BIT Numer Math (2015) 55:487–513
DOI 10.1007/s10543-014-0509-x

# Stability analysis of Crank–Nicolson and Euler schemes for time-dependent diffusion equations

**Cassio M. Oishi · Jin Y. Yuan · Jose A. Cuminato · David E. Stewart**

Received: 11 October 2012 / Accepted: 23 June 2014 / Published online: 29 July 2014


**Abstract** In this paper, we study the stability of the Crank–Nicolson and Euler schemes for time-dependent diffusion coefficient equations on a staggered grid with explicit and implicit approximations to the Dirichlet boundary conditions. Using the matrix representation for the numerical scheme and boundary conditions it is shown that for implicit boundary conditions the Crank–Nicolson scheme is unrestrictedly stable while it becomes conditionally stable for explicit boundary conditions. Numerical examples are provided illustrating this behavior. For the Euler schemes the results are similar to those for the constant coefficient case. The implicit Euler with implicit or explicit boundary conditions is unrestrictedly stable while the explicit Euler with explicit boundary conditions presents the usual stability restriction on the time step.

Communicated by Anna-Karin Tornberg.

C. M. Oishi
Departamento de Matemática, e Computação, Universidade Estadual Paulista, Presidente Prudente, Brazil
e-mail: oishi@fct.unesp.br

J. Y. Yuan
Departamento de Matemática, Universidade Federal do Paraná, Curitiba, Brazil
e-mail: yuan@dm.ufpr.br

J. A. Cuminato
Departamento de Matemática Aplicada e Estatística, Universidade de São Paulo, São Carlos, Brazil
e-mail: jacumina@icmc.usp.br

D. E. Stewart (✉)
Department of Mathematics, University of Iowa, Iowa City, USA
e-mail: david-e-stewart@uiowa.edu

**Keywords** Stability analysis · Crank–Nicolson scheme · Staggered grids · Boundary conditions · Non-constant coefficient diffusion equations

**Mathematics Subject Classification** 65M06 · 65M12 · 65M99

## 1 Introduction

The time-dependent diffusion equation appears frequently in many applications, for example, in microwave heating [7,12], simulation of flows [3,4], economic problems [10,17], among others.

The stability of the so called $\theta$-method for the numerical discretization of ODEs and PDEs has been studied by many authors (see for instance [1,9]), but to our knowledge, none of those works take into account the effect of the staggered grid [11] on the stability of the method. In this work we will show that the Crank–Nicolson method that is known to be unrestrictedly stable may become unstable if inappropriate boundary conditions are used. This paper is concerned with the study of stability of the Crank–Nicolson scheme for diffusion problems, using the matrix method applied to a model (or *toy*) problem. Sufficient conditions for numerical stability are derived. The use of explicit methods for the simulation of high viscosity fluids imposes a stringent stability condition on the time step. To overcome this restriction, the use of an implicit discretization is recommended. Therefore in our study we chose to concentrate our attention on the famous Crank–Nicolson and Euler schemes, that are well known for being unrestrictedly stable for constant coefficient diffusion problems. In addition we also reproduce the stability constraint for the explicit Euler method for the diffusion time dependent coefficient problem.

In [11] a stability study for the diffusion equation with a constant diffusion coefficient was carried out. It should be pointed out that the introduction of a time dependent diffusion coefficient makes the stability analysis much more complex, as this requires the study of the eigenvalues of an infinite product of matrices, each of which have eigenvalues within the unit circle. The main results of this paper deal with this case, and extend the results of our previous work (see Oishi et al. [11]). As a result of the above difficulty there are not many papers dealing with the stability of numerical methods for time-dependent coefficient diffusion problems. In particular, Tadjeran [15] analyzed the stability of the Crank–Nicolson method using the time dependent one-dimensional heat equation. However, in his paper, Tadjeran [15] did not take into account the influence of the boundary conditions in his stability analysis. Stability studies of finite difference discretizations of some constant (and non-constant) coefficient Partial Differential Equations, including the influence of the boundary conditions, can be found in [13,14,16].

