Results in Applied Mathematics 11 (2021) 100164

Contents lists available at ScienceDirect

Results in Applied Mathematics

journal homepage: www.elsevier.com/locate/results-in-applied-mathematics

Explicit inverse of near Toeplitz pentadiagonal matrices related to higher order difference operators

Bakytzhan Kurmanbek [a], Yogi Erlangga [b], Yerlan Amanbek [a,*]

[a] Nazarbayev University, Department of Mathematics, 53 Kabanbay Batyr Ave, Nur-Sultan 010000, Kazakhstan
[b] Zayed University, Department of Mathematics, Abu Dhabi Campus, P.O. Box 144534, United Arab Emirates

ARTICLE INFO

Article history:
Received 5 April 2021
Received in revised form 6 June 2021
Accepted 8 June 2021
Available online xxxx

Keywords:
Explicit formula
Pentadiagonal matrices
Finite difference
Nonlinear beam equation
Fixed point method
Near Toeplitz

ABSTRACT

This paper analyzes the inverse of near Toeplitz pentadiagonal matrices, arising from a finite-difference approximation to the fourth-order nonlinear beam equation. Explicit non-recursive inverse matrix formulas and bounds of norms of the inverse matrix are derived for the clamped–free and clamped–clamped boundary conditions. The bound of norms is then used to construct a convergence bound for the fixed-point iteration of the form $\boldsymbol{u} = f(\boldsymbol{u})$ for solving the nonlinear equation. Numerical computations presented in this paper confirm the theoretical results.

1. Introduction

Many applications give arise to mathematical problems that involve numerical computations with (near) Toeplitz pentadiagonal matrices, which require their inversion (see [1] and references therein). Even though inversion of a nonsingular pentadiagonal matrix can be done efficiently by a numerical linear algebra software, explicit inverse formulas are useful, for example, in a computer algebra software.

Early results on inverses of banded matrices can be traced as far back as to the work of [2–4] for general band matrices. Results for band Toeplitz matrices are given in [5], with explicit inverse formulas for tridiagonal matrices in [6] and pentadiagonal matrices in [1,7–13]. In addition, properties including determinants of such matrices related to finite difference operators have been investigated, e.g. in [14–17]. The recursive formula for computing the determinants of the general pentadiagonal matrices, including the Toeplitz case, are given in [18]. In [19] there were closed expressions for the determinants of arbitrary pentadiagonal matrices, which can be decomposed as a multiplication of the tridiagonal matrices in terms of Chebyshev polynomials of the second kind. In [20], it was presented efficient computing algorithms for finding the inverse and determinant of the pentadiagonal Toeplitz matrices.

In this study, we focus on the specific pentadiagonal matrices arising in a fixed-point iteration for numerically solving the fourth-order nonlinear beam equation:

$$\frac{d^4\widehat{\phi}}{d\widehat{x}^4} = \alpha_1 e^{-\alpha_2\widehat{\phi}}, \quad \widehat{x} \in \Omega = (0, l),$$

* Corresponding author.
E-mail addresses: bakytzhan.kurmanbek@nu.edu.kz (B. Kurmanbek), yogi.erlangga@zu.ac.ae (Y. Erlangga), yerlan.amanbek@nu.edu.kz (Y. Amanbek).

https://doi.org/10.1016/j.rinam.2021.100164

where $\widehat{\phi}$ is a displacement at $\widehat{x}$. This nonlinear equation finds applications in mechanical and civil engineering, which models, e.g., a cantilever beam subjected to swelling pressure on one side. In the above equation, the right-hand side term is the swelling pressure, which in this form is proposed by Grob [21], based on empirical studies (see, e.g., [22] and the references therein), $l > 0$ is the length of the beam, and $\alpha_1, \alpha_2 > 0$ represents the mechanical property of the beam, which are assumed to be constant.

Scaling the domain to unity using the dimensionless variable $x = \widehat{x}/l$ and setting $\phi = \alpha_2\widehat{\phi}$ yields

$$\frac{d^4\phi}{dx^4} = ke^{-\phi}, \text{ in } \Omega = (0, 1), \tag{1}$$

where $k = \alpha_1\alpha_2 l > 0$. We shall use this formulation throughout. For (1) two types of boundary conditions are employed:

1. Clamped–Free (CF) condition:

$$\phi(0) = \phi'(0) = 0 \quad \text{and} \quad \phi''(1) = \phi'''(1) = 0, \tag{2}$$

2. Clamped–Clamped (CC) condition:

$$\phi(0) = \phi'(0) = 0 \quad \text{and} \quad \phi(1) = \phi'(1) = 0. \tag{3}$$

Since $\dfrac{d^4\phi}{dx^4} = ke^{-\phi} > 0$, obviously, $\phi = 0$ cannot be a solution, even though it satisfies the boundary conditions.

The solution of (1) with the boundary conditions (2) is concave up and an increasing function, which can be deduced from a mixed formulation of (2):

$$\begin{cases} \dfrac{d^2\omega}{dx^2} = ke^{-\phi}, & \omega(1) = \omega'(1) = 0, \\ \dfrac{d^2\phi}{dx^2} = \omega, & \phi(0) = \phi'(0) = 0. \end{cases} \tag{4}$$

From the first part of (4), with $e^{-\phi} > 0$ in Ω, $\omega'' > 0$, and w' increases in Ω. The condition $\omega'(1) = 0$ requires that $w' < 0$ in Ω, which furthermore, together with the condition $\omega(1) = 0$, implies that $\omega > 0$ and decreases. From the second part of (4), we have $\phi'' = \omega > 0$; thus, ϕ' is an increasing function in Ω. Since $\phi'(0) = 0$, $\phi > 0$, which implies $\phi > 0$ and increases. This characterization also holds in the finite-difference setting based on the second-order scheme we use in this paper (c.f. Section 4).

Numerical methods based on finite element methods for (1) have been proposed and studied, e.g., in [23,24], where focus is given on the accurate approximation of the solution. This paper approaches the problem from a different angle, with emphasis put on the convergence of the iteration method of the form

$$\phi = \mathcal{L}^{-1}\left(ke^{-\phi}\right),$$

where $\mathcal{L} = d^4/dx^4$, and the properties of the related iteration matrices involved. Using the second-order finite difference approach, these matrices are pentadiagonal and near Toeplitz.

In this paper, we present explicit formulas for inverses of the specific pentadiagonal matrices and their bounds of norms, which are necessary in the convergence analysis of the fixed-point iteration. As the inverse can be formed explicitly, we are able to construct an exact norm of some of those matrices. The convergence rate for the clamped–free and clamped–clamped problems were derived and then numerical examples were presented for different parameters.

The paper is organized as follows. Section 2 is devoted to the convergence and the inverse of the iteration matrix for problem with the clamp-free condition. Similar discussion for the clamp-clamp condition is given in Section 3. Numerical results are presented in Section 4, followed by some concluding remarks in Section 5.

2. The case with clamped–free boundary conditions

We consider $n + 1$ equidistant grid points on the closed interval $[0, 1]$, with the distance (grid size) $h = 1/n$, at which the solution of (1) is approximated by a finite difference scheme. Each grid point is indexed by $i = 0, \ldots, n$, where $i = 0$ and n correspond to the boundary points. Throughout the paper, we shall consider $n \geq 5$ for A to be a meaningful approximation to the differential operator $\mathcal{L}$, even though $n = 5$ may not be of practical interest.

For the interior nodes, $1 \leq i \leq n-1$, the fourth-order derivative is approximated by the second-order finite difference scheme:

$$\frac{d^4\phi}{dx^4}(x_i) \approx \frac{1}{h^4}(\phi_{i-2} - 4\phi_{i-1} + 6\phi_i - 4\phi_{i+1} + \phi_{i+2}),$$

where $x_i = ih$ and $\phi_i \equiv \phi(x_i)$. For $i = 2$, we impose the boundary condition $\phi(0) \equiv \phi_0 = 0$. For $i = 1$, ϕ_{-1} corresponds to a fictitious point outside the computational domain, which is eliminated using the central scheme approximation to the boundary condition $\phi'(0) = 0$. Similar approaches are used for $i = n - 1$ and n, with the boundary conditions $\phi''(1) = \phi'''(1) = 0$ be approximated by appropriate second-order finite difference schemes.

