Journal of Scientific Computing (2024) 99:66
https://doi.org/10.1007/s10915-024-02491-8

Algorithms for Square Root of Semi-Infinite Quasi-Toeplitz *M*-Matrices

Hongjia Chen[1] · Hyun-Min Kim[2] · Jie Meng[3]

Received: 25 July 2023 / Revised: 2 February 2024 / Accepted: 4 February 2024 /
Published online: 22 April 2024

Abstract

A quasi-Toeplitz M-matrix A is an infinite M-matrix that can be written as the sum of a semi-infinite Toeplitz matrix and a correction matrix. This paper is concerned with computing the square root of invertible quasi-Toeplitz M-matrices which preserves the quasi-Toeplitz structure. We show that the Toeplitz part of the square root can be easily computed through evaluation/interpolation. This advantage allows us to propose algorithms solely for the computation of correction part, whence we propose a fixed-point iteration and a structure-preserving doubling algorithm. Additionally, we show that the correction part can be approximated by solving a nonlinear matrix equation with coefficients of finite size followed by extending the solution to infinity. Numerical experiments showing the efficiency of the proposed algorithms are performed.

Keywords Quasi-Toeplitz matrix · Infinite M-matrix · Square root · Structured-preserving doubling algorithm

Mathematics Subject Classification 15A24 · 65F45 · 15B05

1 Introduction

M-matrices in the context of infinite dimensional spaces are called M-operators, which, to our knowledge, were firstly investigated in [19], since then related theoretical properties have been developed in [1, 17, 19–21, 25]. Quasi-Toeplitz M-matrices are infinite M-matrices with an

✉ Jie Meng
mengjie@ouc.edu.cn

Hongjia Chen
chenhongjia@ncu.edu.cn

Hyun-Min Kim
hyunmin@pusan.ac.kr

1 Department of Mathematics, Nanchang University, Nanchang 330031, China

2 Department of Mathematics, Pusan National University, Busan 46241, Republic of Korea

3 School of Mathematical Sciences, Ocean University of China, Qingdao 266100, China

almost Toeplitz structure, they are encountered in the numerical solution of a quadratic matrix equation [10] involved in 2-dimensional Quasi-Birth-Death (QBD) stochastic processes [23] and are recently studied in [22] in terms of their theoretical and computational properties.

In this paper, we are interested in the quasi-Toeplitz M-matrices that belong to the class $\mathcal{QT}_\infty = \{T(a)+E : a(z) \in \mathcal{W}, E \in \mathcal{K}_d(\ell^\infty)\}$, where $T(a)$ is a semi-infinite Toeplitz matrix associated with the function $a(z) = \sum_{i\in\mathbb{Z}} a_i z^i$ in the sense that $(T(a))_{i,j} = a_{j-i}$, $\mathcal{W}$ is the Wiener algebra, defined as the set $\mathcal{W} = \{a(z) = \sum_{i\in\mathbb{Z}} a_i z^i : z \in \mathbb{T}, \|a\|_\mathcal{W} := \sum_{i\in\mathbb{Z}} |a_i| < \infty\}$, and $\mathcal{K}_d(\ell^\infty) = \{E = (e_{i,j})_{i,j\in\mathbb{Z}^+} : \lim_i \sum_{j=1}^{\infty} |e_{i,j}| = 0\}$. It has been proved in [11, Theorem 2.16] that the class $\mathcal{QT}_\infty$ is a Banach algebra with the infinity matrix norm $\|\cdot\|_\infty$, which turns out to be $\|A\|_\infty = \sup_i \sum_{j=1}^{\infty} |a_{i,j}|$ for $A = (a_{i,j})_{i,j\in\mathbb{Z}^+}$. For $A = T(a) + E \in \mathcal{QT}_\infty$, $T(a)$ is called the Toeplitz part with a symbol a, E is called the correction part. Matrices in the class $\mathcal{QT}_\infty$ have rich and elegant theoretical and computational properties, we refer the reader to [3, 6–13, 18, 24] for more details.

For a quasi-Toeplitz M-matrix $A = T(a) + E_A \in \mathcal{QT}_\infty$, it has been proved in [22] that if A is an (invertible) M-matrix, then $T(a)$ is also an (invertible) M-matrix. Moreover, it shows that if A is invertible, there exists a unique quasi-Toeplitz M-matrix $S = T(s) + E_S \in \mathcal{QT}_\infty$ such that $A = S^2$. Concerning the computation of the matrix S, Binomial iteration and Cyclic Reduction (CR) algorithm have been proposed in [22], where the CR algorithm seems to be better suited in the numerical computations. However, both the Binomial iteration and the CR algorithm exploit the quasi-Toeplitz structure indirectly by performing approximate operations of semi-infinite quasi-Toeplitz matrices in the format. It would be natural to ask whether the quasi-Toeplitz structure can be fully exploited to propose more efficient algorithms.

Suppose $B = T(b) + E_B \in \mathcal{QT}_\infty$ satisfies $(I - B)^2 = A$, where $A = T(a) + E_A$ is a given quasi-Toeplitz M-matrix, then we have for the symbols of the Toeplitz parts that $(1-b(z))^2 = a(z)$. Observe that for a positive integer $n > 0$, there is always a unique Laurent polynomial $\hat{b}(z) = \sum_{i=-n+1}^{n} \hat{b}_i z^i$ that interpolates $b(z)$ at the $2n$ roots of unity. Based on the technic of evaluation/interpolation, we investigate computation of the coefficients b_i of $b(z) = \sum_{i\in\mathbb{Z}} b_i z^i$, so that the Toeplitz part $T(b)$ of the quasi-Toeplitz M-matrix A can be easily obtained.

Concerning the computation of the correction part, we propose a fixed-point iteration with a linear convergence rate, and a structure-preserving doubling algorithm, which is of quadratic convergence rate. Moreover, we show that the correction part can be approximated by extending a finite size matrix to infinity, where the finite size matrix solves a nonlinear matrix equation. Numerical experiments show that the proposed algorithms provide convergence acceleration in terms of CPU times comparing with the Binomial iteration and CR algorithm proposed in [22], both of which keep the whole quasi-Toeplitz matrices in the computations.

This paper is organized as follows. In the remaining part of this introduction, we recall some definitions and properties concerning quasi-Toeplitz matrices and M-matrices. Sections 2 and 3 concern with algorithms that fully exploit the quasi-Toeplitz structure of square root of invertible quasi-Toeplitz M-matrices, in Sect. 2 we show how the Toeplitz part is computed, while in Sect. 3, we design and analyze the convergence of algorithms that are applicable in computing the correction part. In Sect. 4, we show that the correction part can be approximated by extending to infinity of the solution of a nonlinear matrix equation with finite size coefficients. In Sect. 5, we show by numerical examples the efficiency of the proposed algorithms.

1.1 Preliminary Concepts

Let ℓ^∞ be the space of sequences $\{x = (x_1, x_2, \ldots)\}$ such that $\sup_{i\in\mathbb{Z}^+} |x_i| < \infty$, one can see that quasi-Toeplitz M-matrices in the class $\mathcal{QT}_\infty$ are bounded linear operators from ℓ^∞ to ℓ^∞. Denote by $\mathcal{B}(\ell^\infty)$ the Banach space of bounded linear operators from ℓ^∞ to itself, we first recall definition of M-operators on $\mathcal{B}(\ell^\infty)$. For definition of more general M-operators on a real partially ordered Banach space, we refer the reader to [17, 21, 25] and the references therein. M-operators on the Banach space $\mathcal{B}(\ell^\infty)$ are defined as

Definition 1.1 An operator $A \in \mathcal{B}(\ell^\infty)$ is said to be a Z-operator if $A = sI - P$, with $s \geq 0$, $P(\ell^\infty_+) \subseteq \ell^\infty_+$, where $\ell^\infty_+ = \{x = (x_i)_{i\in\mathbb{Z}^+} \in \ell^\infty : x_i \geq 0 \ for\ all\ i\}$. A Z-operator is said to be an M-operator if $s \geq \rho(P)$, where $\rho(P)$ is the spectral radius of P. A is an invertible M-operator if $s > \rho(P)$.

As matrices in $\mathcal{QT}_\infty$ can be represented by matrices of infinite size, we keep using the term M-matrix when referring M-operators in $\mathcal{QT}_\infty$. This way, a matrix $A \in \mathcal{QT}_\infty$ is said to be an M-matrix if $A = \beta I - B$ with $B \geq 0$ and $\beta \geq \rho(B)$, and A is invertible if $\beta > \rho(B)$. Here $B \geq 0$ means that B is an elementwise nonnegative infinite matrix.

The following lemma contains a collection of properties of quasi-Toeplitz matrices and quasi-Toeplitz M-matrices, where properties (i) and (ii) have been proved in [2], while properties (iii–v) can be found from [22].

Lemma 1.1 *If $A = T(a) + E_A \in \mathcal{QT}_\infty$ and $B = T(b) + E_B \in \mathcal{QT}_\infty$, then the following properties hold:*

(i) $AB = T(ab) - H(a^-)H(a^+) \in \mathcal{QT}_\infty$, where $(H(a^-))_{i,j} = (a_{-i-j+1})_{i,j\in\mathbb{Z}^+}$ and $(H(a^+))_{i,j} = (a_{i+j-1})_{i,j\in\mathbb{Z}^+}$;
(ii) it holds that $\|a\|_{\mathcal{W}} = \|T(a)\|_\infty \leq \|A\|_\infty$;
(iii) $T(a) \geq 0$ if $A \geq 0$.
(iv) $\|a\|_{\mathcal{W}} = a(1)$ if $T(a) \geq 0$.
(v) T(a) is an (invertible) M-matrix if A is an (invertible) M-matrix.

