Convergence of finite difference schemes: matrix versus kernel analysis

M. Ben-Artzi J.-P. Croisille and D. Fishelov

Abstract We consider the convergence analysis of a compact finite difference scheme for the equation $u_t + \frac{\partial^4}{\partial x^4} u = 0$. The discrete in space, continuous in time approximation is $\mathfrak{v}_t + \delta_x^4 \mathfrak{v} = 0$, where δ_x^4 is the discrete biharmonic operator (DBO) operator. The error $\mathfrak{e}(t) = u^*(t) - \mathfrak{v}(t)$ is shown to be $O(h^4)$ for sufficiently smooth data . This problem serves as a model to compare an analytic approach, based on functional analysis and a purely matrix approach. The matrix approach benefits from tools of matrix theory of linear algebra and from known results for the solution of a set of ordinary differential equations. The functional analytic approach utilizes the connection to the continuous problem.

1 Biharmonic time dependent problem, continous and semidiscrete

The convergence analysis of finite difference schemes attracts the interest of the numerical analysis community for several decades (see for example [10]). A renew of interest invoking various analytical and discrete frameworks is currently observed, (see [15, 14, 12, 8]). Here we wish to draw attention to the concurrent ”languages” (analysis versus algebra oriented) tools, which may be used to establish the same convergence rate result. We would like to inspect the advantages versus disadvantages for each approach.

M. Ben-Artzi
Institute of Mathematics, The Hebrew University, Jerusalem 91904, Israel e-mail: mbartzi@math.huji.ac.il
J.-P. Croisille
Univ. Lorraine, CNRS, IECL, 57000, Metz, France, e-mail: jean-pierre.croisille@univ-lorraine.fr
and D. Fishelov
Afeka Tel Aviv Academic College for Engineering 38 Mivza Kadesh St. Tel Aviv 69107, Israel e-mail: daliaf@afeka.ac.il

As a model problem we consider the evolution equation (T is a fixed final time),

$$\frac{\partial}{\partial t}u = -\Big(\frac{\partial}{\partial x}\Big)^4 u, \quad x \in \Omega = [0,1],\ t \in [0,T], \tag{1}$$

subject to initial condition and boundary conditions

$$\begin{cases} u(x,0) = u_0(x),\ x \in \Omega = [0,1], \\ u(0,t) = u_x(0,t) = u(1,t) = u_x(1,t) = 0,\ t \in [0,T]. \end{cases} \tag{2}$$

When there is no risk of confusion we shall write $u(t)$ for $u(x,t)$. Note that (1) is well-posed in the space $L^2(\Omega)$. In other words, the semigroup $e^{-t(\frac{\partial}{\partial x})^4}$ is a continuous contraction semigroup in this space. The domain of its generator is $H^4(\Omega) \cap H_0^2(\Omega)$. Moreover, $H^s(\Omega)$ is a persistence space for every $s \geq 0$, where H^s is the Sobolev space of order s. This is readily seen by casting the semigroup $e^{-t(\frac{\partial}{\partial x})^4}$ in terms of Fourier series. The following finite difference operators are involved [1] (see [5, 2]).

- The standard three point Laplacian is δ_x^2 and the centered difference δ_x are defined by

$$(\delta_x^2 \mathfrak{v})_j = \frac{\mathfrak{v}_{j+1} + \mathfrak{v}_{j-1} - 2\mathfrak{v}_j}{h^2},\ \ (\delta_x \mathfrak{v})_j = \frac{\mathfrak{v}_{j+1} - \mathfrak{v}_{j-1}}{2h}, \quad 1 \leq j \leq N-1. \tag{3}$$

- The *Hermitian derivative* is $\tilde{\delta}_x \mathfrak{v}$, defined as a function of $\mathfrak{v}$ by

$$(\sigma_x \tilde{\delta}_x \mathfrak{v})_j = (\delta_x \mathfrak{v})_j, \quad j = 1,...,N-1, \tag{4}$$

where

$$(\sigma_x \mathfrak{w})_j = \frac{1}{6}(\mathfrak{w}_{j-1} + 4\mathfrak{w}_j + \mathfrak{w}_{j+1}), \quad j = 1,...,N-1. \tag{5}$$

Equivalently, $\tilde{\delta}_x = \sigma_x^{-1}\delta_x$. Here we assume that $\mathfrak{w}_0 = \mathfrak{w}_N = (\tilde{\delta}_x \mathfrak{v})_0 = (\tilde{\delta}_x \mathfrak{v})_N = 0$.

