OPTIMAL CONVERGENCE FOR TIME-DEPENDENT STOKES EQUATION: A NEW APPROACH

D. FISHELOV* AND J.-P. CROISILLE*

Abstract. In our book "Navier-Stokes Equations in Planar Domains", Imperial College Press, 2013, we have suggested a fourth-order compact scheme for the Navier-Stokes equations in streamfunction formulation $\partial_t(\Delta\psi) + (\nabla^\perp\psi)\cdot\nabla(\Delta\psi) = \nu\Delta^2\psi$. Here we present a new approach for the analysis of a high-order compact scheme for the Navier-Stokes equations in cases where the convective term vanishes, or in cases where the viscous term dominates the convective term. In these cases the Navier-Stokes equations is replaced by the time-dependent Stokes equation $\partial_t(\Delta\psi) = \nu\Delta^2\psi$. The same type of fourth-order compact schemes that were proposed for the Navier-Stokes equations, may be adopted to the time-dependent Stokes problem. For these methods the truncation error is only of first-order at near-boundary points, but is of fourth order at interior points. We prove that the rate of convergence is actually four, thus the error tends to zero as $O(h^4)$, where h is the size of the mesh.
Keywords: Biharmonic problem, high-order compact scheme, optimal convergence, time dependent Stokes problem, Navier-Stokes equations.

1. Introduction

The 2D incompressible Navier-Stokes (NS) equations in streamfunction formulation, $\partial_t(\Delta\psi)+(\nabla^\perp\psi)\cdot\nabla(\Delta\psi) = \nu\Delta^2\psi$, play an important role in various areas of physics. The streamfunction formulation of NS equations is due to Lagrange (1768); see [25]. In [5] we have suggested a fourth-order compact scheme for this equation, including a suboptimal error analysis.

In this paper we consider the time dependent Stokes equation in streamfunction formulation

$$\partial_t(\Delta\psi) = \nu\Delta^2\psi + f. \tag{1.1}$$

This equation coincides with the NS equation $\partial_t\Delta\psi(\mathbf{x},t) + C(\psi) = \nu\Delta^2\psi(\mathbf{x},t) + f(\mathbf{x},t)$ when the convective term vanishes. Examples of the so-called Taylor flows and generalizations are given in [14, Chap. 4.6]. The Stokes flow is also an approximate version of the Navier-Stokes system in cases of low Reynolds numbers (and therefore the convective term is small compared to the viscous term). Typically it happens in microfluidics and in creeping flows [33]. It is therefore natural to use (1.1) as an intermediate model for the NS equations [3, 26]. In this paper we consider the fourth-order convergence of a compact high-order scheme suggested for the time-dependent Stokes equation.

The finite difference scheme studied here is based on the Discrete Biharmonic Operator (DBO), which is a compact ("3 point") operator of fourth-order accuracy. Compact high-order schemes for the biharmonic equation were suggested by Stephenson [35] for the two-dimensional biharmonic problem. The DBO studied here may be viewed as a one-dimensional analog of Stephenson's scheme. In our approach, the DBO is obtained as a fourth-order derivative of an interpolating polynomial. This polynomial requires not only values of the function at neighboring points, but also a high-order approximation for the first-order derivative. It turns out that if the first-order derivative is approximated by an Hermitian scheme, then an overall fourth-order accuracy for the fourth-order derivative at interior points is achieved. The DBO operator has a special property that it is compact and its truncation error is fourth order accurate at interior points, while it is only first order accurate at near boundary points. This numerical phenomenon is known in various other contexts. In [19, 20] a hyperbolic system of first order and a parabolic problem were analyzed in the case where extra boundary conditions were given in order to "close" the numerical scheme. It was shown that if the accuracy of the extra boundary conditions is one less compared to the accuracy of the inner scheme, then the overall accuracy of the scheme is determined by the accuracy at inner points. Similarly, in [1] it was proved for a parabolic equation that if the scheme is of order $O(h^\alpha)$ at inner points and of order $O(h^{\alpha-s})$ at near boundary points and if $s = 0, 1$, then the accuracy of the scheme is

Date: October 21, 2021.

[1]this is sometimes referred to as a super convergence property

$O(h^\alpha)$. Finally, in [37] the authors consider the wave equation, which is approximated via a summation-by-parts finite difference scheme and is fourth-order accurate at interior points. They show that if the truncation error at interior points is $O(h^4)$ and it is only $O(h^2)$ at near boundary points, then the overall accuracy of the scheme is $O(h^4)$.

In previous papers we studied various compact schemes for the Navier Stokes equations: We describe now the convergence analysis that what was already carried out in previous works. The NS equation in streamfunction form reads

$$\partial_t \Delta\psi(\mathbf{x},t) + C(\psi) = \nu\Delta^2\psi(\mathbf{x},t) + f(\mathbf{x},t), \tag{1.2}$$

where the convective term is the RHS is $C(\psi) = \nabla^\perp\psi\cdot(\nabla\Delta\psi(\mathbf{x},t))$. In [2] and in [5] the authors described a fourth-order accurate finite difference scheme for the 2D Navier-Stokes equations. In [3] we have proved convergence for a second-order compact scheme, where the convective term was replaced by a constant coefficients equation. It was shown in [3] that the error in the discrete l_2 norm (including the discrete l_2 norm of the first order derivatives) is bounded by Ch^2, where h is the mesh size. In [4], we had proven convergence and error bounds for a second-order compact scheme for the full (nonlinear) Navier-Stokes equations. We have shown that the error in the discrete l_2 norm of the first order derivatives is bounded by $Ch^{1.5}$. We don't have yet a convergence proof for the 4th order compact scheme for the NS equations, which we have suggested in [2] and [5]. The convergence analysis carried out in this paper for the Stokes equation may pave out the way for the convergence of a fourth-order scheme for the NS system which was suggested in [5].

The focus of this paper is the way to overcome the first order truncation at near boundary points and prove 4th order error estimate. This numerical "phenomenon" is the main topic of the paper. It is of interest at the computing level, since it gives a direction to better understand "superconvergence" phenomenon observed in scientific computing.

Our objective here is to extend the DBO fourth-order error analysis in the one-dimensional time-independent case, as accomplished in [16, 2], to a discrete-in-space and continuous-in-time approximation of (1.1). The key idea is to represent the error $e(t)$ as an integral over time of the time-evolution operator, which operates on the truncation error. Then, to express $e(t)$ in terms of the DBO error, rather than the truncation error, using the analysis derived in [16, 2]. As a first step we establish the fourth-order convergence for the model problem $\partial_{xxt}u = \partial_x^4 u + f$. Then, we proceed to the convergence analysis for (1.1).

The novelty in this approach is that the exact solution of the system of ordinary differential equations is handled delicately, while postponing to later stages of the proof the application of energy estimates. This is done by dealing with the orders of magnitude of the matrices involved, using their known structure, given in [7, 2, 6]. This approach paves the way for the case of the full NS equations, which may however require an additional analysis which is related to the non-linear convective term. In [4] an error estimate was derived for a second order compact scheme for the NS equations in streamfunction formulation. The error due to the convective term was decomposed to a sum of several terms that may be bounded in terms of h and in terms of the error. Finally, with convergence analysis was concluded using the Gronwall's inequality. We have shown in [4] a sub-optimal error estimate of $O(h^{1.5})$. For the fourth-order compact scheme suggested in [5] a similar decomposition of the error due to the convective term may be applied as well, however the challenge is to prove optimal $O(h^4)$ convergence. This is deferred to a future work.

The quest for fourth order accuracy and beyond for the incompressible Navier-Stokes equations has been an active area of research since more than forty years. In the context of finite differences, we refer to [15, 11, 27, 23]. In the paper of E and Liu [15], the authors suggested a fourth-order essentially compact scheme for the Navier-Stokes equations. The scheme in [15] is designed for the NS equations in vorticity-streamfunction formulation, where the vorticity is evolved in time and the streamfunction is constructed from the vorticity in a separate stage. The boundary condition on the vorticity was chosen via a Briley's formula (see [10], [38]). Its stability was proved in [38]. Moreover, since Briley's formula is formally only third order accurate, a new vorticity boundary formula, formally fourth order accurate on the boundary, has been proposed and analyzed in [39, 40, 28]. Stability and convergence estimates have been provided for the full PDE system. The scheme in [5] uses pure streamfunction formulation, where the streamfunction is evolved directly in time. The boundary conditions are applied solely on the streamfunction and there is no need for vorticity boundary conditions. A convergence analysis for the biharmonic equation in two dimensions was carried out in [30]. It was proved that the standard 13-point finite difference scheme converges to the exact solution with $O(h^2)$ error in case the exact solution u is in $H^4(\Omega)\cap H_0^2(\Omega)$.

In [24] a second-order scheme for the NS equations in streamfunction formulation was designed; numerical results for the driven cavity were performed. In the context of finite element methods (continuous or discontinuous) a primitive formulation (velocity-pressure) of NS equations is invoked. Then, convergence analysis is often available. The precise design of the global discrete operators resulting from the variational form requires a fine tuning, including nontrivial penalization operators. It is possible to use these schemes for the Stokes equation by dropping the convective term. We refer to [13] and the references therein.

The outline of the paper is the following. In Section 2 we introduce the notation of the compact scheme involved on the model equation

$$\partial_{xxt}u = \partial_x^4 u + f. \tag{1.3}$$

Vectors and matrices, which are associated with the compact difference operators are introduced as well. In Section 3 the fourth order convergence for (1.3) is proved. In Section 4, it is shown that a similar result extends to the two dimensional time dependent Stokes problem. In Section 5 optimal bounds, given on norms of several matrices of interest which operate on the truncation vector, are derived. These bounds are needed for the convergence proofs in Sections 3 and 4. Truncation errors, needed for the analysis conduced in Sections 3 and 4, are derived in Section 6. Finally, numerical results are presented in Section 7. They assess the fourth-order accuracy of the scheme.

2. A Compact Scheme for the diffusion equation $\partial_{xxt}u = \partial_x^4 u + f$

In [32] the diffusion equation (1.3) is considered as a model for (1.1). The goal was to analyze the constraints on the solution $u(x,t)$ of (1.3) imposed by the rigid wall conditions $u(0,t) = 0$ and $\partial_x u(0,t) = 0$. The absence of boundary conditions on $\partial_{xx}u(0,t)$ corresponds to the unknown vorticity at the wall for (1.1). In this section, we study equation (1.3) as a model for our convergence analysis. Similar ideas will be applied to (1.1) in Section 4.

2.1. **Grid function notation.**

We consider the problem

$$\begin{cases} \partial_{xxt}u = \partial_x^4 u + f, \quad x \in (0,1), \qquad t \geq 0, \\ \\ u(0,t) = u(1,t) = 0, \quad \partial_x u(0,t) = \partial_x u(1,t) = 0; \qquad u(x,0) = g(x). \end{cases} \tag{2.1}$$

In what follows, we assume that the data, $f(x,t)$, $g(x)$ and the solution $u(x,t)$, are regular functions of $(x,t) \in [0,1] \times [0,t_f)$, where $0 < t_f$ is a fixed time.

The discrete setting in approximating (2.1) is based on the analysis conducted in [2, Part II]. If a uniform grid $x_j = jh, h = 1/N, \quad j = 0,1,...,N$ is laid out on $[0,1]$, then we denote by $\mathfrak{v} = [\mathfrak{v}_0, \mathfrak{v}_1, \ldots, \mathfrak{v}_{N-1}, \mathfrak{v}_N]^\intercal$ the grid function, defined at the points $x_j, \quad j = 0,1,...,N$ only. Using grid functions, it is understood that $h = 1/N$ becomes a small parameter, which tends to zero. Equivalently, the number of grid points $N = 1/h$ tends to $+\infty$. Finite difference operators act naturally on grid functions.

Throughout this paper we assume that the grid functions $\mathfrak{v}$ satisfy $\mathfrak{v}_0 = \mathfrak{v}_N = 0$. The space of such grid functions is called $l^2_{h,0}$, which is equipped with the norm

$$|\mathfrak{v}|_h = \sqrt{h \sum_{j=1}^{N-1} |v_j|^2}. \tag{2.2}$$

To a given function $u(x)$, $0 \leq x \leq 1$, we denote by u^* the grid function defined by

$$u_j^* = u(x_j), \;\; 1 \leq j \leq N-1, \tag{2.3}$$

where $N \geq 1$.

Now, we approximate the solution of Equation (2.1), on the uniform grid $x_j = j/N, \quad j = 0,1,...,N$. The grid function $\mathfrak{v}(t) = [\mathfrak{v}_j(t)], j = 0,1,...,N$, which serves as an approximation in space of $u(x,t)$, is defined as the solution of

$$\begin{cases} (\partial_t \tilde{\delta}_x^2 \mathfrak{v})_j = \delta_x^4 \mathfrak{v}_j + \mathbf{f}_j, \quad j = 1,...,N-1, \\ \\ \mathfrak{v}_0(t) = \mathfrak{v}_N(t) = 0, \quad (\tilde{\delta}_x \mathfrak{v})_0(t) = (\tilde{\delta}_x \mathfrak{v})_N(t) = 0, \qquad \mathfrak{v}_j(0) = g(x_j), \quad j = 0,1,...,N. \end{cases} \tag{2.4}$$

Hence $t \mapsto \mathfrak{v}(t)$ satisfies

$$\partial_t \tilde{\delta}_x^2 \mathfrak{v} = \delta_x^4 \mathfrak{v} + \mathbf{f}^*, \tag{2.5}$$

where $\mathbf{f}^*$ is the grid function defined by $\mathbf{f}_j = f(x_j, t)$. The following finite difference operators are involved.

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

$$(\delta_x^2 \mathfrak{v})_j = \frac{\mathfrak{v}_{j+1} + \mathfrak{v}_{j-1} - 2\mathfrak{v}_j}{h^2}, \quad (\delta_x \mathfrak{v})_j = \frac{\mathfrak{v}_{j+1} - \mathfrak{v}_{j-1}}{2h}, \quad 1 \le j \le N-1. \tag{2.6}$$

- The *Hermitian derivative* is $\tilde{\delta}_x \mathfrak{v} \in l^2_{h,0}$, defined by

$$\tilde{\delta}_x \mathfrak{v}_j = (\sigma_x^{-1} \delta_x) \mathfrak{v}_j, \quad j = 1, ..., N-1, \tag{2.7}$$

where

$$(\sigma_x \mathfrak{w})_j = \frac{1}{6}(\mathfrak{w}_{j-1} + 4\mathfrak{w}_j + \mathfrak{w}_{j+1}), \quad j = 1, ..., N-1. \tag{2.8}$$

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

$$(\delta_x^4 \mathfrak{v})_j = \frac{12}{h^2} \left((\delta_x \tilde{\delta}_x) \mathfrak{v} - \delta_x^2 \mathfrak{v} \right)_j, \quad j = 1, ..., N-1. \tag{2.9}$$

- The operator $\tilde{\delta}_x^2$ is a discrete Laplacian, which has higher accuracy compared with δ_x^2. It is defined by

$$(\tilde{\delta}_x^2 \mathfrak{v})_j = 2(\delta_x^2 \mathfrak{v})_j - \left((\delta_x \tilde{\delta}_x) \mathfrak{v} \right)_j = (\delta_x^2 \mathfrak{v})_j - \frac{h^2}{12} (\delta_x^4 \mathfrak{v})_j, \quad j = 1, ..., N-1. \tag{2.10}$$

The exact solution $u(x,t)$ of (2.1) is associated with the grid function $u^*(t)$, defined by

$$u_j^*(t) = u(x_j, t), \quad 1 \le j \le N-1. \tag{2.11}$$

The truncation error for the discrete equation (2.5) is the grid function $t \mapsto \mathfrak{r}(t)$, defined by

$$\partial_t \tilde{\delta}_x^2 u^* = \delta_x^4 u^* + \mathbf{f}^* + \mathfrak{r}. \tag{2.12}$$

It is expressed in terms of $u(x,t)$ as

$$\mathfrak{r}(t) = -\left(\delta_x^4 u(.,t)^* - (\partial_x^4 u(.,t))^* \right) + \left(\partial_t (\tilde{\delta}_x^2 u^*(.,t)) - (\partial_t \partial_x^2 u(.,t))^* \right). \tag{2.13}$$

Define the error $\mathfrak{e}(t) = \mathfrak{v}(t) - u^*(t)$. Subtracting (2.12) from (2.5) yields

$$\partial_t (-\tilde{\delta}_x^2 \mathfrak{e}) + \delta_x^4 \mathfrak{e} = \mathfrak{r}. \tag{2.14}$$

In Section 3, we will show an "optimal estimate" for $\mathfrak{e}(t)$ and its derivative $\tilde{\delta}_x \mathfrak{e}(t)$, meaning that $|\mathfrak{e}(t)|_h$ and $|\tilde{\delta}_x \mathfrak{e}(t)|_h$ tend to zero as $O(h^4)$.