The following lemma shows that an invertible M-matrix in the class $\mathcal{QT}_\infty$ admits a unique quasi-Toeplitz M-matrix as a square root.

Lemma 1.2 *[22, Theorem 3.6] Suppose $A = \beta(I - A_1) \in \mathcal{QT}_\infty$ satisfies $\beta > 0$, $A_1 \geq 0$ and $\|A_1\|_\infty < 1$, then there is a unique $B \in \mathcal{QT}_\infty$ such that $B \geq 0$, $\|B\|_\infty < 1$, and $(I - B)^2 = I - A_1$.*

For quasi-Toeplitz M-matrix $A = \gamma(I - A_1) \in \mathcal{QT}_\infty$ such that $A_1 \geq 0$ and $\|A_1\|_\infty < 1$, it can be seen from Lemma 1.2 that it suffices to compute matrix B such that $(I - B)^2 = I - A_1$. In what follows, we propose algorithms for computing the Toeplitz part and the correction part of matrix B.

2 Computing the Toeplitz Part

Observe that the Toeplitz part $T(b)$ is uniquely determined by the coefficients b_j of the symbol $b(z) = \sum_{j\in\mathbb{Z}} b_j z^j$. In this section, we show that $b(z)$ can be approximated by $\hat{b}(z) = \sum_{i=-n+1}^{n} \hat{b}_j z^j$ in the sense that $\|b - \hat{b}\|_{\mathcal{W}} \leq c\epsilon$ for some constant c and a given tolerance ϵ.

Suppose $B = T(b) + E_B$ satisfies $\gamma(I - B)^2 = A$, where $A = \gamma(I - A_1) \in \mathcal{QT}_\infty$ is such that $A_1 \geq 0$ and $\|A_1\|_\infty < 1$. Suppose $T(a)$ is the Toeplitz part of A, we have from property (i) of Lemma 1.1 that $\gamma(1 - b(z))^2 = a(z)$, that is,

$$a(z)/\gamma = b(z)^2 - 2b(z) + 1, \tag{2.1}$$

from which we obtain $b(z) = 1 \pm \sqrt{a(z)/\gamma}$. Since $A_1 \geq 0$, in view of properties (ii)-(iv) of Lemma 1.1, we have $a_1(1) = \|a_1\|_{\mathcal{W}} \leq \|A_1\|_\infty < 1$, where $a_1(z)$ is the symbol of the Toeplitz part of A_1, hence we deduce that $a(1) = \gamma(1 - a_1(1)) > 0$. On the other hand, it follows from $B \geq 0$ that $b(1) = \|b\|_{\mathcal{W}} = \|T(b)\|_\infty \leq \|B\|_\infty < 1$, which, together with $\sqrt{a(1)/\gamma} > 0$, implies that $b(1) = 1 - \sqrt{a(1)/\gamma}$ and therefore $b(z) = 1 - \sqrt{a(z)/\gamma}$.

Let $n > 0$ be a positive integer, set $m = 2n$, then there is always a unique Laurent series $\hat{b}(z) = \sum_{j=-n+1}^{n} \hat{b}_j z^j$ such that $\hat{b}(\omega_m^\ell) = b(\omega_m^\ell)$, $\ell = -n+1, \ldots, n$, where ω_m is the principal m-th root of 1, that is, $\omega_m = \cos\frac{2\pi}{m} + \mathbf{i}\sin\frac{2\pi}{m}$. Based on the evaluation/interpolation technique, where the interpolation can be done by the means of the Fast Fourier Transform (FFT), an approximation $\hat{b}_i$, $i = -n+1, ..., n$, to the coefficients b_i of $b(z)$ can be obtained. Since $B \geq 0$, we have from property (iii) of Lemma 1.1 that $T(b) \geq 0$, so that $b(z) = \sum_{i\in\mathbb{Z}} b_i z^i$ has nonnegative coefficients. If in addition $b''(z) \in \mathcal{W}$, the following lemma provides a bound to $|\hat{b}_i - b_i|$.

Lemma 2.1 *[10, Lemma 3.1] For $g(z) = \sum_{i\in\mathbb{Z}} g_i z^i \in \mathcal{W}$ with nonnegative coefficients, 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., $g(w_m^i) = \hat{g}(w_m^i)$ for $i = -n+1, \ldots, n$, where $m = 2n$. If $g''(z) \in \mathcal{W}$, then $g''(1) \geq 0$ and*

$$g''(1) - \hat{g}''(1) \geq 2n\Big(\sum_{j<-n+1} g_j + \sum_{j>n} g_j\Big).$$

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

For $\hat{b}(z) = \sum_{j=-n+1}^{n} \hat{b}_j z^j$ interpolating $b(z)$ at ω_m^i for $i = -n+1, \ldots, n$, suppose $b''(z) \in \mathcal{W}$ and $b''(1) > 0$, we have from Lemma 2.1 that

$$b''(1) - \hat{b}''(1) \geq 2n\Big(\sum_{j<-n+1} b_j + \sum_{j>n} b_j\Big), \tag{2.2}$$

and

$$|\hat{b}_j - b_j| \leq \frac{1}{2n}(b''(1) - \hat{b}''(1)),\ \ j = -n+1, \ldots, n. \tag{2.3}$$

If $b''(1) - \hat{b}''(1) < \epsilon$ for a given tolerance $\epsilon > 0$, we have from (2.3) that $|b_j - \hat{b}_j| \leq \epsilon/(2n)$ for $j = -n+1, \ldots, n$, which together with (2.2) implies that

$$\begin{aligned}\|b - \hat{b}\|_{\mathcal{W}} &= \sum_{j=-n+1}^{n} |b_j - \hat{b}_j| + \sum_{j<-n+1} b_j + \sum_{j>n} b_j \\ &\leq \epsilon + \frac{1}{2n}(b''(1) - \hat{b}''(1)) \\ &\leq (1 + \frac{1}{2n})\epsilon.\end{aligned}$$

Hence, in the computation of $\hat{b}_j$, $j = -n+1, \dots, n$, under the evaluation/interpolation scheme, the approximation is accurate enough if $b''(1) - \hat{b}''(1) < \epsilon$. Actually, the values of $b''(1) - \hat{b}''(1)$ can be easily obtained. Indeed, once the coefficients $\hat{b}_j$ of $\hat{b}(z) = \sum_{j=-n+1}^{n} \hat{b}_j z^j$ are computed, one can easily obtain $\hat{b}''(1) = \sum_{j=-n+1}^{n} j(j-1)\hat{b}_j$. On the other hand, we have from Eq. (2.1) that

$$b'(z) = \frac{a'(z)}{2\gamma(b(z)-1)} \text{ and } b''(z) = \frac{a''(z) - 2\gamma(b'(z))^2}{2\gamma(b(z)-1)},$$

from which we easily obtain $b'(1)$ and $b''(1)$.

Observe that equation (2.1) is a special case of the quadratic Eq.

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

where $a_i(z)$ for $i = -1, 0, 1$ are known functions in the class $\mathcal{W}$ and $g(z)$ is the function to be determined. Algorithms for computing the approximations of the coefficients of $g(z)$ has been proposed in [10], based on which we propose the following Algorithm 1 that is more efficient in computing the coefficients $\hat{b}_j$ of the Laurent series $\hat{b}(z) = \sum_{j=-n+1}^{n} \hat{b}_j z^j$, so that we get an approximation $T(\hat{b})$ to the Toeplitz part $T(b)$ in the sense that $\|T(b) - T(\hat{b})\|_\infty = \|b - \hat{b}\|_{\mathcal{W}} \le (1 + \frac{1}{2n})\epsilon$ for a given tolerance ϵ.

Algorithm 1 Approximation of $b(z)$

Require: The coefficients of $a(z)$, a scalar γ such that $A = \gamma(I - A_1)$ and a tolerance $\epsilon > 0$.

Ensure: Approximations $\hat{b}_j$, $j = -n+1, \dots, n$, to the coefficients b_j of $b(z)$ such that $|\hat{b}_j - b_j| \le \epsilon/(2n)$.

1: Set n=4, and compute $b(1) = 1 - \sqrt{a(1)/\gamma}$ and $b'(1) = \frac{a'(1)}{2\gamma(b(1)-1)}$ and $b''(1) = \frac{a''(1)-2\gamma(b'(1))^2}{2\gamma(b(1)-1)}$;

2: Set $m = 2n$ and $w_m = \cos\frac{2\pi}{m} + \mathbf{i}\sin\frac{2\pi}{m}$. Evaluate $a(z)$ at $z = w_m^i$ for $i = -n+1, \dots, n$;

3: For $i = -n+1, \dots, n$, compute $s_i = 1 - \sqrt{a(\omega_m^i)/\gamma}$;

4: Interpolate the values s_i, $i = -n+1, \dots, n$, by means of FFT and obtain the coefficients $\hat{b}_j$ of the Laurent polynomial $\hat{b}(z) = \sum_{j=-n+1}^{n} \hat{b}_j z^j$ such that $b(w_m^i) = \hat{b}(w_m^i)$, $i = -n+1, \dots, n$;

5: Compute $\hat{b}''(1) = \sum_{j=-n+1}^{n} j(j-1)\hat{b}_j$ and $\delta_m = b''(1) - \hat{b}''(1)$;

6: If $\delta_m < \epsilon$ then exit, else set $n = 2n$ and compute from Step 2.