The resultant system of nonlinear equations is

$$A\boldsymbol{u} = h^4 k \exp(-\boldsymbol{u}), \tag{5}$$

where $\boldsymbol{u} = (u_1, \ldots, u_n)^T \in \mathbb{R}^n$, with $u_i \approx \phi(x_i)$, and

$$A := \begin{bmatrix} 7 & -4 & 1 & 0 & \cdots & & 0 \\ -4 & 6 & -4 & \ddots & \ddots & & \vdots \\ 1 & -4 & \ddots & \ddots & & & \\ 0 & \ddots & \ddots & & \ddots & 1 & 0 \\ & \ddots & & \ddots & 6 & -4 & 1 \\ \vdots & & & 1 & -4 & 5 & -2 \\ 0 & \cdots & & 0 & 2 & -4 & 2 \end{bmatrix}. \tag{6}$$

Here, $A \in \mathbb{R}^{n\times n}$ is a nonsymmetric, nondiagonally dominant pentadiagonal matrix.

Our first result on A is that it is nonsingular. In fact, we have the following theorem of the explicit inverse of matrix

Theorem 2.1. *Let $B = [b_{i,j}]_{i,j=1,n} \in \mathbb{R}^{n\times n}$ such that*

$$\begin{aligned} b_{i,j} &= \frac{3ij^2 + j - j^3}{6}, \quad \forall j \le i, i \in \{1, 2, \ldots, n\}, j \in \{1, 2, \ldots, n-1\}, \\ b_{i,n} &= \frac{1}{2} b_{n,i}, \\ b_{n,n} &= \frac{1}{12} n(2n^2 + 1), \\ b_{i,j} &= b_{j,i}, \quad i, j \in \{1, 2, \ldots, n-1\}. \end{aligned}$$

Then B is the inverse of A, where $A = [a_{i,j}]_{i,j=1,n}$ is given in (6).

Proof. The proof is done by the direct computation. Let D be matrix such that $D = AB$. We want to show that the product

$$d_{i,j} := \begin{bmatrix} a_{i,1} & a_{i,2} & \ldots & a_{i,n} \end{bmatrix} \begin{bmatrix} b_{1,j} \\ b_{2,j} \\ \vdots \\ b_{n,j} \end{bmatrix} = \begin{cases} 1, & i = j, \\ 0, & i \ne j. \end{cases}$$

In other words, D is the $n \times n$ identity matrix.

(i) The case $3 \le i \le n-2$ and $1 \le j \le n$.
In this case, $a_{i,i-2} = 1$, $a_{i,i-1} = -4$, $a_{i,i} = 6$, $a_{i,i+1} = -4$, $a_{i,i+2} = 1$, while the others are 0. Therefore,

$$d_{i,j} = b_{i-2,j} - 4b_{i-1,j} + 6b_{i,j} - 4b_{i+1,j} + b_{i+2,j}. \tag{7}$$

If $i = j$, then $b_{i-2,i} = b_{i,i-2} = (2i^3 - 6i^2 + i + 6)/6$, $b_{i-1,i} = b_{i,i-1} = (2i^3 - 3i^2 + i)/6$, $b_{ii} = (2i^3 + i)/6$, $b_{i+1,i} = (2i^3 + 3i^2 + i)/6$, and $b_{i+2,i} = (2i^3 + 6i^2 + i)/6$, yielding $d_{i,i} = 1$.
For $i \ne j$, we consider several cases.

(a) $j \le i - 2$; Then $b_{i-2,j} = (3ij^2 + j - 6j^2 - j^3)/6$, $b_{i-1,j} = (3ij^2 + j - 3j^2 - j^3)/6$, $b_{i,j} = (3ij^2 + j - j^3)/6$, $b_{i+1,j} = (3ij^2 + j + 3j^2 - j^3)/6$, $b_{i+2,j} = (3ij^2 + j + 6j^2 - j^3)/6$, yielding $d_{i,j} = 0$.
(b) $j = i - 1$; Then $b_{i-2,j} = (2i^3 - 9i^2 + 13i - 6)/6$, $b_{i-1,j} = (2i^3 - 6i^2 + 7i - 3)/6$, $b_{i,j} = (2i^3 - 3i^2 + i)/6$, $b_{i+1,j} = (2i^3 - 5i + 3)/6$, $b_{i+2,j} = (2i^3 + 3i^2 - 11i + 6)/6$, yielding $d_{i,i-1} = 0$;
(c) $j = i + 1$; Then $b_{i-2,j} = b_{j,i-2} = (2i^3 - 3i^2 - 11i + 18)/6$, $b_{i-1,j} = b_{j,i-1} = (2i^3 - 5i + 3)/6$, $b_{i,j} = b_{j,i} = (2i^3 + 3i^2 + i)/6$, $b_{i+1,j} = (2i^3 + 6i^2 + 7i + 3)/6$, $b_{i+2,j} = (2i^3 + 9i^2 + 13i + 6)/6$, yielding $d_{i,i+1} = 0$.
(d) $j \ge i - 1$; Then $b_{i-2,j} = b_{j,i-2} = (3j(i-2)^2 + (i-2) - (i-2)^3)/6$, $b_{i-1,j} = b_{j,i-1} = (3j(i-1)^2 + (i-1) - (i-1)^3)/6$, $b_{i,j} = b_{j,i} = \frac{3ji^2 + i - i^3}{6}$, $b_{i+1,j} = (3j(i+1)^2 + (i+1) - (i+1)^3)/6$, $b_{i+2,j} = (3j(i+2)^2 + (i+2) - (i+2)^3)/6$, yielding $d_{i,j} = 0$.

(ii) The case $i = 1$.
For $j = 1$, $b_{1,1} = 3/6$, $b_{2,1} = 1$, and $b_{3,1} = 3/2$; Thus, $d_{i,j} = d_{1,1} = 7b_{1,1} - 4b_{2,1} + b_{3,1} = 1$.
For $j > 1$, we have $b_{1,j} = b_{j,1} = j/2$, $b_{2,j} = b_{j,2} = 2j - 1$, and $b_{3,j} = b_{j,3} = \frac{9j}{2} - 4$; Thus, $d_{i,j} = 7b_{1,j} - 4b_{2,j} + b_{3,j} = 0$.

(iii) The case $i = 2$, with $d_{i,j} = -4b_{1,j} = 6b_{2,j} - 4b_{3,j} + b_{4,j}$.
For $j = 2$, we have $b_{1,2} = b_{2,1} = 1$, $b_{2,2} = 3$, $b_{3,2} = 5$, $b_{4,2} = 7$; Thus, $d_{2,2} = 1$.
For $j \neq 2$, then $b_{1,j} = j/2$, $b_{2,j} = 2j - 1$, $b_{3,j} = \frac{9j}{2} - 4$, and $b_{4,j} = 8j - 10$. We have $d_{i,j} = 0$.
(iv) For the case $i \in \{n-1, n\}$, similar computations using (7) complete the proof. □

From now on, we shall use $a_{i,j}^{-1}$ to denote the (i, j)-entry of A^{-1}, the inverse of A; thus, $a_{i,j}^{-1} = b_{i,j}$.
The following corollary is a consequence of Theorem 2.1.

Corollary 2.2. *The inverse of A is a positive matrix; i.e., $A^{-1} > 0$, implying $a_{i,j}^{-1} > 0$.*

Proof. By Theorem 2.1 it follows that $a_{n,n}^{-1} = n(2n^2+1)/12$ is positive. Notice that, for $i \geq j$, $a_{i,j}^{-1} = \frac{3ij^2+j-j^3}{6} \geq \frac{3j^3+j-j^3}{6} > 0$. Consequently, entries determined by the other 2 parts of Theorem 2.1 are also positive. □

The above positivity result is important in the context of the fixed-point iteration we devise to solve the nonlinear system (5). Consider the iteration

$$\boldsymbol{u}^{\ell} = h^4 k A^{-1} \exp(-\boldsymbol{u}^{\ell-1}), \ell = 1, 2. \ldots \tag{8}$$

Since $A^{-1} > 0$ (Corollary 2.2), the recipe (8) generates a sequence of positive vectors $\{\boldsymbol{u}^{\ell}\}$, if started with $\boldsymbol{u}^0 > \boldsymbol{0}$. As the solution of this type of boundary-value problem is a nonnegative function (c.f., Section 1; see also later for the finite-difference equation case), if the above iteration converges, it converges to a positive solution.