2.2. **Vector notation.** Assuming that the grid size N is fixed, then grid functions and finite difference operators reduce to vectors and matrices. Matrix analysis (with fixed dimension $N-1$) is therefore the main tool in what follows. We represent (informally) by $\hookrightarrow$, the correspondence between finite difference operators and matrices and between grid functions and vectors. The matrices involved are T and K, of order $(N-1) \times (N-1)$.

$$T_{i,m} = \begin{cases} 2, & m = i \\ -1, & |m-i| = 1 \\ 0, & |m-i| \ge 2 \end{cases}, \quad K_{i,m} = \begin{cases} 0, & m = i \\ 1, & m - i = 1 \\ -1, & m - i = -1 \end{cases} \tag{2.15}$$

- We have

$$(-\delta_x^2) \hookrightarrow \tilde{T} \triangleq T/h^2, \quad \delta_x \hookrightarrow K/2h. \tag{2.16}$$

- The matrices corresponding to $\tilde{\delta}_x$ and to σ_x are

$$\tilde{\delta}_x \hookrightarrow 3P^{-1}K/h, \quad \sigma_x \hookrightarrow P/6, \tag{2.17}$$

where $P = 6I - T$.

- We have $\delta_x^4 \hookrightarrow B$ where B is the matrix

$$B = \frac{6}{h^4} P^{-1} T^2 + \frac{36}{h^4}(V_1 V_1^\intercal + V_2 V_2^\intercal), \tag{2.18}$$

where $V = [V_1, V_2] \in \mathbb{M}_{N-1,2}$ is defined by

$$\begin{cases} V_1 = (\alpha-\beta)^{1/2} P^{-1}(\frac{\sqrt{2}}{2} e_1 - \frac{\sqrt{2}}{2} e_{N-1}) \\ V_2 = (\alpha+\beta)^{1/2} P^{-1}(\frac{\sqrt{2}}{2} e_1 - \frac{\sqrt{2}}{2} e_{N-1}) \end{cases}, \quad \begin{cases} \alpha = 2(2 - e_1^\intercal P^{-1} e_1) \\ \beta = 2 e_{N-1}^\intercal P^{-1} e_1. \end{cases} \tag{2.19}$$

- The matrix D such that $(-\tilde{\delta}_x^2) \hookrightarrow D$ is

$$D = \frac{1}{h^2}\Big(T + \frac{1}{2} P^{-1} T^2\Big) + \frac{3}{h^2}(V_1 V_1^\intercal + V_2 V_2^\intercal) = \tilde{T} + \frac{h^2}{12} B. \tag{2.20}$$

- For a fixed time $t > 0$, $E(t), R(t) \in \mathbb{R}^{N-1}$ are defined by

$$\mathfrak{e}(t) \hookrightarrow E(t), \quad \mathfrak{r}(t) \hookrightarrow R(t). \tag{2.21}$$

2.3. **The optimal convergence theorem for the discrete biharmonic operator.** The DBO (discrete biharmonic operator) δ_x^4 has been introduced in [3], based on [35]. It appears as the main tool to solve the NS equations in streamfunction form. This discrete operator has a fourth order truncation error except at near boundary points, where the truncation drops to first order. Nevertheless, it has been proved in [16, 2] that this does not prevent a fourth order error estimate. A precise statement of this "optimal convergence theorem" for the DBO operator (2.9), both in grid function and vector formulations, is given in Theorems 2.1 and 2.2.

Theorem 2.1 (Optimal convergence theorem for the DBO)**.** *Assume that the vector $\mathfrak{r}$, which contains the truncation errors, satisfy the following estimates:*

$$\begin{aligned} &|(\sigma_x \mathfrak{r})_j| \le Ch^4, \quad j = 2, ..., N-2, \\ &|(\sigma_x \mathfrak{r})_1| \le Ch, \quad |(\sigma_x \mathfrak{r})_{N-1}| \le Ch. \end{aligned} \tag{2.22}$$

The DBO operator δ_x^4 is invertible and its inverse is denoted by

$$\delta_x^{-4} = (\delta_x^4)^{-1}. \tag{2.23}$$

Then, the vector $\delta_x^{-4}\mathfrak{r} = [(\delta_x^{-4}\mathfrak{r})_1, \ldots, (\delta_x^{-4}\mathfrak{r})_{N-1}]^\intercal$ satisfies

$$|(\delta_x^{-4}\mathfrak{r})_j| \le Ch^4, \qquad j = 1, ..., N-1. \tag{2.24}$$

Therefore, the following max norm estimate holds

$$\max_{1 \le j \le N-1} |(\delta_x^{-4}\mathfrak{r})_j| \le Ch^4. \tag{2.25}$$

The following Theorem translates Theorem 2.1 into vector form.

Theorem 2.2 (Vector form of the optimal convergence theorem for the DBO)**.** *Assume that $R \in \mathbb{R}^{N-1}$ satisfies*

$$PR = [O(h), O(h^4), \ldots, O(h^4), O(h)]^\intercal, \tag{2.26}$$

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

$$|(B^{-1}R)_j| \le Ch^4, \quad j = 1, \ldots, N-1. \tag{2.27}$$

In addition,

$$\begin{cases} |(P^{-1}B^{-1}R)_j| \le Ch^4, \quad j = 2, \ldots, N-2, \\ \\ |(P^{-1}B^{-1}R)_j| \le Ch^5, \quad j = 1, N-1. \end{cases} \tag{2.28}$$

Let us recall the main steps in the proof presented in [16, 2]. From (2.18) we deduce that the matrix PBP is expressed as

$$PBP = \frac{6}{h^4} T^2 P + \frac{36}{h^4} JJ^\intercal, \text{ where } J = PV. \tag{2.29}$$

This may be rewritten as

$$PBP = GH^{-1}, \tag{2.30}$$

where

$$G = I + 6JJ^{\intercal}P^{-1}T^{-2}, \quad H = \frac{h^4}{6}P^{-1}T^{-2}. \tag{2.31}$$

Thus,

$$(PBP)^{-1} = HG^{-1}. \tag{2.32}$$

Therefore,

$$B^{-1} = PHG^{-1}P, \tag{2.33}$$

which results in

$$P^{-1}B^{-1}R = HG^{-1}PR. \tag{2.34}$$

We have shown in [16] and [2, chap. 10.7] that

$$G^{-1}PR = [O(h^2), O(h^4), ..., O(h^4), O(h^2)]^{\intercal}. \tag{2.35}$$

Furthermore, the matrix H may be diagonalized using the spectral basis Z^k of the matrix T, where

$$Z_j^k = \Big(\frac{2}{N}\Big)^{1/2} \sin\frac{kj\pi}{N}, \;\; 1 \leq k, j \leq N-1. \tag{2.36}$$

It results that (see [2, Sect. 10.7, pp. 174 sqq])

$$H_{ij} = \begin{cases} O(h), & \text{if } 2 \leq i, j \leq N-2 \\ O(h^2), & \text{if } i \in \{1, N-1\} \text{ or } j \in \{1, N-1\} \\ O(h^3), & \text{if } (i,j) \in \big\{(1,1), (1,N-1), (N-1,1), (N-1,N-1)\big\}. \end{cases} \tag{2.37}$$

Therefore

$$P^{-1}B^{-1}R = H(G^{-1}PR) = [O(h^5), O(h^4), ..., O(h^4), O(h^5)]^{\intercal}. \tag{2.38}$$

Thus, the estimate (2.28) holds. It yields

$$B^{-1}R = P[O(h^5), O(h^4), ..., O(h^4), O(h^5)]^{\intercal} = [O(h^4), O(h^4), ..., O(h^4), O(h^4)]^{\intercal}. \tag{2.39}$$

Hence, we conclude that (2.27) is valid.

3. Convergence analysis for the equation $\partial_{xxt}u = \partial_x^{(4)}u + f$

In this section we state and prove the fourth-order converge of our scheme (2.4) to the exact solution of the diffusion equation (2.1). We have defined the error $\mathfrak{e}(t) = \mathfrak{v}(t) - u^*(t)$ and obtained that (see Equation (2.14))

$$\partial_t(-\tilde{\delta}_x^2\mathfrak{e}) + \delta_x^4\mathfrak{e} = \mathfrak{r}. \tag{3.1}$$

Our goal is to prove the following bound on $\mathfrak{e}(t)$, assuming a priori regularity on $u(x,t)$.

Theorem 3.1. *Let $t_f > 0$ be a fixed time.*
(i) Suppose that u is a solution to the problem (2.1) having spatial derivatives up to order 8, then the error $\mathfrak{e}(t)$, satisfying Equation (3.1), is bounded by

$$\max_{0\leq t\leq t_f} |(-\tilde{\delta}_x^2)^{1/2}\mathfrak{e}(t)|_h \leq C(t_f)h^4. \tag{3.2}$$

(ii) The following error estimates holds

$$\max_{0\leq t\leq t_f} |e(t)|_h \leq C(t_f)h^4, \tag{3.3}$$

and

$$\max_{0\leq t\leq t_f} |\tilde{\delta}_x e(t)|_h \leq C(t_f)h^4. \tag{3.4}$$

In (3.2)-(3.3)-(3.4), $C(t_f)$ denotes a constant depending only on $\partial_t^{k_0}\partial_x^{k_1}u(x,t)$, $0 \leq k_0 \leq 2$, $4 \leq k_1 \leq 8$ $(x,t) \in [0,1]\times[0,t_f]$.

Proof. (i) Our approach for estimating the error $\mathfrak{e}(t)$ relies on vectors and matrices formulations in $\mathbb{R}^{N-1}$ and $\mathbb{M}_{N-1}(\mathbb{R})$, respectively. Consider the vector function $t \in [0, t_f] \mapsto E(t) \in \mathbb{R}^{N-1}$, where $E(t) \hookrightarrow \mathfrak{e}(t)$. We have $E(t) = [\mathfrak{e}_1(t), ..., \mathfrak{e}_{N-1}(t)]^\intercal$ at time t. The claim in Theorem 3.1 is equivalent to

$$\max_{0\leq t\leq t_f} \|D^{1/2}E(t)\|_2 \leq C(t_f)h^{3.5}. \tag{3.5}$$

We invoke the matrices B and D given in (2.18) and (2.20), where the matrix B represents the biharmonic operator δ_x^4 and the matrix

D corresponds to the operator $-\tilde{\delta}_x^2$. In vector form, Equation (2.14) may be written as

$$\partial_t(DE(t)) + B\ E(t) = R(t). \tag{3.6}$$

The vector $R(t) = [\mathfrak{r}_1(t), ..., \mathfrak{r}_{N-1}(t)]^\intercal \in \mathbb{R}^{N-1}$ is such that $R(t) \hookrightarrow \mathfrak{r}(t))$, where $\mathfrak{r}(t)$ is the truncation error in (2.13). It is obtained by applying matrices of fixed size $N-1$ to the vectors $U(t) \hookrightarrow [u(x_1, t), \ldots, u(x_{N-1}, t)]^\intercal$ and $U^k(t) \hookrightarrow [\partial_x^{(k)}(x_1, t), \ldots, \partial_x^{(k)}(x_{N-1}, t)]^\intercal$. Due to the a priori regularity hypothesis on $u(x, t)$, the solution of (2.1), the function $t \mapsto R(t)$ is regular in time, with $R(0) = 0$.

Define

$$F(t) = D^{1/2}E(t), \tag{3.7}$$

or equivalently

$$E(t) = D^{-1/2}F(t), \tag{3.8}$$

then

$$\partial_t\big(D^{1/2}F(t)\big) + BD^{-1/2}\ F(t) = R(t). \tag{3.9}$$

Multiplying both sides by $D^{-1/2}$ we have

$$\partial_t F(t) + D^{-1/2}BD^{-1/2}F(t) = D^{-1/2}R(t). \tag{3.10}$$

Let $\tilde{B}$ be the symmetric matrix defined by

$$\tilde{B} = D^{-1/2}BD^{-1/2}, \tag{3.11}$$

then, the time dependent equation for $F(t)$ is

$$\partial_t F(t) + \tilde{B}F(t) = D^{-1/2}R(t). \tag{3.12}$$

Note that the matrix $\tilde{B}$ stands for an approximation of a second order derivative. Since $F(0) = 0$, the Duhamel formula for (3.12) is

$$F(t) = \int_0^t e^{-\tilde{B}(t-\rho)}D^{-1/2}R(\rho)d\rho, \tag{3.13}$$

or equivalently

$$F(t) = \int_0^t e^{-\rho\tilde{B}}D^{-1/2}R(t-\rho)d\rho. \tag{3.14}$$

For $\rho \in [0, t]$, $t \leq T$, we rewrite the integrand of (3.14) as

$$e^{-\rho\tilde{B}}D^{-1/2}R(t-\rho) = e^{-\rho\tilde{B}}\tilde{B}\ \tilde{B}^{-1}D^{1/2}R(t-\rho) \quad = e^{-\rho\tilde{B}}\tilde{B}\ D^{1/2}B^{-1}R(t-\rho). \tag{3.15}$$

Therefore, (3.14) is expressed as

$$F(t) = \int_0^t (e^{-\rho\tilde{B}}\tilde{B})\ \Big(D^{1/2}B^{-1}R(t-\rho)\Big)\, d\rho. \tag{3.16}$$

Here the main point is that the integrand involves the error $B^{-1}R$, which is $O(h^4)$ and not the truncation error R. Since $\tilde{B}$ is a symmetric definite positive, it may be diagonalized via an orthogonal matrix Q. We denote the eigenvalues of $\tilde{B}$ by

$$0 < \lambda_1 < \cdots < \lambda_{N-1}. \tag{3.17}$$

Let $\tilde{\Lambda}(\rho)$ be

$$\tilde{\Lambda}(\rho) = \text{diag}\left\{e^{-\rho\lambda_1}\lambda_1, ..., e^{-\rho\lambda_{N-1}}\lambda_{N-1}\right\}, \tag{3.18}$$

where $\lambda_i, i = 1, ..., N-1$ are the eigenvalues of $\tilde{B} = D^{-1/2}BD^{-1/2}$. Then,

$$e^{-\rho\tilde{B}}\tilde{B} = Q\ \tilde{\Lambda}(\rho)\ Q^{\intercal}. \tag{3.19}$$

Inserting Equation (3.19) in (3.16), we have

$$F(t) = \int_0^t Q\ \tilde{\Lambda}(\rho)\ Q^{\intercal}\ D^{1/2}B^{-1}R(t-\rho)d\rho. \tag{3.20}$$