It can be seen that the overall computational cost of Algorithm 1 is $O(n \log n)$ arithmetic operations. Now the Toeplitz part of matrix B is approximated by $T(\hat{b})$, it remains to compute the correction part of B in order to complete the computation of the square root. We show this subject in next section.

3 Computing the Correction Part

Suppose $A = \beta(I - A_1) \in \mathcal{QT}_\infty$, where $A_1 \ge 0$ and $\|A_1\|_\infty < 1$, then for $B = T(b) + E_B \ge 0$ and $\|B\|_\infty < 1$ such that $(I - B)^2 = I - A_1$, we design and analyze the convergence of a fixed-point iteration and a structure-preserving doubling algorithm that can be used for the computation of E_B.

3.1 Fixed-Point Iteration

Consider the nonlinear matrix equation

$$(I - T(b) - X)^2 = I - A_1$$

which can be equivalently written as

$$X^2 - (I - T(b))X - X(I - T(b)) + Q = 0, \tag{3.1}$$

where $Q = A_1 + T(b)^2 - 2T(b)$. It is clear that E_B solves Eq. (3.1). On the other hand, it follows from Lemma 1.2 that $I - A_1$ allows a unique quasi-Toeplitz M-matrix as a square root, so that E_B is the unique solution of Eq. (3.1) such that $T(b)+E_B \geq 0$ and $\|T(b)+E_B\|_\infty < 1$.

Observe that Eq. (3.1) can be equivalently written as $X = (2I - T(b) - X)^{-1}(Q + XT(b))$, from which we propose the following iteration

$$X_{k+1} = (2I - T(b) - X_k)^{-1}(Q + X_k T(b)) \tag{3.2}$$

with $X_0 = 0$. We show that the sequence $\{X_k\}$ converges to E_B. To this end, we first show the following result.

Theorem 3.1 *Let $A = \beta(I - A_1) \in \mathcal{QT}_\infty$ with $A_1 \geq 0$ and $\|A_1\|_\infty < 1$. Suppose $B = T(b) + E_B \in \mathcal{QT}_\infty$ is the unique quasi-Toeplitz matrix such that $B \geq 0$, $\|B\|_\infty < 1$, and $(I - B)^2 = I - A_1$. Then, the sequence $\{X_k\}$ generated by iteration (3.2) satisfies*

(i) the sequence $\{X_k\}$ is well defined;
(ii) $T(b) + X_k \geq 0$ and $\|T(b) + X_k\|_\infty < 1$.

Proof Concerning item (i), observe that X_{k+1} is well defined as long as $2I - T(b) - X_k$ is invertible. It follows from [16, Lemma 3.1.5] that $2I - T(b) - X_k$ is invertible if $\|T(b) + X_k\|_\infty < 2$, which can be verified if item (ii) is true. Hence, it suffices to prove item (ii).

We prove item (ii) by induction. For $k = 0$, we have $T(b) + X_0 = T(b) \geq 0$, where the inequality follows from property (iii) of Lemma 1.1 and the fact $B \geq 0$. On the other hand, we have from property (ii) of Lemma 1.1 that $\|T(b)+X_0\|_\infty \leq \|B\|_\infty < 1$. For the inductive step, assume that $T(b) + X_k \geq 0$ and $\|T(b) + X_k\|_\infty < 1$, we show that $T(b) + X_{k+1} \geq 0$ and $\|T(b) + X_{k+1}\|_\infty < 1$.

Observe that

$$\begin{aligned} X_{k+1} &= (2I - T(b) - X_k)^{-1}(A_1 - (2I - T(b) - X_k)T(b)) \\ &= (2I - T(b) - X_k)^{-1}A_1 - T(b), \end{aligned}$$

from which we have

$$T(b) + X_{k+1} = (2I - T(b) - X_k)^{-1}A_1.$$

On the other hand, we have the following Neumann series expansion

$$(2I - T(b) - X_k)^{-1} = \frac{1}{2}\sum_{i=0}^{\infty}\left(\frac{1}{2}(T(b) + X_k)\right)^i,$$

so that $(2I - T(b) - X_k)^{-1} \geq 0$ since $T(b) + X_k \geq 0$. Recall that $A_1 \geq 0$, we thus have $(2I - T(b) - X_k)^{-1}A_1 \geq 0$, that is, $T(b) + X_{k+1} \geq 0$.

It remains to show $\|T(b)+X_{k+1}\|_\infty < 1$. Observe that

$$\begin{aligned}\|T(b)+X_{k+1}\|_\infty &= \|(2I-T(b)-X_k)^{-1}A_1\|_\infty \\ &\le \|(2I-T(b)-X_k)^{-1}\|_\infty\|A_1\|_\infty \\ &\le \frac{\|A_1\|_\infty}{2-\|T(b)+X_k\|_\infty},\end{aligned}$$

where the last inequality holds since

$$\begin{aligned}\|(2I-T(b)-X_k)^{-1}\|_\infty &\le \frac{1}{2}\sum_{i=0}^{\infty}\big(\frac{1}{2}\|T(b)+X_k\|_\infty\big)^i \\ &= \frac{1}{2-\|T(b)+X_k\|_\infty}.\end{aligned} \tag{3.3}$$

Recall that $\|T(b)+X_k\|_\infty < 1$ and $\|A_1\|_\infty < 1$, one can check that

$$\frac{\|A_1\|_\infty}{2-\|T(b)+X_k\|_\infty} < 1,$$

that is, $\|T(b)+X_{k+1}\|_\infty < 1$. □

The following result shows the convergence of sequence $\{X_k\}$.

Theorem 3.2 *Let $A=\beta(I-A_1)\in \mathcal{QT}_\infty$ with $A_1\ge 0$ and $\|A_1\|_\infty<1$. Suppose $B=T(b)+E_B\in\mathcal{QT}_\infty$ is the unique quasi-Toeplitz matrix such that $B\ge 0$, $\|B\|_\infty<1$ and $(I-B)^2=I-A_1$. Then the sequence $\{X_k\}$ generated by iteration (3.2) converges to E_B in the sense that $\lim_{k\to\infty}\|E_B-X_k\|_\infty=0$.*

Proof Let $W_k=E_B-X_k$, a direct computation yields

$$W_{k+1}=(2I-T(b)-X_k)^{-1}W_kB,$$

which, together with (3.3), yields

$$\|W_{k+1}\|_\infty \le \frac{\|B\|_\infty}{2-\|T(b)+X_k\|_\infty}\|W_k\|_\infty. \tag{3.4}$$

Since $\|T(b)+X_k\|_\infty<1$, it follows that $\frac{\|B\|_\infty}{2-\|T(b)+X_k\|_\infty}<\|B\|_\infty$, so that

$$\|W_{k+1}\|_\infty \le \|B\|_\infty\|W_k\|_\infty \le \|B\|_\infty^k\|W_0\|_\infty.$$

Since $\|B\|_\infty<1$, it implies that $\lim_{k\to\infty}\|E_B-X_k\|_\infty=0$. □

We may observe from inequality (3.4) that the sequence $\{X_k\}$ generated by iteration (3.2) satisfies $\|X_{k+1}-E_B\|_\infty \le \frac{\|B\|_\infty}{2-\|T(b)+X_k\|_\infty}\|X_k-E_B\|_\infty$. The fact $\frac{\|B\|_\infty}{2-\|T(b)+X_k\|_\infty}<\|B\|_\infty$ may provide some insights to say that the fixed-point iteration (3.2), which is used for the computation of the correction part, converges faster than the Binomial iteration [22] in the computation of the whole square root, as the sequence $\{Y_k\}$ generated by the Binomial iteration $Y_{k+1}=\frac{1}{2}(A_1+Y_k^2)$ with $Y_0=0$ satisfies that $\|Y_{k+1}-B\|_\infty\le\|B\|_\infty\|Y_k-B\|_\infty$.

3.2 Structure-Preserving Doubling Algorithm

We show that a structure-preserving doubling algorithm (SDA) is applicable in the computation of E_B such that $(I - T(b) - E_B)^2 = A$, where A is an invertible quasi-Toeplitz M-matrix. This method has been motivated by the ideas in [5], where the SDA that enables refining an initial approximation is applied to solve quadratic matrix equations with quasi-Toeplitz coefficients. We fist recall the design and convergence analysis of SDA. For more details of SDA, we refer the reader to [5], [4, Chapter 5] and [15].