Let $p \in \{1, 2, \infty\}$. Our starting point for the convergence analysis is the relation, with $\boldsymbol{u}^0 > \boldsymbol{0}$,

$$\begin{aligned}\|\boldsymbol{u}^{\ell} - \boldsymbol{u}^{\ell-1}\|_p &= \|h^4 k A^{-1}(\exp(-\boldsymbol{u}^{\ell-1}) - \exp(-\boldsymbol{u}^{\ell-2}))\|_p \\ &= h^4 k \|A^{-1}\left(\exp(-\boldsymbol{u}^{\ell-2}) + G(\boldsymbol{u}^{\ell-1} - \boldsymbol{u}^{\ell-2}) - \exp(-\boldsymbol{u}^{\ell-2})\right)\|_p \\ &= h^4 k \|A^{-1} G(\boldsymbol{u}^{\ell-1} - \boldsymbol{u}^{\ell-2})\|_p,\end{aligned}$$

where $G = -\text{diag}(\exp(-\xi_1), \ldots, \exp(-\xi_n))$, such that the vector $\boldsymbol{\xi} = [\xi_i]_{i=1,n} \in \mathcal{B} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{u}^{\ell-2}\|_p < \|\boldsymbol{u}^{\ell-1} - \boldsymbol{u}^{\ell-2}\|_p\}$. Since $\{\boldsymbol{u}^{\ell}\}$ is a sequence of positive vectors, $\boldsymbol{\xi}$ is also a positive vector, and consequently the diagonal entries of G are strictly less than 1. Thus, $\|G\|_p < 1$, and

$$\begin{aligned}\|\boldsymbol{u}^{\ell} - \boldsymbol{u}^{\ell-1}\|_p &\leq h^4 k \|A^{-1}\|_p \|G\|_p \|(\boldsymbol{u}^{\ell-1} - \boldsymbol{u}^{\ell-2})\|_p \\ &< h^4 k \|A^{-1}\|_p \|\boldsymbol{u}^{\ell-1} - \boldsymbol{u}^{\ell-2}\|_p.\end{aligned} \tag{9}$$

We define L_p to be

$$L_p = h^4 k \|A^{-1}\|_p. \tag{10}$$

Convergence guarantee of the fixed point iteration (8) requires $L_p < 1$, which in turn, for given k and chosen h, requires that

$$\|A^{-1}\|_p < 1/(h^4 k). \tag{11}$$

Lemma 2.3. *For the inverse of A in Theorem 2.1, the following holds true:*

$$a_{i_1,j}^{-1} > a_{i_2,j}^{-1}, \quad \forall i_1 > i_2 > j, \text{ with } i_1, i_2, j \in \{1, 2, \ldots, n\}.$$

Proof. From Theorem 2.1 it follows that $a_{i_1,j}^{-1} = (3i_1 j^2 + j - j^3)/6$ and $a_{i_2,j}^{-1} = (3i_2 j^2 + j - j^3)/6$. Thus, one can notice that $a_{i_1,j}^{-1} > a_{i_2,j}^{-1}$, for $i_1 > i_2 > j$. □

Theorem 2.4. *Let $A \in \mathbb{R}^{n \times n}$ be given in (6), with $n \geq 5$. Then*

$$\|A^{-1}\|_p = \begin{cases} (n^4 - n^2)/8, & \text{if } p = 1, \\ (n^4 + n^2)/8, & \text{if } p = \infty. \end{cases}$$

Proof. For $p = 1$ case, it follows from Lemma 2.3 that

$$\|A^{-1}\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{i,j}^{-1}| = \max \left\{ \sum_{i=1}^{n} |a_{i,n-1}^{-1}|, \sum_{i=1}^{n} |a_{i,n}^{-1}| \right\}.$$

We have

$$\sum_{i=1}^{n} |a_{i,n-1}^{-1}| = \sum_{i=1}^{n} |a_{n-1,i}^{-1}| = \sum_{i=1}^{n} \frac{3(n-1)i^2 + i - i^3}{6} = \frac{n^4 - n^2}{8}.$$

We can now proceed similarly:

$$\sum_{i=1}^{n}|a_{i,n}^{-1}| = \frac{1}{2}\sum_{i=1}^{n}|a_{n,i}^{-1}| = \frac{1}{2}\sum_{i=1}^{n}\frac{3ni^2+i-i^3}{6} = \frac{3n^4+4n^3+3n^2+2n}{48}.$$

From the above results,

$$\sum_{i=1}^{n}|a_{i,n-1}^{-1}| - \sum_{i=1}^{n}|a_{i,n}^{-1}| = \frac{3n^4-4n^3-9n^2-2n}{48} > 0$$

for $n \geq 5$. Therefore,

$$\max_{1\leq j\leq n}\sum_{i=1}^{n}|a_{i,j}^{-1}| \leq \sum_{i=1}^{n}|a_{i,n-1}^{-1}| = \frac{n^4-n^2}{8} = \|A^{-1}\|_1.$$

Next for $p=\infty$ using Lemma 2.3,

$$\begin{aligned}\|A^{-1}\|_\infty &= \max_{1\leq i\leq n}\sum_{j=1}^{n}|a_{i,j}^{-1}| = \sum_{j=1}^{n}|a_{n,j}^{-1}| = \sum_{j=1}^{n-1}|a_{n,j}^{-1}| + a_{n,n}^{-1}\\ &= \sum_{j=1}^{n-1}\frac{3nj^2+j-j^3}{6} + \frac{n(2n^2+1)}{12}\\ &= \frac{n^4+n^2}{8}. \quad \square\end{aligned}$$

Using Hölder's inequality,

$$\|A^{-1}\|_2 \leq \sqrt{\|A^{-1}\|_1\|A^{-1}\|_\infty} = \frac{1}{8}\sqrt{n^8-n^4} \leq \frac{1}{8}n^4.$$

We conclude this section by the characterization of the finite difference solution of the system (5). Because $A\boldsymbol{u} = h^4k\exp(-\boldsymbol{u}) > \boldsymbol{0}$, for the last row of the system,

$$2u_{n-2} - 4u_{n-1} + 2u_n > 0 \quad\Longrightarrow\quad u_n - u_{n-1} > u_{n-1} - u_{n-2}. \tag{12}$$

From the $(n-1)$th row, with $u_{n-3} - 4u_{n-2} + 5u_{n-1} - 2u_n > 0$, we have

$$u_{n-1} - u_{n-2} > u_{n-2} - u_{n-3} + 2u_{n-2} - 4u_{n-1} + 2u_n > u_{n-2} - u_{n-3}, \tag{13}$$

after using the inequality (12). Furthermore, this row leads to

$$4(u_{n-1} - u_{n-2}) > u_n - u_{n-3} + u_n - u_{n-1} > u_n - u_{n-3} + u_{n-1} - u_{n-2},$$

after again using (12), which in turn yields

$$3(u_{n-1} - u_{n-2}) > u_n - u_{n-3}. \tag{14}$$

We then have the following lemma:

Lemma 2.5. *For the inequality $A\boldsymbol{u} > \boldsymbol{0}$, with A given by (6), the following inequalities hold, with $j = i+2$ and $i = 3\ldots, n-1$ the rows of A:*

$$u_{j+2} - u_{j+1} > u_{j+1} - u_j, \quad 3(u_{j+2} - u_{j+1}) > u_{j+3} - u_j.$$

