.4}$$

By taking the second derivative for $z = 1$, we get

$$\begin{aligned} a_1''(1)g(1)^2 &+ 2a_1'(1)g'(1)g(1) + 2a_1'(1)g'(1)g(1) + 2a_1(1)g'(1)^2 \\ &+ 2a_1(1)g''(1)g(1) + (a_0(1) - 1)g''(1) + a_0'(1)g'(1) \\ &+ a_0''(1)g(1) + a_0'(1)g'(1) + a_{-1}''(1) = 0 \end{aligned}$$

which yields

$$\begin{aligned} g''(1) = \big[ & a_{-1}''(1) + a_0''(1)g(1) + a_1''(1)g(1)^2 + 2a_1(1)g'(1)^2 \\ & + 2g'(1)(2g(1)a_1'(1) + a_0'(1)) \big] / (1 - 2a_1(1)g(1) - a_0(1)). \end{aligned} \tag{3.5}$$

Lemma 3.1 provides an *a posteriori* bound to the error in the approximation of the Laurent series $g(z)$ together with a stop condition for the following evaluation interpolation algorithm for computing the coefficients of $g(z)$.

**Algorithm 3.1** Approximation of $g(z)$

**Require:** The coefficients of $a_i(z)$, $i = -1, 0, 1$ and a tolerance $\epsilon > 0$.

**Ensure:** Approximations $\hat{g}_i$, $i = -n+1, \ldots, n$, to the coefficients $g_i$ of $g(z)$ such that $\hat{g}_i - g_i \leq \epsilon/(2n)$.

1: Set $n = 4$, and compute $g(1) = \min(1, a_{-1}(1)/a_1(1))$, $g'(1)$ and $g''(1)$ by means of (3.4) and (3.5);
2: Set $m = 2n$, $\omega_m = \cos\frac{2\pi}{m} + \mathbf{i}\sin\frac{2\pi}{m}$, and evaluate $a_{-1}(z),\ a_0(z),\ a_1(z)$ at $z = \omega_m^i$, $i = -n+1, \ldots, n$;
3: For $i = -n+1, \ldots, n$, compute the solution $\lambda_i$ of minimum modulus of the quadratic equation (2.4), where $z = \omega_m^i$;
4: Interpolate the values $\lambda_i$, $i = -n+1, \ldots, n$ by means of FFT and obtain the coefficients $\hat{g}_i$ of the Laurent polynomial $\hat{g}(z) = \sum_{i=-n+1}^{n} \hat{g}_i z^i$ such that $g(\omega_m^i) = \hat{g}(\omega_m^i)$, $i = -n+1, \ldots, n$;
5: Compute $\delta_m = g''(1) - \hat{g}''(1)$, where $\hat{g}''(1) = \sum_{i=-n+1}^{n} i(i-1)\hat{g}_i$;
6: If $\delta_m/m \leq \epsilon$ then exit, else set $n = 2n$ and continue from Step 2.

Observe that the error bound converges to zero at least as $O(1/n)$. If the function $g(z)$ is analytic in a neighborhood of the unit circle, then its coefficients decay exponentially to zero [12] so that also the the bound on the error converges exponentially to zero. It is also interesting to observe that, for $n \to \infty$, the convergence of the coefficients of $\hat{g}(z)$ to the corresponding coefficients of $g(z)$ is monotonic. Finally observe that the overall computational cost of this algorithm is $O(n \log n)$ arithmetic operations. This way, the Toeplitz part of $G$ is approximated by the banded Toeplitz matrix $T(\hat{g})$, i.e., the approximation of the Toeplitz part is obtained by truncating to finite size the bandwidth of $T(g)$.

In order to complete the computation of $G$ it remains to approximate the correction $E_g$. In view of the fact that $E_g$ has entries $e_{i,j}^{(g)}$ such that $v_i = \sum_j |e_{i,j}^{(g)}|$ is finite and $\lim_i v_i = 0$, we can approximate $E_g$ with a finite number of parameter within an error bound $\epsilon$. This computation is performed by means of functional iteration and is analyzed in the next section.

Finally, we point out that, while $G = T(g) + E_g$ has always nonnegative entries and $g(z)$ has nonnegative coefficients, the entries of the correction $E_g$ may have any sign in general.

**4. Fixed point iterations.** In this section we analyze the convergence of sequences generated by a functional iteration of the kind $X_{k+1} = F(X_k)$, $k = 0, 1, \ldots$, where $F(X)$ is a matrix function such that $G = F(G)$ where $G$ is the minimal nonnegative solution of (1.1). More precisely, we will consider the following cases

$$\begin{aligned} F_1(X) &= A_{-1} + A_0 X + A_1 X^2, \\ F_2(X) &= (I - A_0)^{-1}(A_{-1} + A_1 X^2), \\ F_3(X) &= (I - A_0 - A_1 X)^{-1} A_{-1}, \end{aligned} \tag{4.1}$$

while Newton iteration is considered in the next section.

In the case where $A_{-1}, A_0, A_1$ are finite matrices, the convergence analysis of the sequences generated by the functions (4.1) has been performed in [21]. Here, we

extend the results of [21] to the case of matrices of infinite size belonging to $\mathcal{L}_\infty$. We need the following

LEMMA 4.1. *Let* $A_{-1}\mathbf{1} > A_1\mathbf{1}$, *and*

$$\sigma = 1 - \min(a_{-1}(1) - a_1(1), b_{-1}(1) - b_1(1)) < 1. \tag{4.2}$$

*Let* $H_1 = A_0 + A_1 + A_1G$. *Then* $\|H_1\|_\infty \le \sigma$, *so that* $\|A_0\|_\infty \le \sigma$, $\|A_0 + A_1G\|_\infty \le \sigma$. *Therefore* $I - A_0$ *and* $I - A_0 - A_1G$ *are invertible and, for* $H_2 = (I - A_0)^{-1}(A_1 + A_1G)$, $H_3 = (I - A_0 - A_1G)^{-1}A_1$ *we have*

$$\|H_3\|_\infty \le \|H_2\|_\infty \le \|H_1\|_\infty \le \sigma < 1.$$

*Proof.* We have $\|A_0 + A_1 + A_1G\|_\infty = \|(A_0 + A_1 + A_1G)\mathbf{1}\|_\infty$. Moreover, since by Theorem 2.2 we have $G\mathbf{1} = \mathbf{1}$ and $\mathbf{1} = (A_{-1} + A_0 + A_1)\mathbf{1}$, then $(A_0 + A_1 + A_1G)\mathbf{1} = (A_0 + A_1 + A_1)\mathbf{1} = \mathbf{1} - (A_{-1} - A_1)\mathbf{1} \le \sigma\mathbf{1}$, by definition of $\sigma$. This implies $\|A_0 + A_1 + A_1G\|_\infty \le \sigma$. Since $A_0, A_1, G$ are nonnegative, then $\|A_0\|_\infty \le \|A_0 + A_1 + A_1G\|_\infty \le \sigma$ and $\|A_0 + A_1G\|_\infty \le \sigma$. By following the same arguments used in the proof of Lemma 2.1 we find that the matrices $I - A_0$ and $I - A_0 - A_1G$ are invertible in $\mathcal{L}_\infty$. Concerning $H_2$ we have $\|H_2\|_\infty = \|H_2\mathbf{1}\|_\infty$. Moreover,

$$\begin{aligned} H_2\mathbf{1} &= (I - A_0)^{-1}(A_1 + A_1G)\mathbf{1} = (I - A_0)^{-1}(I - A_0 - (A_{-1} - A_1G))\mathbf{1} \\ &= \mathbf{1} - (I - A_0)^{-1}(A_{-1} - A_1G)\mathbf{1} \le \mathbf{1} - (A_{-1} - A_1G)\mathbf{1} \\ &= (A_0 + A_1 + A_1G)\mathbf{1} = H_1\mathbf{1}, \end{aligned}$$

where we used the properties $(A_{-1} - A_1G)\mathbf{1} > 0$ and $(I - A_0)^{-1} \ge I$. Concerning $H_3$, since $A_1\mathbf{1} = (I - A_0 - A_{-1})\mathbf{1} = (I - A_0 - A_1G - (A_{-1} - A_1G))\mathbf{1}$ we have

$$\begin{aligned} H_3\mathbf{1} &= (I - A_0 - A_1G)^{-1}A_1\mathbf{1} = \mathbf{1} - (I - A_0 - A_1G)^{-1}(A_{-1} - A_1G)\mathbf{1} \\ &\le \mathbf{1} - (I - A_0)^{-1}(A_{-1} - A_1G)\mathbf{1} = H_2\mathbf{1}, \end{aligned}$$

where we used the fact that $G\mathbf{1} = \mathbf{1}$, $(A_{-1} - A_1G)\mathbf{1} > 0$ and $(I - A_0 - A_1G)^{-1} \ge (I - A_0)^{-1}$. ∎

Observe that, from Lemma 2.1 and from (2.2), it follows that $H_3 = W^{-1}RW$ and $\|H_3\|_\infty = \|W^{-1}RW\|_\infty \le \gamma$ where $\gamma < 1$ is defined in (2.3). This provides a different bound on the norm of $H_3$. Therefore we have

$$\|H_3\|_\infty \le \tau, \quad \tau = \min\{\gamma, \sigma\}, \tag{4.3}$$

with $\gamma$ and $\sigma$ defined in (2.3) and (4.2), respectively.

We are ready to prove the following result which shows that the three sequences generated by (4.1) starting with $X_0 = 0$ monotonically converge to $G$, convergence holds in the infinity norm and is linear. Moreover, the convergence of the third iteration is faster than that of the second one, while the convergence of the second iteration is faster than that of the first one.