In the finite dimensional space, the design of SDA is based on a linear pencil $M - \lambda N$, where M and N are $2n \times 2n$ matrices of the form

$$M = \begin{bmatrix} E & O \\ -P & I \end{bmatrix}, \quad N = \begin{bmatrix} I & -Q \\ O & F \end{bmatrix}, \tag{3.5}$$

where E, F, P, Q are $n \times n$ matrices, I and O are, respectively, the $n \times n$ identity matrix and the zero matrix. Suppose there are $n \times n$ matrices X and W such that

$$M \begin{bmatrix} I \\ X \end{bmatrix} = N \begin{bmatrix} I \\ X \end{bmatrix} W,$$

The columns of $\begin{bmatrix} I \\ X \end{bmatrix}$ is said to span a graph deflating subspace of the pencil $M - \lambda N$ associated with the eigenvalues of W [5]. Consider the problem of computing the matrix X, which is equivalent to compute a graph deflating subspace of the pencil $M - \lambda N$ associated with the eigenvalues of W, a new pencil $M_k - \lambda N_k$ such that

$$M_k \begin{bmatrix} I \\ X \end{bmatrix} = N_k \begin{bmatrix} I \\ X \end{bmatrix} W^{2^k}, \tag{3.6}$$

is constructed, where the matrix sequences $\{M_k\}$ and $\{N_k\}$ are generated such that for $k = 0, 1, 2, \ldots$, $\det N_k \neq 0$, $N_{k+1}^{-1} M_{k+1} = (N_k^{-1} M_k)^2$ with $N_0 = N$, $M_0 = M$. If M and N have the form as in (3.5), it follows from [4, page 148] that

$$M_k = \begin{bmatrix} E_k & O \\ -P_k & I \end{bmatrix}, \quad N_k = \begin{bmatrix} I & -Q_k \\ O & F_k \end{bmatrix},$$

where $E_0 = E$, $F_0 = F$, $P_0 = P$, $Q_0 = Q$, and

$$\begin{aligned} E_{k+1} &= E_k (I - Q_k P_k)^{-1} E_k, \\ P_{k+1} &= P_k + F_k (I - P_k Q_k)^{-1} P_k E_k, \\ F_{k+1} &= F_k (I - P_k Q_k)^{-1} F_k, \\ Q_{k+1} &= Q_k + E_k (I - Q_k P_k)^{-1} Q_k F_k. \end{aligned} \tag{3.7}$$

If $\rho(W) < 1$ and the sequence $\{N_k\}$ is uniformly bounded, then it can be seen from (3.6) that $\lim_{k\to\infty} M_k \begin{bmatrix} I \\ X \end{bmatrix} = 0$, from which we obtain $\lim_{k\to\infty} P_k = X$. The algorithm based on the above technique for computing the matrix X is known as SDA. That is, SDA consists in computing the sequences defined in (3.7), and under suitable convergence properties, as shown in Lemma 3.1, the sequence $\{P_k\}$ converges to the matrix X.

We mention that the scheme (3.7) is quite related to the forms of matrices M and N in (3.5), which is called the standard structured form-I. For different forms, say the standard structured form-II (see [4, Chapter 5]), different schemes can be obtained.

Concerning the convergence results of SDA, it has been proved in [5] that

Lemma 3.1 *[5, Theorem 2] Let X, Y, W, V be $n \times n$ matrices such that*

$$M \begin{bmatrix} I \\ X \end{bmatrix} = N \begin{bmatrix} I \\ X \end{bmatrix} W, \quad M \begin{bmatrix} Y \\ I \end{bmatrix} V = N \begin{bmatrix} Y \\ I \end{bmatrix},$$

and it satisfies that $\rho(W) \leq 1$, $\rho(V) \leq 1$, $\rho(W)\rho(V) < 1$. If the scheme (3.7) can be carried out with no breakdown, then $\lim_k \|X - P_k\|^{1/2^k} \leq \rho(W)\rho(V)$ and $\lim_k \|Y - Q_k\|^{1/2^k} \leq \rho(W)\rho(V)$.

Concerning the feasibility of SDA in the infinite dimensional spaces, it has been shown in [5, page 11] that the convergence results of SDA still hold when matrices belong to the Banach algebra $\mathcal{QT}_\infty$. We are ready to show how SDA can be applied in the computation of E_B.

Suppose $A = I - A_1 \in \mathcal{QT}_\infty$ is such that $A_1 \geq 0$ and $\|A_1\|_\infty < 1$, we have from Lemma 1.2 that the matrix equation

$$(I - X)^2 = I - A_1 \tag{3.8}$$

has a unique nonnegative solution $B \in \mathcal{QT}_\infty$ satisfying $\|B\|_\infty < 1$. Observe that equation (3.8) can be equivalently written as

$$X^2 - 2X + A_1 = 0, \tag{3.9}$$

so that B solves Eq. (3.9) and is the unique solution such that $B \geq 0$ and $\|B\|_\infty < 1$. Let $V = (2I - B)^{-1}$, it is easy to check that V solves the quadratic matrix equation

$$A_1 Y^2 - 2Y + I = 0. \tag{3.10}$$

Moreover, we have $V = \frac{1}{2} \sum_{i=0}^{\infty} (\frac{1}{2} B)^i \geq 0$ and $\|V\|_\infty \leq \frac{1}{2} \sum_{i=1}^{\infty} (\frac{1}{2} \|B\|_\infty)^i = \frac{1}{2 - \|B\|_\infty} < 1$.

Suppose $T(b)$ with $b \in \mathcal{W}$ is the Toeplitz part of B, replacing X by $T(b) + H$ in Eq. (3.9) results in the following quadratic matrix equation

$$H^2 + (T(b) - 2I)H + HT(b) + R = 0, \tag{3.11}$$

where $R = T(b)^2 - 2T(b) + A_1$. Then, equation (3.11) can be equivalently written as

$$\widetilde{M} \begin{bmatrix} I \\ H \end{bmatrix} = \widetilde{N} \begin{bmatrix} I \\ H \end{bmatrix} B,$$

where $\widetilde{M} = \begin{bmatrix} T(b) & I \\ -R & 2I - T(b) \end{bmatrix}$ and $\widetilde{N} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$

On the other hand, according to [5, Theorem 3], the pencil $\widetilde{M} - \lambda \widetilde{N}$ can be transformed into the pencil $\mathcal{M} - \lambda \mathcal{N}$, where $\mathcal{M}$ and $\mathcal{N}$ are of the form

$$\mathcal{M} = \begin{bmatrix} SA_1 & 0 \\ -SR & I \end{bmatrix}, \quad \mathcal{N} = \begin{bmatrix} I & -S \\ 0 & S \end{bmatrix},$$

where $S = (2I - T(b))^{-1}$. It can be seen that $\mathcal{M}$ and $\mathcal{N}$ are of the same forms as those in (3.5), and we have

$$\mathcal{M} \begin{bmatrix} I \\ H \end{bmatrix} = \mathcal{N} \begin{bmatrix} I \\ H \end{bmatrix} B,$$

so that SDA can be applied to compute the matrix H, which consists of computing the sequences as defined in the scheme (3.7) by setting

$$P_0 = SR, \quad E_0 = P_0 + T(b), \quad \text{and} \quad Q_0 = F_0 = S.$$

On the other hand, it can be verified that the matrices $\mathcal{M}$ and $\mathcal{N}$ also satisfy

$$\mathcal{M}\begin{bmatrix} Y \\ I \end{bmatrix} Z = \mathcal{N}\begin{bmatrix} Y \\ I \end{bmatrix}, \tag{3.12}$$

where $Y = V(I - T(b)V)^{-1}$, $Z = (I - T(b)V)V(I - T(b)V)^{-1}$. It can be seen that Z has the same spectrum as V so that $\rho(Z) = \rho(V) \leq \|V\|_\infty < 1$, we then have from the fact $\rho(B) \leq \|B\|_\infty < 1$ that $\rho(B)\rho(Z) < 1$. Hence, according to Lemma 3.1, we obtain the following convergence result of SDA in solving Eq. (3.11).

Theorem 3.3 *For $A = I - A_1 \in \mathcal{QT}_\infty$ such that $A_1 \geq 0$ and $\|A_1\|_\infty < 1$, suppose $I - B$ with $B = T(b) + E_B \in \mathcal{QT}_\infty$ is the unique quasi-Toeplitz M-matrix such that $(I - B)^2 = A$. If the scheme (3.7) can be carried out with no breakdown, then the sequence $\{P_k\}$ converges to E_B and it satisfies $\lim_k \|E_B - P_k\|^{1/2^k} \leq \rho(B)\rho(Z)$, where $Z = (I - T(b)V)(2I - B)^{-1}(I - T(b)V)^{-1}$ and $V = (2I - B)^{-1}$.*

Actually, according to the ideas in [5], the scheme (3.7) allows to refine a given initial approximation to E_B, that is, if $E_B = \tilde{E}_B + D$, where $\tilde{E}_B$ is given and it satisfies $\|T(b) + \tilde{E}_B\|_\infty < 1$, then SDA can be used to compute D. Indeed, if H in Eq. (3.11) is replaced by $\tilde{E}_B + D$, it yields

$$D^2 + (T(b) + \tilde{E}_B - 2I)D + D(T(b) + \tilde{E}_B) + \tilde{R} = 0, \tag{3.13}$$

where $\tilde{R} = (T(b) + \tilde{E}_B)^2 - 2(T(b) + \tilde{E}_B) + A_1$. Analogously to the analysis above, we obtain the matrix pencil $\widehat{\mathcal{M}} - \lambda\widehat{\mathcal{N}}$ such that

$$\widehat{\mathcal{M}} = \begin{bmatrix} \tilde{S}A_1 & 0 \\ -\tilde{S}\tilde{R} & I \end{bmatrix}, \quad \widehat{\mathcal{N}} = \begin{bmatrix} I & -\tilde{S} \\ 0 & \tilde{S} \end{bmatrix},$$

where $\tilde{S} = (2I - T(b) - \tilde{E}_B)^{-1}$, and it holds

$$\widehat{\mathcal{M}}\begin{bmatrix} I \\ D \end{bmatrix} = \widehat{\mathcal{N}}\begin{bmatrix} I \\ D \end{bmatrix} B, \quad \widehat{\mathcal{M}}\begin{bmatrix} \tilde{Y} \\ I \end{bmatrix} \tilde{Z} = \widehat{\mathcal{N}}\begin{bmatrix} \tilde{Y} \\ I \end{bmatrix},$$

where $\tilde{Y} = V(I - (T(b) + \tilde{E}_B)V)^{-1}$, $\tilde{Z} = (I - (T(b) + \tilde{E}_B)V)V(I - (T(b) + \tilde{E}_B)V)^{-1}$.