- The Discrete Biharmonic Operator (DBO) δ_x^4 is defined by, [4, 13]

$$(\delta_x^4 \mathfrak{v})_j = \frac{12}{h^2}(\delta_x \tilde{\delta}_x \mathfrak{v} - \delta_x^2 \mathfrak{v})_j, \quad j = 1,...,N-1. \tag{6}$$

The equation (1) is approximated in space on a uniform grid $x_j = jh, h = 1/N, \quad j = 0,1,...,N$ by $t \mapsto \mathfrak{v}(t) = [\mathfrak{v}_1(t), \ldots, \mathfrak{v}_{N-1}(t)]^T$. The gridfunction $\mathfrak{v}(t)$ is solution of the discrete analog to Equation (1) is

$$\mathfrak{v}_t = -\delta_x^4 \mathfrak{v}, \quad t \in [0,T], \tag{7}$$

[1] Gridfunctions are denoted by the fraktur font

The function $\mathfrak{v}(t)$ depends smoothly on $t \in [0,T]$ and is subject to the initial condition $\mathfrak{v}(0) = u_0^*$, and boundary condition $\mathfrak{v}_0(t) = (\mathfrak{v}_x)_0(t) = \mathfrak{v}_N(t) = (\mathfrak{v}_x)_N(t) = 0,\ t \in [0,T]$ [2] .

The error $\mathfrak{e}(t) = \mathfrak{v}(t) - u^*(t)$ satisfies

$$\partial_t \mathfrak{e} + \delta_x^4 \mathfrak{e} = \tau, \tag{8}$$

where $\tau(t)$ is the truncation error. We adopt the notation $\mu \hookrightarrow M$ to express that the gridfunction (resp. finite difference operator) μ corresponds to the vector (resp. matrix) M. Here $\mathfrak{e}(t) \hookrightarrow E(t) \in \mathbf{R}^{N-1}$, $\delta_x^4 \hookrightarrow B \in \mathbf{R}^{(N-1)\times(N-1)}$ and $\tau(t) \hookrightarrow R(t) \in \mathbf{R}^{N-1}$. The vector form of (8) is

$$\partial_t E(t) + BE(t) = R(t). \tag{9}$$

We now consider the convergence of $\mathfrak{v}(t)$ to $u(t)$ as $N \to \infty$. In the rest of the paper, we establish the convergence to 0 of the error $\mathfrak{e}(t)$, or equivalently of $E(t)$, for $t \in [0,T]$ when $h \to 0$. We will need the following

Claim 1 *Let*

$$\left(\frac{\partial}{\partial x}\right)^4 u = f, \quad x \in \Omega = [0,1],$$

and assume that $f \in C^4(\Omega)$. Then operating on the grid function u^* by the discrete operator δ_x^4 leads to the following truncation errors.

$$(\delta_x^4 u^*)_j = f_j^* + O(h^4), \quad \forall j \in \{2,...,N-2\}, \tag{10}$$

and on near-boundary points

$$(\delta_x^4 u^*)_1 = f_1^* + O(h), \qquad (\delta_x^4 u^*)_{N-1} = f_{N-1}^* + O(h). \tag{11}$$

2 Error analysis for the Discrete Biharmonic Operator (DBO)

We consider the equation

$$\left(\frac{\partial}{\partial x}\right)^4 u = f, \quad x \in \Omega = [0,1]. \tag{12}$$

It is well-known that the kernel of $\left(\frac{\partial}{\partial x}\right)^{-4}$ is given by the following claim [4, Claim 5.1].

Claim 2 *The solution of* (12) *is given by*

$$u(x) = \int_0^1 K(x,y) f(y) dy, \tag{13}$$

[2] For $u(x)$ a given function, u^* is the gridfunction defined by $u^* = [u(x_1), \dots u(x_{N-1})]$

where

$$K(x,y)=\begin{cases}\frac{1}{6}(1-x)^2y^2[2x(1-y)+x-y], & y<x\\ \frac{1}{6}x^2(1-y)^2[2y(1-x)+y-x], & x<y\end{cases}. \tag{14}$$

Now consider the discrete analog

$$\delta_x^4\mathfrak{v}=f^*. \tag{15}$$

We wish to estimate the difference of grid functions $\mathfrak{v}-u^*$. Corollary 5.2 in [4] gives the explicit form for the kernel of the inverse discrete operator $(\delta_x^4)^{-1}:l^2_{h,0}\to l^2_{h,0}$:

Claim 3 *The discrete operator* $(\delta_x^4)^{-1}:l^2_{h,0}\to l^2_{h,0}$ *is represented by a matrix* $\left\{K^h_{i,j}\right\}_{1\le i,j\le N-1}$, *explicitly given by*

$$K^h_{i,j}=hK(x_i,x_j),\quad 1\le i,j\le N-1, \tag{16}$$

where $K(x,y)$ *is the resolvent kernel of* $\left(\frac{\partial}{\partial x}\right)^{-4}$, *as in Equation* (14).