THEOREM 4.2. *Assume that* $A_{-1}\mathbf{1} > A_1\mathbf{1}$. *For* $i \in \{1, 2, 3\}$ *define* $X_{k+1}^{(i)} = F_i(X_k^{(i)})$, $k = 0, 1, 2, \ldots$, *where* $X_0^{(i)} = 0$ *and* $F_i(X)$ *are given in* (4.1). *Then*

1. *the three sequences* $\{X_k^{(i)}\}$ *are well defined,*
2. $0 \le X_k^{(i)} \le X_{k+1}^{(i)} \le G$,

3. *for the error* $\mathcal{E}_k^{(i)} = G - X_k^{(i)}$ *we have* $\|\mathcal{E}_{k+1}^{(i)}\|_\infty \le \|H_i\|_\infty \|\mathcal{E}_k^{(i)}\|_\infty$, *where* $H_i$, $i = 1, 2, 3$ *are the matrices defined in Lemma 4.1, so that* $\lim_{k\to\infty} \|\mathcal{E}_k^{(i)}\|_\infty = 0$.

*Proof.* The first iteration is clearly well defined. The second is well defined since, according to Lemma 4.1, the matrix $I - A_0$ is invertible in $\mathcal{L}_\infty$. The third iteration is well defined as long as the matrix $I - A_0 - A_1X_k$ is invertible. On the other hand if $0 \le X_k \le G$, the latter matrix is invertible in view of Lemma 4.1 since $\|A_0 + A_1X_k\|_\infty \le \|A_0 + A_1G\|_\infty$. In order to prove that $0 \le X_k^{(i)} \le X_{k+1}^{(i)} \le G$ we use an induction argument. We prove it for the first iteration, i.e., for $i = 1$, the same technique can be used for the other iterations. For notational simplicity we omit the superscript and write $X_k$ in place of $X_k^{(1)}$. Since for $X_0 = 0$ we have $X_1 = A_{-1}$, so that $0 \le X_0 \le X_1$ and $G - X_1 = G - A_{-1} = (A_0 + A_1G)G \ge 0$. For the inductive step, assume that $0 \le X_{k-1} \le X_k \le G$. We first show that $0 \le X_k \le X_{k+1}$. From $X_{k+1} = A_1X_k^2 + A_0X_k + A_{-1}$ and from the property $0 \le X_{k-1} \le X_k$ we get

$$X_{k+1} \ge A_1X_{k-1}^2 + A_0X_{k-1} + A_{-1} = X_k.$$

Now consider

$$\begin{aligned} G - X_{k+1} &= A_1(G^2 - X_k^2) + A_0(G - X_k) = \\ &= A_1((G - X_k)G + X_k(G - X_k)) + A_0(G - X_k). \end{aligned} \tag{4.4}$$

Since $G - X_k \ge 0$ then also $G - X_{k+1} \ge 0$.

Concerning the norm bounds to $\mathcal{E}_k$, for $\mathcal{E}_k = \mathcal{E}_k^{(1)}$ for the sequence defined by $F_1$, from (4.4) we obtain

$$\mathcal{E}_{k+1} = A_1\mathcal{E}_kG + A_1X_k\mathcal{E}_k + A_0\mathcal{E}_k. \tag{4.5}$$

Since $\mathcal{E}_k \ge 0$ for any $k$, then $\|\mathcal{E}_k\|_\infty = \|\mathcal{E}_k\mathbf{1}\|_\infty$, so that

$$\begin{aligned} \|\mathcal{E}_{k+1}\|_\infty = \|\mathcal{E}_{k+1}\mathbf{1}\|_\infty &\le \|A_1\mathcal{E}_k\mathbf{1} + A_1X_k\mathcal{E}_k\mathbf{1} + A_0\mathcal{E}_k\mathbf{1}\|_\infty \\ &\le \|(A_0 + A_1 + A_1G)\mathcal{E}_k\mathbf{1}\|_\infty \le \|H_1\|_\infty\|\mathcal{E}_k\|_\infty. \end{aligned}$$

Similarly, concerning $F_2$ we obtain

$$\mathcal{E}_{k+1} = (I - A_0)^{-1}A_1(\mathcal{E}_kG + X_k\mathcal{E}_k), \tag{4.6}$$

whence

$$\begin{aligned} \|\mathcal{E}_{k+1}\|_\infty = \|\mathcal{E}_{k+1}\mathbf{1}\|_\infty &\le \|(I - A_0)^{-1}A_1(I + X_k)\mathcal{E}_k\mathbf{1}\|_\infty \\ &\le \|(I - A_0)^{-1}(A_1 + A_1G)\mathcal{E}_k\mathbf{1}\|_\infty \le \|H_2\|_\infty\|\mathcal{E}_k\|_\infty. \end{aligned}$$

Concerning $F_3$, we have

$$\mathcal{E}_{k+1} = (I - A_0 - A_1X_k)^{-1}A_1\mathcal{E}_kG, \tag{4.7}$$

whence

$$\|\mathcal{E}_{k+1}\mathbf{1}\|_\infty \le \|(I - A_0 - A_1G)^{-1}A_1\mathcal{E}_k\mathbf{1}\|_\infty \le \|H_3\|_\infty\|\mathcal{E}_k\|_\infty. \qquad \square$$

Observe that the reduction of the error per step of the $i$th iteration is bounded from above by $\|H_i\|_\infty$ which are in turn bounded by $\sigma$ for $i = 1, 2$ and by $\tau$ for $i = 3$. These constants are explicitly computable by means of (4.2) and (4.3).

The following result shows convergence properties in the case where $X_0$ is a stochastic matrix, say $X_0 = I$.

THEOREM 4.3. *Assume that* $A_{-1}\mathbf{1} > A_1\mathbf{1}$. *For* $i \in \{1,2,3\}$ *define* $X_{k+1}^{(i)} = F_i(X_k^{(i)})$, $k = 0,1,2,\ldots$, *where* $X_0^{(i)} \geq 0$, $X_0^{(i)}\mathbf{1} = \mathbf{1}$ *and* $F_i(X)$ *are given in* (4.1). *Then*

1. *the three sequences* $\{X_k^{(i)}\}$ *are well defined,*
2. $X_k^{(i)} \geq 0$, *and* $X_k^{(i)}\mathbf{1} = \mathbf{1}$,
3. *for the error* $\mathcal{E}_k^{(i)} = G - X_k^{(i)}$ *we have the following property:* $\mathcal{E}_k^{(i)}\mathbf{1} = 0$, *and for any other eigenvector* $w \neq \mathbf{1}$ *of* $G$ *such that* $Gw = \lambda w$, *the sequence* $w_k^{(i)} = \mathcal{E}_k^{(i)} w$ *satisfies* $\|w_{k+1}^{(i)}\|_\infty \leq \|H_i(\lambda)\|_\infty \|w_k^{(i)}\|_\infty$, *for* $i = 1,2$ *where* $H_1(\lambda) = (|\lambda|+1)A_1 + A_0$, $H_2(\lambda) = (|\lambda|+1)(I - A_0)^{-1}A_1$. *Moreover,* $\|w_{k+1}^{(3)}\|_\infty \leq |\lambda| \cdot \|(I - A_0 - A_1X_k)^{-1}A_1\|_\infty \|w_k^{(3)}\|_\infty$, *and* $\limsup_k \frac{\|w_{k+1}^{(3)}\|_\infty}{\|w_k^{(3)}\|_\infty} \leq |\lambda|$.

*Proof.* We show that if $X \geq 0$ and $X\mathbf{1} = \mathbf{1}$ then $F_i(X) \geq 0$ and $F_i(X)\mathbf{1} = \mathbf{1}$. For $F_1$ this property can be easily checked. For $F_2$, since $X \geq 0$ and $(I - A_0)^{-1} \geq 0$ then $F_2(X) \geq 0$. Moreover $F_2(X)\mathbf{1} = (I - A_0)^{-1}(A_{-1} + A_1)\mathbf{1} = (I - A_0)^{-1}(I - A_0)\mathbf{1} = \mathbf{1}$. Concerning $F_3$, since $X\mathbf{1} = \mathbf{1}$ then $\|A_0 + A_1X\|_\infty = \|A_0 + A_1G\|_\infty$ so that in light of Lemma 4.1 the matrix $I - A_0 - A_1X$ is invertible and has nonnegative inverse. This implies that $F_3(X) \geq 0$. Moreover, since $X\mathbf{1} = \mathbf{1}$ then $(I - A_0 - A_1X - A_{-1})\mathbf{1} = 0$ so that $F_3(X)\mathbf{1} = (I - A_0 - A_1X)^{-1}A_{-1}\mathbf{1} = \mathbf{1}$. From (4.5) we obtain $w_{k+1}^{(1)} = (\lambda A_1 + A_1X_k + A_0)w_k^{(1)}$ so that $\|w_{k+1}^{(1)}\|_\infty \leq \|\lambda A_1 + A_1X_k + A_0\|_\infty \|w_k^{(1)}\|_\infty$. On the other hand $\|\lambda A_1 + A_1X_k + A_0\|_\infty \leq \||\lambda|A_1 + A_1X_k + A_0\|_\infty = \|(|\lambda|A_1 + A_1X_k + A_0)\mathbf{1}\|_\infty = \|(|\lambda|A_1 + A_1 + A_0)\mathbf{1}\|_\infty = \|H_1(\lambda)\|_\infty$. Similarly, we proceed with $F_2$ relying on (4.6). Concerning $F_3$, from (4.7) we have $w_{k+1}^{(3)} = \lambda(I - A_0 - A_1X_k)^{-1}A_1w_k^{(3)}$, whence $\|w_{k+1}^{(3)}\|_\infty \leq |\lambda|\|(I - A_0 - A_1X_k)^{-1}A_1\|_\infty \|w_k^{(3)}\|_\infty$. Since $A_1\mathbf{1} < A_{-1}\mathbf{1}$ then $\|(I - A_0 - A_1X_k)^{-1}A_1\|_\infty = \|(I - A_0 - A_1X_k)^{-1}A_1\mathbf{1}\|_\infty \leq \|(I - A_0 - A_1X_k)^{-1}A_{-1}\mathbf{1}\|_\infty = 1$. Taking the limsup for $k \to \infty$ we obtain $\limsup_k \|(I - A_0 - A_1X_k)^{-1}A_1\|_\infty \leq \|(I - A_0 - A_1G)^{-1}A_{-1}\|_\infty = \|G\|_\infty = 1$. This completes the proof. ∎