Now we set

$$\widehat{\mathcal{M}}_k = \begin{bmatrix} \tilde{E}_k & O \\ -\tilde{P}_k & I \end{bmatrix}, \quad \widehat{\mathcal{N}}_k = \begin{bmatrix} I & -\tilde{Q}_k \\ O & \tilde{F}_k \end{bmatrix},$$

where $\tilde{P}_0 = \tilde{S}\tilde{R}$, $\tilde{E}_0 = \tilde{S}A_1$, $\tilde{Q}_0 = \tilde{F}_0 = \tilde{S}$, and

$$\begin{aligned} \tilde{E}_{k+1} &= \tilde{E}_k(I - \tilde{Q}_k\tilde{P}_k)^{-1}\tilde{E}_k \\ \tilde{P}_{k+1} &= \tilde{P}_k + \tilde{F}_k(I - \tilde{P}_k\tilde{Q}_k)^{-1}\tilde{P}_k\tilde{E}_k; \\ \tilde{F}_{k+1} &= \tilde{F}_k(I - \tilde{P}_k\tilde{Q}_k)^{-1}\tilde{F}_k; \\ \tilde{Q}_{k+1} &= \tilde{Q}_k + \tilde{E}_k(I - \tilde{Q}_k\tilde{P}_k)^{-1}\tilde{Q}_k\tilde{F}_k. \end{aligned} \tag{3.14}$$

Then we obtain a new pencil $\widehat{\mathcal{M}}_k - \lambda\widehat{\mathcal{N}}_k$ such that

$$\widehat{\mathcal{M}}_k \begin{bmatrix} I \\ D \end{bmatrix} = \widehat{\mathcal{N}}_k \begin{bmatrix} I \\ D \end{bmatrix} B^{2^k}.$$

Since $\rho(B) \le \|B\|_\infty < 1$, if in addition the sequence $\{\widehat{N}_k\}$ is uniformly bounded, we have $\lim_{k\to\infty} \widehat{\mathcal{M}}_k \begin{bmatrix} I \\ D \end{bmatrix} = 0$, from which we obtain $\lim_{k\to\infty} \tilde{P}_k = D$.

Hence, SDA can be applied to solve Eq. (3.13), which consists in computing the sequences defined in (3.14). Observe that $\rho(\tilde{Z}) = \rho(V) < 1$, then according to Lemma 3.1 it holds that $\lim_k \|\tilde{P}_k - D\|_\infty^{1/2^k} < \rho(B)\rho(V) < 1$, that is, the sequence $\{\tilde{P}_k\}$ converges to D, so that $E_B = \tilde{E}_B + D$ is computed.

One alternative is to set $\tilde{E}_B = (b(1)\mathbf{1} - T(b)\mathbf{1})e_1^T$, where $\mathbf{1} = (1, 1, \ldots)^T$ and $e_1 = (1, 0, \ldots)^T$, then $T(b) + \tilde{E}_B$ is a nonnegative substochastic matrix such that $(T(b) + \tilde{E}_B)\mathbf{1} = b(1)\mathbf{1}$. Numerical experiments in Sect. 5 shows that there are cases where a reduction in CPU time occurs when setting $\tilde{E}_B = (b(1)\mathbf{1} - T(b)\mathbf{1})e_1^T$ and applying iteration (3.14) for computing D.

We mention that when applying the fixed-point iteration and SDA to compute the correction part of a quasi-Toeplitz M-matrix, the computations rely on the package CQT-Toolbox of [9] which implements the operations of semi-infinite quasi-Toeplitz matrices. In next section, we show that the fixed-point iteration and SDA can be applied to a finite dimensional nonlinear matrix equation, whose solution after extending to infinity is a good approximation to E_B.

4 Truncation to a Finite Dimensional Matrix Equation

Recall that the correction part of a quasi-Toeplitz matrix $A = T(a) + E \in \mathcal{QT}_\infty$ satisfies $\lim_i \sum_{j=1}^\infty |e_{i,j}| = 0$ for $E = (e_{i,j})_{i,j\in\mathbb{Z}^+}$. Denote by $E^{(k)}$ the infinite matrix that coincides with the leading principal $k \times k$ submatrix of E and is zero elsewhere, it follows form [11, Lemma 2.9] that there is a matrix $E^{(k)}$ such that $\lim_{k\to\infty} \|E - E^{(k)}\|_\infty = 0$.

For an invertible M-matrix $A = I - A_1 \in \mathcal{QT}_\infty$, suppose $(I - T(b) - E_B)^2 = A$, then for E_B and a given $\epsilon > 0$, there is a sufficiently large k such that

$$\|E_B^{(k)} - E_B\|_\infty < \epsilon. \tag{4.1}$$

If we partition E_B into $E_B = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}$, where E_{11} is the principal $k \times k$ submatrix of E_B, $E_{12} \in \mathbb{R}^{k\times\infty}$, $E_{21} \in \mathbb{R}^{\infty\times k}$ and $E_{22} \in \mathbb{R}^{\infty\times\infty}$, it follows from $\|E_B^{(k)} - E_B\|_\infty < \epsilon$ that $\|E_{12}\|_\infty < \epsilon$, $\|E_{21}\|_\infty < \epsilon$ and $\|E_{22}\|_\infty < \epsilon$.

Let $W = 2T(b) - A_1 - T(b)^2$, then $T(b)$ and W can be partitioned into $T(b) = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}$ and $W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}$, where T_{11} and W_{11} are, respectively, the principal $k \times k$ submatrices of $T(b)$ and W. Substituting E_B, $T(b)$ and W into the equation $(I - T(b) - E_B)^2 = I - A_1$, we get

$$E_{11}^2 - (I_k - T_{11})E_{11} - E_{11}(I_k - T_{11}) = W_{11} - E_{12}E_{21} - E_{12}T_{21} - T_{12}E_{21}, \tag{4.2}$$

where I_k is the identity matrix of size k.

Consider the matrix equation

$$G^2 - (I_k - T_{11})G - G(I_k - T_{11}) = W_{11}, \tag{4.3}$$

which is equivalent to

$$(I_k - T_{11} - G)^2 = I - A_{11} - T_{12}T_{21}, \tag{4.4}$$

where A_{11} is the principal $k \times k$ submatrix of A_1. Observe that $A_{11} \geq 0$ and $T_{12}T_{21} \geq 0$, if in addition $\rho(A_{11} + T_{12}T_{21}) < 1$, which can be verified if $\|A_{11} + T_{12}T_{21}\|_\infty < 1$, then $I - A_{11} - T_{12}T_{21}$ is a nonsingular M-matrix. In what follows we assume $\|A_{11} + T_{12}T_{21}\|_\infty < 1$, then $I - A_{11} - T_{12}T_{21}$ admits a unique M-matrix as a square root (see [14, Theorem 6.18]), so that Eq. (4.4), as well as Eq. (4.3), has a unique solution G such that $T_{11} + G \geq 0$ and $\rho(T_{11} + G) < 1$. In fact, analogously to [22, Theorem 3.1], it is can be seen that $\|T_{11} + G\|_\infty < 1$.

Subtracting Eq. (4.2) form Eq. (4.3) yields

$$G^2 - E_{11}^2 - (G - E_{11})(I_k - T_{11}) - (I_k - T_{11})(G - E_{11}) = \Delta W, \tag{4.5}$$

where $\Delta W = E_{12}E_{21} + E_{12}T_{21} + T_{12}E_{21}$. It can be seen that

$$\begin{aligned} \|\Delta W\|_\infty &= \|E_{12}E_{21} + E_{12}T_{21} + T_{12}E_{21}\|_\infty \\ &\leq \epsilon^2 + \|T_{21}\|_\infty \epsilon + \|T_{12}\|_\infty \epsilon \\ &\leq (2\|b\|_{\mathcal{W}} + \epsilon)\epsilon, \end{aligned} \tag{4.6}$$

where the last inequality holds as $\|T_{12}\|_\infty \leq \|T(b)\|_\infty = \|b\|_{\mathcal{W}}$ and $\|T_{21}\|_\infty \leq \|T(b)\|_\infty = \|b\|_{\mathcal{W}}$.