Thus, in this paper we attempt to bridge this gap by investigating the influence of the boundary conditions on the numerical stability of the Crank–Nicolson scheme applied to one dimension variable coefficient diffusion problems. The stability analysis to be carried out in this work supposes the use of a staggered grid discretization. It might seem unnatural to use a staggered grid for solving a one dimensional diffusion problem with Dirichlet boundary conditions. However, when solving the Navier–

Stokes equations by finite differences it is recommended to use a staggered grid to cope with oscillations. Thus, the natural simplification of the Navier–Stokes on a staggered grid is the heat equation discretized on a staggered grid.

## 2 Problem statement

*(a) Mathematical model and discretization*

In this paper, we deal with the following model problem

$$u_t = d(t)u_{xx} + q(x,t), \quad x \in [0,1] \quad \text{and} \quad t \in [0,T], \tag{2.1}$$

$$u(0,t) = u(1,t) = 0, \quad t \in [0,T], \tag{2.2}$$

$$u(x,0) = u_0, \quad x \in [0,1], \tag{2.3}$$

where $d(t) \geq 0$ is the diffusion coefficient which is a bounded time-dependent function and $q(x,t)$ is the source term.

The finite difference discretization of the heat Eq. (2.1) by the $\theta$-method can be written as

$$\begin{aligned} u_i^{n+1} - \theta \frac{d^{n+\theta}\delta t}{\delta x^2}\left(u_{i-1}^{n+1} - 2u_i^{n+1} + u_{i+1}^{n+1}\right) \\ = u_i^n - (\theta - 1)\frac{d^{n+\theta}\delta t}{\delta x^2}\left(u_{i-1}^n - 2u_i^n + u_{i+1}^n\right) + \delta t q_i^{n+\theta}, \end{aligned} \tag{2.4}$$

where $\delta x$ and $\delta t$ are the space and time steps, respectively, and $u_i^n$ denotes an approximation to $u(x_i, t_n)$. The diffusion coefficient $d(t)$ and the source term are calculated according to the value of $\theta$ on the gridpoints $t_{n+\theta} = (n+\theta)\delta t$. In this work, we consider the Crank–Nicolson method ($\theta = \frac{1}{2}$), the implicit ($\theta = 1$) and explicit ($\theta = 0$) Euler schemes.

*(b) Staggered grid and time approximation of the boundary conditions*

When problem (2.1)–(2.3) is approximated on a staggered grid, we discretize the interval $[0,1]$ by a set of equally spaced points $x_i = (i - 1/2)\delta x$, $i = 1, \ldots, m$ where $\delta x = 1/m$. The difference Eq. (2.4) is solved at the internal points $x_1, x_2, \ldots, x_m$ while $x_0$ and $x_{m+1}$ are external ghost points used to impose the boundary conditions. For the staggered grid (see Fig. 1), the points $x_0$ and $x_{m+1}$ do not coincide with the

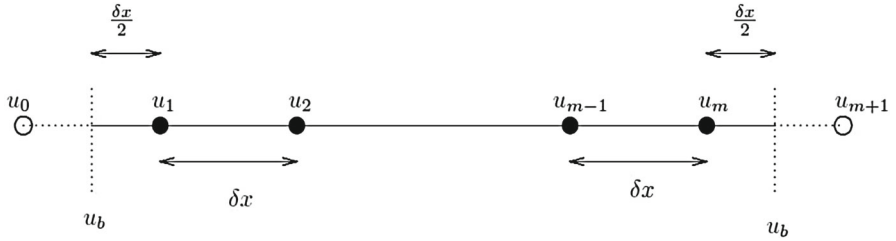


**Fig. 1** Staggered grid for solving (2.1) with $u(0,t) = u(1,t) = u_b$. In particular, we take $u_b = 0$

end points of the interval [0, 1]. Thus, we used linear interpolation to eliminate the unknown values of $u_0$ and $u_{m+1}$ from Eq. (2.4). For this, we consider the interpolating polynomial of degree one through the points $(x_0, u_0^r)$ and $(x_1, u_1^r)$ given by:

$$P_1(x) = \frac{1}{\delta x}\left((x - x_0)u_1^r - (x - x_1)u_0^r\right), \tag{2.5}$$

where $r$ is a generic time level. Now using the boundary condition $u(0, t) = 0$, we have:

$$P_1(0) = 0 = \frac{1}{\delta x}\left(\frac{\delta x}{2}(u_1^r + u_0^r)\right) = \frac{1}{2}(u_0^r + u_1^r). \tag{2.6}$$