Note that the matrix K^h can be written as

$$K^h_{i,j}=K^h_{j,i}=\frac{h}{6}\psi(x_i)\cdot\theta(x_j),\quad 1\le i\le j\le N-1, \tag{17}$$

where the product in the right-hand side is the scalar product in $\mathbf{R}^2$ of the two vector functions:

$$\psi(x)=x^2\Big(-x,3\Big),$$
$$\theta(y)=(1-y)^2\Big(2y+1,y\Big).$$

Thus, the matrix K^h is quasi-separable of rank 2 [6, Section 4.2]. It simplifies (computationally) the evaluation of the matrix. Based on these facts, we can prove a convergence theorem for a general continuous function $f(x)$.

Theorem 1 *Let* $f\in C[0,1]$. *Let* $u(x)$ *be the solution to* (12) *and let* $\mathfrak{v}$ *be the solution to the discrete equation* (15). *Then*

$$\lim_{h\to 0}|\mathfrak{v}-u^*|_\infty=0. \tag{18}$$

Proof In view of Claim 2 and Claim 3 we have

$$\mathfrak{v}_i-u_i^*=\sum_{j=1}^{N-1}hK(x_i,x_j)f_j^*-\int_0^1K(x_i,y)f(y)dy,\quad i=1,2,\ldots,N-1. \tag{19}$$

The first term in the right hand side is the Riemann sum of the second term, so that

$$\max\{|\mathfrak{v}_i-u_i^*|,\ i=1,2,\ldots,N-1\}\le C\omega(h), \tag{20}$$

where $\omega(h)$ is the supremum (over $x \in [0,1]$) of the oscillation functions $K(x,y)f(y)$ as functions of $y \in [0,1]$. □

We can now state our optimal convergence estimate, in the case where the function f is sufficiently smooth. This estimate was first obtained in [7] by using a detailed analysis of the matrix associated with the inverse of the DBO. The proof presented here is much shorter and gives a better insight into the analytical character of the DBO.

Proposition 1 *Suppose that* $f \in C^4[0,1]$ *and let* $u(x)$ *be the solution to* (12)*. Let* $\mathfrak{v}$ *be the solution to the discrete equation* (15)*. Then there exists a constant* $C > 0$ *depending on* f *but not on* h, j *such that*

$$|\mathfrak{v}_j - u_j^*| \leq Ch^4, \quad j = 1,2,\dots,N-1. \tag{21}$$

Proof Let $\mathfrak{e} = \mathfrak{v} - u^*$ and $\mathfrak{r} = \delta_x^4 \mathfrak{e}$. In view of Claim 1

$$\mathfrak{r}_j = O(h^4), \quad \forall j \in \{2,...,N-2\}, \quad \mathfrak{r}_1 = O(h), \qquad \mathfrak{r}_{N-1} = O(h). \tag{22}$$

In view of Claim 3 we have

$$\mathfrak{e}_i = \sum_{j=1}^{N-1} hK(x_i,x_j)\mathfrak{r}_j, \quad i = 1,2,\dots,N-1. \tag{23}$$

By Equation (14) we have (uniformly in $i \in \{1,2,\dots,N-1\}$)

$$K(x_i,x_j) = \begin{cases} O(h^2), & j = 1,N-1, \\ O(1), & j \in \{2,3,\dots,N-2\}. \end{cases} \tag{24}$$

Combining (22) and (24) we conclude

$$|\mathfrak{e}_i| \leq Ch^4 + C\sum_{j=2}^{N-2} h \cdot h^4 \leq Ch^4, \quad i = 1,2,\dots,N-1. \tag{25}$$

which gives (30). □

3 Convergence of the discrete time-evolution solution by purely matrix method

Here we prove optimal convergence in the error $\mathfrak{e}(t)$ (Equation (8)). For this we invoke a slightly different version of Proposition 1, (see [7, 2]).