On the other hand, a direct computation of Eq. (4.5) yields

$$(2I_k - T_{11} - G)(G - E_{11}) - (G - E_{11})(T_{11} + E_{11}) = -\Delta W. \tag{4.7}$$

Observe that $2I_k - T_{11} - G$ is a nonsingular M-matrix as $T_{11} + G \geq 0$ and $\rho(T_{11} + G) < 1$. Moreover, we have $\|T_{11} + E_{11}\|_\infty < 1$ as $T_{11} + E_{11}$ is the principal $k \times k$ submatrix of $T(b) + E_B$ and $\|T(b) + E_B\|_\infty < 1$. Then one can check that

$$G - E_{11} = -\sum_{j=0}^{\infty} (2I_k - T_{11} - G)^{-j-1} \Delta W (T_{11} + E_{11})^j \tag{4.8}$$

is well defined and it solves Eq. (4.7).

Let $\alpha = \|(2I_k - T_{11} - G)^{-1}\|_\infty$ and $\beta = \|T_{11} + E_{11}\|_\infty$, we have $\alpha = \frac{1}{2}\| \sum_{j=0}^{\infty} (\frac{1}{2}(T_{11} + G))^j \|_\infty \leq \frac{1}{2 - \|T_{11}+G\|_\infty}$, so that $\alpha\beta \leq \frac{\beta}{2 - \|T_{11}+G\|_\infty} < 1$ since $\|T_{11} + G\|_\infty < 1$ and $\beta < 1$. Then we deduce from (4.6) and (4.8) that

$$\begin{aligned} \|G - E_{11}\|_\infty &\leq \sum_{j=0}^{\infty} (\alpha\beta)^j \alpha \|\Delta W\|_\infty \\ &\leq \frac{\alpha}{1 - \alpha\beta}(2\|b\|_w + \epsilon)\epsilon. \end{aligned} \tag{4.9}$$

Let E_G be the matrix that coincides in the leading principal $k \times k$ submatrix with G and is zero elsewhere, then we have from (4.1) and (4.9) that

$$\begin{aligned} \|E_G - E_B\|_\infty &\leq \|E_G - E_B^{(k)}\|_\infty + \|E_B^{(k)} - E_B\|_\infty \\ &\leq \|G - E_{11}\|_\infty + \epsilon \end{aligned}$$

$$\leq (1 + \frac{\alpha}{1-\alpha\beta}(2\|b\|_w + \epsilon))\epsilon. \tag{4.10}$$

Hence, we can see from (4.10) that for a given $\epsilon > 0$ and sufficiently large k, if $\alpha\beta \leq c < 1$ for some constant c, then E_G may serve as a good approximation to E_B. This implies that the correction part E_B can be approximated by firstly computing the numerical solution of Eq. (4.3) and then extending the computed solution to infinity.

It is not difficult to see that the fixed-point iteration (3.2) and SDA can be applied to Eq. (4.3) for computing the solution G. Numerical experiments in next section show that when the size k is small, it is efficient to approximate the correction part E_B by computing the solution of Eq. (4.3) and extending it to infinity, while when k is large, that is, the coefficients are large-scale matrices, both fixed-point iteration and SDA lose the effectiveness.

We provide some insight on how to select integer k such that the matrix G of size $k \times k$, after extending to infinity, is approximate enough to E_B. Observe that the substitution of E_G into the equation $(I - T(b) - X)^2 = A$ yields

$$A - (I - T(b) - E_G)^2 = \begin{pmatrix} 0 & GT_{12} - W_{12} \\ T_{21}G - W_{21} & -W_{22}, \end{pmatrix},$$

from which we see that E_G is a good approximation to E_B if $\|GT_{12} - W_{12}\|_\infty < c\epsilon$, $\|T_{21}G - W_{21}\|_\infty < c\epsilon$ and $\|W_{22}\|_\infty < c\epsilon$ for some constant c and a given $\epsilon > 0$. It can be seen that these inequalities hold if

$$\|GT_{12}\| < c_1\epsilon, \tag{4.11}$$

$$\|T_{21}G\|_\infty < c_2\epsilon, \tag{4.12}$$

and

$$\max\{\|W_{12}\|, \|W_{21}\|, \|W_{22}\|_\infty\} < c_3\epsilon, \tag{4.13}$$

for some constants c_1, c_2 and c_3. Hence, we can choose k such that inequalities (4.11)–(4.13) are satisfied.

Actually, since W is a correction matrix, one can check that inequality (4.13) holds if we choose k such that $\|W - W^{(k)}\|_\infty < \epsilon$, where $W^{(k)}$ is the infinite matrix that coincides with the leading principal $k \times k$ submatrix of W and is zero elsewhere. Hence, if the matrix W has a nonzero part of size $n_1 \times n_2$, we can choose k such that $k > \max\{n_1, n_2\}$.

We next show how to choose k such that inequalities (4.11) and (4.12) hold. Observe that for $\epsilon > 0$, there is $N \in \mathbb{Z}^+$ such that $\|E_B - E_B^{(n)}\|_\infty < \epsilon$ for any $n \geq N$. Set $k > N$ and $G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} \in \mathbb{R}^{k\times k}$, where $G_{11} \in \mathbb{R}^{N\times N}$, $G_{12} \in \mathbb{R}^{N\times(k-N)}$, $G_{21} \in \mathbb{R}^{(k-N)\times N}$ and $G_{22} \in \mathbb{R}^{(k-N)\times(k-N)}$. Observe that

$$\|E_G - E_B^{(N)}\|_\infty \leq \|E_G - E_B\|_\infty + \|E_B - E_B^{(N)}\|_\infty,$$

which, together with inequality (4.10) and the fact $\|E_B - E_B^{(N)}\|_\infty < \epsilon$, implies that $\|E_G - E_B^{(N)}\|_\infty < \tilde{c}_1\epsilon$ for some constant $\tilde{c}_1$. On the other hand, observe that $E_G - E_B^{(N)}$ coincides in the leading principal $k \times k$ submatrix with $\begin{pmatrix} * & G_{12} \\ G_{21} & G_{22} \end{pmatrix}$ and is zero elsewhere, where $*$ is an $N \times N$ matrix, we thus have $\|G_{12}\|_\infty < \tilde{c}_1\epsilon$, $\|G_{21}\|_\infty < \tilde{c}_1\epsilon$ and $\|G_{22}\|_\infty < \tilde{c}_1\epsilon$.

Suppose $b(z) = \sum_{j=-q}^{p} b_j z^j$, then from the partition of $T(b)$ we know that $T_{12} = \begin{pmatrix} O \\ \tilde{T} \end{pmatrix}$, where O is a zero matrix of size $(k-p)\times\infty$ and $\tilde{T}$ is a $p\times\infty$ matrix with a $p\times p$ nonzero submatrix located in the bottom leftmost corner. If k is selected such that $k-N > p$, we have from

$\|G_{12}\|_\infty < \tilde{c}_1\epsilon$ and $\|G_{21}\|_\infty < \tilde{c}_1\epsilon$ that $\|GT_{12}\|_\infty \le \max\{\|G_{12}\|_\infty, \|G_{21}\|_\infty\}\|\tilde{T}\|_\infty < c_1\epsilon$ for some constant c_1. Similarly, if $k - N > q$, inequality (4.12) holds.

The above analysis indicates that if the matrix W has a nonzero part of size $n_1 \times n_2$ and the symbol $b(z)$ of $T(b)$ is a Laurent series $b(z) = \sum_{j=-q}^{p} b_j z^j$, then we can choose k such that

$$k - N > p,\ K - N > q \text{ and } k > \max\{n_1, n_2\}. \tag{4.14}$$

Observe that the value of N in (4.14) is unknown, hence we can obtain a necessary condition for determining k, that is, $k > \max\{p, q, n_1, n_2\}$. In our numerical experiments, we have set $k = 3\max\{p, q, n_1, n_2\}$ and it seems sufficient.

Note that equation (4.2) is a special case of the following equation

$$X^2 - AX - XA = B,$$

where A is a large-scale nonsingular M-matrix with an almost Toeplitz structure, and B is a low-rank matrix. It seems interesting to investigate whether there are more efficient algorithms for computing the solution by exploiting the quasi-Toeplitz structure of A and the low-rank structure of matrix B. We leave this as a future consideration.

5 Numerical Experiments

In this section, we show by numerical experiments the effectiveness of the fixed-point iteration (3.2) and SDA. The computations of semi-infinite quasi-Toeplitz matrices rely on the package CQT-Toolbox [9], which can be downloaded at https://github.com/numpi/cqt-toolbox, while computation of the solution of Eq. (4.3) is implemented relying on the standard finite size matrix operations. The tests were performed in MATLAB/version R2019b on the Dell Precision 5570 with an Intel Core i9-12900 H and 64 GB main memory. We set the internal precision in the computations to `threshold = 1.e-15`. For each experiment, the iteration is terminated if

$$\|(I - T(b) - X)^2 - A\|_\infty / \|A\|_\infty \le \mathtt{1.e-13}.$$

The codes are available at https://github.com/JieMeng00/structured_sqrtm_square_root_m-matrices.

We recall that a quasi-Toeplitz matrix $A = T(a) + E_A$ is representable in MATLAB relying on the CQT-toolbox [9] by `A=cqt(an,ap,E)`, where the vectors `an` and `ap` contain the coefficients of the symbol $a(z)$ with non negative and non positive indices, respectively, and E is a finite matrix representing the non zero part of the correction E_A.