The interpolation at the end $x = 1$ is obtained in a similar manner resulting in the equation $\frac{1}{2}(u_{m+1}^r + u_m^r) = 0$. Hence the equations for the unknowns $u_0^r$ and $u_{m+1}^r$ become

$$u_0^r = -u_1^r \quad \text{and} \quad u_{m+1}^r = -u_m^r. \tag{2.7}$$

In this work we will study the cases where $r$ take the values $n$ or $n + 1$. When implicit boundary conditions are imposed on the Crank–Nicolson scheme, we use

$$u_0^{n+1} = -u_1^{n+1} \quad , \quad u_{m+1}^{n+1} = -u_m^{n+1} \quad , \quad u_0^n = -u_1^n \quad \text{and} \quad u_{m+1}^n = -u_m^n. \tag{2.8}$$

An alternative way of approximating the boundary conditions is by an explicit formulation as

$$u_0^{n+1} = -u_1^n \quad , \quad u_{m+1}^{n+1} = -u_m^n \quad , \quad u_0^n = -u_1^n \quad \text{and} \quad u_{m+1}^n = -u_m^n. \tag{2.9}$$

This slight difference between the approximations in Eqs. (2.8) and (2.9) is fundamental for the stability of the Crank–Nicolson scheme.

## 3 Spectral Analysis for time dependent problems

In order to study the stability of the Crank–Nicolson and Euler schemes applied to the time dependent problem (2.1)–(2.3), the matrix form of the numerical method will be employed. For time dependent problems the coefficient matrix will be non-constant and hence the stability analysis becomes much harder. In this work we shall present this analysis in detail. We will first concentrate our attention to the stability study of the Crank–Nicolson method, followed by the same study for the Euler scheme.

The matrix form for the Crank–Nicolson method is:

$$\mathsf{A}(\sigma^n)\mathbf{u}^{n+1} = \mathsf{B}(\sigma^n)\mathbf{u}^n + \mathbf{c}^{n+\frac{1}{2}}, \tag{3.1}$$

where

$$\sigma^n = (d^{n+\frac{1}{2}})\,\delta t/(\delta x)^2, \tag{3.2}$$

$\mathsf{A}(\sigma^n)$ and $\mathsf{B}(\sigma^n)$ are $\sigma^n$-dependent matrices with dimensions $m \times m$, $\mathbf{u} = (u_1, u_2, \ldots, u_m)^T$ and $\mathbf{c} = (c_1, c_2, \ldots, c_m)^T$ are $m \times 1$ vectors.

We can rewrite Eq. (3.1) as

$$\mathbf{u}^{n+1} = \mathsf{M}(\sigma^n)\mathbf{u}^n + \mathsf{A}^{-1}(\sigma^n)\mathbf{c}^{n+\frac{1}{2}}, \tag{3.3}$$

where $n = 0, \ldots, index_{max}$, since the problem is solved on finite time interval $t \in [0, T]$ with $t_n = n\,\delta t$ and $T = index_{max}\delta t$.

Therefore, from Eq. (3.3), the iteration matrix is given by

$$\mathsf{M}(\sigma^n) = \mathsf{A}^{-1}(\sigma^n)\mathsf{B}(\sigma^n). \tag{3.4}$$

As a number of different definitions of stability appear in the literature, we adopted the following definition.

For stability, a small perturbation to the initial data must not be amplified along the process. That is let $\tilde{\boldsymbol{u}}^0 = \mathbf{u}^0 + \varepsilon$ then we have

$$\begin{aligned}
\tilde{\boldsymbol{u}}^1 &= \mathsf{M}(\sigma^1)(\mathbf{u}^0 + \varepsilon) + \mathsf{A}^{-1}(\sigma^1)\mathbf{c}^{\frac{1}{2}} = \mathbf{u}^1 + \mathsf{M}(\sigma^1)\varepsilon,\\
\tilde{\boldsymbol{u}}^2 &= \mathsf{M}(\sigma^2)(\mathbf{u}^1 + \mathsf{M}(\sigma^1)\varepsilon) + \mathsf{A}^{-1}(\sigma^2)\mathbf{c}^{\frac{3}{2}} = \mathbf{u}^2 + \mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1)\varepsilon,\\
&\vdots\\
\tilde{\boldsymbol{u}}^{n+1} &= \mathbf{u}^{n+1} + \mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1})\ldots\mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1)\varepsilon.
\end{aligned} \tag{3.5}$$