The above theorem provides an estimate of the rate of convergence of the projection $w_k$ of the error $\mathcal{E}_k^{(i)}$ along each eigenvector $w$ of $G$. In particular, $w_k = 0$ for $w = \mathbf{1}$. In the finite dimensional case, this property, together with a compactness argument, implies convergence of $\|\mathcal{E}_k^{(i)}\|_\infty$ to 0 [21]. In the infinite dimensional case, the compactness arguments cannot be used to prove convergence of $\|\mathcal{E}_k^{(i)}\|_\infty$ to 0. Concerning the speed of convergence of the sequences defined by $F_1(X)$ and $F_2(X)$, since $\|H_1(\lambda)\|_\infty \leq \|H_1\|_\infty$ and $\|H_2(\lambda)\|_\infty \leq \|H_2\|_\infty$, then the projection $w_k$ of the error $\mathcal{E}_k^{(i)}$, $i = 1,2$, along each eigenvector of $G$ converges faster than the infinity norm of the error obtained with $X_0 = 0$. Concerning the sequence defined by $F_3(X)$, if $\lim_k \|\mathcal{E}_k^{(3)}\|_\infty = 0$, then $\|w_{k+1}^{(3)}\|_\infty / \|w_k^{(3)}\|_\infty \leq |\lambda| \lim_k \|(I - A_0 - A_1X_k)^{-1}A_1\|_\infty = |\lambda| \cdot \|H_3\|_\infty$. Since $|\lambda| \leq 1$, this inequality shows that the projection $w_k$ of the error $\mathcal{E}_k^{(i)}$ along each eigenvector of $G$ converges faster than the infinity norm of the error obtained with $X_0 = 0$. The convergence of the sequence $w_k$ is faster the smaller $\sup\{|\lambda| : \lambda \neq 1,\ \lambda \text{ eigenvalue of } G\}$. It is important to point out that, while in the finite dimensional case, under mild conditions, 1 is the only eigenvalue of maximum modulus of $G$ [2], in the infinite dimensional case the situation is more involved. Indeed, in some situations 1 is the only eigenvalue of modulus 1 and is isolated, in other situations 1 is an accumulation point of an infinite set of eigenvalues [17]. In this latter case there is not advantage in choosing a stochastic matrix as starting approximation.

**4.1. Implementation issues.** Let $g(z)$ be the solution of minimum modulus of (2.4) and consider the sequences generated by the fixed point iterations (4.1) obtained starting with $X_0 = T(g) + C$, where $C$ is any correction. Denote by $X_k$ any one of these three sequences so that we have $X_k = T(g) + E_k$, $E_0 = C$, where $E_k$ is the correction part. In this section we aim to explicit the equation which relates $E_{k+1}$ to $E_k$.

Consider the first iteration $X_{k+1} = A_1X_k^2 + A_0X_k + A_{-1}$ and denote by $F$ the correction matrix such that $T(g) + F = A_1T(g)^2 + A_0T(g) + A_{-1}$. Subtracting the latter equation from the former and performing formal manipulations yields

$$E_{k+1} = F + (A_1E_k + S)E_k + A_1E_kT(g), \quad S = A_0 + A_1T(g). \tag{4.8}$$

This equation provides a more efficient way to implement the first iteration since it involves multiplications of QT matrices and correction matrices and avoids the multiplication of QT matrices having a nonzero symbol. In this version, the precomputation of $g$, $S$ and $F$ is needed.

Consider the second iteration $X_{k+1} = (I - A_0)^{-1}(A_1X_k^2 + A_{-1})$. Subtract it from the equation $T(g) = (I - A_0)^{-1}(A_1T(g)^2 + A_{-1}) - (I - A_0)^{-1}F$ and performing formal manipulations yields

$$\begin{aligned} &E_{k+1} = \widehat{S}((T(g) + E_k)E_k + E_kT(g)) + \widetilde{S}, \\ &\widehat{S} = (I - A_0)^{-1}A_1, \ \widetilde{S} = (I - A_0)^{-1}F. \end{aligned} \tag{4.9}$$

Also in this case the precomputation of $T(g)$, $\widehat{S}$, $F$ and $\widetilde{S}$ is needed.

Concerning the third iteration and proceeding similarly we arrive at the recursion

$$\begin{aligned} &E_{k+1} = \widehat{V}E_k(I - A_1(T(g) + E_k) - A_0)^{-1}A_{-1} + \widetilde{V}, \\ &\widehat{V} = (I - A_1T(g) - A_0)^{-1}A_1, \ \widetilde{V} = (I - A_1T(g) - A_0)^{-1}F. \end{aligned} \tag{4.10}$$

Also in this case the precomputation of $T(g)$, $\widehat{V}$, and $\widetilde{V}$ is needed. However, at each step the inverse of a QT matrix must be computed.

The iterations (4.8)–(4.10) can be started with $X_0 = T(g)$, that is, $E_0 = 0$. Alternatively, they can be started with $X_0 = T(g) + ve_1^T$, where $e_1^T = (1, 0, \ldots)$ and $v$ is chosen in such a way that $X_0$ is row-stochastic so that convergence is faster. This can be accomplished by setting $E_0 = ve_1^T$.

**5. Newton's method.** Rewrite equation $A_1X^2 + A_0X + A_{-1} = X$ as

$$L(X) = 0, \quad L(X) := A_1X^2 + (A_0 - I)X + A_{-1}. \tag{5.1}$$

Newton's method applied to equation (5.1) generates the sequence

$$X_{k+1} = X_k - Z_k, \quad k = 0, 1, \ldots, \tag{5.2}$$

where the matrix $Z_k$ solves the equation $L'(Z_k) = L(X_k)$, and $L'(H) = A_1XH + A_1HX + (A_0 - I)H$ is the Fréchet derivative of $L(X)$ at $X$ applied to the matrix $H$. More specifically, $Z_k$ solves the following Sylvester equation

$$(A_1X_k + A_0 - I)Z_k + A_1Z_kX_k = L(X_k). \tag{5.3}$$

Observe that we may write

$$Z_k = -\sum_{i=0}^{\infty} S_k^i(I - A_0 - A_1X_k)^{-1}L(X_k)X_k^i, \quad S_k = (I - A_0 - A_1X_k)^{-1}A_1, \tag{5.4}$$

provided that $\|(I-A_0-A_1X_k)^{-1}\|_\infty$ is bounded from above, $\|S_k\|_\infty < 1$ and $\|X_k\|_\infty \leq 1$ so that we have

$$\|Z_k\|_\infty \leq \frac{\|L(X_k)\|_\infty \|(I - A_0 - A_1X_k)^{-1}\|_\infty}{1 - \|S_k\|_\infty}. \tag{5.5}$$

Moreover, we have

$$L(X_{k+1}) = A_1 Z_k^2. \tag{5.6}$$

The latter equality can be proved by observing that in general, for $Y = X - H$, we have $L(Y) = A_1(X-H)^2 + (A_0 - I)(X - H) + A_{-1} = L(X) - L'(H) + A_1H^2$ so that, if $H$ is such that $L'(H) = L(X)$ as in a Newton step, then $L(Y) = A_1H^2$.

Another useful property is the following. Equation (5.2) can be rewritten as $Z_k = \mathcal{E}_{k+1} - \mathcal{E}_k$, where $\mathcal{E}_k = G - X_k$, and $L(X_k)$ can be rewritten as $L(X_k) = L(X_k) - L(G) = -A_1(\mathcal{E}_k G + X_k\mathcal{E}_k) - (A_0 - I)\mathcal{E}_k$. Replace these two representations for $Z_k$ and $L(X_k)$ in (5.3), and get

$$(I - A_0 - A_1X_k)\mathcal{E}_{k+1} - A_1\mathcal{E}_{k+1}X_k = A_1\mathcal{E}_k^2.$$

By following the same arguments used to arrive at (5.4), we may rewrite the above equation as

$$\mathcal{E}_{k+1} - (I - A_0 - A_1X_k)^{-1}A_1\mathcal{E}_{k+1}X_k = (I - A_0 - A_1X_k)^{-1}A_1\mathcal{E}_k^2$$

and get

$$\mathcal{E}_{k+1} = \sum_{i=0}^{\infty} S_k^{i+1}\mathcal{E}_k^2\, X_k^i. \tag{5.7}$$

The following result extends to QT matrix coefficients the convergence results valid in the finite dimensional case [16]:

THEOREM 5.1. *Assume that $A_{-1}\mathbf{1} > A_1\mathbf{1}$. Let $X_k$, $k = 0, 1, \ldots$, be the sequence generated by (5.2) and (5.3) starting with $X_0 = 0$. Then, for any $k = 0, 1, 2, \ldots$,*

1. *equation (5.3) has a solution $Z_k$ such that $\|Z_k\|_\infty \leq \beta$, where*

$$\beta = \frac{2\|W^{-1}\|_\infty}{1 - \|W^{-1}A_1\|_\infty},$$

   *for $W = I - A_0 - A_1G$, so that $X_{k+1}$ is well defined;*
2. *$Z_k \leq 0$, $L(X_{k+1}) \geq 0$ and $0 \leq X_k \leq X_{k+1} \leq G$;*
3. *$\lim_{k\to\infty}(\mathcal{E}_k)_{i,j} = 0$ for any $i, j \geq 1$, where $\mathcal{E}_k = G - X_k$;*
4. *$\|\mathcal{E}_{k+1}\|_\infty \leq \frac{\tau}{1-\tau}\|\mathcal{E}_k\|_\infty^2$, $\|Z_{k+1}\|_\infty \leq \frac{\tau}{1-\tau}\|Z_k\|_\infty^2$, where $\tau = \min\{\gamma, \sigma\}$, with $\gamma$ and $\sigma$ defined in (2.3) and (4.2), respectively.*