Proof. We have proved the inequalities for $j = n-3$, which comes from the $(n-1)$th row of $A\boldsymbol{u} > \boldsymbol{0}$. Now suppose that they hold also for $j = n-3, n-4, \ldots, k+1$. Associated with $j = k$ is the inequality $u_k - 4u_{k+1} + 6u_{k+2} - 4u_{k+3} + u_{k+4} > 0$ from the $(k+2)$th row of $A\boldsymbol{u} > \boldsymbol{0}$, which gives

$$\begin{aligned}u_{k+2} - u_{k+1} &> 3u_{k+1} - u_k - 5u_{k+2} + 4u_{k+3} - u_{k+4}\\ &= u_{k+1} - u_k + [4(u_{k+3} - u_{k+2}) + u_{k+1} - u_{k+4} + u_{k+1} - u_{k+2}]\\ &> u_{k+1} - u_k + [3(u_{k+3} - u_{k+2}) + u_{k+1} - u_{k+4}]\\ &> u_{k+1} - u_k\end{aligned}$$

by assumption. Next, note that $u_k - 4u_{k+1} + 6u_{k+2} - 4u_{k+3} + u_{k+4} = 3(u_{k+2} - u_{k+1}) + u_k - u_{k+3} - [3(u_{k+3} - u_{k+2}) + u_{k+1} - u_{k+4}] > 0$. Thus, $3(u_{k+2} - u_{k+1}) + u_k - u_{k+3} > 3(u_{k+3} - u_{k+2}) + u_{k+1} - u_{k+4} > 0$, by assumption. $\square$

Theorem 2.6. *The solution of the finite difference system (5) is a nonnegative vector $\boldsymbol{u}$, with increasing u_i.*

Proof. On the nodes $i = 0, 1$, approximation to the differential term leads to

$$u_{i-2} - 4u_{i-1} + 6u_i - 4u_{i+1} + u_{i+2} > 0,$$

which is of the same structure as the $i = 3, \ldots, n-2$ rows of A. By Lemma 2.5,

$$u_2 - u_1 > u_1 - u_0, \quad u_1 - u_0 > u_0 - u_{-1}.$$

Therefore,

$$u_n - u_{n-1} > u_{n-1} - u_{n-2} > \cdots > u_2 - u_1 > u_1 - u_0 > u_0 - u_{-1}.$$

With $u_0 = 0$ (from the boundary condition $\phi(0) = \phi_0 = 0$) and $u_{-1} = u_1$ (from using central finite differencing on $\phi'(0) = 0$), from the most right inequality, we get $u_1 > 0 = u_0$. Also, $u_2 - u_1 > u_1 - u_0 > 0$; thus $u_2 > u_1$. In general, we have $u_{i+1} > u_i$, $i = 1, \ldots, n-1$. □

3. The case with clamped–clamped boundary conditions

In this section, we consider the case with the boundary conditions (3). Conditions at $x = 1$ are treated in the same way as at $x = 0$, leading to (5), but now with $\boldsymbol{u} = (u_1, \ldots, u_{n-1})^T \in \mathbb{R}^{n-1}$ and $A \in \mathbb{R}^{(n-1)\times(n-1)}$ given by

$$A = \begin{bmatrix} 7 & -4 & 1 & 0 & & \cdots & 0 \\ -4 & 6 & -4 & \ddots & \ddots & & \vdots \\ 1 & -4 & \ddots & \ddots & & \ddots & \\ 0 & \ddots & \ddots & & \ddots & \ddots & 0 \\ & \ddots & & & \ddots & -4 & 1 \\ \vdots & & \ddots & \ddots & -4 & 6 & -4 \\ 0 & \cdots & & 0 & 1 & -4 & 7 \end{bmatrix}. \tag{15}$$

However, to simplify our notation, we shall consider the case where $\boldsymbol{u} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ in the subsequent analysis; in this case, $h = 1/(n+1)$.

In contrast to (6), the matrix (15) is centrosymmetric and near Toeplitz. Furthermore, it admits the rank-2 decomposition as follows:

$$A = T^2 + UU^t, \tag{16}$$

where $T = \text{tridiag}_n(-1, 2, -1)$ is an $n \times n$ tridiagonal symmetric Toeplitz matrix, and

$$U = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & \sqrt{2} \end{bmatrix} \in \mathbb{R}^{n\times 2}. \tag{17}$$

T is a symmetric M-matrix, with positive inverse given explicitly by (see, e.g., [10])

$$[T^{-1}]_{ij} = \begin{cases} \frac{j}{n+1}(n - (i-1)), & i \geq j, \\ \frac{i}{n+1}(n - (j-1)), & i < j. \end{cases} \tag{18}$$

A is symmetric positive definite because $T^2 = T^T T$ (and UU^t) is symmetric positive (semi) definite. The inverse of A can be computed by applying the Sherman–Morrison formula on (16):

$$\begin{aligned} A^{-1} &= T^{-2} - T^{-2}U(I_2 + U^t T^{-2} U)^{-1} U^t T^{-2} \\ &= T^{-1}(I - T^{-1}U(I_2 + U^t T^{-2}U)^{-1}U^t T^{-1})T^{-1} \\ &= T^{-t}(I - T^{-1}U(I_2 + U^t T^{-2}U)^{-1}(T^{-1}U)^t)T^{-1}. \end{aligned} \tag{19}$$

Because A^{-1} is symmetric positive definite, the middle term on the right-hand side $I - T^{-1}U(I_2 + U^T T^{-2}U)(T^{-1}U)^T$ is also symmetric positive definite. Rewriting (16) as

$$A = T^2 + UU^t = T^t(I + T^{-1}U(T^{-1}U)^t)T,$$

clearly

$$(I + T^{-1}U(T^{-1}U)^t)^{-1} = I - T^{-1}U(I_2 + U^t T^{-2}U)^{-1}(T^{-1}U)^t =: M.$$

Note that, with (17) and (18),

$$T^{-1}U = \frac{\sqrt{2}}{n+1}\begin{bmatrix} n & 1 \\ n-1 & 2 \\ \vdots & \vdots \\ 2 & n-1 \\ 1 & n \end{bmatrix}. \tag{20}$$

Direct computation yields

$$\begin{aligned} I_2 + U^t T^{-2} U &= \frac{2}{(n+1)^2}\begin{bmatrix} \frac{(n+1)^2}{2} + \sum_{k=1}^{n} k^2 & \sum_{k=1}^{n}(n-(k-1))k \\ \sum_{k=1}^{n}(n-(k-1))k & \frac{(n+1)^2}{2} + \sum_{k=1}^{n} k^2 \end{bmatrix} \\ &= \frac{1}{\gamma}\begin{bmatrix} \gamma+\tau & \gamma n - \tau \\ \gamma n - \tau & \gamma + \tau \end{bmatrix}, \end{aligned}$$

where $\tau = \frac{2n^3+3n^2+n}{6}$ and $\gamma = \frac{(n+1)^2}{2}$. Its inverse is given by

$$(I_2 + U^T T^{-2} U)^{-1} = \frac{1}{\delta}\begin{bmatrix} \gamma+\tau & -\gamma n + \tau \\ -\gamma n + \tau & \gamma + \tau \end{bmatrix}, \tag{21}$$

where $\det(I_2 + U^T T^{-2} U) = \frac{1}{3}(n^2+2n+3) > 0$ and $\delta = (n+1)(2\tau + \gamma(1-n)) = \frac{1}{6}(n+1)^2(n^2+2n+3)$.

Let $M = [m_{ij}]_{i,j=1,n}$. Using (20) and (21), we have, for $i \neq j$,

$$\begin{aligned} m_{ij} &= -\frac{2}{\delta(n+1)^2}[n-(i-1) \quad i]\begin{bmatrix} \gamma+\tau & -\gamma n+\tau \\ -\gamma n+\tau & \gamma+\tau \end{bmatrix}\begin{bmatrix} n-(j-1) \\ j \end{bmatrix} \\ &= q_0(n) + q_1(n)(i+j) + q_2(n)ij, \end{aligned}$$

where

$$\begin{aligned} q_0(n) &= -\frac{4n^2+8n+6}{(n+1)(n^2+2n+3)}, \\ q_1(n) &= \frac{6}{n^2+2n+3}, \\ q_2(n) &= -\frac{12}{(n+1)(n^2+2n+3)}. \end{aligned}$$

One can verify that m_{ij} change signs. Thus M is not an M-matrix.