Proposition 2 (Optimal convergence theorem for the DBO)

Assume that the vector $\mathfrak{r}$*, which contains the truncation errors, satisfy the following estimates:*

$$\begin{aligned} |(\sigma_x \mathfrak{r})_j| &\le Ch^4, \quad j = 2, ..., N-2, \\ |(\sigma_x \mathfrak{r})_1| &\le Ch, \quad |(\sigma_x \mathfrak{r})_{N-1}| \le Ch. \end{aligned} \tag{26}$$

The DBO operator δ_x^4 is invertible and its inverse is denoted by $\delta_x^{-4} = (\delta_x^4)^{-1}$. Then, the vector $\delta_x^{-4}\mathfrak{r} = [\delta_x^{-4}\mathfrak{r}_1, \dots, \delta_x^{-4}\mathfrak{r}_{N-1}]^T$ satisfies

$$|(\delta_x^{-4}\mathfrak{r})_j| \le Ch^4, \qquad j = 1, ..., N-1. \tag{27}$$

The following Lemma results from the optimal theorem above (see also [2, 9, 8]).

Claim 4 (Vector form of the optimal convergence theorem for the DBO)
Assume that $R \in \mathbf{R}^{N-1}$ satisfies

$$PR = [O(h), O(h^4), \dots, O(h^4), O(h)]^T, \tag{28}$$

then the vector $B^{-1}R$ satisfies

$$|(B^{-1}R)_j| \le Ch^4, \quad j = 1, \dots, N-1. \tag{29}$$

We state and prove now the following theorem.

Theorem 2 *Suppose that u is a solution to the problem (1) so that $u(t) \in C^8(\mathbf{R})$ continuously in t, then the error $\mathfrak{e}(t)$, satisfying Equation (7) is bounded by*

$$\max_{0<t<T} |\mathfrak{e}(t)|_h \le C(T)h^4, \tag{30}$$

where $|\mathfrak{g}(t)|_h = \sqrt{\sum_{j=1}^{N-1} h|\mathfrak{g}_j(t)|^2}$ and C is depends only on u_0 and T.

Proof Here we consider a proof by a matrix approach. Consider (9). Since $E(0) = 0$, the Duhamel formula gives

$$E(t) = \int_0^t e^{-(t-\rho)B} R(\rho) d\rho = \int_0^t e^{-\rho B} R(t-\rho) d\rho. \tag{31}$$

The goal is to obtain an error estimate of the form

$$\|E(t)\|_2 \le C(T)h^{3.5} \tag{32}$$

This is implied by the point-wise estimate $E_j(t) = O(h^4)$, $j = 1, ..., N-1$. Rewriting the integrand in (31) as $e^{-\rho B}R(t-\rho) = e^{-\rho B}BB^{-1}R(t-\rho)$, the error $E(t)$ may be expressed as

$$E(t) = \int_0^t \left(e^{-\rho B}B\right)\left(B^{-1}R(t-\rho)\right) d\rho. \tag{33}$$

We denote the eigenvalues of B by

$$0 < \lambda_1 < \cdots < \lambda_{N-1}, \tag{34}$$

and denote $\tilde{\Lambda}(\rho) = \mathrm{diag}\{e^{-\rho\lambda_1}\lambda_1, ..., e^{-\rho\lambda_{N-1}}\lambda_{N-1}\}$. Let $Q \in \mathbf{R}^{(N-1)\times(N-1)}$ be orthogonal such that $B = Q\,\mathrm{diag}\{\lambda_1, \dots, \lambda_{N-1}\}Q^T$. Then,

$$e^{-\rho B}B = Q\,\tilde{\Lambda}(\rho)\,Q^T. \tag{35}$$

Inserting Equation (35) in (33), we have

$$E(t) = \int_0^t Q\,\tilde{\Lambda}(\rho)\,Q^T\,B^{-1}R(t-\rho)d\rho. \tag{36}$$

Denoting $\Lambda(\rho) = \mathrm{diag}\{e^{-\rho\lambda_1},...,e^{-\rho\lambda_{N-1}}\}$, we have

$$\tilde{\Lambda}(\rho) = -\frac{d}{d\rho}\Lambda(\rho). \tag{37}$$

Inserting (37) into (36), and integrating by parts, yields

$$\begin{aligned} E(t) &= -\int_0^t \frac{d}{d\rho}(Q\Lambda(\rho)\,Q^T)B^{-1}R(t-\rho)d\rho \\ &= \underbrace{\left[-(Q\Lambda(\rho)\,Q^T)B^{-1}R(t-\rho)\right]_0^t}_{E^{(1)}(t)} + \underbrace{\int_0^t (Q\Lambda(\rho)\,Q^T)B^{-1}(\frac{d}{d\rho}R(t-\rho))d\rho}_{E^{(2)}(t)}. \end{aligned} \tag{38}$$