Thus we need that

$$\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1})\ldots\mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1) \to 0 \quad \text{as} \quad n \to \infty. \tag{3.6}$$

In order to prove (3.6) we need to study the eigenvalues of the iteration matrix $\mathsf{M}(\sigma^j)$, $j = 1, \ldots, n$. A convenient result for calculating the exact formulae for the eigenvalues of certain tridiagonal matrix was proposed by Yueh [18]. In that paper, Yueh demonstrated the following Theorem.

**Theorem 3.1** *Consider the tridiagonal matrix of the form*

$$\mathsf{T} = \begin{bmatrix}
-\alpha + b & c & 0 & 0 & \ldots\\
a & b & c & 0 & \ldots\\
0 & \ddots & \ddots & \ddots & 0\\
\ldots & 0 & a & b & c\\
\ldots & 0 & 0 & a & -\beta + b
\end{bmatrix}_{m\times m}$$

*The eigenvalues* $\lambda_i^{\mathsf{T}}$ *of* $\mathsf{T}$ *are given by*

$$\lambda_i^{\mathsf{T}} = \begin{cases} b + 2\sqrt{ac}\cos(\frac{i\pi}{m+1}), & i = 1, \ldots, m, \quad \text{if } \alpha = \beta = 0, \\ b + 2\sqrt{ac}\cos(\frac{i\pi}{m}), & i = 1, \ldots, m, \quad \text{if } \alpha = \beta = \sqrt{ac} \neq 0, \\ b + 2\sqrt{ac}\cos(\frac{(i-1)\pi}{m}), & i = 1, \ldots, m, \quad \text{if } \alpha = \beta = -\sqrt{ac} \neq 0. \end{cases} \tag{3.7}$$

## 4 Stability study of numerical methods

In order to prove our results on the stability of the Crank–Nicolson scheme on a staggered grid for the time-dependent coefficient diffusion equation we consider the implicit and explicit approximations for the boundary conditions.

In the remainder of the section, we prove stability results for various combinations of methods for solving the differential equations (Crank–Nicolson, explicit and implicit Euler methods), and explicit and implicit approximations for the boundary conditions. A summary of the results of this paper is shown in Table 1.

While some of the results (conditional convergence for explicit Euler and unconditional convergence for implicit Euler with implicit boundary conditions) are intuitively reasonable, we give comprehensive proofs of this. However, the use of explicit boundary conditions can reduce the stability of methods, so that the Crank–Nicolson method with explicit boundary conditions is only conditionally convergent while the implicit Euler method with explicit boundary conditions is unconditionally stable. Furthermore, the results for the explicit boundary conditions hold for time-dependent diffusion functions provided $\sigma^n = d(t_n)(\delta t)/(\delta x)^2$ has variation bounded by $\gamma$:

$$\sum_{k=0}^{n-1} \left| \sigma^{k+1} - \sigma^k \right| \leq \gamma. \tag{4.1}$$

The condition (4.1) shows that there is considerable latitude in changing diffusion (or time-steps $\delta t$). The only exception to this may be for implicit Euler method with explicit boundary conditions, where there is no natural upper bound for the $\sigma^n$; large values of $\delta t/(\delta x)^2$ could then amplify the effect of changes in $d(t)$. However, rapid

**Table 1** Summary of results

| | ODE method | | |
|---|---|---|---|
| | Crank–Nicolson | Implicit Euler | Explicit Euler |
| Boundary conditions | | | |
| Explicit | Conditional | Unconditional | Conditional |
| | $0 < \sigma < 2$ | | $0 < \sigma < \frac{1}{2}$ |
| Implicit | Unconditional | Unconditional | Not covered |

oscillation in $d(t)$ or the size of the time-steps may cause numerical instability, although the authors do not have definitive proof of this.