Notice that

$$\tilde{\Lambda}(\rho) = -\frac{d}{d\rho}\Lambda(\rho), \tag{3.21}$$

where

$$\Lambda(\rho) = \mathrm{diag}\{e^{-\rho\lambda_1}, ..., e^{-\rho\lambda_{N-1}}\}. \tag{3.22}$$

Inserting (3.21)-(3.22) into (3.20), we have

$$F(t) = -\int_0^t \frac{d}{d\rho}(Q\Lambda(\rho)\ Q^{\intercal})D^{1/2}B^{-1}R(t-\rho)d\rho. \tag{3.23}$$

Integration by parts yields

$$F(t) = \big[-(Q\Lambda(\rho)\ Q^{\intercal})D^{1/2}B^{-1}R(t-\rho)\big]_{\rho=0}^{t} + \int_0^t (Q\Lambda(\rho)\ Q^{\intercal})D^{1/2}B^{-1}\Big(\frac{d}{d\rho}R(t-\rho)\Big)d\rho. \tag{3.24}$$

We decompose $F(t)$ as $F(t) = F^{(1)}(t) + F^{(2)}(t)$, where

$$\begin{cases} F^{(1)}(t) = -\Big[(Q\Lambda(\rho)\ Q^{\intercal})D^{1/2}B^{-1}R(t-\rho)\Big]_{\rho=0}^{t}, \\ F^{(2)}(t) = \displaystyle\int_0^t (Q\Lambda(\rho)Q^{\intercal})D^{1/2}B^{-1}\big(\frac{d}{d\rho}R(t-\rho)\big)d\rho. \end{cases} \tag{3.25}$$

The first term is

$$\begin{aligned} F^{(1)}(t) &= (Q\ \Lambda(0)\ Q^{\intercal})D^{1/2}B^{-1}R(t) - (Q\Lambda(t)\ Q^{\intercal})D^{1/2}B^{-1}R(0) \\ &= (Q\ \Lambda(0)\ Q^{\intercal})D^{1/2}B^{-1}R(t). \end{aligned} \tag{3.26}$$

Refer now to Section 6, Corollary 6.6, which shows that $PR(t) = [O(h), O(h^4), \ldots, O(h^4), O(h)]^{\intercal}$ with explicit representation of each component in terms of $x \mapsto u(.,t)$, $x \mapsto \partial_t^{k_0}\partial_x^{k_1}u(.,t)$, with $0 \le k_0 \le 1$, $4 \le k \le 8$. Then, it will result from Lemma 5.2 that the components of $D^{1/2}\ B^{-1}\ R(t)$ are $O(h^4)$. This gives

$$\|D^{1/2}\ B^{-1}\ R(t)\|_2 \le C_1(u,t)h^{3.5}, \tag{3.27}$$

where $C_1(u,t)$ can be expressed as

$$C_1(u,t) = C' \max_{\substack{0\le k_0\le 1\\ 4\le k_1\le 8}} \max_{x\in[0,1]} |\partial_t^{k_0}\partial_x^{k_1}u(x,t)| \tag{3.28}$$

and C' a universal constant. In addition, $\|Q\|_2 = \|Q^{\intercal}\|_2 = 1$ and $\|\Lambda(\rho)\|_2 \le 1$, for $\rho \ge 0$. Thus, we will have

$$\|F^{(1)}(t)\|_2 \le \|D^{1/2}B^{-1}R(t)\|_2 \le C_1(u,t)h^{3.5}. \tag{3.29}$$

We turn now to $F^{(2)}(t)$. We have

$$\|F^{(2)}(t)\|_2 \le \max_{0\le\rho\le t} \|D^{1/2}\ B^{-1}\ \frac{dR}{d\rho}(\rho)\|_2. \tag{3.30}$$

By continuity, there exists $\bar{\rho} \in [0,t]$ such that $\max_{0\le\rho\le t}\|D^{1/2}\ B^{-1}\ \frac{dR}{d\rho}(\rho)\|_2 = \|D^{1/2}\ B^{-1}\ \frac{dR}{d\rho}(\bar{\rho})\|_2$. Therefore

$$\|F^{(2)}\|_2 \le C_1(\partial_t u, \bar{\rho})h^{3.5} \tag{3.31}$$

Combining Equations (3.24), (3.29) and (3.31), we conclude that

$$\|F(t)\|_2 \le \|F^{(1)}(t)\|_2 + \|F^{(2)}(t)\|_2 \le C(t)h^{3.5}, \tag{3.32}$$

where

$$C(t) = C'' \max_{\substack{0\le k_0\le 2\\ 4\le k_1\le 8}} \max_{x\in[0,1]} |\partial_t^{k_0}\partial_x^{k_1} u(x,t)|. \tag{3.33}$$

Using $F(t) = D^{1/2}E(t)$, we have $\|D^{1/2}E(t)\|_2 \le C(t)h^{3.5}$. Hence we conclude that

$$|(-\tilde{\delta}_x^2)^{1/2}\mathfrak{e}(t)|_h \le C(t)h^4. \tag{3.34}$$

This gives (3.2), by taking the maximum over $t \in [0, t_f]$.

(ii) The function $t \in \mathbb{R}_+ \mapsto t^{1/2}$ is operator monotone (see [9, chap.5]). This means that for positive definite matrices A and B, we have

$$A \succ B \Rightarrow A^{1/2} \succ B^{1/2}, \tag{3.35}$$

where $\succ$ stands for the ordering of positive definite matrices (see [21, chap. 7.7]). By (2.20)

$$D = \tilde{T} + \frac{h^2}{12}B \succ \tilde{T} \succ I. \tag{3.36}$$

Thus, using (3.35), we have

$$D^{1/2} \succ \tilde{T}^{1/2} \succ I. \tag{3.37}$$

We have used that the eigenvalues of $\tilde{T}$ are $\tilde{\lambda}_k = 4\sin^2(k\pi/2N)/h^2 > 1$, $1 \le k \le N-1$. Therefore,

$$|\mathfrak{e}(t)|_h \le |(-\delta_x^2)^{1/2}\mathfrak{e}(t)|_h \le |(-\tilde{\delta}_x^2)^{1/2}\mathfrak{e}(t)|_h \le C(t)h^4. \tag{3.38}$$

Consider next the bound on the discrete derivative $\tilde{\delta}_x$. By the definition of $\tilde{\delta}_x$ in (2.7) and by the definitions of the one-sided discrete first-order derivatives $\delta_x^+ e_j = (e_{j+1} - e_j)/2$ and $\delta_x^- e_j = (e_j - e_{j-1})/2$, we have

$$\sigma_x\tilde{\delta}_x\mathfrak{e}(t) = \delta_x\mathfrak{e}(t) = \frac{1}{2}\Big(\delta_x^+\mathfrak{e}(t) + \delta_x^-\mathfrak{e}(t)\Big). \tag{3.39}$$

Then,

$$\begin{aligned} |\sigma_x\tilde{\delta}_x\mathfrak{e}(t)|_h^2 &= \frac{1}{4}|\delta_x^+\mathfrak{e}(t) + \delta_x^-\mathfrak{e}(t)|_h^2 \le \frac{1}{2}(|\delta_x^+\mathfrak{e}(t)|_h^2 + |\delta_x^-\mathfrak{e}(t)|_h^2) \\ &= (-\delta_x^2\mathfrak{e}(t), \mathfrak{e}(t))_h = |(-\delta_x^2)^{1/2}\mathfrak{e}(t)|_h^2 \\ &\le C(t)^2h^8. \end{aligned} \tag{3.40}$$

In the last inequality we have invoked (3.38). In addition, σ_x^{-1} is a uniformly (with h) bounded operator, hence

$$|\tilde{\delta}_x\mathfrak{e}(t)|_h \le |\sigma_x^{-1}|_h|\sigma_x\tilde{\delta}_x\mathfrak{e}(t)|_h \le C(t)h^4. \tag{3.41}$$

The inequalities (3.3)-(3.4) are deduced by taking the maximum of all constants of the form $C(t)$ over $t \in [0, t_f]$. ■

4. Convergence analysis for the equation $\partial_t\Delta u = \Delta^2 u + f$

Consider the time dependent Stokes problem

$$\begin{cases} \partial_t\Delta u = \Delta^2 u + f, \quad (x,y) \in [0,1]\times[0,1], \qquad t \ge 0, \\ u(0,y,t) = u(1,y,t) = 0, \quad u_x(0,y,t) = u_x(1,y,t) = 0, \\ u(x,0,t) = u(x,1,t) = 0, \quad u_y(x,0,t) = u_y(x,1,t) = 0, \\ u(x,y,0) = g(x,y), \quad 0 \le x,y \le 1. \end{cases} \tag{4.1}$$

As in Section 2, we assume that the functions $f(x,y,t)$, $g(x,y)$ and the solutions $u(x,y,t)$ are regular functions of their variables $(x,y,t) \in [0,1]\times[0,1]\times[0,t_f)$, where $0 < t_f$ is a fixed time.

Define the grid function $t \mapsto \mathfrak{v}_{j,k}(t), j,k = 0,1,...,N$, which serves as an approximation of u, to be the solution of

$$\begin{cases} \partial_t \tilde{\Delta}_h \mathfrak{v}_{j,k} = \tilde{\Delta}_h^2 \mathfrak{v}_{j,k} + \mathbf{f}_{j,k}, \quad j,k = 1,...,N-1, \\ \mathfrak{v}_{0,k}(t) = \mathfrak{v}_{N,k}(t) = 0, \quad (\tilde{\delta}_x \mathfrak{v})_{0,k}(t) = (\tilde{\delta}_x \mathfrak{v})_{N,k}(t) = 0, \quad k = 0,...,N, \\ \mathfrak{v}_{j,0}(t) = \mathfrak{v}_{j,N}(t) = 0, \quad (\tilde{\delta}_y \mathfrak{v})_{j,0}(t) = (\tilde{\delta}_y \mathfrak{v})_{j,N}(t) = 0, \quad j = 0,...,N, \\ \mathfrak{v}_{j,k}(0) = g_{j,k}, \qquad j,k = 0,...,N. \end{cases} \tag{4.2}$$

Here, for $1 \leq j,k \leq N-1$,

$$\begin{cases} (\delta_x^4 \mathfrak{v})_{j,k} = \dfrac{12}{h^2}(\delta_x \tilde{\delta}_x \mathfrak{v} - \delta_x^2 \mathfrak{v})_{j,k}, \quad (\delta_y^4 \mathfrak{v})_{j,k} = \dfrac{12}{h^2}(\delta_y \tilde{\delta}_y \mathfrak{v} - \delta_y^2 \mathfrak{v})_{j,k}, \\ (\sigma_x \tilde{\delta}_x \mathfrak{v})_{j,k} = (\delta_x \mathfrak{v})_{j,k}, \quad (\sigma_y \tilde{\delta}_y \mathfrak{v})_{j,k} = (\delta_y \mathfrak{v})_{j,k}, \\ (\sigma_x \mathfrak{w})_{j,k} = \dfrac{1}{6}(\mathfrak{w}_{j-1,k} + 4\mathfrak{w}_{j,k} + \mathfrak{w}_{j+1,k}), \quad (\sigma_y \mathfrak{w})_{j,k} = \dfrac{1}{6}(\mathfrak{w}_{j,k-1} + 4\mathfrak{w}_{j,k} + \mathfrak{w}_{j,k+1}). \end{cases} \tag{4.3}$$

In addition,

$$\tilde{\Delta}_h^2 \mathfrak{v} = \delta_x^4 \mathfrak{v} + \delta_y^4 \mathfrak{v} + 2\Big(\delta_x^2 \delta_y^2 - \frac{h^2}{12}(\delta_x^4 \delta_y^2 + \delta_y^4 \delta_x^2)\Big)\mathfrak{v}, \tag{4.4}$$

and

$$\tilde{\Delta}_h \mathfrak{v} = (\tilde{\delta}_x^2 + \tilde{\delta}_y^2)\mathfrak{v}, \qquad \tilde{\delta}_x^2 \mathfrak{v} = 2\delta_x^2 \mathfrak{v} - \delta_x \tilde{\delta}_x \mathfrak{v}, \qquad \tilde{\delta}_y^2 \mathfrak{v} = 2\delta_y^2 \mathfrak{v} - \delta_y \tilde{\delta}_y \mathfrak{v}. \tag{4.5}$$

The exact solution u satisfies

$$\partial_t \tilde{\Delta}_h u^*_{j,k} = \tilde{\Delta}_h^2 u^*_{j,k} + \mathbf{f}_{j,k} + \mathfrak{r}_{j,k}. \tag{4.6}$$

where $\mathfrak{r}$ is the truncation error. By Taylor expansions, the components of the truncation error $\mathfrak{r}$ may be written as (see [2] Proposition 10.8)

$$\begin{cases} \sigma_x \mathfrak{r}_{j,k} = O(h^4), \quad \sigma_y \mathfrak{r}_{j,k} = O(h^4) \quad j,k = 2,...,N-2, \\ \sigma_x \mathfrak{r}_{1,k} = O(h), \quad \sigma_x \mathfrak{r}_{N-1,k} = O(h), \quad k = 1,...,N, \\ \sigma_y \mathfrak{r}_{j,1} = O(h), \quad \sigma_y \mathfrak{r}_{j,N-1} = O(h), \quad j = 1,...,N. \end{cases} \tag{4.7}$$

with explicit representations similar to the ones in Section 3 (dimension 1). Define the error

$$\mathfrak{e} = \mathfrak{v} - u^*. \tag{4.8}$$

By subtracting (4.6) from (4.2), we have

$$\partial_t(-\tilde{\Delta}_h)\mathfrak{e} + \tilde{\Delta}_h^2 \mathfrak{e} = \mathfrak{r}. \tag{4.9}$$

We relate the grid function $\mathfrak{v}_{j,k}, j,k = 1,...,N-1$ with the column vector $V = \mathbf{vec}(\mathfrak{v})$ defined by

$$V = \big[\mathfrak{v}_{1,1},...,\mathfrak{v}_{N-1,1},\mathfrak{v}_{1,2},...\mathfrak{v}_{N-1,2},...,\mathfrak{v}_{1,N-1},...,\mathfrak{v}_{N-1,N-1}\big]^{\intercal} \in \mathbb{R}^{(N-1)^2}. \tag{4.10}$$

The bottom ordering of the vector $V \in \mathbb{R}^{(N-1)^2}$ is obtained by letting the index j vary first while keeping k fixed, then vary the index k [2]. Then, we relate the two-dimensional finite difference operators with matrix operators of size $(N-1)\times(N-1)$ (for $N \geq 2$) acting on a vector V. Most of those operators are obtained as Kronecker products of $(N-1)\times(N-1)$ matrices. Recall that the Kronecker product of the matrices $G \in \mathbb{M}_{m,n}$ and $H \in \mathbb{M}_{p,q}$ is the matrix $G \otimes H \in \mathbb{M}_{mp,nq}$ defined by

$$G \otimes H = \begin{bmatrix} g_{1,1}H & g_{1,2}H & ... & g_{1,n}H \\ ... & & & \\ ... & & & \\ g_{m,1}H & g_{m,2}H & ... & g_{m,n}H \end{bmatrix}. \tag{4.11}$$

[2] This is the standard definition of the operator **vec**, see [22, Ch. 4.2, p. 244]

Let the matrix B represent the biharmonic discrete operator δ_x^4 in one dimension and the matrix $\tilde{T}$ represent $-\delta_x^2$ (or $-\delta_y^2$) in one dimension. Then, $I \otimes B$ and $B \otimes I$ represent the biharmonic operators δ_x^4 and δ_y^4, respectively. Similarly, $I \otimes \tilde{T}$ and $\tilde{T} \otimes I$ represents the operator $-\delta_x^2$ and $-\delta_y^2$, respectively, in the 2D setting. In addition,

$$R(t) = \left[\mathfrak{r}_{1,1}, ..., \mathfrak{r}_{N-1,1}, \mathfrak{r}_{1,2}, ..., \mathfrak{r}_{N-1,2}, ..., \mathfrak{r}_{1,N-1}, ..., \mathfrak{r}_{N-1,N-1}\right]^\intercal \in \mathbb{R}^{(N-1)^2} \tag{4.12}$$

is related to the truncation error.