Since $R(0) = 0$, we have

$$E^{(1)}(t) = (Q\,\Lambda(0)\,Q^T)B^{-1}R(t). \tag{39}$$

Using Claim 4 and [9, 8], we have that

$$\|B^{-1}\,R(t)\|_2 \le C_1\|B^{-1}\,R(t)\|_2 \le C(t)h^{3.5}.$$

In addition, $\|Q\|_2 = \|Q^T\|_2 = 1$ and $\|\Lambda(\rho)\|_2 \le 1$, for $\rho \ge 0$. Thus, we will have

$$\|E^{(1)}(t)\|_2 \le C(t)h^{3.5}. \tag{40}$$

We turn now to $E^{(2)}(t)$. The components of $P\frac{d}{d\rho}R(t-\rho)$ are of the same order (as powers of h) as the components of $PR(\rho)$. This will yield (using Claim 4) that

$$\|B^{-1}\,\frac{d}{d\rho}R(t-\rho)\|_2 \le C(t)h^{3.5}. \tag{41}$$

Using again $\|Q\|_2 = \|Q^T\|_2 = 1$ and $\|\Lambda(\rho)\|_2 \le 1$, for $0 \le \rho \le t$, we have for some $\bar{\rho} \in [0,t]$,

$$\begin{aligned} \|E^{(2)}(t)\|_2 &= C(t)\max_{0\le\rho\le t}\|\,B^{-1}\,\frac{dR}{d\rho}(\rho)\|_2 = C(t)\|\,B^{-1}\,\frac{dR}{d\rho}(\bar{\rho})\|_2 \\ &\le C_1(t)\|\,B^{-1}\,\frac{dR}{d\rho}(\bar{\rho})\|_2 \end{aligned} \tag{42}$$

Therefore,

$$\|E^{(2)}(t)\|_2 \leq C(t)h^{3.5}. \tag{43}$$

Combining Equations (38), (40) and (43), we conclude that

$$\|E(t)\|_2 \leq C(t)h^{3.5} \text{ or equivalently } |\mathfrak{e}(t)|_h \leq C(t)h^4. \tag{44}$$

We comment now how (30) can be extended to the case of a nonlinear perturbation. Consider the equation

$$\frac{\partial}{\partial t}u = -\Big(\frac{\partial}{\partial x}\Big)^4 u - \alpha(u) \quad x \in \Omega = [0,1],\, t \in [0,T], \tag{45}$$

together with zero boundary conditions on $u, \partial_x u$. Here we assume that the perturbation $\alpha(u)$ is such that (45) has a regular solution $u(x,t)$ on $\Omega \times [0,T]$. The numerical scheme is

$$\mathfrak{v}_t = -\delta_x^4 \mathfrak{v} - \alpha(\mathfrak{v}), \quad t \in [0,T],\;\; t \geq 0. \tag{46}$$

It is assumed that $t \mapsto \mathfrak{v}(t)$ is a regular function of time existing on $[0,T]$ for all value of h. The error $\mathfrak{e}(t) = \mathfrak{v}(t) - u^*(t)$ evolves for $t \in [0,T]$ as

$$\partial_t \mathfrak{e} + \delta_x^4 \mathfrak{e} = \tau_1 + \tau_2, \tag{47}$$

where (the time dependence is dropped) $\tau_1 = (\partial_x^4 u)^* - \delta_x^4 u^*$ and $\tau_2 = \alpha(u^*) - \alpha(\mathfrak{v})$. In matrix/vector form, (9) becomes

$$\partial_t E(t) + BE(t) = R_1(t) + R_2(t), \tag{48}$$

where $\tau_{1,2} \hookrightarrow R_{1,2}$. The term R_1 represents the truncation error of δ_x^4. It depends only on u^*. The term R_2 involves u^* and $\mathfrak{v}$. The Duhamel formula is expressed as

$$E(t) = \underbrace{\int_0^t e^{-\rho B} R_1(t-\rho)d\rho}_{E_a(t)} + \underbrace{\int_0^t e^{-\rho B} R_2(t-\rho)d\rho}_{E_b(t)}. \tag{49}$$

so that

$$\|E(t)\|_2 \leq \|E_a(t)\|_2 + \int_0^t \|e^{-(t-\rho)B}\|_2 \|R_2(\rho)\|_2 d\rho \tag{50}$$

We next proceed along the lines of [11, 1]. Fix $h > 0$. Using $E(0) = 0$ and the continuity of $t \mapsto E(t)$, define $t_0(h)$ by

$$t_0(h) = \sup\{t > 0 \; s.t. \; \|E(t)\|_2 < 1\}. \tag{51}$$

This implies $|u_j^*(t) - \mathfrak{v}_j(t)| < 1$ for $1 \leq j \leq N-1$, $0 \leq t < t_0(h)$ and

$$\max_{\xi \in [u_j^*(t), \mathfrak{v}_j(t)]} |\alpha'(\xi)| \leq \max_{|\xi - u_j^*(t)| < 1} |\alpha'(\xi)| \leq C'(T). \tag{52}$$

where $C'(T)$ depends only on $u(x,t)$. Therefore, for $0 \leq t < t_0(h)$,

$$|R_{2,j}(t)| = |\alpha(u_j^*(t)) - \alpha(\mathfrak{v}_j(t))| \leq C'(T)|E_j(t)|,\ \ 1 \leq j \leq N-1. \tag{53}$$