For $i = j$,

$$\begin{aligned} m_{ii} &= 1 - \frac{2}{\delta(n+1)^2}[n-(i-1) \quad i]\begin{bmatrix} \gamma+\tau & -\gamma n+\tau \\ -\gamma n+\tau & \gamma+\tau \end{bmatrix}\begin{bmatrix} n-(i-1) \\ i \end{bmatrix} \\ &= \frac{n^3-n^2-3n-3}{(n+1)(n^2+2n+3)} + \frac{12}{n^2+2n+3}i - \frac{12}{(n+1)(n^2+2n+3)}i^2 \\ &> 0, \end{aligned}$$

for $n \geq 1$.

Theorem 3.1. *The inverse of A given by (15) is a positive matrix. Furthermore, let $\alpha = n+1-i$, $\beta = j\alpha/(6(n+1)(n^2+2n+3))$, and $\varepsilon = 3(1+\alpha(n+1))(1+(i-j)j)$. The entries of A^{-1} are*

- $a_{ij}^{-1} = \beta(\varepsilon + (j^2-1)(2\alpha^2+1))$, *for* $i \geq j$
- $a_{ij}^{-1} = a_{ji}^{-1}$, *otherwise.*

Proof. Let $A^{-1} = [a_{ij}^{-1}]$, with $A^{-1} = T^{-1}MT^{-1}$. Denote by $\mathbf{y}_j = [y_{k,j}]_{k=1,n} = MT_j^{-1}$, the product of M and the jth column of T^{-1}. For $i \geq j$,

$$\mathbf{y}_j = \frac{1}{n+1}\begin{bmatrix} (n+1-j)(m_{1,1}+2m_{1,2}+\cdots+jm_{1,j})+j(m_{1,j+1}(n-j)+\cdots+m_{1,n}) \\ \vdots \\ (n+1-j)(m_{j,1}+2m_{j,2}+\cdots+jm_{j,j})+j(m_{j,j+1}(n-j)+\cdots+m_{j,n}) \\ \vdots \\ (n+1-j)(m_{n,1}+2m_{n,2}+\cdots+jm_{n,j})+j(m_{n,j+1}(n-j)+\cdots+m_{n,n}) \end{bmatrix}.$$

Using $m_{i,j}$, for $i \leq j$, we have, with $m^*_{i,j} = q_0 + q_1(i+j) + q_2ij$,

$$\begin{aligned} y_{i,j} &= \frac{(n-(j-1))i}{n+1} \\ &+ \frac{n+1-j}{n+1}(m^*_{i,1}+2m^*_{i,2}+\cdots+jm^*_{i,j}) + \frac{j}{n+1}(m^*_{i,j+1}(n-j)+\cdots+m^*_{i,n}) \\ &= \frac{(n-(j-1))i}{n+1} \\ &+ \frac{q_0+q_1i}{n+1}((n+1-j)(1+\cdots+j)+j(n-j+\cdots+1)) \\ &+ \frac{q_1+q_2i}{n+1}((n+1-j)(1^2+\cdots+j^2)+j((n-j)(j+1)+(n-j-1)(j+2)+\cdots+n)) \\ &= \frac{1}{n+1}\left((n-(j-1))i+r_0+r_1i\right), \end{aligned}$$

where

$$r_0 = -\frac{j(n+1-j)((n+1)(n+1-j)+1)}{n^2+2n+3}$$

and

$$r_1 = \frac{j(n+1-j)(n+1-2j)}{n^2+2n+3}.$$

Using similar calculation for $i > j$, we get

$$y_{i,j} = \frac{1}{n+1}\begin{cases} r_0+r_1i+(n+1-j)i = r_0+(r_1-j)\,i+(n+1)i, & i \leq j, \\ r_0+r_1i+(n+1-i)j = r_0+(r_1-i)\,i+(n+1)j, & i > j; \end{cases}$$

hence,

$$\mathbf{y}_j = \frac{1}{n+1}\left(r_0\begin{bmatrix}1\\ \vdots \\ 1\\ 1\\ \vdots\\ 1\end{bmatrix} + (r_1-j)\begin{bmatrix}1\\ \vdots\\ j-1\\ j\\ \vdots\\ n\end{bmatrix} + (n+1)\begin{bmatrix}1\\ \vdots\\ j-1\\ j\\ \vdots\\ j\end{bmatrix}\right).$$

Consider the ith row of T^{-1}:

$$T_i^{-t} = \frac{1}{n+1}\begin{bmatrix} n+1-i & 2(n+1-i) & \cdots & i(n+1-i) & i(n-i) & \cdots & i\end{bmatrix}.$$

We have

$$\begin{aligned} a_{ij}^{-1} &= T_i^{-t}MT_j^{-1} = T_i^{-t}\mathbf{y}_j \\ &= r_0\frac{i(n+1-i)}{2} + \left(r_1-\frac{j}{n+1}\right)\frac{i(n+1-i)(n+1+i)}{6} \\ &+ \frac{j(n+1-i)(3i(n+1)+1-j^2)}{6(n+1)} \\ &= \beta(\varepsilon+(j^2-1)(2\alpha^2+1)), \end{aligned}$$

where

$$\alpha = n+1-i,$$

$$\beta = \frac{j(n+1-i)}{6(n+1)(n^2+2n+3)},$$
$$\varepsilon = 3(1+\alpha+n\alpha)(1+ij-j^2).$$

Notice that $\alpha, \beta, \varepsilon > 0$, $\forall i, j = 1, \ldots, n$. With $i \geq j$ and $j^2 > j^2 - 1$,

$$\begin{aligned} a_{ij}^{-1} &> \beta(j^2-1)(2\alpha^2+1) \\ &\geq 0. \quad \square \end{aligned}$$

By Theorem 3.1, starting from $\boldsymbol{u}^0 > \mathbf{0}$, the fixed-point iteration (8) is guaranteed to generate a sequence of positive vectors.

In the sequel, we present two ways of constructing an estimate for norms of the inverse of A. The first approach is based on the factorization $A^{-1} = T^{-t}M^{-1}T^{-1}$ in (19). The result is presented in the next theorem.

Theorem 3.2. *For $p \in \{1, 2, \infty\}$,*

$$\|A^{-1}\|_p \leq (n+1)^4/32.$$

Proof.

$$\|A^{-1}\|_p \leq \|T^{-1}\|_p \|M\|_p \|T^{-1}\|_p = \|T^{-1}\|_p^2 \|M\|_p.$$

Note that $\|T^{-1}\|_1 = \|T^{-1}\|_\infty$, due to symmetry. Thus, we shall consider only $\|T^{-1}\|_1$. Using (18),

$$\begin{aligned} \sum_{j=1}^{n} |T_{ij}^{-1}| &= \frac{1}{n+1}\left[(n-(i-1))\sum_{j=1}^{i-1} j + i \sum_{j=1}^{n-(i-1)} j\right] \\ &= \frac{1}{2(n+1)}\left[(n+1)^2 i - (n+1)i^2\right]. \end{aligned}$$

The maximum of the rowsum is then attained for $i = (n+1)/2$. Thus,

$$\|T^{-1}\|_1 = \max_{1\leq i\leq n} \sum_{j=1}^{n} |T_{ij}^{-1}| \leq \frac{(n+1)^2}{8}, \tag{22}$$

with equality holding when n is odd.