Example 5.1 Let $A = I - S$ with $S = \tilde{S}/(\|\tilde{S}\|_\infty + 1)$, where the construction of $\tilde{S}$ in MATLAB is done as $\tilde{\mathtt{S}} = \mathtt{cqt}(\mathtt{s_n}, \mathtt{s_p}, \mathtt{E}_{\tilde{\mathtt{S}}})$. We set $\mathtt{s_n} = \mathtt{rand(32,1)}$, $\mathtt{s_p} = \mathtt{rand(30,1)}$, $\mathtt{s_{n(1)}} = \mathtt{s_{p(1)=1}}$, for the first test, we set $\mathtt{E}_{\tilde{\mathtt{S}}} = \mathtt{0}$, while for the second test, we set $\mathtt{E}_{\tilde{\mathtt{S}}} = \mathtt{rand(1000,1000)}$.

Suppose $B = T(b) + E_B$ is such that $(I - B)^2 = A$, we first compute by Algorithm 1 an approximation $\hat{b}(z) = \sum_{j=-n+1}^{n} \hat{b}_j z^j$ to the symbol $b(z)$ of $T(b)$, then we apply the fixed-point iteration (3.2) and SDA to compute E_B. In Fig. 1 we show the graph of the computed coefficients $\hat{b}_j$, $j = -n+1, \dots, n$. In Fig. 2, we show the correction part $E_B = (e_{i,j})_{i,j \in \mathbb{Z}^+}$ in logarithmic scale, which is obtained by the fixed-point iteration. The number of iterations, CPU times required in the computations and the relative residuals are reported in Table 1.

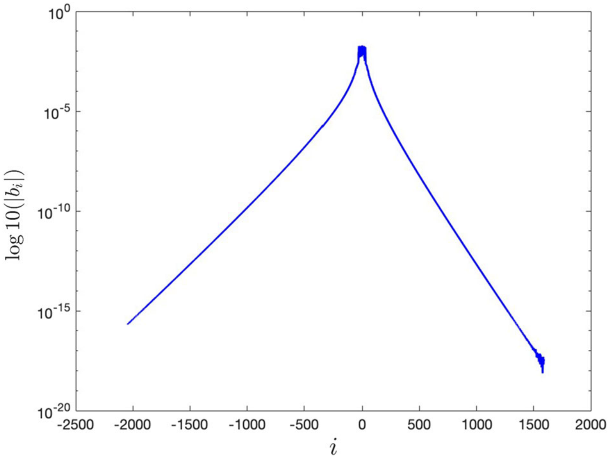
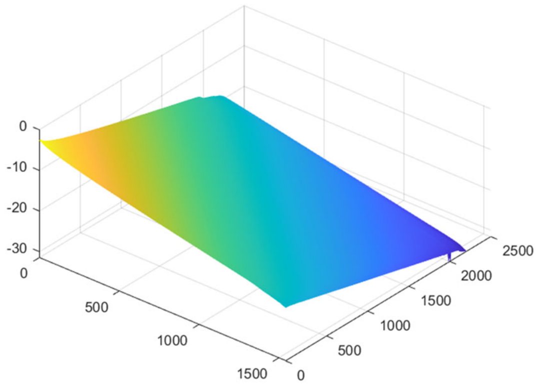


Fig. 1 Toeplitz part of the computed $B = T(b)+E_B$ in Test 1: the log-scale of the absolute value of coefficients b_i of the symbol $b(z) = \sum_{i\in\mathbb{Z}} b_i z^i$. The coefficients are computed by Algorithm 1

Fig. 2 The correction part E_B in Test 1: absolute value of E_B in log scale, where E_B is computed by the fixed-point iteration (3.2)



In Table 2 we report the features of the computed matrix $B = T(b) + E_B$, where E_B is computed by the fixed-point iteration, including band of the Toeplitz part, the rank and the number of the nonzero rows and columns of the correction part.

It can be seen from Table 1 that the number of iterations required by SDA is much less than the number of iterations required by the fixed-point iteration. Concerning the CPU time, we can see that the fixed-point iteration takes less time than SDA in Test 1, while in Test 2, the CPU time taken by SDA is about 1/3 of that taken by the fixed-point iteration. Together with the results in Table 2, it seems that the rank of the correction part concerns a lot, that is, when the rank of the correction part is small, it seems that the fixed-point iteration is faster than SDA, but when the correction part is large, SDA is more efficient.

Moreover, in test 1, when applying SDA to compute matrix D such that $E_B = D + \tilde{E}_B$, where $\tilde{E}_B = (s(1)\mathbf{1} - T(s)\mathbf{1})e_1^T$, it takes 119.56s, which provides a reduction in CPU time comparing with the case where SDA is applied directly for the computation of E_B.

The computation of square root of invertible quasi-Toeplitz M-matrices has been implemented in [22] by the Binomial iteration and CR algorithm, respectively. Numerical tests show that the CR algorithm appears to be better suited for quasi-Toeplitz matrices. The fol-

Table 1 Relative residual, number of iterations, CPU time in seconds in the computation E_B. FPI means the fixed-point iteration

	Test 1			Test 2		
Iterations	res	iter	Time	res	iter	Time
FPI	7.02e−14	55	$1.15 \cdot 10^2$	9.62e−14	54	$2.52 \cdot 10^2$
SDA	4.42e−14	6	$1.70 \cdot 10^2$	6.61e−14	6	$8.10 \cdot 10^1$

Table 2 Features of matrix $B = T(b) + E_B$ in Test 1 which is computed by FPI, including the band of the Toeplitz part $T(b)$, number of nonzero rows and columns, and rank of the correction of the computed E_B

	Test 1	Test 2
Band	4200	376
Rows	2799	1296
Columns	1319	1162
Rank	80	1026

Table 3 Different values of the parameters s_0, m, n, p and q

Test	s_0	m	n	p	q
1	0.1	100	1000	1	100
2	0.5	100	1500	2	100
3	0.9	100	2000	2	100

lowing example shows that the fixed-point iteration (3.2) and the SDA have their advantages in computing the square root when decompose the task into the computation of the Toeplitz part and the correction part.

Example 5.2 Let $A = I - S$ with $S = T(s) + E_S \in \mathcal{QT}_\infty$, where $T(s) = s_0 I$ with $s_0 < 1$ and E_B is the correction matrix with a $(p+m+n) \times (p+m+n)$ leading submatrix E_S^P and zero elsewhere. Here, $E_S^P = \begin{pmatrix} V_p & & \\ & O_m & \\ & & -s_0 I_n \end{pmatrix}$, where O_m is the zero matrix of size $m \times m$, I_n is the identity matrix of size n, and the matrix $V_p = \begin{pmatrix} U_{p\times q} \\ O_{(q-p)\times q} \end{pmatrix}$ is a $q \times q$ block matrix with

$$U_{p\times q} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1p} & \cdots & u_{1q} \\ 0 & u_{22} & \cdots & u_{2p} & \cdots & u_{2q} \\ \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{pp} & \cdots & u_{pq} \end{pmatrix}_{p\times q}.$$

where $u_{ii} = -s_0$ for $i = 1, \ldots, p$, and $u_{i,j} \geq 0$ for $i = 1, 2, \ldots, p$ and $j = i+1 \ldots, q$. Moreover, for $i = 1, 2, \ldots, p$, it satisfies that $\sum_{j=i+1}^{q} u_{ij} < 1$.

For different values of the parameters s_0, m, n, p and q as listed in Table 3, we apply the fixed-point iteration (3.2) and SDA to compute the matrix E_B such that $(I - T(b) - E_B)^2 = A$. It can be seen that the symbol $b(z)$ satisfies $(1 - b(z))^2 = 1 - s_0$, which, together with the fact that $\|b\|_{\mathcal{W}} = \|T(b)\|_\infty < 1$, implies $b(z) = 1 - \sqrt{1 - s_0}$, so that $T(b)$ is a diagonal matrix with diagonal elements being $1 - \sqrt{1 - s_0}$.

Table 4 Comparison of the fixed-point iteration (3.2) and SDA in computing E_B with the Binomial iteration and CR algorithm in computing B: the CPU time in seconds and relative residual in the computations

	Test 1		Test 2		Test 3	
Algorithms	Time	res	Time	res	Time	res
FPI	$2.77 \cdot 10^0$	1.01e−15	$1.94 \cdot 10^1$	3.00e−15	$3.43 \cdot 10^1$	3.41e−14
SDA	$8.44 \cdot 10^0$	1.40e−15	$2.53 \cdot 10^1$	2.35e−15	$4.48 \cdot 10^1$	6.79e−14
CR	$1.13 \cdot 10^1$	1.83e−15	$4.50 \cdot 10^1$	4.59e−15	$9.59 \cdot 10^1$	7.48e−14
BI	$1.44 \cdot 10^1$	7.65e−16	$4.84 \cdot 10^1$	2.02e−15	$1.12 \cdot 10^2$	5.22e−14

In this example, we observe that E_B can be obtained by the fixed-point iteration as well as SDA in just one or two steps. We also implement the Binomial iteration (BI) and the CR in [22] for computing the whole matrix $B = T(b) + E_B$, the CPU time and residual error are compared with the fixed-point iteration and SDA in the computation of E_B, and are reported in Table 4. We mention that the residual error for BI and CR is obtained by $r = \|(I - \hat{Y})^2 - A\|_\infty / \|A\|_\infty$, where $I - \hat{Y}$ is the computed square root of A.