The spectrum of B satisfies $\mathrm{spec}(B) \subset [\lambda_{\min}, \lambda_{\max}]$. We have with $\lambda_{\min} = O(1)$ and $\lambda_{\max} = O(1/h^4)$. It results from [4, Theorem 7.7] that $\lambda_{\min} > b$, with some constant $b > 0$ uniform in h . By (44), $\|E_a(t)\|_2 \leq C(T)h^{3.5}$. This gives in (50),

$$\|E(t)\|_2 \leq C(T)h^{3.5} + C'(T)\int_0^t e^{-b(t-\rho)}\|E(\rho)\|_2 d\rho. \tag{54}$$

Using Gronwall's inequality, we obtain that for $0 \leq t < t_0(h)$,

$$\|E(t)\|_2 \leq C(T)h^{3.5}\exp\big(\int_0^t e^{-b(t-\rho)}d\rho\big) \leq C''(T)h^{3.5}. \tag{55}$$

where $C''(T)$ depends only on $u(x,t)$. It results from (55) that there exists h_0 small enough such that $C''(T)h_0^{3.5} < 1/2$. Thus, for all $h < h_0$, we have that $t_0(h) > T$ (proceed by contradiction). Therefore for h small enough, $\max_{0\leq t\leq T}\|E(t)\|_2 \leq C''(T)h^{3.5}$ and (30) holds for (45).

4 Convergence of the discrete time-evolution solution by kernel analysis

As in the elliptic case discussed in the previous sections we can study this issue either under "minimal regularity" assumptions or "high regularity" leading to "optimal" fourth-order convergence. We begin with the latter case, analogous to Proposition 1. The "minimal regularity" case is postponed to Theorem 3. The following proposition improves the "almost optimal" estimate obtained in [3]. The same result has been obtained in [7] using matrix techniques. We recall that $H^9(\Omega)$ is the Sobolev space of order 9.

Proposition 3 *Assume that $u_0(x) \in C^9(\Omega)$ and*

$$u_0(0) = u_0'(0) = u_0(1) = u_0'(1) = 0.$$

Let $u(t)$ be the solution to (1), and $\mathfrak{v}(t)$ the solution to (7). There exists a constant $C > 0$ depending on u_0, T but not on h, j such that

$$|\mathfrak{v}_j(t) - u_j^*(t)| \leq Ch^4, \quad j = 1,2,\ldots,N-1,\ t \in [0,T]. \tag{56}$$

Proof As observed in Section 1, the Sobolev space H^9 is a persistence space for the solution. Thus for $u_0 \in C^9(\Omega) \subseteq H^9(\Omega)$ the function $u(t)$ is continuous on $[0,T]$ into H^9. The Sobolev embedding theorem implies that it is also continuous into $C^8(\Omega)$. Thus, let $\mathfrak{e}(t) = \mathfrak{v}(t) - u^*(t)$ be the error function and

$$\mathfrak{r}(t) = \Big[(\frac{\partial}{\partial x})^4 u\Big]^* - \delta_x^4 u^*(t).$$

The error function satisfies the equation

$$\mathfrak{e}_t(t) + \delta_x^4 \mathfrak{e}(t) = \mathfrak{r}(t). \tag{57}$$

Since the function $t \to u(x,t) \in C^8(\Omega)$ is continuous the family $\{u(t), t \in [0,T]\} \subseteq C^8(\Omega)$ is compact and Claim 1 can be applied uniformly to this family. It follows that, uniformly in $t \in [0,T]$,

$$\mathfrak{r}_j(t) = O(h^4), \quad \forall j \in \{2,...,N-2\}, \quad \mathfrak{r}_1(t) = O(h), \qquad \mathfrak{r}_{N-1}(t) = O(h). \tag{58}$$

Equation (57) can be rewritten as

$$\frac{d}{dt}\Big[(\delta_x^4)^{-1}\mathfrak{e}(t)\Big] + \mathfrak{e}(t) = (\delta_x^4)^{-1}\mathfrak{r}(t). \tag{59}$$

In light of (25) we have

$$\mathfrak{w}(t) := (\delta_x^4)^{-1}\mathfrak{r}(t) = O(h^4) \tag{60}$$

uniformly in $t \in [0,T]$.