As in Section 3, the function $t \mapsto R(t) = \mathbf{vec}(\mathfrak{r}(t))$ is regular, since it is obtained by applying linear operators to the function $t \mapsto u(.,t)$ and its derivatives up to order 8, (see Section 6). Invoking Theorem 2.2 we have

Corollary 4.1. *Let $R(t) = R^{(1)}(t) + R^{(2)}(t) \in \mathbb{R}^{(N-1)^2}$, where $R^{(1)} = \left[R_1^{(1)}; ...; R_{N-1}^{(1)}\right]$ and $R^{(2)} = \left[R_1^{(2)}; ...; R_{N-1}^{(2)}\right]$, respectively. Here*

$$\begin{array}{ll} R_1^{(1)} = [\mathfrak{r}_{1,1}, 0, ..., 0, \mathfrak{r}_{N-1,1}]^\intercal, & R_1^{(2)} = [0, \mathfrak{r}_{2,1}, ..., \mathfrak{r}_{N-2,1}, 0]^\intercal, \\ R_j^{(1)} = [\mathfrak{r}_{1,j}, ..., \mathfrak{r}_{N-1,j}]^\intercal,\ j = 2, ..., N-2, & R_j^{(2)} = [0, ..., 0]^\intercal,\ j = 2, ..., N-2, \\ R_{N-1}^{(1)} = [\mathfrak{r}_{1,N-1}, 0, ..., 0, \mathfrak{r}_{N-1,N-1}]^\intercal, & R_{N-1}^{(2)} = [0, \mathfrak{r}_{2,N-1}, ..., \mathfrak{r}_{N-2,N-1}, 0]^\intercal. \end{array} \tag{4.13}$$

Then,

$$\max_{1 \le m \le (N-1)^2} |\left((I \otimes B^{-1}) R^{(1)}(t)\right)_m| \le Ch^4, \quad 0 < t < t_f, \tag{4.14}$$

where $I \otimes B^{-1}$ represents the operator δ_x^{-4}, and

$$\max_{1 \le m \le (N-1)^2} |\left((B^{-1} \otimes I) R^{(2)}(t)\right)_m| \le Ch^4, \quad 0 < t < t_f, \tag{4.15}$$

where $(B^{-1} \otimes I)$ represents the operator δ_y^{-4}.

Proof. Using the definition of a Kronecker product, we have

$$I \otimes B = \begin{bmatrix} B & \mathbf{0} & ... & ... & \mathbf{0} \\ \mathbf{0} & B & \mathbf{0} & ... & \mathbf{0} \\ ... & & & & \\ \mathbf{0} & \mathbf{0} & ... & \mathbf{0} & B \end{bmatrix}, \qquad (I \otimes B)^{-1} = \begin{bmatrix} B^{-1} & \mathbf{0} & ... & ... & \mathbf{0} \\ \mathbf{0} & B^{-1} & \mathbf{0} & ... & \mathbf{0} \\ ... & & & & \\ \mathbf{0} & \mathbf{0} & ... & \mathbf{0} & B^{-1} \end{bmatrix}. \tag{4.16}$$

Therefore, $(I \otimes B^{-1}) R^{(1)}(t) = \left[B^{-1} R_1^{(1)}(t), B^{-1} R_2^{(1)}(t), ..., B^{-1} R_{N-2}^{(1)}(t), B^{-1} R_{N-1}^{(1)}(t)\right]^\intercal$. By the optimal convergence Theorem 2.1

$$\max_{1 \le m \le (N-1)^2} |\left((I \otimes B^{-1}) R^{(1)}(t)\right)_m| \le Ch^4, \quad 0 < t < t_f. \tag{4.17}$$

Hence (4.14) holds. By a similar proof (4.15) holds. ■

We return now to the error analysis.

Theorem 4.2. *Suppose that u, the solution to the system (4.1), has derivatives up to order 8, then the error $\mathfrak{e}$ is bounded by*

$$\max_{0 \le t \le t_f} |(-\tilde{\Delta}_h)^{1/2} \mathfrak{e}(t)|_h \le C(t_f) h^4, \quad 0 < t < t_f. \tag{4.18}$$

Here $-\tilde{\Delta}_h \mathbf{g}(t) = -(\tilde{\delta}_x^2 + \tilde{\delta}_y^2)\mathbf{g}(t)$ is a fourth-order approximation to the minus Laplacian operator, and $C(t_f)$ denotes a constant depending only on $\partial_t^{k_0} u$, $\partial_x^{k_1} u$, $\partial_y^{k_2} u$, $t \in [0, t_f]$, $(x, y) \in [0,1] \times [0,1]$, with $0 \le k_0 \le 2$, $4 \le k_1, k_2 \le 8$.

Proof. Let $R(t)$ represents the vector of the truncation error. Define $E(t) \triangleq \mathbf{vec}(\mathfrak{e}(t))$ as the vector containing the components of the error at time t

$$\mathfrak{e} = \left[\mathfrak{e}_{1,1}, ...\mathfrak{e}_{N-1,1}; \mathfrak{e}_{1,2}, ...\mathfrak{e}_{N-1,2}; ...; \mathfrak{e}_{1,N-1}, ...\mathfrak{e}_{N-1,N-1}\right]^\intercal \in \mathbb{R}^{(N-1)^2}. \tag{4.19}$$

The operator $\tilde{\Delta}_h^2$ may be represented by the matrix A of size $(N-1)^2 \times (N-1)^2$ (refer to Section 2.2 for the matrix notation.)

$$A = I \otimes B + B \otimes I + 2\Big[(\tilde{T} \otimes I)(I \otimes \tilde{T}) + \frac{h^2}{12}(B \otimes I)(I \otimes \tilde{T}) + \frac{h^2}{12}(I \otimes B)(\tilde{T} \otimes I)\Big] \succ 0. \tag{4.20}$$

We turn now to the matrix representation of the error. The operator $-\tilde{\Delta}_h$ may be represented by the matrix M of size $(N-1)^2 \times (N-1)^2$ (see [7]), where

$$M = I \otimes \tilde{T} + \tilde{T} \otimes I + \frac{h^2}{12}(I \otimes B) + \frac{h^2}{12}(B \otimes I). \tag{4.21}$$

Therefore, in vector notation, Equation (4.9) may be written as

$$\partial_t(ME(t)) + A\ E(t) = R(t). \tag{4.22}$$

Define

$$F(t) = M^{1/2}E, \tag{4.23}$$

Proceeding parallel to (3.6-3.12), we obtain that

$$F(t) = \int_0^t e^{-\tilde{A}\rho}\ M^{-1/2}\ R(t-\rho)d\rho. \tag{4.24}$$

Multiplying $M^{-1/2}R(\rho)$ from the left by $\tilde{A}\ (\tilde{A})^{-1}$, we have

$$F(t) = \int_0^t [e^{-\tilde{A}\rho}\ \tilde{A}]\ [\tilde{A}^{-1}\ M^{-1/2}\ R(t-\rho)]d\rho. \tag{4.25}$$

where the matrix $\tilde{A} = M^{-1/2}AM^{-1/2}$ is a symmetric positive-definite matrix. Therefore, $e^{-\tilde{A}\rho}\ \tilde{A}$ is also a symmetric matrix positive-definite matrix. It may be diagonalized by a a unitary matrix U. Thus,

$$e^{-\tilde{A}\rho}\ \tilde{A} = U\ \tilde{\Lambda}(\rho)\ U^{\intercal}, \tag{4.26}$$

where

$$\tilde{\Lambda}(\rho) = \mathrm{diag}\{e^{-\rho\lambda_1}\lambda_1, ..., e^{-\rho\lambda_{(N-1)^2}}\lambda_{(N-1)^2}\}. \tag{4.27}$$

Here, $\lambda_k,\ k = 1, ..., (N-1)^2$ are the eigenvalues of $\tilde{A}$, which are positive. Therefore,

$$F(t) = \int_0^t U\tilde{\Lambda}(\rho)\ U^{\intercal}\ [\tilde{A}^{-1}\ M^{-1/2}\ R(t-\rho)]d\rho. \tag{4.28}$$

Decomposing $R(t)$ into $R(t) = R^{(1)}(t) + R^{(2)}(t)$, where $R^{(1)}(t)$ and $R^{(2)}(t)$ are defined in (4.13)), we have

$$F(t) = \int_0^t [U\tilde{\Lambda}(\rho)\ U^{\intercal}]\ [\tilde{A}^{-1}\ M^{-1/2}]\ \Big[R^{(1)}(t-\rho) + R^{(2)}(t-\rho)\Big]d\rho. \tag{4.29}$$

Invoking $\tilde{A}^{-1} = M^{1/2}A^{-1}M^{1/2}$ and decomposing $F(t)$ in the sum $F(t) = F^{(1)}(t) + F^{(2)}(t)$ we have

$$\begin{cases} F^{(1)}(t) = \displaystyle\int_0^t [U\tilde{\Lambda}(\rho)\ U^{\intercal}]\ \Big[M^{1/2}A^{-1}\ R^{(1)}(t-\rho)\Big]d\rho, \\ F^{(2)}(t) = \displaystyle\int_0^t [U\tilde{\Lambda}(\rho)\ U^{\intercal}]\ \Big[M^{1/2}A^{-1}\ R^{(2)}(t-\rho)\Big]d\rho. \end{cases} \tag{4.30}$$

We show that $\|F^{(1)}(t)\|_2 \le Ch^3$. The equality $\|F^{(2)}(t)\|_2 \le Ch^3$ follows similarly. Notice that

$$\tilde{\Lambda}(\rho) = -\frac{d}{d\rho}\Lambda(\rho), \tag{4.31}$$

where

$$\Lambda(\rho) = \mathrm{diag}\{e^{-\rho\lambda_1}, ..., e^{-\rho\lambda_{(N-1)^2}}\}. \tag{4.32}$$

Inserting (4.31)-(4.32) into (4.30), we have

$$F^{(1)}(t) = -\int_0^t \frac{d}{d\rho}(U\Lambda(\rho)\ U^{\intercal})\big[M^{1/2}\ A^{-1}\ R^{(1)}(t-\rho)\big]d\rho. \tag{4.33}$$

Integration by parts yields

$$F^{(1)}(t) = [-(U\Lambda(\rho)\ U^{\intercal})\ M^{1/2}\ A^{-1}\ R^{(1)}(t-\rho)]_0^t + \int_0^t (U\Lambda(\rho)\ U^{\intercal})\ M^{1/2}\ A^{-1}\ \frac{d}{d\rho}R^{(1)}(t-\rho)d\rho. \tag{4.34}$$

We decompose $F^{(1)}(t)$ to $F^{(1,1)}(t) + F^{(1,2)}(t)$, where

$$\begin{cases} F^{(1,1)}(t) = \Big[-(U\Lambda(\rho)\ U^{\intercal})\ M^{1/2}\ A^{-1}\ R^{(1)}(t-\rho)\Big]_{\rho=0}^t \\ F^{(1,2)}(t) = \displaystyle\int_0^t (U\Lambda(\rho)\ U^{\intercal})\ M^{1/2}\ A^{-1}\ \frac{d}{d\rho}R^{(1)}(t-\rho)d\rho. \end{cases} \tag{4.35}$$

We have

$$\begin{aligned} F^{(1,1)}(t) &= (U\Lambda(0)\ U^{\intercal})\ M^{1/2}\ A^{-1}\ R^{(1)}(t) - (U\Lambda(t)\ U^{\intercal})\ M^{1/2}\ A^{-1}\ R^{(1)}(0) \\ &= (U\Lambda(0)\ U^{\intercal})\ M^{1/2}\ A^{-1}\ R^{(1)}(t). \end{aligned} \tag{4.36}$$

We will show in Lemma 5.3 that $\|M^{1/2}\ A^{-1}\ R^{(1)}(t-\rho)\|_2 \le C(t)h^3$. In addition $\|\Lambda(\rho)\|_2 \le 1$ for $\rho \ge 0$ and $\|U\|_2 = \|U^{\intercal}\|_2 = 1$. Therefore,

$$\|F^{(1,1)}(t)\|_2 \le C(t)h^3. \tag{4.37}$$

We turn now to $F^{(1,2)}(t)$. Notice that $R^{(1)}(\rho)$ (and therefore $\frac{d}{d\rho}R^{(1)}(t-\rho)$) contains fifth and eighth-order derivatives of u with respect to x, y at some intermediate points $(\bar{x}, \bar{y}, \rho)$, where $(\bar{x}, \bar{y}) \in [0,1] \times [0,1]$. Therefore, the time-derivative $\frac{d}{d\rho}R(t-\rho)$ contains first order time-derivative of fifth and eighth-order spatial derivatives of u, where they are all assumed to be bounded.

Since the components of $\frac{d}{d\rho}R^{(1)}(t-\rho)$ are of the same order (as powers of h) as the components of $R^{(1)}(t-\rho)$, we may apply Lemma 5.3 to $\tilde{R} = \frac{d}{d\rho}R(t-\rho)$ and obtain

$$\|M^{1/2}\ A^{-1}\ \frac{d}{d\rho}R^{(1)}(t-\rho)\|_2 \le C(t)h^3.$$

In addition, since $\|U\|_2 = \|U^{\intercal}\|_2 = 1$ and $\|\Lambda(\rho)\|_2 \le 1$, for $0 \le \rho \le t$, therefore

$$\|F^{(1,2)}\|_2 \le t \max_{0\le\rho\le t} \|M^{1/2}\ A^{-1}\ \frac{d}{d\rho}R^{(1)}(\rho)]\|_2 \le C(t)h^3. \tag{4.38}$$

From (4.37)-(4.38), we conclude that

$$\|F^{(1)}(t)\|_2 \le C(t)h^3. \tag{4.39}$$

Similarly,

$$\|F^{(2)}(t)\|_2 \le C(t)h^3. \tag{4.40}$$

Therefore, $\|F(t)\|_2 \le C(t)h^3$. Noting that $F(t) = M^{-1/2}E(t)$, we conclude that

$$|(-\tilde{\Delta}_h)^{1/2}\mathfrak{e}(t)|_h \le C(t)h^4. \tag{4.41}$$

■

5. Fourth order estimates for the matrices $D^{1/2}B^{-1}R$ and $M^{1/2}A^{-1}\bar{R}$

In this section we provide fourth-order estimates for $\tilde{T}^{1/2}B^{-1}R$, $\tilde{T}^{1/2}B^{-1}R$ and $M^{1/2}A^{-1}\bar{R}$, where R and $\bar{R}$ is the vector which contains the truncation errors in 1D and 2D, respectively. In subsections 5.1, 5.2 and 5.3 we provide a bounds for $\tilde{T}^{1/2}B^{-1}R$, $D^{1/2}B^{-1}R$ and $M^{1/2}A^{-1}\bar{R}$, respectively.