We now estimate the 1-norm of M. Let $\widetilde{m}_{ij} = q_0(n) + q_1(n)(i+j) + q_2(n)ij$, $\forall i, j = 1, \ldots, n$ and consider $\sum_{j=1}^{n} |\widetilde{m}_{ij}|$. For a fixed i, $\widetilde{m}_{ij}$ can be viewed as a linear function of j. $\sum_{j=1}^{n} |\widetilde{m}_{ij}|$ can then be viewed as the rectangular rules that approximate the area made by the function $\widetilde{m}_{ij}$ and the j-axis. In this case, treating $j \in [0, n+1] \subset \mathbb{R}$,

$$\sum_{j=1}^{n} |\widetilde{m}_{ij}| \leq \int_{j=0}^{n+1} |\widetilde{m}_{ij}| dj = \frac{1}{2}(|\widetilde{m}_{i,0}| + |\widetilde{m}_{i,n+1}|)(n+1),$$

where $\widetilde{m}_{i,0} = -(4n^2+6n(1-i)+8-6i)/[(n+1)(n^2+2n+3)]$ and $\widetilde{m}_{i,n+1} = (2n^2+6n-2-6i(n+1))/[(n+1)(n^2+2n+3)]$.

Since the matrix $\widetilde{M} = [\widetilde{m}_{ij}]$ is persymmetric, we just need to consider $i = 1, \ldots, (n+1)/2$. Then,

$$\begin{aligned} \sum_{j=1}^{n} |\widetilde{m}_{ij}| &\leq \max_i \frac{1}{2}(|\widetilde{m}_{i,0}| + |\widetilde{m}_{i,n+1}|)(n+1) = \frac{1}{2}\frac{2(n^2+5)}{(n+1)(n^2+2n+3)}(n+1) \\ &= \frac{n^2+5}{n^2+2n+3}. \end{aligned}$$

Now,

$$\begin{aligned} \sum_{j=1}^{n} |m_{ij}| &= \sum_{j=1, j\neq 1}^{n} |m_{ij}| + |m_{ii}| = \sum_{j=1, j\neq 1}^{n} |\widetilde{m}_{ij}| + |1 + \widetilde{m}_{ii}| \\ &\leq 1 + |\widetilde{m}_{ii}| + \sum_{j=1, j\neq i}^{n} |\widetilde{m}_{ij}| \\ &= 1 + \sum_{j=1}^{n} |\widetilde{m}_{ij}|. \end{aligned}$$

Thus, for $n \geq 1$,

$$\begin{aligned}\|M\|_1 = \max_i \sum_{j=1}^{n} |m_{ij}| &\leq 1 + \max_i \sum_{j=1}^{n} |\widetilde{m}_{ij}| \\ &\leq 1 + \frac{n^2+5}{n^2+2n+3} \\ &\leq 2,\end{aligned}$$

since $n^2 + 5 < n^2 + 2n + 3$ for $n \geq 1$.

Combining with $\|T^{-1}\|_1$, we get the desired result. Furthermore, using Hölder's inequality, $\|A^{-1}\|_2 \leq \sqrt{\|A^{-1}\|_1 \|A^{-1}\|_\infty} \leq (n+1)^4/32$. □

The second approach uses the knowledge of the entries of A^{-1} in Theorem 3.1. Tedious calculation results in exact norms in some cases, and, hence, much stronger estimates than the previous estimates.

Theorem 3.3. *For $p \in \{1, 2, \infty\}$,*

$$\|A^{-1}\|_p \leq (n+1)^2 \left((n+1)^2 + 8\right)/384.$$

If n is odd, then the equality holds for $p \in \{1, \infty\}$.

Proof. We shall first consider the case $p = \infty$. In this case, by using $a_{i,j}^{-1} > 0$,

$$\|A^{-1}\|_\infty = \max_i \sum_{j=1}^{n} |a_{i,j}^{-1}| = \max_i \sum_{j=1}^{n} a_{i,j}^{-1}$$

For $i = 1, \ldots, n$,

$$\sum_{j=1}^{n} a_{i,j}^{-1} = \sum_{j=1}^{i} a_{i,j}^{-1} + \sum_{j=i+1}^{n} a_{i,j}^{-1} = \sum_{j=1}^{i} a_{i,j}^{-1} + \sum_{k=1}^{n-i} a_{k,i}^{-1},$$

because of the centrosymmetry of A^{-1}. Calculating each sum using the formula for the entries a_{ij}^{-1}, we get

$$\begin{aligned}\sum_{j=1}^{i} a_{i,j}^{-1} &= \widehat{\delta}^{-1} \left[C_1^i \sum_{j=1}^{i} j + C_2^i \sum_{j=1}^{i} j^2 + C_3^i \sum_{j=1}^{i} j^3 \right] \\ &= \widehat{\delta}^{-1} \left[C_1^i \frac{i^2+i}{2} + C_2^i \frac{2i^3+3i^2+i}{6} + C_3^i \frac{i^4+2i^3+i^2}{4} \right],\end{aligned}$$

where $\widehat{\delta} = 6(n+1)(n^2+2n+3)$ and

$$\begin{aligned}C_1^i &= n^3 + 3n^2 - 3i^2 n + 5n + 2i^3 - 3i^2 - 2i + 3, \\ C_2^i &= 3in^3 - 6i^2 n^2 + 9in^2 + 3i^3 n - 12i^2 n + 12in + 3i^3 - 9i^2 + 6i, \\ C_3^i &= -n^3 - 3n^2 + 3i^2 n - 5n - 2i^3 + 3i^2 + 2i - 3.\end{aligned}$$

Also,

$$\begin{aligned}\sum_{k=1}^{n-i} a_{k,i}^{-1} &= \widehat{\delta}^{-1} \left[C_1^k \sum_{k=1}^{n-i} k + C_2^k \sum_{k=1}^{n-i} k^2 + C_3^k \sum_{k=1}^{n-i} k^3 \right] \\ &= \widehat{\delta}^{-1} \left[C_1^k \frac{(n-i)^2+(n-i)}{2} + C_2^k \frac{2(n-i)^3+3(n-i)^2+(n-i)}{6} \right] \\ &\quad + \widehat{\delta}^{-1} C_3^k \frac{(n-i)^4+2(n-i)^3+(n-i)^2}{4},\end{aligned}$$

where

$$\begin{aligned}C_1^k &= 3i^2 n - 2i^3 + 3i^2 + 2i, \\ C_2^k &= 3i^2 n^2 - 3i^3 n + 6i^2 n + 3in - 3i^3 + 3i, \\ C_3^k &= -3i^2 n + 2i^3 - 3i^2 - 2i.\end{aligned}$$

Assuming that $i \in [1, n] \subset \mathbb{R}$, the maximum of the rowsum is obtained from the condition $\frac{d}{di}\sum_{j=1}^{n} a_{i,j}^{-1} = 0$. In this regard, we have

$$\frac{d}{di}\sum_{j=1}^{n} a_{i,j}^{-1} = \widehat{\delta}^{-1}\left[C_0' + C_1' i + C_2' i^2 + C_3' i^3\right] = 0,$$

where

$$\begin{aligned}
C_0' &= \frac{1}{2}n^4 + 2n^3 + 4n^2 + 4n + \frac{3}{2},\\
C_1' &= \frac{1}{2}n^5 + \frac{5}{2}n^4 + 5n^3 + 5n^2 + \frac{1}{2}n - \frac{3}{2},\\
C_2' &= -\frac{3}{2}n^4 - 6n^3 - 12n^2 - 12n - \frac{9}{2},\\
C_3' &= n^3 + 3n^2 + 5n + 3.
\end{aligned}$$

The only acceptable solution of the above equation is $i = (n+1)/2$. The other solutions are rejected: $i = -\frac{1}{2}(\sqrt{n^2+2n+5} - (n+1)) < 0$ and $i = \frac{1}{2}(\sqrt{n^2+2n+5} + (n+1)) > n+1 > n$. One can verify that $i = (n+1)/2$ maximizes the rowsum.