Let

$$\Lambda_h = \left\{\lambda_{1,h} < \lambda_{2,h} < \ldots < \lambda_{N-1,h}\right\} \tag{61}$$

be the eigenvalues [3] of δ_x^4 with corresponding normalized eigenvectors $\left\{\mathfrak{c}_h^1, \ldots, \mathfrak{c}_h^{N-1}\right\}$. We can expand

$$\mathfrak{e}(t) = \sum_{k=1}^{N-1} a_{k,h}(t)\mathfrak{c}_h^k, \quad \mathfrak{w}(t) = \sum_{k=1}^{N-1} b_{k,h}(t)\mathfrak{c}_h^k \tag{62}$$

and projecting Equation (59) on the $k-th$ eigenvector yields

$$\lambda_{k,h}^{-1}\frac{d}{dt}a_{k,h}(t) + a_{k,h}(t) = b_{k,h}(t), \quad k = 1,2,\ldots,N-1. \tag{63}$$

From (60) we get, for a constant $C > 0$ depending only on T (and not on N)

$$|b_{k,h}(t)| \le Ch^4, \quad k = 1,2,\ldots,N-1,\ t \in [0,T].$$

We obtain (since $\mathfrak{e}(0) = 0$)

$$a_{k,h}(t)e^{\lambda_{k,h}t} = \int_0^t \lambda_{k,h}e^{\lambda_{k,h}s}b_{k,h}(s)ds, \quad t \in [0,T].$$

In view of the estimate for $b_{k,h}(s)$ the right-hand side can be estimated by

$$\Big|\int_0^t \lambda_k e^{\lambda_{k,h}s}b_{k,h}(s)ds\Big| \le Ch^4 e^{\lambda_{k,h}t},$$

hence also

[3] The eigenvalues $\lambda_{k,h}$ are identical to the λ_k in (34).

$$|a_{k,h}(t)| \leq Ch^4, \quad k = 1,2,\ldots,N-1,\ t \in [0,T].$$

We now proceed to study the convergence of the discrete solution to the continuous one, when the initial data is just continuous. Of course, such convergence is not expected to be "optimal" as in Proposition 3.

As pointed out already in the beginning of this section, Equation (1) is not well-posed in $C(\Omega)$, due to the lack of a maximum principle. On the other hand the space $H^1(\Omega)$ is a persistence space for the solution. Recall that $H^1(\Omega)$ is the Sobolev space of L^2 functions whose distributional derivatives are also in L^2. The domain of the generator is $H^4(\Omega)\cap H_0^2(\Omega)$. In particular, this space contains all continuous piecewise linear ("zigzag") functions. By the Sobolev embedding theorem $H^1(\Omega) \subseteq C(\Omega)$.

For the discrete semigroup we use the operator notation $e^{-t\delta_x^4}$. We first state the following coercivity property.

$$(\delta_x^4 \mathfrak{z}, \mathfrak{z})_h \geq C\Big(|\mathfrak{z}|_h^2 + |\delta_x^2 \mathfrak{z}|_h^2 + |\delta_x \tilde{\delta}_x \mathfrak{z}|_h^2\Big), \tag{64}$$

valid for any grid function $\mathfrak{z} \in l_{h,0}^2$ such that also $\tilde{\delta}_x \mathfrak{z} \in l_{h,0}^2$.

Theorem 3 *Let $u_0(x) \in C^1(\Omega) \subseteq H^1(\Omega)$ and $u(x,t)$ the solution to (1). Let $\mathfrak{v}(t) = e^{-t\delta_x^4} u_0^*$ be the corresponding discrete solution. Then, uniformly in $t \in [0,T]$,*

$$\lim_{h\to 0} |\mathfrak{v}(t) - u^*(t)|_h = 0, \quad t \in [0,T]. \tag{65}$$

Proof Pick $\varepsilon > 0$. Let $\widetilde{u}(x,t)$ be solution to (1) with initial data $\widetilde{u_0} \in C^9(\Omega)$. For notational simplicity we occasionally designate $\widetilde{u}(t)$ for $\widetilde{u}(x,t)$ and $u(t)$ for $u(x,t)$. Due to the continuity of the solution of (1) in H^1 we can assume that $\widetilde{u}(t) \in H^9(\Omega)$ satisfies

$$\|\widetilde{u}(t) - u(t)\|_{H^1} < \varepsilon, \quad t \in [0,T]. \tag{66}$$

The Sobolev embedding theorem implies

$$\sup_{t\in[0,T]} \|\widetilde{u}(t) - u(t)\|_{C(\Omega)} < \varepsilon. \tag{67}$$