Consider the grid function $\mathfrak{r}(t)$ in (2.14) with its components $R_i(t)$. The vector $R(t)$ is the source term in (3.6). In fact, it appears in the expression $D^{1/2}B^{-1}R(t)$, which is part of the integrand of Equation (3.16) In terms of grid functions, it consists of estimating

$$(-\tilde{\delta}_x^2)^{1/2}(\delta_x^4)^{-1}\mathfrak{r} \tag{5.1}$$

as a function of h. However, Theorem 2.2 provides only the estimate

$$|(\delta_x^4)^{-1}\mathfrak{r}| \le Ch^4, \tag{5.2}$$

which is not sufficient for our purpose. The goal in this section is to prove a sharper estimate

$$|(-\tilde{\delta}_x^2)^{1/2}(\delta_x^4)^{-1}\mathfrak{r}| \le Ch^4. \tag{5.3}$$

5.1. **Estimate for $\tilde{T}^{1/2}B^{-1}R$.** The proof of Theorem 2.3 is based on accurate representations of the vector $B^{-1}R$. Here, instead of estimating $D^{1/2}B^{-1}R$, we consider, as a first step, estimating $\tilde{T}^{1/2}B^{-1}R$.

Lemma 5.1. *Assume that the vector $R \in \mathbb{R}^{N-1}$ satisfies*

$$PR = [Ch, Ch^4, \dots, Ch^4, Ch]^{\intercal}, \tag{5.4}$$

then all components of $(\tilde{T}^{1/2}B^{-1})R$ are $O(h^4)$, i.e.,

$$|(\tilde{T}^{1/2}B^{-1}R)_i| \le Ch^4, \;\; 1 \le i \le N-1. \tag{5.5}$$

Proof. The proof proceeds along the same lines of Theorem 2.2. We replace the matrix B^{-1} by the matrix $\tilde{T}^{1/2}B^{-1}$. By Equation (2.32)

$$P^{-1}B^{-1}P^{-1} = HG^{-1}. \tag{5.6}$$

For $\tilde{T}^{1/2}B^{-1}$ we have

$$P^{-1}\tilde{T}^{1/2}B^{-1}P^{-1} = \tilde{H}G^{-1}, \tag{5.7}$$

where $\tilde{H} = \tilde{T}^{1/2}H \in \mathrm{Span}(T)$, (since $\tilde{T} = T/h^2$). Thus, (see (2.31) for the definition of H)

$$\tilde{H} = \frac{h^3}{6}P^{-1}T^{-3/2}. \tag{5.8}$$

As in [16, 2], we bound the elements of $\tilde{H}$ in terms of h. Similar to Equation (10.140) in [2], we have for $\tilde{H}$

$$\tilde{H}_{i,k} = \sum_{j=1}^{N-1} h^3 \frac{2}{N} \frac{\sin(\frac{ij\pi}{N})\sin(\frac{jk\pi}{N})}{\sin^3(\frac{j\pi}{2N})(6-4\sin^2(\frac{j\pi}{2N}))}. \tag{5.9}$$

In particular

$$\tilde{H}_{i,1} = \sum_{j=1}^{N-1} 2h^4 \frac{\sin(\frac{ij\pi}{N})\sin(\frac{j\pi}{N})}{\sin^3(\frac{j\pi}{2N})(6-4\sin^2(\frac{j\pi}{2N}))}. \tag{5.10}$$

Since for $0 \le x \le \pi/2$, we have $\frac{2}{\pi}|x| \le |\sin x| \le |x|$ and $2 \le 6 - 4\sin^2(\frac{j\pi}{2N}) \le 6$. This gives

$$0 \le \tilde{H}_{1,1} = \tilde{H}_{N-1,N-1} \le C\sum_{j=1}^{N-1} \Big|\frac{h^4(jh)^2}{(jh)^3}\Big| \le Ch^3\big|\ln(h)\big|, \tag{5.11}$$

and

$$|\tilde{H}_{1,N-1}| = |\tilde{H}_{N-1,1}| \le Ch^3|\ln(h)|. \tag{5.12}$$

In addition,

$$|\tilde{H}_{i,1}| \le C\sum_{j=1}^{N-1} \frac{h^4 jh}{(jh)^3} \le Ch^2, \tag{5.13}$$

and

$$|\tilde{H}_{i,k}| \le C\sum_{j=1}^{N-1} \frac{h^4}{(jh)^3} \le Ch, \quad i,k = 2, ..., N-2. \tag{5.14}$$

Therefore,

$$\tilde{H} = \begin{bmatrix} O(h^3|\ln(h)|) & O(h^2) & \dots & O(h^2) & O(h^3|\ln(h)|) \\ O(h^2) & O(h) & \dots & O(h) & O(h^2) \\ \vdots & \vdots & \dots & \dots & \vdots \\ O(h^2) & O(h) & \dots & O(h) & O(h^2) \\ O(h^3|\ln(h))| & O(h^2) & \dots & O(h^2) & O(h^3|\ln(h)|) \end{bmatrix}. \tag{5.15}$$

Invoking the estimates in (10.162), (10.163) in [2] and in [16], we have

$$G^{-1}R(t) = [O(h^2), O(h^4), ..., O(h^4), O(h^2)]^\intercal. \tag{5.16}$$

Therefore,

$$\tilde{H}G^{-1}R(t) = \begin{bmatrix} O(h^3|\ln(h)|) & O(h^2) & \dots & O(h^2 & O(h^3|\ln(h)|) \\ O(h^2) & O(h) & \dots & O(h) & O(h^2) \\ \vdots & \vdots & \dots & \dots & \vdots \\ O(h^2) & O(h) & \dots & O(h) & O(h^2) \\ O(h^3|\ln(h)|) & O(h^2) & \dots & O(h^2) & O(h^3|\ln(h)|) \end{bmatrix} \begin{bmatrix} O(h^2) \\ O(h^4) \\ \vdots \\ O(h^4) \\ O(h^2) \end{bmatrix} = \begin{bmatrix} O(h^5\ln(h)) \\ O(h^4) \\ \vdots \\ O(h^4) \\ O(h^5\ln(h)) \end{bmatrix}. \tag{5.17}$$

Thus, the components of the matrix $\tilde{T}^{1/2}B^{-1}R(t)$ are

$$\left(\tilde{T}^{1/2}B^{-1}R(t)\right)_j = \left(\tilde{H}G^{-1}\tilde{R}(t)\right)_j = O(h^4), \quad 1 \le j \le N-1. \tag{5.18}$$

■

5.2. **Estimate for $D^{1/2}B^{-1}R$.**

Lemma 5.2. *Let D be*

$$D = \tilde{T} + \frac{h^2}{12}B. \tag{5.19}$$

Assume that the vector $R \in \mathbb{R}^{N-1}$ satisfies

$$PR = [Ch, Ch^4, \dots, Ch^4, Ch]^\intercal, \tag{5.20}$$

then

$$\|D^{1/2}B^{-1}R\|_2 \le Ch^{3.5}. \tag{5.21}$$

Proof. Denoting $U = D^{1/2}B^{-1}R$, $U^\intercal U$ is expressed as

$$U^\intercal U = R^\intercal\ B^{-1}\ D\ B^{-1}R. \tag{5.22}$$

Using (5.19)-(5.22) gives

$$U^\intercal U = R^\intercal\ B^{-1}\ \tilde{T}\ B^{-1}R + (h^2/12)\ R^\intercal\ B^{-1}R. \tag{5.23}$$

Define

$$\begin{aligned} J &= R^\intercal\ B^{-1}\ \tilde{T}\ B^{-1}R, \\ K &= \left((h^2/12)\ R\right)^\intercal\ B^{-1}R. \end{aligned} \tag{5.24}$$

We look first at the second term K in $U^\intercal T$. By Theorem 2.2 $B^{-1}R = [O(h^4), O(h^4), ..., O(h^4), O(h^4)]$, we find that

$$\begin{aligned} (h^2/12)R &= [O(h^3), O(h^6), ..., O(h^6), O(h^3)], \\ B^{-1}R(t) &= [O(h^4), O(h^4), ..., O(h^4), O(h^4)]. \end{aligned} \tag{5.25}$$

Therefore,

$$K = O(h^7). \tag{5.26}$$

As for the first term J in $U^\intercal U$, denoting by U_1

$$U_1 = (\tilde{T})^{1/2}\ B^{-1}R, \tag{5.27}$$

we have

$$J = U_1^\intercal U_1 = R^\intercal\ B^{-1}\ \tilde{T}\ B^{-1}R. \tag{5.28}$$

It has been shown in (5.18) that

$$U_1 = [O(h^4), O(h^4), ..., O(h^4), O(h^4)]. \tag{5.29}$$

Therefore,

$$J = U_1^\intercal U_1 = O(h^7). \tag{5.30}$$

Combining (5.26) and (5.29) yields

$$U^\intercal U = J + K = O(h^7). \tag{5.31}$$

■

5.3. **Estimate for $M^{1/2}A^{-1}\tilde{R}$.** We prove now that $\|M^{1/2}A^{-1}\ R^{(1)}(t-\rho)\|_2 \le Ch^3$. To simplify the notation, let us define $\tilde{R} = R^{(1)}(t-\rho)$ [a similar Lemma is true for $\tilde{R} = \frac{d}{d\rho}R^{(1)}(t-\rho)$].

Lemma 5.3. *Let M be (see (4.21))*

$$M = I\otimes D + D\otimes I = I\otimes \tilde{T} + \frac{h^2}{12}(I\otimes B) + \tilde{T}\otimes I + \frac{h^2}{12}(B\otimes I) \tag{5.32}$$

and let $\tilde{R} = R^{(1)}$, where $R^{(1)}$ is defined in (4.13). Then,

$$\|M^{1/2}\ A^{-1}\ \tilde{R}\|_2 \le C(t)h^3. \tag{5.33}$$

Proof. Consider now

$$(\tilde{R})^\intercal A^{-1}M\ A^{-1}\ \tilde{R}. \tag{5.34}$$

Invoking (4.21), we have

$$A^{-1}M\ A^{-1} = A^{-1}\big[I\otimes D + D\otimes I\big]\ A^{-1}. \tag{5.35}$$

Splitting the sum above to two terms $I\otimes D$ and $D\otimes I$, we have

$$A^{-1}M\ A^{-1} = A^{-1}\ (I\otimes D)\ A^{-1} + A^{-1}\ (D\otimes I)\ A^{-1} = P_1 + P_2, \tag{5.36}$$

where

$$\begin{aligned} P_1 &= A^{-1}\ (I\otimes D)\ A^{-1}, \\ P_2 &= A^{-1}\ (D\otimes I)\ A^{-1}. \end{aligned} \tag{5.37}$$

First, we consider P_1. Since the elements of $(I\otimes D)^{1/2}\ A^{-1}\tilde{R}$ are smaller or equal in terms of orders of h compared to $(I\otimes D)^{1/2}\ (I\otimes B^{-1})\tilde{R}$, we will bound the latter. Thus, we want to bound now $\tilde{R}^\intercal\ (I\otimes B^{-1})(I\otimes D)\ (I\otimes B^{-1})\tilde{R}$.

Recall that

$$D = \tilde{T} + \frac{h^2}{12}B. \tag{5.38}$$

We will show that

$$\|(\tilde{R})^\intercal\ P_1\ \tilde{R}\|_2 \le C\|(\tilde{R})^\intercal[(I\otimes B)^{-1}\ (I\otimes D)\ (I\otimes B)^{-1}]\ \tilde{R}\|_2 \le Ch^6. \tag{5.39}$$

This is equivalent to showing that

$$\|\ (I\otimes D)^{1/2}\ (I\otimes B)^{-1}\ \tilde{R}\|_2 \le C(t)h^3. \tag{5.40}$$

It is shown in Lemma 5.2

$$\|D^{1/2}B^{-1}R\|_2 \le Ch^{3.5},$$

therefore

$$\|\ (I\otimes D)^{1/2}\ (I\otimes B)^{-1}\ \tilde{R}\|_2 \le Ch^3. \tag{5.41}$$

Thus, we conclude that

$$(\tilde{R})^\intercal P_1\tilde{R} \le C(t)h^6. \tag{5.42}$$

Second, we consider P_2, defined in (5.37). In Lemma 5.4 we prove that

$$P_2 = (\tilde{R})^\intercal P_2\tilde{R} \le C(t)h^6. \tag{5.43}$$

Therefore, from $A^{-1}M\ A^{-1} = P_1 + P_2$, we will conclude that

$$\|M^{1/2}A^{-1}\tilde{R}\|_2 \leq Ch^3. \tag{5.44}$$

■

Lemma 5.4. *We show that* $(R^{(1)})^\intercal P_2 R^{(1)} \leq Ch^6$, *where* P_2 *is defined*

$$P_2 = A^{-1}(D \otimes I)A^{-1}, \tag{5.45}$$

and $R^{(1)}$ *is defined in (4.13).*

Proof. Here we look at $R^{(1)} = R^{(1,1)} + R^{(1,2)}$, where
$R^{(1,1)} = \left[R_1^{(1,1)}; ...; R_{N-1}^{(1,1)}\right]$ and $R^{(1,2)} = \left[R_1^{(1,2)}; ...; R_{N-1}^{(1,2)}\right]$, respectively, where

$$\begin{array}{ll} R_1^{(1,1)} = [\mathfrak{r}_{1,1}, 0, ..., 0, \mathfrak{r}_{N-1,1}]^\intercal, & R_1^{(1,2)} = [0, 0, ..., 0, 0]^\intercal, \\ R_j^{(1,1)} = [\mathfrak{r}_{1,j}, 0, ..., 0, \mathfrak{r}_{N-1,j}]^\intercal,\ j = 2, ..., N-2, & R_j^{(1,2)} = [0, \mathfrak{r}_{2,j}, ..., \mathfrak{r}_{N-2,j}, 0]^\intercal,\ j = 2, ..., N-2, \\ R_{N-1}^{(1,1)} = [\mathfrak{r}_{1,N-1}, 0, ..., 0, \mathfrak{r}_{N-1,N-1}]^\intercal. & R_{N-1}^{(1,2)} = [0, 0, ..., 0, 0]^\intercal. \end{array} \tag{5.46}$$

First, we consider $(\tilde{R})^\intercal P_2 \tilde{R}$, where we take $\tilde{R} = R^{(1,1)}$. Note that $P_2 = A^{-1}(D \otimes I)A^{-1}$.

Consider now

$$(\tilde{R})^\intercal P_2 \tilde{R} = \tilde{R}^\intercal A^{-1}(D \otimes I)A^{-1}\tilde{R}. \tag{5.47}$$

We want to determine the order with respect to h of the components of $(D^{1/2} \otimes I)A^{-1}\tilde{R}$. Notice that the elements in $\tilde{R}$ related to $i = 2, ..., N-2$ are zero for $j = 1, ..., N-1$. Consider the operator $(D^{1/2} \otimes I)(I \otimes B)^{-1}$. This operator approximates $\partial_y(\delta_x^4)^{-1}$. Thus, for $i = 2, ..., N-2$, $j = 1, ..., N-1$ we have that $(D^{1/2} \otimes I)A^{-1}\tilde{R}$ is zero. Consider now $i = 1, N-1$. The elements of $\tilde{R}$ which are related to $i = 1, N-1$ are $O(h)$ for any $j = 1, .., N-1$. Note that the elements of $A^{-1}\tilde{R}$ are smaller or equal compared to the magnitudes of the elements of $(I \otimes B)^{-1}\tilde{R}$, therefore we consider now $(I \otimes B)^{-1}\tilde{R}$. We have that $(I \otimes B^{-1}) = (I \otimes H)(I \otimes G^{-1})$. We also have that $G_{1,1}^{-1}, G_{N-1,1}^{-1}, G_{1,N-1}^{-1}, G_{N-1,N-1}^{-1}$ are $O(h)$, therefore the elements of $(I \otimes G^{-1})\tilde{R}$ related to $i = 1, N-1$ are $O(h^2)$ for $j = 1, ..., N-1$. In addition, we have that $H_{1,1}, H_{N-1,1}, H_{1,N-1}, H_{N-1,N-1}$ are $O(h^3)$. Therefore, the elements of $(I \otimes B^{-1})\tilde{R}$ which are related to $i = 1, N-1$ are $O(h^5)$ for all $j = 1, ..., N-1$. Operating on $(I \otimes B)^{-1}\tilde{R}$ with $(D^{1/2} \otimes I)$ results in elements which are $O(h^4)$ for $i = 1, N-1$, $j = 1, ..., N-1$.