### 4.1 Crank–Nicolson method with implicit boundary conditions

In this case, we take the boundary conditions as in (2.8), thus the matrix form of the scheme is

$$\mathsf{A}(\sigma^n) = \mathsf{I} + \sigma^n \tilde{\mathsf{A}}, \tag{4.2}$$

and

$$\mathsf{B}(\sigma^n) = \mathsf{I} - \sigma^n \tilde{\mathsf{A}}, \tag{4.3}$$

with $\sigma$ defined in (3.2) and $\tilde{\mathsf{A}}$ given by

$$\tilde{\mathsf{A}} = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} & 0 & 0 & \dots \\ -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \dots & 0 & -\frac{1}{2} & 1 & -\frac{1}{2} \\ \dots & 0 & 0 & -\frac{1}{2} & \frac{3}{2} \end{bmatrix}_{m\times m}. \tag{4.4}$$

In this case, we have $\mathsf{M}(\sigma^n) = (\mathsf{I} + \sigma^n \tilde{\mathsf{A}})^{-1}(\mathsf{I} - \sigma^n \tilde{\mathsf{A}})$.

**Lemma 4.1** *The function $f(x, y) = \frac{1-xy}{1+xy}$ is a decreasing function of $x$ and $y$ and $-1 < f(x, y) < 1$, for all $x, y > 0$.*

*Proof* From the fact that $x, y > 0$, we have for every $x, y$ that:

$$f_x(x, y) < 0 \quad \text{and} \quad f_y(x, y) < 0, \tag{4.5}$$

Therefore $f(x, y) = \frac{1-xy}{1+xy}$ is a decreasing function. □

**Theorem 4.1** *Suppose that $\sigma^j$ is a bounded sequence with $0 \le \sigma^j$ for all $j$, which is bounded away from zero infinitely often (that is, there is an $\varepsilon > 0$ such that $\varepsilon \le \sigma^j$ for infinitely many $j$). Then*

$$\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1}) \cdots \mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1) \to 0$$

*as $n \to \infty$.*

*Proof* To prove this, we write

$$\mathsf{M} = \mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1}) \cdots \mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1). \tag{4.6}$$

Note that for each $j$ we have

$$\mathsf{M}(\sigma^j) = (\mathsf{I} + \sigma^j \tilde{\mathsf{A}})^{-1}(\mathsf{I} - \sigma^j \tilde{\mathsf{A}}) = \mathsf{I} - 2\sigma^j (\mathsf{I} + \sigma^j \tilde{\mathsf{A}})^{-1}\tilde{\mathsf{A}}. \tag{4.7}$$

Since $\tilde{\mathsf{A}}$ is symmetric and positive definite, there is an orthogonal matrix $\mathsf{Q}$ such that $\mathsf{Q}^T\tilde{\mathsf{A}}\mathsf{Q} = \mathsf{D}$ where $\mathsf{D}$ is a diagonal matrix with the eigenvalues of $\tilde{\mathsf{A}}$. Hence Eq. (4.7) can be rewritten as

$$\mathsf{M}(\sigma^j) = \mathsf{Q}\left(\mathsf{I} - 2\sigma^j(\mathsf{I} + \sigma^j\mathsf{D})^{-1}\mathsf{D}\right)\mathsf{Q}^T, \tag{4.8}$$

which implies that $\mathsf{M}(\sigma^j)$ is diagonalizable.

Note that $\mathsf{M}(\sigma^j)\mathsf{M}(\sigma^k) = \mathsf{M}(\sigma^k)\mathsf{M}(\sigma^j)$ for each pair $j, k$ because

$$\begin{aligned}\mathsf{M}(\sigma^j)\mathsf{M}(\sigma^k) &= \mathsf{Q}\left[(\mathsf{I} - 2\sigma^j(\mathsf{I} + \sigma^j\mathsf{D})^{-1}\mathsf{D})(\mathsf{I} - 2\sigma^k(\mathsf{I} + \sigma^k\mathsf{D})^{-1}\mathsf{D})\right]\mathsf{Q}^T \\ &= \mathsf{Q}\left[(\mathsf{I} - 2\sigma^k(\mathsf{I} + \sigma^k\mathsf{D})^{-1}\mathsf{D})(\mathsf{I} - 2\sigma^j(\mathsf{I} + \sigma^j\mathsf{D})^{-1}\mathsf{D})\right]\mathsf{Q}^T \\ &= \mathsf{M}(\sigma^k)\mathsf{M}(\sigma^j)\end{aligned} \tag{4.9}$$