Since $\widetilde{u}(x,t)$ is sufficiently regular, Proposition 3 can be invoked. Let $\widetilde{\mathfrak{v}}(t) = e^{-t\delta_x^4} \widetilde{u_0}^*$ be the corresponding discrete solution. There exists $h_0 > 0$ such that

$$\sup_{t\in[0,T]} |\widetilde{\mathfrak{v}}(t) - \widetilde{u}(t)^*|_\infty < \varepsilon, \quad 0 < h < h_0. \tag{68}$$

Finally, the positivity of δ_x^4 (see (64)) implies that the semigroup $e^{-t\delta_x^4}$ is contractive on $l_{h,0}^2$ hence

$$|\widetilde{\mathfrak{v}}(t) - \mathfrak{v}(t)|_h \leq |\widetilde{u_0}^* - u_0^*|_h < \varepsilon. \tag{69}$$

Combining (67), (68) and (69) we obtain (65). □

5 Comments and perspectives

In summary, the matrix approach benefits from tools consisting of matrix theory of linear algebra and from known results of the solution of a set of ordinary differential equations. It may be extended to nonlinear (see Section 3) and multidimensional problems [8]. In particular, there is no need to derive the appropriate kernel of the discrete problem, which is in general a difficult task. On the contrary, the functional approach utilizes the connection between the discrete and the continuous problem. It may require the knowledge of the discrete kernel. An important aspect of the functional approach is its ability to deal with low regularity data (see Theorem 1). Note finally that an important topic for convergence analysis is the notion of consistency (see also [14]). In our context, this notion requires further studies.

References

1. Ben-Artzi, M., Croisille, J.P., Fishelov, D.: Convergence of a compact scheme for the pure streamfunction formulation of the unsteady Navier–Stokes system. SIAM J. Numer. Anal. **44**, 1997–2024 (2006)
2. Ben-Artzi, M., Croisille, J.P., Fishelov, D.: Navier–Stokes equations in planar domains. Imperial College Press (2013)
3. Ben-Artzi, M., Croisille, J.P., Fishelov, D.: Time evolution of discrete fourth-order elliptic operators. Num. Meth. PDEs **35(4)**, 1429–1457 (2019)
4. Ben-Artzi, M., Katriel, G.: Spline functions, the biharmonic operator and approximate eigenvalues. Numer. Math. **141**(4), 839–879 (2019)
5. Davies, E.: Spectral theory and differential operators. Blackie and Son (1961)
6. Eidelman, Y., Gohberg, I., Haimovici, I.: Separable Type Representations of Matrices and Fast Algorithms, vol. I. Birkhäuser (2014)
7. Fishelov, D., Ben-Artzi, M., Croisille, J.P.: Recent advances in the study of a fourth-order compact scheme for the one-dimensional biharmonic equation. J. Sci. Comput. **53**(1), 55–79 (2012)
8. Fishelov, D., Croisille, J.P.: Optimal convergence for time dependent Stokes equation: A new approach. J. Sci. Comput. **89** (2021)
9. Fishelov, D., Croisille, J.P.: Time dependent two dimensional fourth-order problems: optimal convergence. In: F.J. Vermolen, C. Vuik (eds.) Numerical Mathematics and Advanced Application, ENUMATH 2019, *Lect. N. Comp. Sci. Eng.*, vol. 139, pp. 429–438. Springer (2021)
10. Gustafsson, B.: The convergence rate for difference approximations to mixed initial boundary value prolems. Math. of Comput. **29**, 396–406 (1975)
11. Hou, T., Wetton, B.: Convergence of a finite difference scheme for the Navier–Stokes equations using vorticity boundary conditions stream function-only formulation. SIAM J. Numer. Anal. **29**, 615–639 (1992)
12. Müller, S., Schweiger, F., Süli, E.: Optimal-order finite difference approximation of generalized solutions to the biharmonic equation in a cube. SIAM J. Numer. Anal. **58**, 298–329 (2020, Vol. 58, No. 1, pp. 298-329)
13. Stephenson, J.W.: Single cell discretizations of order two and four for biharmonic problems. J. Comput. Phys. **55**, 65–80 (1984)
14. Svärd, M., Nordström, J.: On the convergence rates of energy-stable finite-difference schemes. J. Comput. Phys. **397**, paper 108819 (2019)
15. Wang, S., Kreiss, G.: Convergence of summation-by-parts finite difference methods for the wave equation. J. Sci. Comput. **71(1)**, 219–245 (2017)