*Proof.* We prove properties 1 and 2 by induction on $k$. For $k = 0$ we have $X_0 = 0$, $Z_0 = -(I - A_0)^{-1}A_{-1} \leq 0$, $L(X_1) = A_1Z_0^2 \geq 0$, $X_1 = -Z_0 \geq X_0$ and $X_1 = (I - A_0)^{-1}A_{-1} \leq G$. Moreover, clearly $\|Z_0\|_\infty \leq \beta$. For the inductive step, assume that properties 1 and 2 are valid for $k$ and prove them for $k+1$. We show that $\|Z_{k+1}\| \leq \beta$. Consider (5.5). Since by induction $L(X_{k+1}) \geq 0$ then $\|L(X_{k+1})\|_\infty = \|L(X_{k+1})\mathbf{1}\|_\infty \leq \|(A_1X_{k+1}^2 + A_0X_{k+1} + A_{-1})\mathbf{1}\|_\infty + \|X_{k+1}\mathbf{1}\|_\infty$. Moreover, since $X_{k+1} \leq G$ then $X_{k+1}\mathbf{1} \leq G\mathbf{1}$ so that $\|L(X_{k+1})\|_\infty \leq 2$. From $X_{k+1} \leq G$ it follows also $A_0 + A_1X_{k+1} \leq A_0 + A_1G$ so that $(I - A_0 - A_1X_{k+1})^{-1} \leq (I - A_0 - A_1G)^{-1} = W^{-1}$,

whence $\|(I - A_0 - A_1X_{k+1})^{-1}\|_\infty \le \|W^{-1}\|_\infty$. Similarly, $\|S_{k+1}\|_\infty \le \|(I - A_0 - A_1G)^{-1}A_1\|_\infty = \|W^{-1}A_1\|_\infty < 1$ in view of Lemma 4.1. From (5.4) and (5.5) we get $\|Z_{k+1}\|_\infty \le \beta$. The property $Z_{k+1} \le 0$ follows from (5.4) since $S_{k+1} \ge 0$, $(I - A_0 - A_1X_{k+1})^{-1} \ge 0$, and $L(X_{k+1}) \ge 0$ by the inductive assumption. The inequality $L(X_{k+2}) \ge 0$ follows from (5.6) since $Z_{k+1} \le 0$. Consequently $X_{k+2} = X_{k+1} - Z_{k+1} \ge X_{k+1}$. The property $X_{k+2} \le G$ follows from (5.7) since $\mathcal{E}_{k+1} \ge 0$, $S_{k+1} \ge 0$ and $X_{k+1} \ge 0$. Given $i, j$ consider the sequence $(\mathcal{E}_k)_{i,j}$ for $k = 0, 1, \ldots$. This sequence is non-increasing and bounded from below by 0 therefore it has a limit. The value of the limit cannot be positive since $G$ is the minimal nonnegative solution to the matrix equation. From the representation (5.7) of $\mathcal{E}_{k+1}$, since $\|X_k\|_\infty \le \|G\|_\infty \le 1$, and $\|S_k\|_\infty \le \|(I - A_0 - A_1G)^{-1}A_1\|_\infty \le \tau < 1$ in view of (4.3), we deduce that

$$\|\mathcal{E}_{k+1}\|_\infty \le \frac{\tau}{1-\tau}\|\mathcal{E}_k\|_\infty^2.$$

Similarly we can do for $\|Z_{k+1}\|_\infty$. ∎

**6. Perturbation results.** For the case where the coefficient matrices are finite, Higham and Kim [13] derived a condition number $\Psi(X)$ for a solvent $X$ of a general quadratic matrix equation of the kind (1.1) namely,

$$\Psi(X) = \|P^{-1}[\alpha(X^2)^T \otimes I_n, \beta X^T \otimes I_n, \gamma I_n^2]\|_2 / \|X\|_F,$$

where $P = I_n \otimes A_1X + X^T \otimes A_1 + I_n \otimes (A_0 - I)$ and $\alpha, \beta, \gamma$ are nonnegative parameters.

However, when the coefficient matrices are semi-infinite, there are cases where $P^{-1}$ does not exist or $\|X\|_F = \infty$, so the definition of $\Psi(X)$ does not apply. In this section, we take into account the structure of the coefficient matrices and derive a structured condition number for the minimal nonnegative solution of equations (1.1) and (1.2). Without loss of generality we consider only equation (1.1).

Consider the perturbed matrix equation obtained from (1.1) by replacing the coefficients $A_i$ by $A_i + \Delta_{A_i}$ where $\Delta_{A_i} = T(\delta_i) + E_{\delta_i} \in \mathcal{QT}$, $A_i + \Delta_{A_i} \ge 0$, for $i = -1, 0, 1$, and $(A_1 + \Delta_{A_1} + A_0 + \Delta_{A_0} + A_{-1} + \Delta_{A_{-1}})\mathbf{1} = \mathbf{1}$. Denote $X + \Delta_X$ a solution of the perturbed equation so that we may write

$$(A_1 + \Delta_{A_1})(X + \Delta_X)^2 + (A_0 + \Delta_{A_0})(X + \Delta_X) + A_{-1} + \Delta_{A_{-1}} = X + \Delta_X. \tag{6.1}$$

The analysis is separated into two parts, that is, the the analysis of the structured condition number of the Toeplitz part and the analysis of the condition number of the whole matrix.

**6.1. Toeplitz part.** In this section we provide a perturbation result for the function $g(z)$ which is the solution of minimum modulus of the scalar equation (2.4).

For the sake of notational simplicity, we omit the variable $z$ from the symbols, say, we write $g$ in place of $g(z)$ and $a_i$ in place of $a_i(z)$.

Under the assumption that the matrix coefficients $A_i + \Delta_{A_i}$ of equation (6.1) still satisfy the condition $A_i + \Delta_{A_i} \ge 0$, $\sum_{i=-1}^{1}(A_i + \Delta_{A_i})\mathbf{1} = \mathbf{1}$, for Theorem 2.2 the minimal nonnegative solution of (6.1) can be written as $G + \Delta_G$ , where $\Delta_G = T(\delta_g) + E_{\delta_g} \in \mathcal{QT}$ and $g + \delta_g$ is the solution of minimum modulus of the equation

$$a_{-1} + \delta_{-1} + (a_0 + \delta_0)\mu + (a_1 + \delta_1)\mu^2 = \mu.$$

Taking the difference of the above equation with (2.4), where we set $\mu = g + \delta_g$ and $\lambda = g$, we obtain

$$\delta_{-1} + \delta_0(g + \delta_g) + \delta_1(g + \delta_g)^2 + (a_0 - 1)\delta_g + a_1((g + \delta_g)^2 - g^2) = 0.$$

Whence, neglecting higher order terms in the perturbations we get

$$\delta_{-1} + \delta_0 g + \delta_1 g^2 + (a_0 - 1)\delta_g + 2a_1 g\delta_g \doteq 0,$$

where $\doteq$ means equality up to higher order terms with respect to the perturbations. This yields

$$\delta_g \doteq \frac{\delta_1 g^2 + \delta_0 g + \delta_{-1}}{1 - 2a_1 g - a_0}. \tag{6.2}$$

Note that $g(z)$, $a_0(z)$ and $a_1(z)$ have nonnegative coefficients so that $\|g\|_w = \sum_{i\in\mathbb{Z}} g_i = g(1) = 1$, and $|2a_1(z)g(z) + a_0(z)| \le 2a_1(1)g(1) + a_0(1) = 2a_1(1) + a_0(1) = 1 - a_{-1}(1) + a_1(1) < 1$, due to (2.5). This way, we have $\|(1 - 2a_1 g - a_0)^{-1}\|_w = (1 - 2a_1(1)g(1) - a_0(1))^{-1} = (a_{-1}(1) - a_1(1))^{-1}$. Whence, from (6.2), we obtain

$$\|\delta_g\|_w \dot{\le} (a_{-1}(1) - a_1(1))^{-1}\|\delta_1 g^2 + \delta_0 g + \delta_{-1}\|_w,$$

where $\dot{\le}$ means inequality up to higher order terms with respect to the perturbations. Therefore we arrive at the bound

$$\|\delta_g\|_w \dot{\le} \frac{1}{a_{-1}(1) - a_1(1)}\big(\|\delta_1\|_w + \|\delta_0\|_w + \|\delta_{-1}\|_w\big). \tag{6.3}$$

If we measure the perturbations by

$$\epsilon = \max\Big\{\frac{\|\delta_i\|_w}{\|a_i\|_w},\ i = -1, 0, 1\Big\},$$

we have $\|\delta_i\|_w \le \epsilon\|a_i\|_w$, moreover, since $\|a_{-1}\|_w + \|a_0\|_w + \|a_1\|_w = a_{-1}(1) + a_0(1) + a_1(1) = 1$, the relative variation of the symbol is bounded by

$$\frac{\|\delta_g\|_w}{\|g\|_w} = \|\delta_g\|_w \le \frac{1}{a_{-1}(1) - a_1(1)}\epsilon + O(\epsilon^2). \tag{6.4}$$

It follows from (6.4) that $\mathrm{cond}_{T(g)} := \frac{1}{a_{-1}(1)-a_1(1)}$ is an upper bound to the condition number of the Toeplitz part of $G$.

**6.2. Whole matrix.** Expanding (6.1), omitting the second and higher order terms in the perturbations, and setting $X = G$ lead to

$$(I - A_1 G - A_0)\Delta_G - A_1\Delta_G G \doteq (\Delta_{A_1} G^2 + \Delta_{A_0} G + \Delta_{A_{-1}}). \tag{6.5}$$

Now we prove some properties that will be useful to estimate the condition number of the whole solution.