The vector $(D^{1/2} \otimes I)(I \otimes B)^{-1}\tilde{R}$ contains only $2(N-1)$ non zero terms, which are $O(h^4)$. Therefore,

$$\|(D^{1/2} \otimes I)(I \otimes B)^{-1}\tilde{R}\|_2 = O(h^3). \tag{5.48}$$

Thus,

$$\tilde{R}^\intercal (I \otimes B)^{-1}(D \otimes I)(I \otimes B)^{-1}\tilde{R} = O(h^6). \tag{5.49}$$

Therefore, we have

$$\tilde{R}^\intercal P_2 \tilde{R} \leq Ch^6. \tag{5.50}$$

As for

$$(\tilde{R}^{(1,2)})^\intercal A^{-1}(D \otimes I)\ A^{-1}\tilde{R}^{(1,2)}, \tag{5.51}$$

since $(D \otimes I)^{1/2}\ A^{-1}$ is a bounded operator, and since the non-zero components of $\tilde{R}^{(1,2)}$ are $O(h^4)$, we conclude that

$$(\tilde{R}^{(1,2)})^\intercal P_2 \tilde{R}^{(1,2)} = O(h^6). \tag{5.52}$$

Combining (5.50)-(5.52), we conclude that

$$(\tilde{R}^{(1)})^\intercal P_2 \tilde{R}^{(1)} \leq C\ h^6. \tag{5.53}$$

■

6. Truncation error representation

Let $u(x,t)$ be a regular function with associated grid function $u^*(t) = [u(x_1,t), u(x_2,t), \dots, u(x_{N-1},t)]$. The truncation error $\mathfrak{r}(u)(t)$, denoted for simplicity also by $\mathfrak{r}(t)$, is expressed as

$$\mathfrak{r}(t) = \big(\mathcal{L}u(.,t)\big)^* - \mathcal{L}_h u^*(t), \tag{6.1}$$

where $\mathcal{L}$ the continuous operator $\mathcal{L} = -\partial_{txx} + \partial_x^4$ and $\mathcal{L}_h$ the semi-discrete operator $\mathcal{L}_h = \partial_t(-\tilde{\delta}_x^2) + \delta_x^4$, with Dirichlet boundary conditions in both cases. Here, our goal is to derive an expression of the form

$$\sigma_x \mathfrak{r}(t) = \big[O(h), O(h^4), \cdots, O(h^4), O(h)\big]^{\mathsf{T}}, \tag{6.2}$$

with an effective *representation* of the $O(h^\alpha)$ components in terms of a spatial multi-point Taylor expansion of the functions $\partial_x^{(k)}u(x,t)$ and $\partial_t\partial_x^{(k)}u(x,t)$, (with fixed time t). Indeed, such a representation is more explicit than the formal identity (6.2). This form is used to accurately handling the estimates (3.29) and (3.31) in Section 3.

The analysis hereafter is elementary and relies on the standard Taylor-Lagrange expansions and convex combinations of derivatives. The following abbreviation is used for convenience. For a given regular function $x \in (0,1) \mapsto u(x)$, we denote

$$\begin{cases} u^k \triangleq \partial_x^{(k)} u, \\ u^{k,*} \triangleq (\partial_x^{(k)}u)^*, \quad u_i^{k,*} \triangleq (\partial_x^{(k)}u)(x_i). \end{cases} \tag{6.3}$$

The parenthesis is omitted when specifying an operation evaluated at a grid point x_i. For example, we denote $\delta_x u_i^* \triangleq \big(\delta_x u^*\big)_i$. The truncation error $\mathfrak{r}(t)$ is decomposed as $\mathfrak{r}(t) = -\mathfrak{r}_a(t) + \mathfrak{r}_b(t)$ with

$$\begin{cases} \mathfrak{r}_a(t) = \delta_x^4(u(.,t)^*) - u^{4,*}(t), \\ \mathfrak{r}_b(t) = \tilde{\delta}_x^2(\partial_t u)^*(.,t) - (\partial_t u)^{2,*}(t). \end{cases} \tag{6.4}$$

Concerning the Hermitian derivative, the following upper bound of the truncation error for the Hermitian derivative has been derived in [2] Lemma 10.1.

$$|\tilde{\delta}_x u_i^* - u_i^{1,*}| \le Ch^4 \|u^5\|_{\infty,(0,1)}, \tag{6.5}$$

with the constant $C = 1/60$. In [8], a representation of this truncation is given using the basis Z^k in (2.36). In the following Lemma, we give a simpler representation of the truncation error.

Lemma 6.1. *Let $u(x)$ be a regular function. The components of the truncation error $\tau(u) = \tilde{\delta}_x u^* - u^{1,*}$ can be expressed in the form*

$$\tau(u)_i = \tilde{\delta}_x u_i^* - u_i^{1,*} = h^4\Big(\gamma_i^1 u^5(\xi_i^1) - \gamma_i^2 u^5(\xi_i^2)\Big), \;\; 1 \le i \le N-1, \tag{6.6}$$

where

- *The scalar ξ_i^1, $\xi_i^2 \in (0,1)$.*
- *The scalars γ_i^1, γ_i^2 are fixed positive constants, independent of $u(x)$, and satisfying for all $i \in [\![1, N-1]\!]$ and N:*

$$\gamma_i^1 \le 0.0547, \quad \gamma_i^2 \le 0.0506. \tag{6.7}$$

Remark 6.2. *The scalars ξ_i^1 and ξ_i^2 are not necessarily close to x_i. This corresponds to the fact that $\mathfrak{u} \mapsto \tilde{\delta}_x \mathfrak{u}$ is a "non local" operator. They depend on $u(x)$.*

Proof. The 4th order Hermitian derivative $\tilde{\delta}_x u_j$, $1 \le j \le N-1$, is defined in (2.7). It is a non local finite difference operator. This non-locality gives that the truncation error results in a multi-point Taylor representation. The multi-point can be compacted using convex combination of derivatives into a two point representation. The proof is as follows. In vector form, we have $\sigma_x^{-1} \hookrightarrow [\alpha_{i,j}] \in \mathbb{M}_{N-1}$ with

$$[\alpha_{i,j}] = \begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & \dots & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & \dots & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & \dots & 0 & \frac{1}{6} & \frac{2}{3} \end{bmatrix}^{-1}. \tag{6.8}$$

The matrix $[\alpha_{i,j}]$ is the inverse of a tridiagonal Toeplitz positive definite matrix. According to [29], the coefficients $\alpha_{i,j}$ are

$$\alpha_{i,j} = (-1)^{i+j}\beta_{i,j}, \quad \beta_{i,j} = |\alpha_{i,j}|, \quad 1 \le i, j \le N-1, \tag{6.9}$$

where

$$\begin{cases} \beta_{i,j} = 6\dfrac{(\lambda^i - \mu^i)(\lambda^{N-j} - \mu^{N-j})}{(\lambda-\mu)(\lambda^N - \mu^N)}, \quad 1 \le i \le j \le N-1, \\ \beta_{i,j} = \beta_{j,i}, \quad 1 \le i \le j \le N-1. \end{cases} \tag{6.10}$$

The constants λ and μ are $\lambda = 2+\sqrt{3} \simeq 3.73, \quad \mu = 2-\sqrt{3} \simeq 0.28$. The matrix $[\alpha_{i,j}]$ is full. The $\beta_{i,j}$ behave as

$$0 < \beta_{i,1} < \beta_{i,2} < \ldots, \beta_{i,i-1} < \beta_{i,i} > \beta_{i,i+1} \cdots > \beta_{i,N-1} > 0, \quad 1 \le i \le N-1, \tag{6.11}$$

with

$$\begin{cases} 0 < \beta_{i,i} < \dfrac{6}{\lambda-\mu} = \sqrt{3}, \\ \dfrac{\beta_{i,j}}{\beta_{i,j+1}} > \lambda > 1, \quad i \le j \le N-1, \quad 1 \le i \le N-1, \\ \dfrac{\beta_{i,j}}{\beta_{i,j-1}} > \lambda > 1, \quad 1 \le j \le i-1, \quad 1 \le i \le N-1. \end{cases} \tag{6.12}$$

The truncation error τ of $\tilde{\delta}_x \mathfrak{u}$ satisfies

$$(\sigma_x \tau)_i = \delta_x u_i^* - \sigma_x u_i^{1,*}, \quad 1 \le i \le N-1. \tag{6.13}$$

Taylor-Lagrange expansions of u_{i+1}^* and u_{i-1}^* give that

$$\delta_x u_i^* = u_i^{1,*} + h^2 \frac{1}{3!} u_i^{3,*} + h^4 \frac{1}{5!} u^5(\eta_i^1), \quad \eta_i^1 \in (x_{i-1}, x_{i+1}). \tag{6.14}$$

By definition of σ_x is $\sigma_x u_i^{1,*} = u_i^{1,*} + \frac{h^2}{3!}\delta_x^2 u_i^{1,*}$. Furthermore, by Taylor expansion

$$\delta_x^2 u_i^{1,*} = u_i^{3,*} + h^2 \frac{2}{4!} u^5(\eta_i^2), \quad \eta_i^2 \in (x_{i-1}, x_{i+1}). \tag{6.15}$$

It results from (6.13)-(6.15) that

$$\begin{aligned} (\sigma_x \tau)_i &= \delta_x u_i - \sigma_x u_i^{1,*} \\ &= u_i^{1,*} + h^2 \frac{1}{3!} u_i^{3,*} + h^4 \frac{1}{5!} u^5(\eta_i^1) - \Big(I + h^2 \frac{1}{3!}\delta_x^2\Big) u_i^{1,*} \\ &= h^4 \Big(\frac{1}{5!} u^5(\eta_i^1) - \frac{1}{2\,3!^2} u^5(\eta_i^2) \Big). \end{aligned} \tag{6.16}$$

Multiplying the vector $[(\sigma_x\tau)_1, \ldots, (\sigma_x\tau)_{N-1}]^T$ by the matrix $[\alpha_{i,j}]$ in (6.8) yields

$$\tau_i = h^4 \sum_{j=1}^{N-1} (-1)^{i+j} \beta_{i,j} \Big(\frac{1}{5!} u^5(\eta_j^1) - \frac{1}{2\,3!^2} u^5(\eta_j^2) \Big). \tag{6.17}$$

The sum in (6.17) is the difference of two sums, both with positive coefficients. Defining the two sets of indices $J_{\text{even}}(i) = \{1 \le j = i + 2p \le N-1\}$ and $J_{\text{odd}}(i) = \{1 \le j = i + 2p + 1 \le N-1\}$, $p \in \mathbb{Z}$, one has

$$\begin{aligned} \tau_i = h^4 \Big[\frac{1}{5!} \sum_{j \in J_{\text{even}}(i)} \beta_{i,j} u^5(\eta_j^1) + \frac{1}{2\,3!^2} \sum_{j \in J_{\text{odd}}(i)} \beta_{i,j} u^5(\eta_j^2) \Big] \\ - h^4 \Big[\frac{1}{5!} \sum_{j \in J_{\text{odd}}(i)} \beta_{i,j} u^5(\eta_j^1) + \frac{1}{2\,3!^2} \sum_{j \in J_{\text{even}}(i)} \beta_{i,j} u^5(\eta_j^2) \Big]. \end{aligned} \tag{6.18}$$

The first and second term can be collected in one single term by convex combination of the derivatives. Define

$$\begin{cases} \gamma_i^1 = \dfrac{1}{5!} \displaystyle\sum_{j\in J_{\rm even}(i)} \beta_{i,j} + \dfrac{1}{2\,3!^2} \sum_{j\in J_{\rm odd}(i)} \beta_{i,j}, \\ \gamma_i^2 = \dfrac{1}{5!} \displaystyle\sum_{j\in J_{\rm odd}(i)} \beta_{i,j} + \dfrac{1}{2\,3!^2} \sum_{j\in J_{\rm even}(i)} \beta_{i,j}. \end{cases} \tag{6.19}$$

Due to (6.12), the scalars γ_i^1, γ_i^2 are well defined and uniformly bounded (independently of i and N). This gives that there exists $\xi_i^1, \xi_i^2 \in (0,1)$ such that (6.6) holds. ■

We next derive representation formulas for $\mathfrak{r}_a(t)$ and $\mathfrak{r}_b(t)$ in (6.4) having a form similar to (6.6: they use a set of points in $(0,1)$ depending of the function $u(x)$.

Lemma 6.3 (Truncation term τ_a)**.** *For any regular function $u(x)$, consider the truncation error $\mathfrak{r}_a(u)$ defined by*

$$\mathfrak{r}_a(u) = \delta_x^4 u^* - u^{4,*}. \tag{6.20}$$

The components of $\sigma_x(\mathfrak{r}_a)$ are of the following orders of h

$$\sigma_x \mathfrak{r}_a = [O(h), O(h^4), \dots O(h^4), O(h)]^\intercal. \tag{6.21}$$

The structure of (6.21) is as follows. There exists positive constants β_i^k uniformly bounded independently of i and N, such that

- *At the internal points x_i, $2 \le i \le N-2$,*

$$(\sigma_x \mathfrak{r}_a)_i = h^4\Big(\beta_i^1 u^8(\xi_i^1) - \beta_i^2 u^8(\xi_i^2)\Big), \tag{6.22}$$

  with constants $\beta_i^1 = \frac{2^6}{7!} \simeq 0.0127$ and $\beta_i^2 = \frac{71}{7!} \simeq 0.0141$ and for some $\xi_i^1 \in (x_{i-2}, x_{i+2})$, $\xi_i^2 \in (x_{i-1}, x_{i+1})$.
- *At the near boundary point x_1, we have*

$$(\sigma_x \mathfrak{r}_a)_i = h\Big(\beta_1^3 u^5(\xi_i^3) - \beta_1^4 u^5(\xi_i^4)\Big) + h^2\Big(\beta_1^5 u^6(\xi_i^5) - \beta_1^6 u^6(\xi_i^6)\Big), \tag{6.23}$$

  for some ξ_i^3, ξ_i^4, ξ_i^5, $\xi_i^6 \in (0, x_3)$. A similar expression holds at the point x_{N-1}.

Remark 6.4. *Note that (6.22) and (6.23) contain a finite number of terms, which are uniformly bounded in terms of derivatives of $u(x)$. The points ξ_i^j depend on the function $u(x)$.*

Proof. For the truncation analysis at interior points $2 \le i \le N-2$, we refer to [2, chap. 10.4, pp. 160 sqq]. It is based on Taylor-Lagrange expansions at point x_i of

$$(\sigma_x \delta_x^4)\mathfrak{v}_i = (\delta_x^2)^2 \mathfrak{v}_i = \frac{1}{h^4}\big(\mathfrak{v}_{i+2} - 4\mathfrak{v}_{i+1} + 6\mathfrak{v}_i - 4\mathfrak{v}_{i-1} + \mathfrak{v}_{i-2}\big), \quad 2 \le i \le N-2, \tag{6.24}$$

and of

$$\sigma_x u_i^{4,*} = \frac{1}{6} u_{i-1}^{4,*} + \frac{2}{3} u_i^{4,*} + \frac{1}{6} u_{i+1}^{4,*}, \quad 2 \le i \le N-2. \tag{6.25}$$

Subtracting these two expansions results in (6.22).