Let n be odd. With $i = (n+1)/2$,

$$\begin{aligned}
\|A^{-1}\|_\infty &= \max_i \sum_{j=1}^{n} |a_{i,j}^{-1}| = \sum_{j=1}^{n} a_{(n+1)/2,j}^{-1} = \sum_{j=1}^{\frac{n+1}{2}} a_{(n+1)/2,j}^{-1} + \sum_{k=1}^{\frac{n-1}{2}} a_{k,(n+1)/2}^{-1}\\
&= \left(n^4 + 4n^3 + 14n^2 + 20n + 9\right)/384\\
&= (n+1)^2((n+1)^2 + 8)/384.
\end{aligned}$$

If n is even, then $i = (n+1)/2$ is not a row of the matrix A; the maximum of the rowsum will then be attained at $i = \lceil (n+1)/2 \rceil$ or $i = \lfloor (n+1)/2 \rfloor$. Either case satisfies

$$\|A^{-1}\|_\infty \le (n+1)^2((n+1)^2 + 8)/384.$$

Symmetry of A^{-1} leads to $\|A^{-1}\|_1 = \|A^{-1}\|_\infty$. Using Hölder's inequality, the above inequality holds also for $p = 2$. □

4. Numerical results

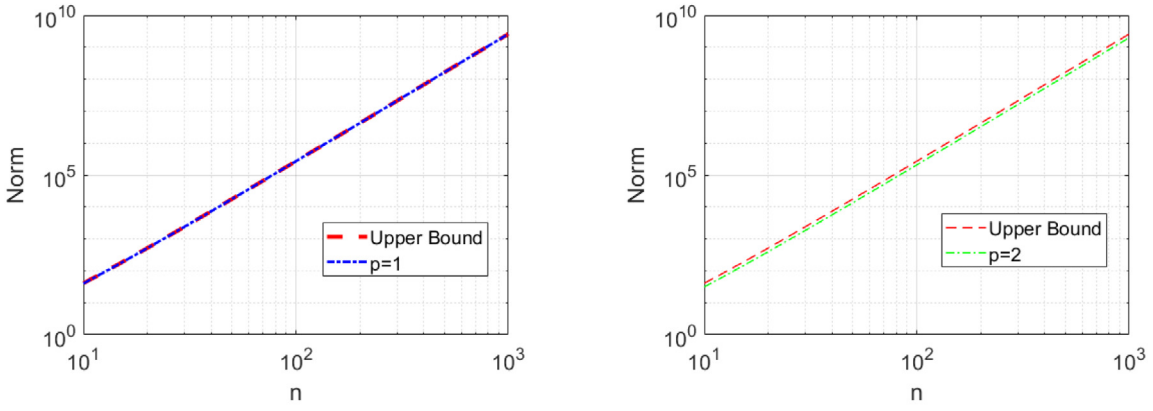
In this section, we provide numerical examples to verify the theoretical results. Table 1 shows the computed norms of the inverse and compares them with the estimate given by the upper bound in Theorem 3.3. For odd n and $p \in \{1, \infty\}$ the norms are exact. For even n, Theorem 3.3 gives an estimate that leads to a small gap. This gap relative to the estimate becomes negligible with an increase in n. To support this statement, the reader is referred to Figs. 1 and 2 in log scales. The numerical tests are performed for all even n from 10 to 1000. The relative error is computed as $|\|A^{-1}\|_p - UBound|/\|A^{-1}\|_p$, where $UBound = (n+1)^2\left((n+1)^2 + 8\right)/384$ from Theorem 3.3. As shown in Fig. 2 (left), the relative error decreases as n increases for $p = 1$ or $p = \infty$. On the other hand, according to the numerical observation the difference between $\|A\|_2$ and the upper bound become constant relative to the norm as n increases, see Fig. 2 (right).

For $n > 5$, we note that the factor$(1 + 8/(n+1)^2)/384$ is lower than $11/3474$. So, alternatively, if k satisfies the condition, $k < 384(n+1)^2/((n+1)^2 + 8)$, we can have a simpler bound: $L_p < 11/3474$. This factor approaches $1/384$ from above as $n \to \infty$. Since the latter is slightly less than the former, for a fixed k, one can expect a slight improvement of convergence by increasing n.

Next, we present numerical results from solving (1) with (2) or (3) using the fixed point method (8). We compare the observed convergence with the theoretical bound given by (10) and Theorem 2.4 (for the clamped–free case) or Theorem 3.3 (for the clamped–clamped case). The fixed point method (8) is declared to have reached a convergence

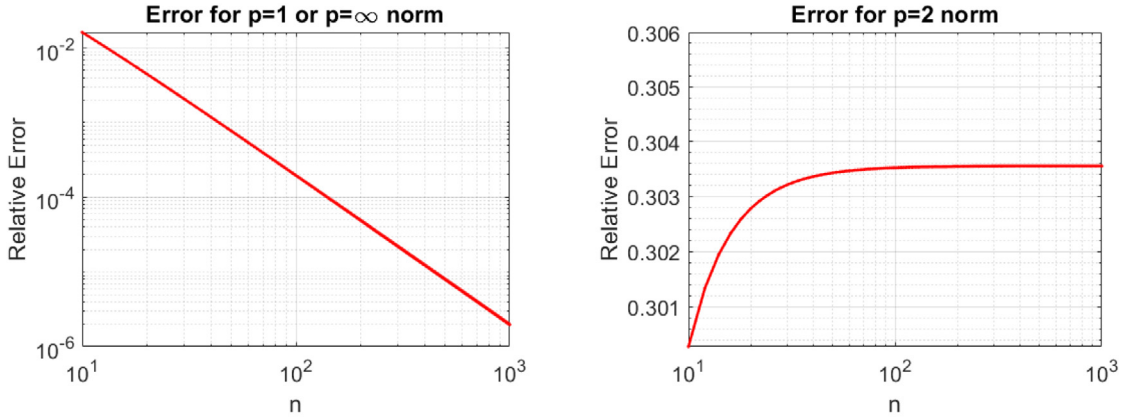
Table 1
Computed $\|A^{-1}\|_p$ and the estimates, for the clamped–clamped case.

n	$p =$			Upper bound from
	1	2	∞	Theorem 3.3
49	16,328	12,527	16,328	16,328
50	17,658	13,558	17,658	17,672
99	260,625	199,939	260,625	260,625
100	271,150	208,055	271,150	271,203
150	1,354,225	1,038,976	1,354,225	1,354,343



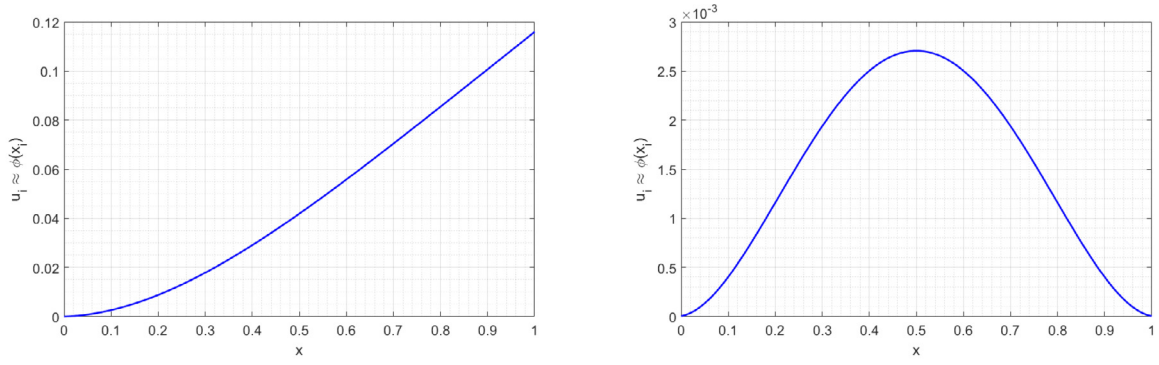
(a) $p = 1$ or $p = \infty$(left) (b) $p = 2$(right)

Fig. 1. The upper bound and actual norm $p = 1$ or $p = \infty$(left) and $p = 2$(right) in log scale.



(a) $p = 1$ or $p = \infty$(left) (b) $p = 2$(right)

Fig. 2. The relative errors in log scale.



(a) clamped-free boundary condition(left) (b) clamped-clamped boundary condition(right)

Fig. 3. Solution at convergence with $k = 1$ and $n = 100$.

if $\|\boldsymbol{u}^{\ell+1} - \boldsymbol{u}^{\ell}\|_p < 10^{-6}$, where $p \in \{1, 2, \infty\}$. Solutions at convergence are shown in Fig. 3 for the clamped–free and clamped–clamped case, with $k = 1$.