Let $\Delta_A := \Delta_{A_1} G^2 + \Delta_{A_0} G + \Delta_{A_{-1}}$. According to (2.2), equation (6.5) can be written as

$$F(\Delta_G) \doteq W^{-1}\Delta_A \tag{6.6}$$

where $F : \mathcal{QT} \to \mathcal{QT}$ is the map defined by

$$F(Y) = Y - (W^{-1}RW)YG.$$

Now, we prove that the map $F(Y)$ is invertible in $\mathcal{L}_\infty$, that is, $F^{-1}$ has bounded infinity norm. By Lemma 2.1, we have $\|W^{-1}RW\|_\infty \leq \gamma < 1$ and $\|G\|_\infty = 1$ so that the series $\sum_{k=0}^{\infty}(W^{-1}RW)^k VG^k$ is convergent for any $V \in \mathcal{L}_\infty$. Therefore, if $V = F(Y)$ then $Y = \sum_{k=0}^{\infty}(W^{-1}RW)^k VG^k$. Thus, we get

$$\Delta_G = F^{-1}(W^{-1}\Delta_A) = \sum_{k=0}^{\infty}(W^{-1}RW)^k(W^{-1}\Delta_A)G^k.$$

Since $\sum_{k=0}^{\infty}\|W^{-1}RW\|_\infty^k = 1/(1-\|W^{-1}RW\|_\infty)$, taking norms in the above expression and applying Lemma 2.1 yields

$$\|\Delta_G\|_\infty \leq \frac{\|W^{-1}\|_\infty}{1-\|W^{-1}RW\|_\infty}\|\Delta_A\|_\infty \leq \frac{1}{\theta(1-\gamma)}\|\Delta_A\|_\infty. \tag{6.7}$$

Whence we conclude with the following

THEOREM 6.1. *If $A_{-1}\mathbf{1} > A_1\mathbf{1}$, then for the perturbation $\Delta_G$ we have*

$$\begin{aligned}\|\Delta_G\|_\infty \leq & \frac{\|W^{-1}\|_\infty}{1-\|W^{-1}RW\|_\infty}\|\Delta_A\|_\infty \\ \leq & \frac{1}{\theta(1-\gamma)}\|\Delta_A\|_\infty \leq \frac{1}{\theta(1-\gamma)}(\|\Delta_{A_{-1}}\|_\infty + \|\Delta_{A_0}\|_\infty + \|\Delta_{A_1}\|_\infty),\end{aligned} \tag{6.8}$$

*where $\theta$ and $\gamma$ are defined in (2.3) and $\Delta_A = \Delta_{A_1}G^2 + \Delta_{A_0}G + \Delta_{A_{-1}}$.*

From the above result it turns out that $\|W^{-1}\|_\infty/(1-\|W^{-1}RW\|_\infty)$ is an estimate of the conditioning of the problem, while $1/(\theta(1-\gamma))$ provides an upper bound. Since $1 - a_1(1)/a_{-1}(1) = (a_{-1}(1) - a_1(1))/a_{-1}(1)$, we may rewrite the upper bound to the conditioning in the following form which is closer to the expression obtained for the Toeplitz part of $G$ in Section 6.1.

$$\frac{1}{\theta(1-\gamma)} = \max\left(\frac{a_{-1}(1)}{a_{-1}(1)-a_1(1)}, \frac{b_{-1}(1)}{b_{-1}(1)-b_1(1)}\right)\frac{1}{\min(a_{-1}(1), b_{-1}(1))}.$$

It is interesting to observe that if $a_{-1}(1) \leq b_{-1}(1)$ and $\frac{a_{-1}(1)}{b_{-1}(1)} \leq \frac{a_1(1)}{b_1(1)}$, which in turn is verified if $a_{-1}(1) \leq b_{-1}(1)$ and $b_1(1) \leq a_1(1)$, then

$$\frac{1}{\theta(1-\gamma)} = \frac{1}{a_{-1}(1)-a_1(1)},$$

that is, the conditioning of the Toeplitz part coincides with the conditioning of the whole problem.

It is interesting to point out that, since $RW = A_1$ (see equation (2.2)), the estimate of the condition number $\|W^{-1}\|_\infty/(1-\|W^{-1}RW\|_\infty)$, appears in the uniform bound $\beta$ to the norm of the Newton correction $Z_k$ introduced in Theorem 5.1. Consequently, if the quadratic matrix equation is well conditioned, the uniform bound to the norm of $Z_k$ is smaller.

| Case | $\lambda_1$ | $\lambda_2$ | $\mu_1$ | $\mu_2$ | $p$ | $q$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 1.5 | 2 | 1 | 0 |
| 2 | 1 | 0 | 2 | 1.5 | 1 | 0 |
| 3 | 0 | 1 | 1.5 | 2 | 0 | 1 |
| 4 | 0 | 1 | 2 | 1.5 | 0 | 1 |
| 5 | 1 | 1 | 2 | 2 | 0.1 | 0.8 |
| 6 | 1 | 1 | 2 | 2 | 0.8 | 0.1 |
| 7 | 1 | 1 | 2 | 2 | 0.4 | 0.4 |
| 8 | 1 | 1 | 10 | 10 | 0.5 | 0.5 |
| 9 | 1 | 5 | 10 | 15 | 0.4 | 0.9 |
| 10 | 5 | 1 | 15 | 10 | 0.9 | 0.4 |

TABLE 6.1
*Parameters defining the matrices $A_{-1}, A_0, A_1$ in the 2-node Jackson network of [24]*

**6.3. A simple example.** This example is taken from Example 6.2 in [24], where a continuous time Markov process modeling a two-node Jackson network is considered. Here, the matrices are modified by means of the uniformization technique [19] in order to represent a discrete-time model. In details,

$$A_{-1} = \alpha \begin{pmatrix} (1-q)\mu_2 & q\mu_2 & & \\ & (1-q)\mu_2 & q\mu_2 & \\ & & \ddots & \ddots \end{pmatrix}, A_1 = \alpha \begin{pmatrix} \lambda_2 & & \\ p\mu_1 & \lambda_2 & \\ & \ddots & \ddots \end{pmatrix},$$

$$A_0 = \alpha \begin{pmatrix} -(\lambda_1+\lambda_2+\mu_2) & \lambda_1 & & \\ (1-p)\mu_1 & -(\lambda_1+\lambda_2+\mu_1+\mu_2) & \lambda_1 & \\ & \ddots & \ddots & \ddots \end{pmatrix} + I,$$

where the parameters $\lambda_1, \lambda_2, \mu_1, \mu_2, p, q$ are chosen as in Table 6.1 and $\alpha = (\lambda_1+\lambda_2+\mu_1+\mu_2)^{-1}$. Each case denotes one instance of the two-node Jackson network, formed by two servers and two queues, where customers arrive at nodes 1 and 2 according to independent Poisson processes with rates $\lambda_1$ and $\lambda_2$, respectively. Customers are served according to a first-come-first-served discipline, service times at nodes 1 and 2 are independent and exponentially distributed with means $1/\mu_1$, $1/\mu_2$. After completing service at node 1, customers enter node 2 with probability $p$ or leave the system with probability $1-p$, where $0 < p < 1$. After completing service at node 2, customers enter node 1 with probability $q$ or leave the system with probability $1-q$, where $0 < q < 1$.

In [24], 10 cases defined by the parameters given in Table 6.1 are analyzed. It can be easily seen that the condition $A_{-1}\mathbf{1} > A_1\mathbf{1}$ holds for cases 1, 3, 4, 5, 7, 8, 9, while the same condition holds in the cases 2, 6 and 10 for the flipped problems where phases and levels are exchanged.

For the flipped problem, the coefficient matrices are

$$A_{-1} = \alpha \begin{pmatrix} (1-p)\mu_1 & p\mu_1 & & \\ & (1-p)\mu_1 & p\mu_1 & \\ & & \ddots & \ddots \end{pmatrix}, A_1 = \alpha \begin{pmatrix} \lambda_1 & & \\ q\mu_2 & \lambda_1 & \\ & \ddots & \ddots \end{pmatrix},$$

$$A_0 = \alpha \begin{pmatrix} -(\lambda_1+\lambda_2+\mu_1) & \lambda_2 & & \\ (1-q)\mu_2 & -(\lambda_1+\lambda_2+\mu_1+\mu_2) & \lambda_2 & \\ & \ddots & \ddots & \ddots \end{pmatrix} + I.$$

| Problem | Conditioning | $\|\delta_g\|_w$ | $\delta_g$-bound | $\|\Delta_G\|_\infty$ | $\Delta_G$-bound |
|---|---|---|---|---|---|
| 1 | 9.0000 | 3.0601e-09 | 1.3211e-08 | 2.3583e-09 | 1.9817e-08 |
| 2* | 4.5000 | 1.5565e-09 | 4.8528e-09 | 1.5649e-09 | 4.8528e-09 |
| 3 | 4.5000 | 2.9937e-09 | 9.3015e-09 | 4.2850e-09 | 2.0256e-08 |
| 4 | 9.0000 | 4.7978e-09 | 2.0256e-08 | 2.5845e-09 | 1.8119e-08 |
| 5 | 7.5000 | 5.9099e-09 | 1.2456e-09 | 1.1106e-09 | 5.9334e-09 |
| 6* | 7.5000 | 4.4110e-10 | 8.2441e-09 | 3.8574e-10 | 8.3809e-09 |
| 7 | 30.0000 | 5.4766e-09 | 6.8300e-08 | 5.4809e-09 | 8.1645e-08 |
| 8 | 5.5000 | 4.9898e-10 | 4.0333e-09 | 5.9838e-10 | 4.9549e-09 |
| 9 | 5.1667 | 7.7413e-10 | 2.8017e-09 | 7.7413e-10 | 3.1621e-09 |
| 10* | 5.1667 | 6.3154e-10 | 3.8820e-09 | 5.3199e-10 | 4.4169e-09 |

TABLE 6.2

*Conditioning of the matrix equation for the 2-node Jackson network of [24]: actual perturbations in the solution and in the Toeplitz part and related upper bounds.*

If we perturb the parameters $\lambda_1, \lambda_2, \mu_1, \mu_2$ by

$$\tilde{\lambda}_i = \lambda_i(1+\epsilon_i^{\lambda}) \quad \tilde{\mu}_i = \mu_i(1+\epsilon_i^{\mu}), \quad i = 1,2,$$
$$\epsilon_i^{\lambda}, \epsilon_i^{\mu} \in [10^{-8}, 2\cdot 10^{-8}],$$

where the perturbations are randomly chosen, then we get the perturbed matrices $A_{-1}+\Delta_{A_{-1}}$, $A_0+\Delta_{A_0}$ and $A_1+\Delta_{A_1}$.