Next consider x_1 and x_2. We have

$$\begin{aligned} (\sigma_x \mathfrak{r}_a)_1 &= \sigma_x \delta_x^4 u_1^* - \sigma_x u_1^{4,*} \\ &= \frac{2}{3}(\delta_x^4 u_1^* - u_1^{4,*}) + \frac{1}{6}(\delta_x^4 u_2^* - u_2^{4,*}). \end{aligned} \tag{6.26}$$

We have

$$(\delta_x^4 u^*)_1 = \frac{12}{h^2}\Big((\delta_x \tilde{\delta}_x) u_1^* - \delta_x^2 u_1^*\Big), \qquad (\delta_x^4 u^*)_2 = \frac{12}{h^2}\Big(\delta_x \tilde{\delta}_x u_2^* - \delta_x^2 u_2^*\Big). \tag{6.27}$$

For all regular functions $v(x)$, the Taylor-Lagrange formula for all $1 \le i \le N-1$ gives

$$\delta_x^2 v_i^* = v_i^{2*} + h^2 \frac{2}{4!} v_i^{4,*} + h^4 \frac{2}{6!} v^6(\eta_i^9), \quad \eta_i^9 \in (x_{i-1}, x_{i+1}). \tag{6.28}$$

In particular, for $i=1$,

$$\delta_x^2 u_1^* = u_1^{2,*} + h^2\frac{2}{4!}u_1^{4,*} + h^4\frac{2}{6!}u^6(\eta_1^9), \quad \eta_1^9 \in (x_0,x_2). \tag{6.29}$$

Consider in (6.27) the term $(\delta_x\tilde{\delta}_x)u_1^*$. By Lemma 6.1 and by (6.14) (recalling that $u_0^{1,*}=\tilde{\delta}_x u_0^*=0$), we have

$$\begin{aligned}(\delta_x\tilde{\delta}_x)u_1^* &= \frac{1}{2h}(\tilde{\delta}_x u_2^* - \tilde{\delta}_x u_0^*) = \frac{1}{2h}\tilde{\delta}_x u_2^* = \frac{1}{2h}\Big(u_2^{1,*} - u_0^{1,*} + h^4\big(\gamma_2^1 u^5(\xi_2^1) - \gamma_2^2 u^5(\xi_2^2)\big)\Big)\\ &= u_1^{2,*} + h^2\frac{1}{3!}u_1^{4,*} + h^3\frac{1}{2}\Big(\gamma_2^1 u^5(\xi_2^1) - \gamma_2^2 u^5(\xi_2^2)\Big) + h^4\frac{1}{5!}u^6(\eta_1^{10}), \quad \xi_2^1,\xi_2^2 \in (0,1),\ \ \eta_1^{10}\in(x_0,x_2).\end{aligned} \tag{6.30}$$

By (6.27), it results that

$$\delta_x^4 u_1^* - u_1^{4,*} = 6h\Big(\gamma_2^1 u^5(\xi_2^1) - \gamma_2^2 u^5(\xi_2^2)\Big) + h^2\Big(\frac{12}{5!}u^6(\eta_1^{10}) - \frac{24}{6!}u^6(\eta_1^9)\Big). \tag{6.31}$$

For $i=2$, we have for some $\eta_2^{11}\in(x_1,x_3)$, $\delta_x^2 u_2^* = u_2^{2,*} + h^2\frac{2}{4!}u_2^{4,*} + h^4\frac{2}{6!}u^6(\eta_2^{11})$ and for some $\eta_2^{12}\in(x_1,x_3)$,

$$(\delta_x\tilde{\delta}_x)u_2^* = u_2^{2,*} + h^2\frac{1}{3!}u_2^{4,*} + h^3\frac{1}{2}\Big((\gamma_3^1 u^5(\xi_3^1) + \gamma_1^1 u^5(\xi_1^1)) - (\gamma_3^2 u^5(\xi_3^2) + \gamma_1^2 u^5(\xi_1^2))\Big) + h^4\frac{1}{5!}u^6(\eta_2^{12}). \tag{6.32}$$

Thus, by (6.27)

$$\delta_x^4 u_2^* - u_2^{4,*} = 6h\Big((\gamma_1^1 u^5(\xi_1^1) + \gamma_3^1 u^5(\xi_3^1)) - (\gamma_1^2 u^5(\xi_1^2) + \gamma_3^2 u^5(\xi_3^2))\Big) + h^2\Big(\frac{12}{5!}u^6(\eta_2^{12}) - \frac{24}{6!}u^6(\eta_2^{11})\Big). \tag{6.33}$$

It results from (6.31) and (6.33) that $(\sigma_x\mathfrak{r}_a)_1$, given in (6.26), satisfies (6.23) with $\beta_1^3 = \gamma_1^1 + 4\gamma_2^1 + \gamma_3^1$, $\beta_1^4 = \gamma_1^2 + 4\gamma_2^2 + \gamma_3^2$, $\beta_1^5 = 1/12$, $\beta_1^6 = 1/36$. The proof for $i = N-1$ is similar. ■

Consider now the term $\mathfrak{r}_b(t)$ in (6.4). Denoting $v(x) = \partial_t u(x,t)$, we have

$$\mathfrak{r}_b(t) = \tilde{\delta}_x^2 v^*(t) - v^{2,*}(t). \tag{6.34}$$

Lemma 6.5 (Truncation term τ_b)**.** *For fixed $t>0$, the truncation error $\mathfrak{r}_b(t)$ satisfies*

$$\sigma_x\mathfrak{r}_b = [O(h^2), O(h^4), \dots O(h^4), O(h^2)]^\intercal. \tag{6.35}$$

There exists positive constants $\beta_i^{\prime,k}$ uniformly bounded with respect to N such that the components of $\sigma_x\mathfrak{r}_b$ are represented as follows.

- *At internal points x_i, i.e., at x_i, for $2\le i\le N-2$,*

$$\sigma_x\mathfrak{r}_{b,i} = h^4\Big(\beta_i^{\prime,1} v^6(\xi_i^{\prime,1})\Big) + h^6\Big(\beta_i^{\prime,2} v^8(\xi_i^{\prime,2}) - \beta_i^{\prime,3} v^8(\xi_i^{\prime,3}))\Big), \tag{6.36}$$

  for some $\xi_i^{\prime,1}, \xi_2^{\prime,2}\in(x_{i-2},x_{i+2})$, $\xi_i^{\prime,3}\in(x_{i-1},x_{i+1})$.
- *At near boundary points x_i with at $i=1$ or $i=N-1$, we have*

$$\sigma_x\mathfrak{r}_{b,i} = h^2\Big(\beta_i^{\prime,4} v^4(\xi_i^{\prime,4})\Big) + h^3\Big(\beta_i^{\prime,5} v^5(\xi_i^{\prime,5}) - \beta_i^{\prime,6} v^5(\xi_i^{\prime,6})\Big) + h^4\Big(\beta_i^{\prime,7} v^6(\xi_i^{\prime,7}) - \beta_i^{\prime,8} v^6(\xi_i^{\prime,8})\Big) \tag{6.37}$$

  for some $\xi_i^{\prime,4}$, $\xi_i^{\prime,5}$, $\xi_i^{\prime,6}$, $\xi_i^{\prime,7}\in(0,1)$.

We skip the proof of Lemma 6.5. Note that in (6.36) and (6.37), the points $\xi_i^{\prime,j}$ depend on $v(x)$.

Consider now a regular function $u(x,t)$ and $v(x,t)=\partial_t u(x,t)$. Consider a fixed time $\bar{t}>0$, and the truncation $\mathfrak{r}(\bar{t}) \triangleq \mathfrak{r}(u(x,\bar{t}))$. It results from Lemma 6.3 and Lemma 6.5 that $\sigma_x\mathfrak{r}(\bar{t})$ can be expressed as follows.

Corollary 6.6. *Suppose that u is a solution to the problem (2.1) such that $x\mapsto u^k(.,t), (\partial_t u)^k(.,t)$ are continuous for $0\le k\le 8$, then the truncation error $\mathfrak{r}(t)$ satisfies an identity of the form*

$$\sigma_x\mathfrak{r}(\bar{t}) = [O(h), O(h^4), \dots O(h^4), O(h)]^\intercal \tag{6.38}$$

with Taylor representation expressed as

$$(\sigma_x\mathfrak{r}(\bar{t}))_i = h^4\Big[\sum_{j=0}^{2} C_{i,j}u^8(y_{i,j},\bar{t}) + D_{i,j}v^6(z_{i,j},\bar{t})\Big] + h^6\Big[\sum_{j=0}^{2} C'_{i,j}v^8(z'_{i,j},\bar{t})\Big] = O(h^4), \quad i = 2,...,N-2,$$

$$\begin{aligned}(\sigma_x\mathfrak{r}(\bar{t}))_1 &= h\Big[\sum_{j=0}^{2} C_{1,j}u^5(y_{1,j},\bar{t})\Big] + h^2\Big[\sum_{j=0}^{2} C'_{1,j}u^6(y'_{1,j},\bar{t}) + D'_{1,j}v^4(z'_{1,j},\bar{t})\Big] \\ &\quad + h^3\Big[\sum_{j=0}^{2} D''_{1,j}v^5(z''_{1,j},\bar{t})\Big] + h^4\Big[\sum_{j=0}^{2} D'''_{1,j}v^6(z'''_{1,j},\bar{t})\Big] = O(h),\end{aligned} \tag{6.39}$$

with a similar expression for $(\sigma_x\mathfrak{r}(\bar{t}))_{N-1}$. *All the constants* $C_{i,j}, D_{i,j}, ...$ *are fixed. The points* $y_{i,j}, y'_{i,j}, ...$ *belong to the interval* $(0,1)$. *They a priori depend on the function* $x \mapsto u(x,\bar{t})$.

Coming back to (3.26), we consider the particular case where $u(x,t)$ is the solution of (2.1). The vector $R(t)$ in (3.26) satisfies $PR(\bar{t}) \hookrightarrow \sigma_x\mathfrak{r}(\bar{t})$ with components given in (6.39) with $\bar{t} = t$. From Corollary 6.6, we conclude that the vector $PR(t)$ satisfies the hypothesis (5.20) of Lemma 5.1.

Similarly, we have

$$P\frac{d}{dt}R(\bar{t}) = [O(h), O(h^4), \ldots, O(h^4), O(h)]. \tag{6.40}$$

with an analog explicit representation.

Remark 6.7. *The truncation analysis for the 2D case proceeds in a similar fashion.*

7. Numerical Examples

In the previous sections we have considered compact approximations in space of the equations (2.1) and (4.1). In this section we display numerical results for the equations above, using the fourth-order discrete operators for the approximation in the spatial direction. The temporal integration of these equations is carried out via the method of lines. We invoke an IMEX2 scheme of Crank-Nicolson type. For example, for the two-dimensional time-dependent Stokes equation $\partial_t\Delta u = \Delta^2 u + f$ we invoke the following IMEX scheme

$$\frac{(\tilde{\Delta}_h u_{i,j})^{n+1/2} - (\tilde{\Delta}_h u_{i,j})^n}{\Delta t/2} = \tfrac{1}{2}\big(\tilde{\Delta}_h^2 u_{i,j}^{n+1/2} + \tilde{\Delta}_h^2 u_{i,j}^n\big) + f_{i,j}^{n+1/4}, \tag{7.1}$$

$$\frac{(\tilde{\Delta}_h u_{i,j})^{n+1} - (\tilde{\Delta}_h u_{i,j})^n}{\Delta t} = +\tfrac{1}{2}\big(\tilde{\Delta}_h^2 u_{i,j}^{n+1} + \tilde{\Delta}_h^2 u_{i,j}^n\big) + f_{i,j}^{n+1/2}. \tag{7.2}$$

This scheme is if second order in time, therefore in order to achieve fourth-order overall accuracy in time and space, we pick $\Delta t = Ch^2$.

7.1. **Example 1.** Consider the solution $u(x,t) = e^{-t}\sin(\pi x)/\pi^2$ of the problem

$$\begin{cases} \partial_{xxt}u = \partial_x^4 u + f(x,t), \quad 0 < x < 1, \quad t > 0, \\ u(0,t) = 0, \quad u_x(0,t) = e^{-t}/\pi, \quad t > 0, \\ u(1,t) = 0, \quad u_x(1,t) = -e^{-t}/\pi, \quad t > 0, \\ u(x,0) = \sin(\pi x)/\pi^2, \quad 0 \le x \le 1. \end{cases} \tag{7.3}$$

Here u and u_x are given at the two boundary points and $f(x,t)$ is chosen such that $u(x,t) = e^{-t}\sin(\pi x)/\pi^2$ is the exact solution of the problem above. The numerical results are shown in Table 1. They calibrate the fourth-order accuracy of the scheme.

mesh	9	Rate	17	Rate	33	Rate	65
$\|e\|_h$	4.9254(-6)	4.02	3.0327(-7)	4.01	1.8884(-8)	4.00	1.1791(-9)
$\|e_x\|_h$	3.2671(-6)	4.01	2.0308(-7)	4.00	1.2671(-8)	4.00	7.9157(-10)

Table 1. Compact scheme for $\partial_{xxt}u = \partial_x^4 u + f$ with exact solution: $u(x,t) = e^{-t}\sin(\pi x)/\pi^2$ on $x \in [0,1], t > 0$. We display e and e_x, the l_2 errors for the u and for u_x. Here $\Delta t = h^2$ and $t = 0.5$.

7.2. **Example 2.** Consider the solution $u(x,t) = 16x^2(1-x)^2\sin(1/((x-0.5)^2+\varepsilon)\sin(2\pi t)$, with $\varepsilon = 0.05$ of the problem

$$\begin{cases} \partial_{xxt}u = \partial_x^4 u + f(x,t), \quad 0 < x < 1, \quad t > 0, \\ u(0,t) = 0, \quad u_x(0,t) = 0, \quad t > 0, \\ u(1,t) = 0, \quad u_x(1,t) = 0, \quad t > 0, \\ u(x,0) = 0, \quad 0 \le x \le 1. \end{cases} \tag{7.4}$$

Here u and u_x are given at the two boundary points and $f(x,t)$ is chosen such that $16x^2(1-x)^2\sin(1/((x-0.5)^2+\varepsilon)\sin(2\pi t)$ is the exact solution of the problem above. Notice that the solution is very oscillatory near the center of the interval $[0,1]$. The numerical results are shown in Table 2. They calibrate the fourth-order accuracy of the scheme.

mesh	33	Rate	65	Rate	129	Rate	257
$\|e\|_h$	0.0523(0)	7.03	3.997(-4)	4.28	2.0579(-5)	4.07	1.2255(-6)
$\|e_x\|_h$	1.4598(0)	4.41	0.0683(0)	4.03	0.0040(0)	4.04	2.4242(-4)

TABLE 2. Compact scheme for $\partial_{xxt}u = \partial_x^4 u + f$ with exact solution: $16x^2(1-x)^2\sin(1/((x-0.5)^2+\varepsilon)\sin(2\pi t)$ with $\varepsilon = 0.05$ on $x \in [0,1], t > 0$. We display e and e_x, the l_2 errors for the u and for u_x. Here $\Delta t = h^2$ and $t = 0.25$.