Table 2
Observed maximum convergence rate for clamped–free case, with $n = 50$. In brackets there are the theoretical rates, which are upper bounds of L_p based on Theorem 2.4.

k	$p =$					
	1		2		∞	
1/8	0.010	[0.016]	0.010	[0.016]	0.010	[0.017]
1	0.074	[0.125]	0.074	[0.125]	0.074	[0.125]
8	0.400	[1.000]	0.400	[1.000]	0.402	[1.000]

Table 3
Observed maximum convergence rate for clamped–free case, with $n = 99$. In brackets there are the theoretical rates, which are upper bounds of L_p based on Theorem 2.4.

k	$p =$					
	1		2		∞	
1/8	0.010	[0.016]	0.010	[0.016]	0.010	[0.017]
1	0.074	[0.125]	0.074	[0.125]	0.074	[0.125]
8	0.400	[1.000]	0.400	[1.000]	0.402	[1.000]

Table 4
Observed maximum convergence rate for the clamped–clamped case, with $n = 49$. In brackets there are the theoretical rates, which are upper bounds of L_p based on Theorem 3.3.

k	$p =$					
	1		2		∞	
1/8	0.0003	[0.0033]	0.0003	[0.00033]	0.0003	[0.00033]
1	0.0020	[0.0026]	0.0020	[0.0026]	0.0020	[0.0026]
8	0.0158	[0.0209]	0.0159	[0.0209]	0.0161	[0.0209]
32	0.0615	[0.0836]	0.0619	[0.0836]	0.0627	[0.0836]
128	0.2223	[0.3344]	0.2237	[0.3344]	0.2262	[0.3344]

Table 5
Observed maximum convergence rate for the clamped–clamped case, with $n = 100$. In brackets there are the theoretical rates, which are upper bounds of L_p based on Theorem 3.3.

k	$p =$					
	1		2		∞	
1/8	0.0002	[0.00033]	0.0002	[0.00033]	0.0003	[0.00033]
1	0.0020	[0.0026]	0.0020	[0.0026]	0.0020	[0.0026]
8	0.0157	[0.0208]	0.0159	[0.0208]	0.0160	[0.0208]
32	0.0614	[0.0834]	0.0618	[0.0834]	0.0625	[0.0834]
128	0.2218	[0.3336]	0.2232	[0.3336]	0.2257	[0.3336]

For both cases, the actual convergence rates (L_p) are lower than the estimate (Tables 2–5), with increasing gaps between the two as k increases. As $\|A^{-1}\|_p$ is exact, except for $p = 2$, (due to the explicit inverse of A), this suggests that the gap in the convergence rate is mainly due to the estimate $\|G\|_p < 1$. The numerical experiments suggest that the simple fixed-point method (8) can be used for a wider range of k than suggested by the theoretical results. For instance, with $k = 386$ and $n = 99$, we have $L_p = 1.006$. The method still however converges to the solution at the maximum rate of 0.5278.

5. Conclusion

The explicit inverse formula for pentadiagonal matrices arising in the fourth-order nonlinear beam boundary value problem were constructed. The explicit formula helped computing some norms of their inverse, used to estimate the convergence of a fixed-point iteration for solving the nonlinear system of equations. Further research on the convergence upper bounds is necessary to extend our knowledge of the convergence rate in the fixed point method.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

YA wishes to acknowledge the research grant, No AP08052762, from the Ministry of Education and Science of the Republic of Kazakhstan and the Nazarbayev University Faculty Development Competitive Research Grant (NUFDCRG), Kazakhstan, Grant No 110119FD4502.

References

[1] Wang Chaojie, Li Hongyi, Zhao Di. An explicit formula for the inverse of a pentadiagonal Toeplitz matrix. J Comput Appl Math 2015;278:12–8.
[2] Allgower EL. Exact inverses of certain band matrices. Numer Math 1973;21(4):279–84.
[3] Barrett Wayne W, Feinsilver Philip J. Inverses of banded matrices. Linear Algebra Appl 1981;41:111–30.
[4] Rehnqvist Lars. Inversion of certain symmetric band matrices. BIT Numer Math 1972;12(1):90–8.
[5] Dow Murray. Explicit inverses of Toeplitz and associated matrices. ANZIAM J 2002;44:E185–215.
[6] Barrett Wayne W. A theorem on inverse of tridiagonal matrices. Linear Algebra Appl 1979;27:211–7.
[7] Lv Xiao-Guang, Huang Ting-Zhu, Le Jiang. A note on computing the inverse and the determinant of a pentadiagonal Toeplitz matrix. Appl Math Comput 2008;206(1):327–31.
[8] Rózsa P. On the inverse of band matrices. Integral Equations Operator Theory 1987;10(1):82–95.
[9] Zhao Xi-Le, Huang Ting-Zhu. On the inverse of a general pentadiagonal matrix. Appl Math Comput 2008;202(2):639–46.
[10] Diele F, Lopez L. The use of the factorization of five-diagonal matrices by tridiagonal Toeplitz matrices. Appl Math Lett 1998;11(3):61–9.
[11] Kurmanbek Bakytzhan, Erlangga Yogi, Amanbek Yerlan. Inverse properties of a class of seven-diagonal (near) Toeplitz matrices. 2021, ArXiv Preprint arXiv:2103.09868.
[12] Elouafi Mohamed. Explicit inversion of band Toeplitz matrices by discrete fourier transform. Linear Multilinear Algebra 2018;66(9):1767–82.
[13] Barrera Mauricio, Böttcher Albrecht, Grudsky Sergei M, Maximenko Egor A. Eigenvalues of even very nice Toeplitz matrices can be unexpectedly erratic. In: The Diversity and Beauty of Applied Operator Theory. Springer; 2018, p. 51–77.
[14] Andjelić Milica, da Fonseca Carlos M. Some determinantal considerations for pentadiagonal matrices. Linear Multilinear Algebra 2020;1–9.
[15] Amanbek Yerlan, Du Zhibin, Erlangga Yogi, Da Fonseca Carlos M, Kurmanbek Bakytzhan, Pereira António. Explicit determinantal formula for a class of banded matrices. Open Math 2020;18(1):1227–9.
[16] Kurmanbek Bakytzhan, Amanbek Yerlan, Erlangga Yogi. A proof of Andjelić-Fonseca conjectures on the determinant of some Toeplitz matrices and their generalization. Linear Multilinear Algebra 2020;1–8.
[17] Shitov Yaroslav. The determinants of certain (0, 1) Toeplitz matrices. Linear Algebra Appl 2021;618:150–7.
[18] Sweet Roland A. A recursive relation for the determinant of a pentadiagonal matrix. Commun ACM 1969;12(6):330–2.
[19] Marr Robert B, Vineyard George H. Five-diagonal toeplitz determinants an their relation to Chebyshev polynomials. SIAM J Matrix Anal Appl 1988;9(4):579–86.
[20] Lv Xiao-Guang, Huang Ting-Zhu. A note on inversion of Toeplitz matrices. Appl Math Lett 2007;20(12):1189–93.
[21] Grob H. Schwelldruck im belchentunnel, In: Proc. Int. Symp. Für Untertagebau, Luzern, 1972, p. 99–119.
[22] Von Wolffersdorff PA, Fritzsche S. Laboratory swell tests on overconsolidated clay and diagenetic solidified clay rocks, In: Proc. Int. Symp. Geotechnical Measurements and Modelling (GTMM), 2003; Karlsruhe, Germany: p. 407–412.
[23] Skrzypacz Piotr, Nurakhmetov Daulet, Wei Dongming. Generalized stiffness and effective mass coefficients for power-law Euler–Bernoulli beams. Acta Mech Sinica 2020;36(1):160–75.
[24] Wei Dongming, Liu Yu, Zhang Dichuan, Ko Match Wai Lun, Kim Jong R. Numerical analysis for retaining walls subjected to swelling pressure. In: 2016 International Conference on Architecture, Structure and Civil Engineering. 2016.