Note that $a_{-1}(1) = b_{-1}(1)$, $a_{-1}(1) - a_1(1) < b_{-1}(1) - b_1(1)$ holds true for both the original and the flipped problems, it follows $\frac{1}{\theta(1-\gamma)} = \frac{1}{a_{-1}(1)-a_1(1)}$, that is, the conditioning of the Toeplitz part coincides with the conditioning of the whole problem.

We denote by "Conditioning" the upper bound on the condition number for the minimal nonnegative solution $G$, that is, $\frac{1}{\theta(1-\gamma)}$. Moreover, we denote by $\delta_g$-bound and $\Delta_G$-bound, respectively, the perturbation bound (6.3) on $\|\delta_g\|_w$ and the bound (6.8) on $\|\Delta_G\|_\infty$. In Table 6.2, we report the upper bound on the condition number for the minimal nonnegative solution $G$ of equation (1.1), and we compare the perturbation bound (6.3) on the Toeplitz part of $G$ and the bound on the solution $G$ with the corresponding perturbation errors.

It can be seen from Table 6.2 that the upper bound $\frac{1}{\theta(1-\gamma)}$ can serve as a very good estimate of the conditioning of the problem. Inequalities (6.3) and (6.8) provide very sharp and revealing perturbation bounds to the Toeplitz part and to the solution $G$ with respect to small perturbations on the coefficients.

**7. Computational issues and numerical experiments.** Observe that, since the class $\mathcal{QT}$ is an algebra, then all the matrices $X_k^{(i)}$ generated by the fixed point iterations of Section 4 belong to $\mathcal{QT}$ and each fixed point iteration can be easily implemented in Matlab relying on the CQT-Toolbox of [6]. Concerning Newton iteration, a crucial role is played by the solution of the Sylvester-like equation (5.3). In the case where the coefficients have finite size $n$, this equation can be solved by the Bartels-Stewart algorithm [1] in $O(n^3)$ arithmetic operations. The case of infinite size coefficients is much more complicated. For quasi-Toeplitz matrices, the problem has been analyzed in [28] by using rational Krylov subspaces techniques where it is proved that $Z_k \in \mathcal{QT}$ under suitable assumptions that are satisfied by the condition $A_{-1}\mathbf{1} > A_1\mathbf{1}$. In the numerical experiments reported in this section we have used the implementation of [28] to solve (5.3).

In the implementation of the algorithms we have exploited the decomposition of $G$ as the sum of a Toeplitz part and a correction. We have implemented the computation

| | 0 | $I$ | $T(g)$ | $T(g)+ve_1^T$ |
|---|---|---|---|---|
| $F_1$ | 58.6 | 54.7 | 65.7 | 69.1 |
| $F_2$ | 37.2 | 34.8 | 40.0 | 28.9 |
| $F_3$ | 59.5 | 56.2 | 73.3 | 52.2 |

| | 0 | $I$ | $T(g)$ | $T(g)+ve_1^T$ |
|---|---|---|---|---|
| $F_1$ | 735 | 654 | 668 | 472 |
| $F_2$ | 466 | 416 | 421 | 297 |
| $F_3$ | 242 | 215 | 217 | 152 |

TABLE 7.1

*Two-node Jackson network for Problem 7 of Table 6.1: CPU time in seconds (left) and number of steps (right) required by the three fixed point iteration to arrive at a residual error at most $5.0 \cdot 10^{-14}$ starting with different values of $X_0$.*

of the Toeplitz part, i.e., of the coefficients of $g(z)$, relying on the evaluation and interpolation strategy at the roots of 1 with automatic handling of the number of interpolation points described in Algorithm 3.1. Concerning the computation of the correction part, we have applied the three fixed point iterations analyzed in Section 4 by following two strategies: the implementation in the standard version (4.1) and in the version which only computes the correction part, see (4.8), (4.9) and (4.10). In the standard version, we have used the pre-computation of the Toeplitz part by setting $X_0 = T(g)$ or $X_0 = T(g) + ve_1^T$ where $e_1 = (1, 0, \ldots)^T$ and $v$ is chosen so that $X_0$ is row-stochastic. Since there is not much difference in the performances of the version based on computing only the correction and the version where the whole matrix is computed, we report only the results concerning the latter version.

For each fixed point iteration, we have also tested different starting approximations, namely, $X_0 = 0$, $X_0 = I$.

We have compared the three fixed point iterations to Newton's iteration and to the algorithm of cyclic reduction (CR) analyzed in [4] which, in the case of finite matrices, is the method of choice commonly used in practice. For each experiment, we report the number of iterations, and CPU time needed to reach the bound $\|A_1X^2 + (A_0 - I)X + A_{-1}\|_\infty \leq \epsilon$ to the residual error where $\epsilon = 5.0 \cdot 10^{-14}$. We have considered some test problems modeling real world networks. More precisely, the "Two-node Jackson network" of Example 6.2 in [24], reported in Section 6.3, and the model "Assistance from an idle server" of [24] with different choices of the parameters, together with a general random walk in the quarter plane where the assigned probabilities have been chosen in such a way to have long queues in the system.

**7.1. Two-node Jackson network.** The model that we consider has been described in Section 6.3. Among the 10 problems in the list of Table 6.1, we report the results of Problem 7 which is the most ill-conditioned in the list, together with the case obtained with different values of the parameters $\lambda_i, \mu_i$, $p$ and $q$ which make the matrix $G$ numerically very large so that the computational effort is substantially large.

Figure 7.1 concerns Problem 7 in the list of Table 6.1. The first three graphs report the residual errors at each step for different values of $X_0$. The fourth graph compares the residual errors of the three iterations for $X_0 = T(g) + ve_1^T$. In Table 7.1 it is reported the CPU time in seconds (left) together with the number of steps (right) required by the three iterations to arrive at a residual error at most $5.0 \cdot 10^{-14}$. The computation of the symbol $g(z)$ is very inexpensive since the coefficients $g_i$ are computed in 0.003 seconds. For this problem, cyclic reduction provides the solution in just 8 steps and in 1.97 seconds, while Newton iteration requires 8 steps but takes a larger amount of seconds, i.e., 256.8.

We may observe that in this case CR is the most efficient algorithm and that the results of Theorems 4.2 and 4.3 are respected. In fact the iteration given by $F_3$ with

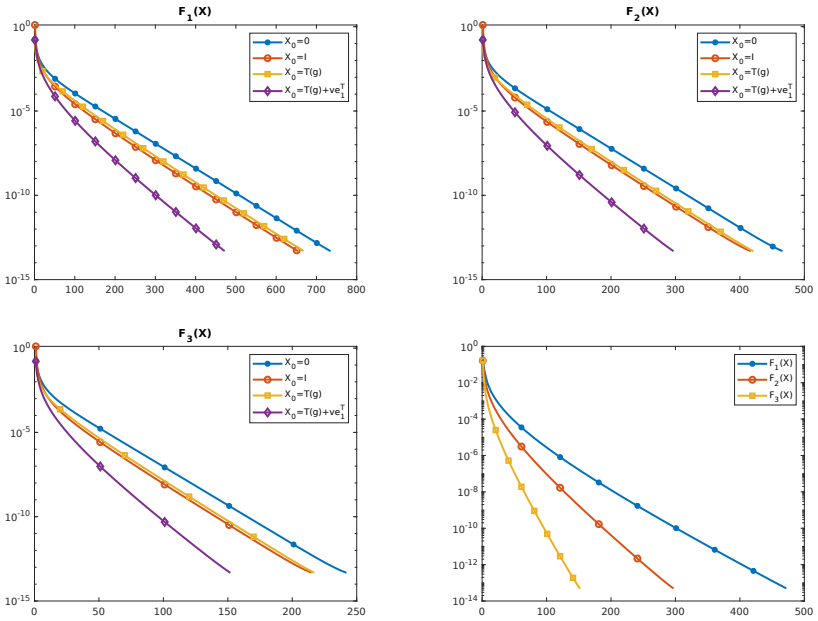


FIG. 7.1. *Two-node Jackson network for Problem 7 of Table 6.1: Residual error per step in the three functional iterations $F_1, F_2, F_3$ for different values of the initial matrix $X_0$. In the fourth graph, comparisons of the errors for the three iterations with $X_0 = T(g) + ve_1^T$.*

| | 0 | $I$ | $T(g)$ | $T(g) + ve_1^T$ |
|---|---|---|---|---|
| $F_1$ | 102.8 | 96.0 | 17.5 | 16.0 |
| $F_2$ | 49.0 | 46.6 | 9.8 | 9.8 |
| $F_3$ | 4981.0 | 4502.0 | 997.0 | 913.0 |

| | 0 | $I$ | $T(g)$ | $T(g) + ve_1^T$ |
|---|---|---|---|---|
| $F_1$ | 806 | 738 | 103 | 100 |
| $F_2$ | 310 | 285 | 47 | 46 |
| $F_3$ | 169 | 149 | 37 | 35 |

TABLE 7.2

*Two-node Jackson network for $\lambda_1 = 5$, $\lambda_2 = 0.7$, $\mu_1 = 2$, $\mu_2 = 2$, $p = 0.5$, $q = 0.5$: CPU time in seconds (left) and number of steps (right) required by the three fixed point iteration to arrive at a residual error at most $5.0 \cdot 10^{-14}$ starting with different values of $X_0$. Cyclic reduction requires 8 steps for the overall CPU time of 70.8 seconds.*

$X_0 = T(g) + ve_1^T$ is the fastest one in terms of number of steps. While concerning the CPU time, the iteration $F_2$ with $X_0 = T(g) + ve_1^T$ is the fastest.