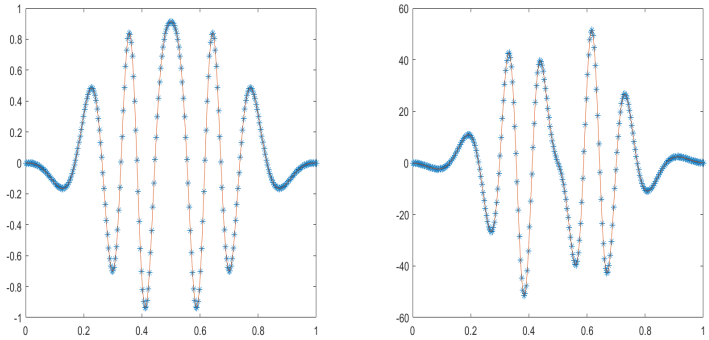

FIGURE 1. Example 2. Left: Approximation for $u(x,t)$. Right: Approximation for $\partial_x u(x,t)$

7.3. **Example 3.** Consider the solution $u(x,y,t) = (1-x^2)^2(1-y^2)^2e^{-t}$ of the problem

$$\begin{cases} \partial_t(\Delta u) = \Delta^2 u + f(x,y,t), \quad -1 < x,y < 1, \quad t > 0, \\ u(-1,y,t) = u(1,y,t) = 0, \quad u_x(-1,y,t) = \quad u_x(1,y,t), \quad t > 0, \\ u(x,-1,t) = u(x,1,t) = 0, \quad u_y(x,-1,t) = \quad u_y(x,1,t) = 0, \quad t > 0, \\ u(x,0) = (1-x^2)^2(1-y^2)^2, \quad -1 \le x,y \le 1. \end{cases} \tag{7.5}$$

Here u and u_x are given at the two boundary points and $f(x,t)$ is chosen such that $u(x,y,t) = (1-x^2)^2(1-y^2)^2e^{-t}$ is the exact solution of the problem above. The numerical results are shown in Table 3. They assess the fourth-order accuracy of the scheme.

mesh	9×9	Rate	17×17	Rate	33×33	Rate	65×65
$\|e\|_h$	1.2386(-4)	4.00	7.7408(-6)	4.00	4.8376(-7)	4.00	3.0235(-8)
$\|e_x\|_h$	2.0259(-4)	3.99	1.2750(-5)	4.00	7.9731(-7)	4.00	4.9834(-8)

TABLE 3. Compact scheme for $\partial_t(\Delta u) = \Delta^2 u + f$ with exact solution: $u(x,y,t) = (1-x^2)^2(1-y^2)^2e^{-t}$ on $(x,y) \in [-1,1]\times[-1,1], t > 0$. We display e and e_x, the l_2 errors for the u and for u_x. Here $\Delta t = h^2$ and $t = 0.25$.

7.4. **Example 4.** Consider the solution $u(x,y,t) = -0.5e^{-2t}\sin^2 x \cdot \sin^2 y$ of the problem

$$\begin{cases} \partial_t(\Delta u) = \Delta^2 u + f(x,y,t), \quad 0 < x,y < \pi \quad t > 0, \\ u(0,y,t) = u(\pi,y,t) = 0, \quad u_x(0,y,t) = \quad u_x(\pi,y,t), \quad t > 0, \\ u(x,0,t) = u(x,\pi,t) = 0, \quad u_y(x,0,t) = \quad u_y(x,\pi,t) = 0, \quad t > 0, \\ u(x,0) = -0.5\sin^2 x \cdot \sin^2 y, \quad 0 \le x,y \le \pi. \end{cases} \tag{7.6}$$

Here u and u_x are given at the two boundary points and $f(x,t)$ is chosen such that $u(x,y,t) = -0.5e^{-2t}\sin^2 x \cdot \sin^2 y$ is the exact solution of the problem above. The numerical results are shown in Table 4. They calibrate the fourth-order accuracy of the scheme.

mesh	9×9	Rate	17×17	Rate	33×33	Rate	65×65
$\|e\|_h$	1.9508(-3)	3.96	1.2527(-4)	3.99	7.9061(-6)	4.00	4.9542(-7)
$\|e_x\|_h$	2.6996(-3)	3.98	1.7134(-4)	3.99	1.0775(-5)	4.00	6.7459(-7)

TABLE 4. Compact scheme for $\partial_t(\Delta u) = \Delta^2 u + f$ with exact solution: $u(x,y,t) = -0.5e^{-2t}\sin^2 x \cdot \sin^2 y$ on $(x,y) \in [0,\pi]\times[0,\pi], t > 0$. We present e and e_x, the l_2 errors for the u and for u_x. Here $\Delta t = h^2$ and $t = 0.6168$.

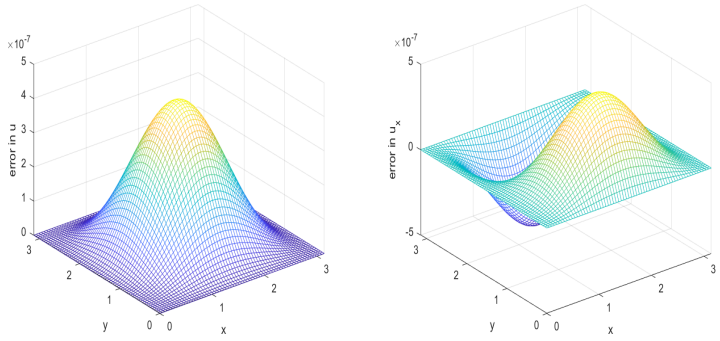

FIGURE 2. Example 4. Left: Error for $u(x,y,t)$. Right: Error for $\partial_x u(x,y,t)$

7.5. **Example 5.** Consider the solution $u(x,y,t) = (1-x^2)^3(1-y^2)^3 e^{-t}$ of the NS problem in the square $[-1,1]\times[-1,1]$.

$$\partial_t \Delta u + C(u) = \nu\Delta^2 u + f(x,y,t), \quad [-1,1]\times[-1,1], \quad t > 0, \tag{7.7}$$

where $C(u) = \nabla^\perp u.\nabla\Delta u$ is the convective term of the NS equations. The boundary and initial data deduced from the exact solution. The viscosity is $\nu = 1$. The final time is $t_f = 1$. The numerical results are shown in Table 5.

mesh	9×9	Rate	17×17	Rate	33×33	Rate	65×65
$\|e\|_h$	1.9373(-3)	4.00	1.2072(-4)	4.00	7.5424(-6)	4.00	4.7138(-7)
$\|e_x\|_h$	1.9886(-3)	4.02	1.2255(-4)	4.00	7.6527(-6)	4.00	4.7827(-7)

TABLE 5. Compact scheme for $\partial_t(\Delta u) = -C(\psi) + \nu\Delta^2 u + f$ with exact solution: $u(x,y,t) = (1-x^2)^3(1-y^2)^3 e^{-t}$ in the square $[-1,1]\times[-1,1]$. We present e and e_x, the l_2 errors for the u and for u_x. Here $\Delta t = Ch^2$ and $t_f = 1$.

8. Conclusion

In [17], a $O(h^{3.5})$ error estimate was proved for (1.3) using energy methods. Here we establish that the error is in fact $O(h^4)$, for (1.3) and (1.1), despite the fact the the truncation error drops at interior points from $O(h^4)$ to $O(h)$ at near boundary points. Note also that in [3] we have proved second order convergence for a compact second order scheme for the linear equation close to (1.1) was established. In [4] we have proved that the error for the second order compact scheme for the full nonlinear Navier-Stokes equations is $O(h^{1.5})$. Our goal in a future research is to prove that fourth-order convergence can be extended to the Navier-Stokes system.

References

[1] S. Abarbanel, A. Ditkowski and B. Gustafsson, "On error bounds of finite difference approximations to partial differential equations temporal behavior and rate of convergence", J. Sci. Comput., **15** (2000), pp. 79-116.

[2] M. Ben-Artzi, J.-P. Croisille and D. Fishelov, *"Navier-Stokes Equations in Planar Domains"*, Imperial College Press, 2013.

[3] M. Ben-Artzi, J.-P. Croisille and D. Fishelov and S. Trachtenberg, "A pure compact scheme for the streamfunction formulation of the Navier-Stokes equations", J. Comp. Phys., **205** (2005), pp. 640-664.

[4] M. Ben-Artzi, J.-P. Croisille and D. Fishelov, "Convergence of a compact scheme for the pure streamfunction formulation of the unsteady Navier-Stokes system", SIAM J. Numer. Anal., **44** (2006), pp. 1997-2024.

[5] M. Ben-Artzi, J.-P. Croisille and D. Fishelov, "A High Order Compact Scheme for the Pure-Streamfunction Formulation of the Navier-Stokes Equations", J. Sci. Comput., **42** (2010), pp. 216-250.

[6] M. Ben-Artzi, J.-P. Croisille and D. Fishelov, "Time evolution of discrete fourth-order elliptic operators", Num. Meth. Part. Diff. Eqns., **35** (2019), pp. 1429-1457).

[7] M. Ben-Artzi, J.-P. Croisille and D. Fishelov, "A fast direct solver for the biharmonic problem in a rectangular grid", SIAM J. Sci. Comput., **31** (2008) , pp. 303-333.

[8] M. Ben-Artzi, J.-P. Croisille and D. Fishelov, "A Cartesian compact scheme for the Navier-Stokes equations in streamfunction formulation in irregular domains", J. Sci. Comp., **81** (2019), pp. 1386-1408.

[9] R. Bhatia, *"Matrix analysis"* , GTM 169, Springer, 1997.

[10] W. R. Briley, "A numerical study of laminar separation bubbles using the Navier-Stokes equations", J. Flu. Mech., **47**, (1971). pp. 713-736.

[11] G.F. Carey and W.F. Spotz, "Extension of High-Order Compact Schemes to Time Dependent Problems, Numer. Meth. Part. Diff. Eqs., **17** (2001), pp. 657-672.

[12] M. H. Carpenter and D. Gottlieb and S. Abarbanel, "The stability of numerical boundary treatments for compact high-order schemes finite difference schemes", J. Comput. Phys., **108** (1993), pp. 272-295.

[13] X. Chen and Y. Li and C. Drapaca and J. Cimbala, "A unified framework of continuous and discontinuous Galerkin methods for solving the incompressible Navier-Stokes equation", J. Comp. Phys, 422, (2020), doi: 10.1016/j.jcp.2020.109799

[14] P. Drazin and N. Riley, *The Navier-Stokes Equations - A classification of Flows and Exact Solutions*, London Math. Soc. Lecture Note Series 334, Cambridge Univ. Press, 2006

[15] W. E and J-G. Liu, "Essentially compact schemes for unsteady viscous incompressible flows", J. Comput. Phys., **126(1)**, (1996), pp. 122-138.

[16] D. Fishelov and M. Ben-Artzi and J.-P. Croisille, "Recent advances in the study of a fourth-order compact scheme for the one-dimensional biharmonic equation", J. Sci. Comput., **53(1)**, (2012), pp. 55-79.

[17] D. Fishelov, "Semi-discrete time-dependent fourth-order problems on an interval: error estimate", "Numerical Mathematics and Advanced Applications - ENUMATH 2013", in Assyr Abdulle, Simone Deparis, Daniel Kressner, Fabio Nobile, and Marco Picasso, Lect. Notes in Comp. Sci. and Eng., Vol. 103, pp. 133-142, Springer, 2015.

[18] M. M. Gupta and R.P.Manohar and J.W.Stephenson, "Single cell high order scheme for the convection-diffusion equation with variable coefficients", Int. J. Numer. Meth. Fluids, **4**, (1984), pp. 641-651.

[19] B. Gustafsson, "The convergence rate for difference approximations to mixed initial boundary value problems", Math. of Comput., **29** (1975), pp. 396-406.

[20] B. Gustafsson, "The convergence rate for difference approximations to general mixed initial boundary value problems", SIAM J. Numer. Anal., **18** (1981), pp. 179-190.

[21] R. A. Horn and C. R. Johnson, Topics in Matrix analysis, vol. 1, Cambridge U. Press, (1985).

[22] R. A. Horn and C. R. Johnson, Topics in Matrix analysis, vol. 2, Cambridge U. Press, (1991).

[23] H. Johnston and R. Krasny, "Fourth-order finite difference simulation of a differentially heated cavity", Int. J. Numer. Meth. Fluids, **40**, (2002), pp 1031-1037.

[24] J. C. Kalita and M. M. Gupta, "A streamfunction-velocity approach for 2D transient incompressible viscous flows", Int. J. Numer. Meth. Fluids, **62** (2010), pp. 237-266.

[25] H. Lamb, *"Hydrodynamics"*, 6th ed., Dover, 1932.

[26] J. Lequeurre and A. Munnier, "Vorticity and Stream Function Formulations for the 2D Navier-Stokes Equations in a Bounded Domain", J. Math. Fluid Mech., **22:15**, (2020). https://doi.org/10.1007/s00021-019-0479-5.

[27] M. Li and T. Tang, "A compact fourth-order finite difference scheme for unsteady viscous incompressible flows", J. Sci. Comput., **103 (1)** (2001), pp. 29-45.

[28] J.-G. Liu, C. Wang, H. Johnston, "A fourth order scheme for incompressible Boussinesq equations", Journal of Scientific Computing, vol. 18 (2), 2003, 253-285.

[29] G. Meurant, "A review on the inverse of symmetric tridiagonal and block tridiagonal matrices", SIAM J. Matrix. Anal. App, **13(3)**, (1992), pp. 707-728.
[30] S. Müller and F. Schweiger and E. Süli, "Optimal-Order Finite Difference Approximation of Generalized Solutions to the Biharmonic Equation in a Cube", SIAM J. Numer. Anal., **58(1)**, (2020), pp. 298-329.
[31] K. Mattsson and J. Nordström, "Summation by parts operators for finite difference approximations of second derivatives", J. Comput. Phys., **199** (2004), pp. 503-540.
[32] M. J. Naughton, "On numerical boundary conditions for the Navier-Stokes equations", PhD Thesis, Cal. Inst. Tech, 1986.
[33] C. Pozrikidis, "Creeping flow in two-dimensional channels", J. Fluid. Mech., **180** (1987), pp. 495-514.
[34] A. Sifounakis and S. Lee and Y. You, "A conservative finite volume method for incompressible Navier–Stokes equations on locally refined nested Cartesian grids", J. Comput. Phys.,**326** (2016), pp. 845-861.
[35] J. W. Stephenson, "Single cell discretizations of order two and four for biharmonic problems", J. Comput. Phys., **55** (1984), pp. 65-80.
[36] K. Svärd and J. Nordström, "On the convergence rates of energy-stable finite-difference schemes", J. Comput. Phys., **397** (2019), 108819.
[37] S. Wang and G. Kreiss, "Convergence of Summation-by-Parts Finite Difference Methods for the Wave Equation", J. Sci. Comput., **71** (2017), pp. 219-245.
[38] C. Wang, J.-G. Liu, "Analysis of finite difference schemes for unsteady Navier-Stokes equations in vorticity formulation", Numerische Mathematik, vol. 91, 2002, 543-576.
[39] C. Wang, J.-G. Liu, "Fourth order convergence of compact difference solver for incompressible flow", Communications in Applied Analysis, vol. 7 (2), 2003, 171-191.
[40] C. Wang, J.-G. Liu, H. Johnston, "Analysis of a fourth order finite difference method for incompressible Boussinesq equations", Numerische Mathematik, vol. 97 (3), 2004, 555-594.

Dalia Fishelov: Afeka - Tel-Aviv Academic College of Engineering, 38 Mivza Kadesh St., Tel-Aviv 69107, Israel
Email address: daliaf@afeka.ac.il

Jean-Pierre Croisille: Department of Mathematics, IECL, UMR 7502, Univ. de Lorraine, Metz 57045, France
Email address: jean-pierre.croisille@univ-lorraine.fr