By using Eqs. (4.8) and (4.9), we can rewrite Eq. (4.6) as

$$\begin{aligned}\mathsf{M} &= \prod_{j=1}^{n}\mathsf{M}(\sigma^j) = \prod_{j=1}^{n}\mathsf{Q}\left(\mathsf{I} - 2\sigma^j(\mathsf{I} + \sigma^j\mathsf{D})^{-1}\mathsf{D}\right)\mathsf{Q}^T \\ &= \mathsf{Q}\left[\prod_{j=1}^{n}\left(\mathsf{I} - 2\sigma^j(\mathsf{I} + \sigma^j\mathsf{D})^{-1}\mathsf{D}\right)\right]\mathsf{Q}^T = \mathsf{Q}\left[\prod_{j=1}^{n}(\mathsf{I} + \sigma^j\mathsf{D})^{-1}(\mathsf{I} - \sigma^j\mathsf{D})\right]\mathsf{Q}^T.\end{aligned} \tag{4.10}$$

Note that the matrix in Eq. (4.10) is similar to a diagonal matrix; thus the eigenvalues of $\mathsf{M}$ are given by

$$\lambda_i^{\mathsf{M}} = \prod_{j=1}^{n}\frac{1 - \sigma^j\lambda_i^{\tilde{\mathsf{A}}}}{1 + \sigma^j\lambda_i^{\tilde{\mathsf{A}}}}, \quad i = 1, \ldots, m. \tag{4.11}$$

Moreover, we can calculate the eigenvalues of matrix $\tilde{\mathsf{A}}$ by Theorem 3.1 to give

$$\lambda_i^{\tilde{\mathsf{A}}} = 1 + \cos\left(\frac{(i-1)\pi}{m}\right), \quad i = 1, \ldots, m, \tag{4.12}$$

implying that $0 < \lambda_i^{\tilde{\mathsf{A}}} \leq 2$.

Let $\sigma_{max} = \sup_{j\geq 1}\sigma^j$. Choose $0 < \varepsilon \leq 1$ so that $\sigma^j \geq \varepsilon$ for infinitely many $j$. Whenever $\sigma^j \geq \varepsilon$,

$$-1 < f(\sigma_{max}, 2) \leq f(\sigma^j, \lambda_i^{\tilde{\mathsf{A}}}) \leq f(\varepsilon, \lambda_1^{\tilde{\mathsf{A}}}) < 1, \tag{4.13}$$

where $f$ is the function defined in Lemma 4.1. Thus whenever $\sigma^j \geq \varepsilon$,

$$\left|\frac{1-\sigma^j\lambda_i^{\tilde{\mathsf{A}}}}{1+\sigma^j\lambda_i^{\tilde{\mathsf{A}}}}\right| \leq \mu \quad \text{where} \quad \mu = \max\left(\left|\frac{1-\varepsilon\lambda_1^{\tilde{\mathsf{A}}}}{1+\varepsilon\lambda_1^{\tilde{\mathsf{A}}}}\right|, \left|\frac{1-2\sigma_{max}}{1+2\sigma_{max}}\right|\right).$$

If $0 \leq \sigma^j < \varepsilon$ we still have

$$\left|\frac{1-\sigma^j\lambda_i^{\tilde{\mathsf{A}}}}{1+\sigma^j\lambda_i^{\tilde{\mathsf{A}}}}\right| \leq 1.$$

Let $k(n) = |\left\{ j \mid \varepsilon \leq \sigma^j,\ 1 \leq j \leq n\right\}|$, the number of $j \leq n$ where $\varepsilon \leq \sigma^j$. Note that $k(n) \to \infty$ as $n \to \infty$ since there are infinitely many $j$ where $\sigma^j \geq \varepsilon$. Therefore

$$0 < |\lambda_i^{\mathsf{M}}| \leq \mu^{k(n)}, \tag{4.14}$$

which results in

$$\|\mathsf{M}\|_2 = \rho(\mathsf{M}) \leq \mu^{k(n)}. \tag{4.15}$$

The first equality holds since $\mathsf{M}$ is symmetric, as is evident from (4.10).