Table 7.2 concerns the case of a Two-node Jackson network with the choice of parameters given by $\lambda_1 = 5$, $\lambda_2 = 0.7$, $\mu_1 = 2$, $\mu_2 = 2$, $p = 0.5$ and $q = 0.5$. In this model, seen as a random walk in the quarter plane, the overall probability to move right is higher than the overall probability to move left, while the probability to move down is higher than the probability to move up. The table reports the CPU time in seconds (left) together with the number of steps required by the three iterations (right) to arrive at a residual error at most $5.0 \cdot 10^{-14}$. The computation of the symbol $g(z)$ remains almost inexpensive even though the numerical length of the coefficient vector of the symbol $g(z)$ is quite large, in fact the coefficients $g_i$ are computed in less than 0.007 seconds, and the size of the coefficient vector is 31 for the coefficients of the negative powers of $z$ and 8424 for the coefficients of the positive powers, respectively. The numerical size of the correction is $28 \times 6937$.

For this problem, cyclic reduction provides the solution in 8 steps and in 70.8 seconds, while Newton iteration requires 8 steps and takes 782 seconds. In this case,

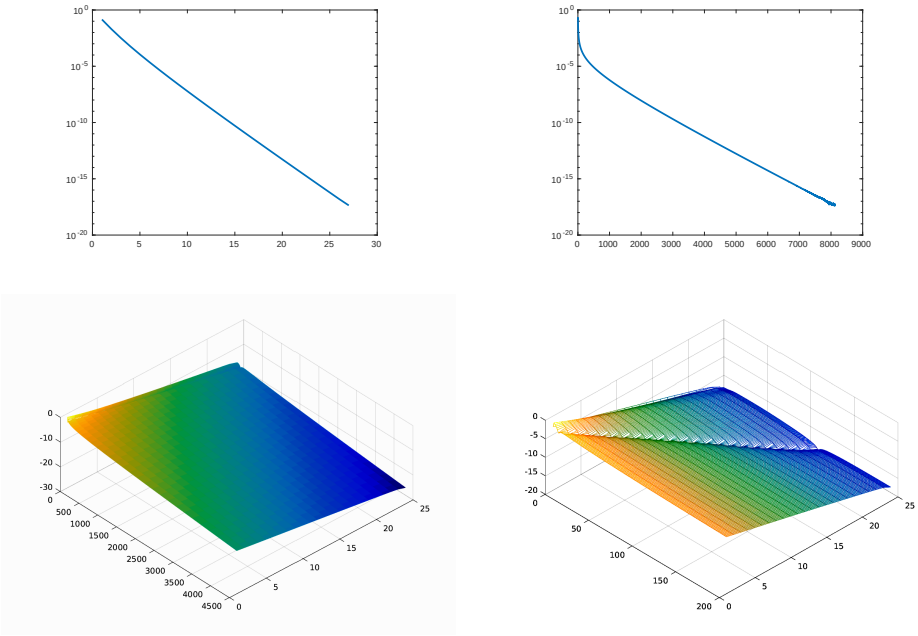

FIG. 7.2. *Two-node Jackson network for $\lambda_1 = 5$, $\lambda_2 = 0.7$, $\mu_1 = 2$, $\mu_2 = 2$, $p = 0.5$, $q = 0.5$: Solution $G$. In the upper part, the log-scale graph of the coefficients $g_i$ of the symbol for $i \leq 0$ (left) and for $i \geq 0$ (right). In the lower part the log-scale graph of the absolute value of the whole correction (left) together with a zoom (right).*

due to the large size of the matrices involved in the computation of matrix inverses, CR takes a much larger time than the simple functional iterations given by $F_1$ and $F_2$ which either do not involve inversion or require just only one matrix inversion. While iteration given by $F_2$ takes less than 10 seconds, the iteration given by $F_3$, even though is the fastest in terms of number of steps, needs a large CPU time. In fact, similarly to CR, it requires a matrix inversion at each step which becomes more expensive as the approximation approaches the limit. Newton iteration has the same convergence features as CR, however, the larger cost of solving a Sylvester equation makes this iteration much slower than the other ones at least for this problem.

In this model, increasing the values of $\lambda_1$ makes the size of the output much larger. In particular, with values $\lambda_1 \geq 6$, cyclic reduction breaks down for memory overflow while $F_2(X)$ provides the solution with a slight increase of the CPU time.

Figure 7.2 provides the solution $G$ in log scale where the Toeplitz part and the correction parts are separately represented.

**7.2. Assistance from idle server.** Here we consider a class of queueing models for a system with two servers and two queues. Arrivals to queues 1 and 2 occur as independent Poisson processes with parameters $\lambda_1$ and $\lambda_2$, respectively. The service times of servers 1 and 2 are exponentially distributed with parameters $\mu_1$ and $\mu_2$ respectively. Each server serves its own queue according to a first-come-first-served discipline. If one of the queues is empty, the server for that queue assists the other server, doubling the latter's service rate. If there is an arrival to a queue while its server is assisting the other queue, the server immediately ceases assisting and serves

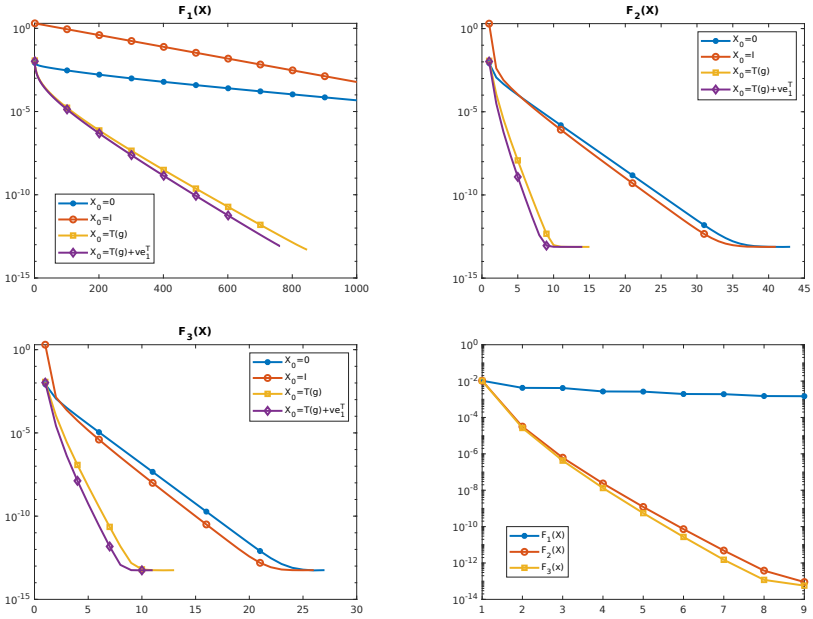


FIG. 7.3. *Assistance from idle server model from [24] for $\lambda_1 = 0.01$, $\lambda_2 = 2.9$, $\mu_1 = 0.03$, $\mu_2 = 2.0$: Residual error per step in the three functional iterations $F1, F_2, F_3$ for different values of the initial matrix $X_0$. In the fourth graph, comparisons of the errors for the three iterations with $X_0 = T(g) + ve_1^T$.*

| | 0 | $I$ | $T(g)$ | $T(g) + ve_1^T$ |
|---|---|---|---|---|
| $F_1$ | * | * | 204.8 | 191.6 |
| $F_2$ | 16.4 | 13.7 | 4.6 | 3.9 |
| $F_3$ | 405.1 | 402.6 | 218.3 | 183.6 |

| | 0 | $I$ | $T(g)$ | $T(g) + ve_1^T$ |
|---|---|---|---|---|
| $F_1$ | * | * | 844 | 782 |
| $F_2$ | 42 | 40 | 10 | 9 |
| $F_3$ | 26 | 25 | 9 | 7 |

TABLE 7.3

*Assistance from idle server model from [24] for $\lambda_1 = 0.01$, $\lambda_2 = 2.9$, $\mu_1 = 0.03$, $\mu_2 = 2.0$: CPU time in seconds (left) and number of steps (right) required by the three fixed point iterations to arrive at a residual error at most $5.0 \cdot 10^{-14}$ starting with different values of $X_0$. A "$*$" denotes a number of steps greater than 1000. Cyclic reduction requires 5 steps and 17 seconds of CPU time.*

its own queue. This stochastic process is ergodic if and only if $\rho_1 + \rho_2 < 2$, where $\rho_i = \lambda_i/\mu_i$, $i = 1, 2$.

For this model, the matrices $A_{-1}, A_0, A_1$ are given by $A_{-1} = \text{diag}(2\mu_1, \mu_1, \mu_1, \ldots)$, $A_0 = \text{trid}(\mu_2, -\lambda_1 - \lambda_2 - \mu_1 - \mu_2, \lambda_2) + (\mu_2 - \mu_1)e_1 e_1^T$, $A_1 = \lambda_1 I$.

In this example, we have chosen the values of the parameters in order that the numerical size of the matrix $G$ is substantially large, namely, $\lambda_1 = 0.01$, $\lambda_2 = 2.9$, $\mu_1 = 0.03$, $\mu_2 = 2.0$.

The situation is analogous to that of the second example in Section 7.1 where the methods based on functional iterations perform better than cyclic reduction and Newton iteration. The graphs in Figure 7.3 and the values in Table 7.3 synthesize the behavior of the algorithms. Cyclic reduction takes about 17 seconds of CPU while Newton iteration about 600 seconds. Slightly increasing the value of $\lambda_1$, CR breaks down for memory overflow while functional iterations still compute correctly the solution $G$.

**7.3. Random walk in the quarter plane.** Here we consider an example where the condition $A_{-1}\mathbf{1} > A_1\mathbf{1}$ is satisfied everywhere exept in the first component. The example describes a random walk in the quarter plane where a particle can occupy positions in a grid and we know the probabilities that the particle moves to the neighboring positions. In order to better describe the test problem, we denote $H = (h_{i,j})_{i,j=-1,1}$ the matrix with the probabilities of transition in the inner part of the quarter plane, while we denote $Y = (y_{i,j})$ the $3 \times 2$ matrix with the probabilities of transition in the $y$ axis. These two matrices fully describe the coefficients $A_i$ which can be written as $A_i = T(a_i) + E_i$ where $a_i(z) = \sum_{j=-1}^{1} h_{i,j} z^j$ and $E_i = e_1[y_{i,0} - h_{i,0}, y_{i,1} - h_{i,1}, 0, \ldots]$, compare with (2.1).