As we can see from Table 4, the fixed-point iteration (3.2) and SDA take less CPU time comparing with the Binomial iteration and CR algorithm. Moreover, the fixed-point iteration (3.2), comparing with CR algorithm, has a speed-up in the CPU time by a factor of about 4 in Test 1 and 2.5 in Tests 2 and 3.

Example 5.3 Let $A = cI - T(s)$ with $T(s) = \mathtt{cqt}(\mathtt{s_n}, \mathtt{s_p})$, where c, $\mathtt{s_n}$ and $\mathtt{s_p}$ are constructed in MATLAB as

$$\mathtt{s_p}\mathtt{= rand(p, 1)}, \quad \mathtt{s_n}\mathtt{= rand(q, 1)}, \quad \mathtt{s_{n(1)}}\mathtt{=}\mathtt{s_{p(1)=1}}, \quad \mathtt{c= sum(s_n)+sum(s_p)}.$$

It can be seen that $\|T(s)\|_\infty = \|s\|_\mathcal{W} < c$, so that A is an invertible M-matrix. For different values of p and q, we apply the fixed-point iteration (3.2) and SDA for computing matrix E_B such that $c(I - T(b) - E_B)^2 = A$, where the symbol $b(z)$ is approximated by $\hat{b}(z)$ that is computed by Algorithm 1.

We also apply the fixed-point iteration and SDA to equation (4.3) for computing its solution G, so that E_B can be approximated by extending G to infinity. Table 5 reports the CPU time taken by the fixed-point iteration and SDA when applied to matrix Eq. (4.3), as well as the CPU time needed in the computation of the E_B relying on the operations of quasi-Toeplitz matrices.

We observe from Table 5 that when the values of p and q are both small, say $p = 4$, $q = 2$, it seems that applying the fixed-point iteration (3.2) and SDA to the truncated matrix equation (4.3) takes less CPU time. For different values of p and q listed in Table 5, the rank of the correction matrix is k=501, 1539, 8496 and 3834, respectively, we observe that when k becomes large, the algorithms applied to the truncated matrix Eq. (4.3) take more CPU times, and it can be seen that the algorithms relying on operations of quasi-Toeplitz matrices are more efficient.

6 Conclusions

We have fully exploited the quasi-Toeplitz structure in the computation of the square root of invertible quasi-Toeplitz M-matrices. We propose algorithms for computing the Toeplitz

Table 5 CPU time in seconds, needed by the fixed-point iteration and SDA for computing a $k \times k$ matrix, which, after extending to infinity, is a good approximation to E_B.

(p, q)	FPI	SDA
(4,2)	$2.79 \cdot 10^{-2}$ [$1.09 \cdot 10^{-1}$]	$1.95 \cdot 10^{-2}$ [$1.13 \cdot 10^{-1}$]
(12,10)	$4.19 \cdot 10^{0}$ [$3.48 \cdot 10^{0}$]	$2.45 \cdot 10^{0}$ [$5.61 \cdot 10^{0}$]
(20,2)	$6.90 \cdot 10^{2}$ [$4.63 \cdot 10^{1}$]	$6.56 \cdot 10^{2}$ [$7.17 \cdot 10^{1}$]
(20,20)	$7.52 \cdot 10^{1}$ [$2.49 \cdot 10^{1}$]	$3.53 \cdot 10^{1}$ [$4.23 \cdot 10^{1}$]

For comparison, the CPU time needed by FPI and SDA relying on the operations of quasi-Toeplitz matrices is written between bracket

part and the correction part respectively. The Toeplitz part is computed by Algorithm 1 at the basis of evaluation/interpolation at the $2n$ roots of unique. We propose a fixed-point iteration and a structure-preserving doubling algorithm for the computation of the correction part. Moreover, we show that the correction part can be approximated by extending the solution of a nonlinear matrix equation to infinity. Numerical experiments show that SDA in general takes less CPU time than the fixed-point iteration. There are also cases where the fixed-point iteration is inferior to SDA. There are cases where both the fixed-point iteration and SDA work better than the Binomial iteration and CR algorithm that exploit the quasi-Toeplitz structure indirectly.

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

Funding The work of the first author was partly supported by the National Natural Science Foundation of China under grant Nos. 12001262, 12371379, 12361080 and by Jiangxi Provincial Natural Science Foundation under grant No. 20224BAB211006. The work of the third author was partly supported by the National Natural Science Foundation of China under grant No. 12201591 and Qingdao Natural Science Foundation under project 23-2-1-158-zyyd-jch.

Data Availability All data used in the manuscript is numerically generated using MATLAB. The MATLAB source code used to generate the numerical results is available at https://github.com/JieMeng00/structured_sqrtm_square_root_m-matrices.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Alefeld, G., Schneider, N.: On square roots of M-matrices. Linear Algebra Appl. **42**, 119–132 (1982)
2. Böttcher, A. and Grudsky, S.M.: Spectral Properties of Banded Toeplitz Matrices. SIAM, Philadelphia, PA (2005)
3. Bini, D.A., Iannazzo, B., and Meng, J.: Algorithms for approximating means of semi-definite quasi-Toeplitz matrices, in: International Conference on Geometric Science of Information, GSI 2021: Geometric Science of Information, 2021, pp. 405–414
4. Bini, D.A., Iannazzo, B. and Meini, B.: Numerical Solution of Algebraic Riccati Equations. Volume 9 of Fundamentals of Algorithms. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2012)
5. Bini, D.A. and Meini, B.: A defect-correction algorithm for quadratic matrix equations, with applications to quasi-Toeplitz matrices. arXiv preprint (2022)
6. Bini, D.A., Massei, S., Meini, B.: On functions of quasi Toeplitz matrices. Sb. Math. **208**, 56–74 (2017)

7. Bini, D.A., Massei, S., Meini, B.: Semi-infinite quasi-Toeplitz matrices with applications to QBD stochastic processes. Math. Comp. **87**, 2811–2830 (2018)
8. Bini, D.A., Massei, S., Meini, B., Robol, L.: On quadratic matrix equations with infinite size coefficients encountered in QBD stochastic processes. Numer. Linear Algebra Appl. **25**, e2128 (2018)
9. Bini, D.A., Massei, S., Robol, L.: Quasi-Toeplitz matrix arithmetic: a MATLAB toolbox. Numer. Algorithms **81**, 741–769 (2019)
10. Bini, D.A., Meini, B., Meng, J.: Solving quadratic matrix equations arising in random walks in the quarter plane. SIAM J. Matrix Anal. Appl. **41**, 691–714 (2020)
11. Bini, D.A., Massei, S., Meini, B., Robol, L.: A computational framework for two-dimensional random walks with restarts. SIAM J. Sci. Comput. **42**(4), A2108–A2133 (2020)
12. Bini, D.A., Iannazzo, B., Meng, J.: Geometric mean of quasi-Toeplitz matrices. BIT **63**, 20 (2023)
13. Bini, D.A., Iannazzo, B., Meini, B., Meng, J., Robol, L.: Computing eigenvalues of semi-infinite quasi-Toeplitz matrices. Numer. Algorithms **92**, 89–118 (2023)
14. Higham, N.J.: Functions of Matrices: Theory and Computation. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2008)
15. Huang, T.-M., Li, R.-C. and Lin, W.-W.: Structure-preserving Doubling Algorithms for Nonlinear Matrix Equations. Volume 14 of Fundamentals of Algorithms. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2018)
16. Kadison, R.V. and Ringrose, J.R.: Fundamentals of the Theory of Operator Algebras. Vol. I, volume 100 of Pure and Applied Mathematics. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York, 1983. Elementary theory
17. Kannan, M.R., Sivakumar, K.C.: On certain positivity classes of operators. Numer. Funct. Anal. Optim. **37**, 206–224 (2017)
18. Kim, H.-M., Meng, J.: Structured perturbation analysis for an infinite size quasi-Toeplitz matrix equation with applications. BIT. **61**, 859–879 (2021)
19. Marek, I.: Frobenius theory of positive operators: comparison theorems and applications. SIAM J. Appl. Math. **19**, 607–628 (1970)
20. Marek, I.: On square roots of M-operators. Linear Algebra Appl. **223**(224), 501–520 (1995)
21. Marek, I., Szyld, D.B.: Splittings of M-operators: irreducibility and the index of the iteration operator. Numer. Funct. Anal. Optim. **11**, 529–553 (1990)
22. Meng, J.: Theoretical and computational properties of semi-infinite quasi-Toeplitz M-matrices. Linear Algebra Appl. **653**, 66–85 (2022)
23. Motyer, A.J., Taylor, P.G.: Decay rates for quasi-birth-and-death processes with countably many phases and tridiagonal block generators. Adv. Appl. Prob. **38**, 522–544 (2006)
24. Robol, L.: Rational Krylov and ADI iteration for infinite size quasi-Toeplitz matrix equations. Linear Algebra Appl. **604**, 210–235 (2020)
25. Shivakumar, P.N., Sivakumar, K.C. and Zhang,Y.: Infinite Matrices and Their Recent Applications. Springer International Publishing Switzerland (2016)