The random walk of this example is obtained with the values

$$H = \frac{1}{9}\begin{bmatrix} 1 & 0 & 1 \\ 2 & 0 & 0 \\ 2 & 2 & 1 \end{bmatrix}, \quad Y = \frac{1}{3}\begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

so that the condition $A_{-1}\mathbf{1} > A_1\mathbf{1}$ is satisfied in all the components but the first.

For this problem, there exist two nonnegative solutions $G$ and $\widehat{G}$ to equation (1.1) which satisfy the inequality $G \leq \widehat{G}$. Moreover $\widehat{G}$ is stochastic while $G$ is substochastic. The two solutions have the same symbol $g(z)$ and differ only for the correction part. Starting with $X_0 = 0$ or with $X_0 = T(g)$ the sequences generated by the functional iterations converge to $G$. Starting with $X_0 = I$ or $X_0 = T(g) + ve_1^T$ the sequences converge to $\widehat{G}$, while CR and Newton iteration converge to $G$.

The results of this test are summarized in Table 7.4. Both CR and Newton iteration take 23 steps in order to arrive at numerical convergence. However CR takes more than 8 minutes of CPU time while Newton iteration just 3.4 seconds. The large amount of CPU time taken by CR is due to the fact that the inverse matrices involved at each step of CR are QT matrices with a correction having a size which increases step after step and reaches values larger than $10^6$, while Newton iteration involves QT matrices with corrections having almost the same size of the correction of $G$ which is $126 \times 36$. The growth of the sizes of the correction matrices in the algorithms is an issue which deserves further analysis.

Functional iterations $F_1, F_2$ and $F_3$ with $X_0 = 0$ or with $X_0 = T(g)$ take a large number of steps to converge numerically to $G$ while with $X_0 = I$ or $X_0 = T(g) + ve_1^T$ the number of iterations is much smaller but the limit of the sequences is the stochastic solution $\widehat{G}$ which is not the minimal one.

For this problem, the combination of few steps of CR followed by few steps of Newton iteration provide a substantial acceleration in terms of CPU time. In fact, the first iterations of CR, involving matrices of small size, have a low cost. The last few steps of CR, which have a much higher cost, are replaced by Newton steps.

**8. Conclusions.** We have analyzed quadratic matrix equations encountered in the solution of random walk in the quarter plane where the solution of interest is the minimal nonnegative solution $G$. This class of equations is characterized by matrix coefficients with infinite size which belong to the class $\mathcal{QT}$ of Quasi-Toeplitz matrices. We have provided a perturbation analysis of $G$, introduced some fixed point algorithms for computing $G$ and compared their convergence speed. The algorithms rely on the properties of $\mathcal{QT}$ matrices recently investigated in [6]. Numerical experiments show that in many cases the CPU time and the memory resources required by our approach are significantly inferior to the ones required by the algorithm of cyclic reduction, which is considered as the algorithm of choice for this class of problems. The

| | $F_1$ | $F_2$ | $F_3$ | $F_1$ | $F_2$ | $F_3$ | CR | Newton | CR+Newton |
|---|---|---|---|---|---|---|---|---|---|
| iter | * | * | * | 285 | 205 | 119 | 23 | 23 | 15+10 |
| CPU | * | * | * | 3.1 | 2.3 | 2.6 | 524 | 3.4 | 0.4+1.5 |

TABLE 7.4

*Random walk in the quarter plane: Number of iterations and CPU time in seconds. From left to right: fixed point iterations with $X_0 = T(g)$, fixed point iterations with $X_0 = T(g) + ve_1^T$, cyclic reduction, Newton iteration with $X_0 = 0$, combination of cyclic reduction and Newton iteration. A "$*$" denotes more than 10000 iterations and a CPU larger than 1000 seconds. Starting the iterations with $X_0 = T(g) + ve_1^T$ generates sequences converging to the stochastic solution $\widehat{G}$, while starting with $X_0 = 0$ or applying CR, Newton iteration and their combinations generate sequences converging to the minimal (substochastic) solution $G$.*

effectiveness of Newton iteration depends on the growth of the sizes of the correction part in the QT matrices generated by the algorithm.

**Acknowledgments.** The authors wish to thank the anonymous referees for providing useful comments that helped to improve the presentation.

REFERENCES

[1] R. Bartels and G. Stewart. Solution of the matrix equation $AX + XB = C$: Algorithm 432. *Comm. ACM*, 15:820–826, 1972.

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

[3] D. A. Bini, S. Massei, and B. Meini. Semi-infinite quasi-Toeplitz matrices with applications to QBD stochastic processes. *Math. Comp.*, 87(314):2811–2830, 2018.

[4] D. A. Bini, S. Massei, B. Meini, and L. Robol. On quadratic matrix equations with infinite size coefficients encountered in QBD stochastic processes. *Numer. Linear Algebra Appl.*, 25(6):2128, 12, 2018.

[5] D. A. Bini, S. Massei, B. Meini, and L. Robol. Matrix analytic methods for reflected random walks with restart. *arXiv:1909.11372*, 2019.

[6] D. A. Bini, S. Massei, and L. Robol. Quasi-Toeplitz matrix arithmetic: a MATLAB toolbox. *Numerical Algorithms*, 81(2):741–769, 2019.

[7] D. A. Bini and B. Meini. On the solution of a nonlinear matrix equation arising in queueing problems. *SIAM J. Matrix Anal. Appl.*, 17(4):906–926, 1996.

[8] D. A. Bini and B. Meini. Improved cyclic reduction for solving queueing problems. *Numer. Algorithms*, 15(1):57–74, 1997.

[9] G. Fayolle, R. Iasnogorodski, and V. Malyshev. *Random walks in the quarter plane*, volume 40 of *Probability Theory and Stochastic Modelling*. Springer, Cham, second edition, 2017. Algebraic methods, boundary value problems, applications to queueing systems and analytic combinatorics.

[10] L. Flatto and S. Hahn. Two parallel queues created by arrivals with two demands I. *SIAM Journal on Applied Mathematics*, 44(5):1041–1053, 1984.

[11] L. Haque, Y. Q. Zhao, and L. Liu. Sufficient conditions for a geometric tail in a QBD process with many countable levels and phases. *Stochastic Models*, 21(1):77–99, 2005.

[12] P. Henrici. *Applied and computational complex analysis. Vol. 1*. Wiley Classics Library. John Wiley & Sons, Inc., New York, 1988. Power series—integration—conformal mapping—location of zeros, Reprint of the 1974 original, A Wiley-Interscience Publication.

[13] N. J. Higham and H.-M. Kim. Numerical analysis of a quadratic matrix equation. *IMA J. Numer. Anal.*, 20(4):499–519, 2000.

[14] M. Kobayashi and M. Miyazawa. Revisiting the tail asymptotics of the double QBD process: refinement and complete solutions for the coordinate and diagonal directions. In *Matrix-analytic methods in stochastic models*, volume 27 of *Springer Proc. Math. Stat.*, pages 145–185. Springer, New York, 2013.

[15] D. P. Kroese, W. R. W. Scheinhardt, and P. G. Taylor. Spectral properties of the tandem Jackson network, seen as a quasi-birth-and-death process. *Ann. Appl. Probab.*, 14(4):2057–2089, 2004.

[16] G. Latouche. Newton's iteration for non-linear equations in Markov chains. *IMA J. Numer. Anal.*, 14(4):583–598, 1994.

[17] G. Latouche, S. Mahmoodi, and P. Taylor. Level-phase independent stationary distributions for GI/M/1-type Markov chains with infinitely-many phases. *Performance Evaluation*, 70(9):551–563, 2013.

[18] G. Latouche, G. T. Nguyen, and P. G. Taylor. Queues with boundary assistance: the effects of truncation. *Queueing Syst.*, 69(2):175–197, 2011.

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

[20] G. Latouche and P. Taylor. Truncation and augmentation of level-independent QBD processes. *Stochastic Process. Appl.*, 99(1):53–80, 2002.

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

[22] M. Miyazawa. Light tail asymptotics in multidimensional reflecting processes for queueing networks. *TOP*, 19(2):233–299, 2011.

[23] M. Miyazawa and Y. Q. Zhao. The stationary tail asymptotics in the $GI/G/1$-type queue with countably many background states. *Adv. in Appl. Probab.*, 36(4):1231–1251, 2004.

[24] A. J. Motyer and P. G. Taylor. Decay rates for quasi-birth-and-death processes with countably many phases and tridiagonal block generators. *Adv. Appl. Prob.*, 38:522–544, 2006.

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

[26] T. Ozawa. Stability condition of a two-dimensional QBD process and its application to estimation of efficiency for two-queue models. *Performance Evaluation*, 130:101–118, 2019.

[27] T. Ozawa and M. Kobayashi. Exact asymptotic formulae of the stationary distribution of a discrete-time two-dimensional QBD process. *Queueing Systems*, 90(3):351–403, Dec 2018.

[28] L. Robol. Rational Krylov and ADI iteration for infinite size quasi-Toeplitz matrix equations. *arXiv:1907.02753*, pages 1–24, 2019.

[29] Y. Sakuma and M. Miyazawa. On the effect of finite buffer truncation in a two-node Jackson network. *J. Appl. Probab.*, 42(1):199–222, 2005.

[30] D. Stanford, W. Horn, and G. Latouche. Tri-layered QBD processes with boundary assistance for service resources. *Stochastic Models*, 22(3):361–382, 2006.

[31] Y. Takahashi, K. Fujimoto, and N. Makimoto. Geometric decay of the steady-state probabilities in a quasi-birth-and-death process with a countable number of phases. *Communications in Statistics. Part C: Stochastic Models*, 17(1):1–24, 2001.