Then, as $n \to \infty$, we have $\mu^{k(n)} \to 0$ (since $0 \leq \mu < 1$), which implies $\|\mathsf{M}\|_2 \to 0$ as $n \to \infty$. Therefore, $\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1}) \cdots \mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1) \to 0$ as $n \to \infty$. □

We can now prove our main result for this Section:

**Theorem 4.2** *The Crank–Nicolson method with implicit boundary conditions applied for solving problem (2.1)–(2.3) on a staggered grid is unconditionally stable.*

*Proof* From Theorem 4.1 we verify that the Crank–Nicolson method with implicit boundary conditions satisfy condition (3.6), and this proves the theorem. □

## 4.2 Crank–Nicolson method with explicit boundary conditions

In this case we take the boundary conditions as in (2.9) so that the matrices in (3.1) are

$$\mathsf{A}(\sigma^n) = \mathsf{I} + \sigma^n\widehat{\mathsf{A}}, \tag{4.16}$$

and

$$\mathsf{B}(\sigma^n) = \mathsf{I} + \sigma^n\widehat{\mathsf{B}}, \tag{4.17}$$

where

$$\widehat{\mathsf{A}} = \begin{bmatrix} 1 & -\frac{1}{2} & 0 & 0 & \dots \\ -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & \dots \\ & & & & \\ 0 & \ddots & \ddots & \ddots & 0 \\ \dots & 0 & -\frac{1}{2} & 1 & -\frac{1}{2} \\ \dots & 0 & 0 & -\frac{1}{2} & 1 \end{bmatrix}_{m\times m} \tag{4.18}$$

and

$$\widehat{\mathsf{B}} = \begin{bmatrix} -2 & \frac{1}{2} & 0 & 0 & \dots \\ \frac{1}{2} & -1 & \frac{1}{2} & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \dots & 0 & \frac{1}{2} & -1 & \frac{1}{2} \\ \dots & 0 & 0 & \frac{1}{2} & -2 \end{bmatrix}_{m\times m} \tag{4.19}$$

We shall need the following results:

**Lemma 4.2** *The matrix* $\mathsf{A}(\sigma^n)$ *defined by (4.16) is symmetric and positive definite if* $\sigma^n > -\frac{1}{2}$.

*Proof* Note that the matrix $\widehat{\mathsf{A}}$ is symmetric and satisfies the assumptions of Theorem 3.1; hence its eigenvalues are given by

$$\lambda_i^{\widehat{\mathsf{A}}} = 1 + \cos\left(\frac{i\pi}{m+1}\right), \quad i = 1, \dots, m, \tag{4.20}$$

and the eigenvalues of $\mathsf{A}(\sigma^n)$ are thus given by:

$$\lambda_i^{\mathsf{A}(\sigma^n)} = 1 + \sigma^n \lambda_i^{\widehat{\mathsf{A}}}; \quad i = 1, \dots, m. \tag{4.21}$$

Therefore, the matrix $\mathsf{A}(\sigma^n)$ in (4.16) is symmetric and positive definite if $\sigma^n > -1/2$ for all $t > 0$, which is the case when $d(t) \geq 0$. □

**Lemma 4.3** *If* $\sigma^n = 2$ *then* $\lambda = -1$ *is the minimum eigenvalue of the matrix*

$$\mathsf{M}(\sigma^n) = \mathsf{A}^{-1}(\sigma^n)\mathsf{B}(\sigma^n). \tag{4.22}$$

*Proof* Note that (4.16) and (4.17) give $\mathsf{A} = \mathsf{I} + 2\widehat{\mathsf{A}}$ and $\mathsf{B} = \mathsf{I} + 2\widehat{\mathsf{B}}$ for $\sigma^n = 2$.

Hence if $\lambda$ is an eigenvalue of the matrix $\mathsf{M}$ in (4.22) with eigenvector $\boldsymbol{v} \neq 0$, then

$$(\mathsf{I} + 2\widehat{\mathsf{B}})\boldsymbol{v} = \lambda(\mathsf{I} + 2\widehat{\mathsf{A}})\boldsymbol{v}. \tag{4.23}$$

From the form of the matrices $\widehat{\mathsf{A}}$ in (4.18) and $\widehat{\mathsf{B}}$ in (4.19) we have that $(\mathsf{I}+2\widehat{\mathsf{B}})\mathbf{e}_1 = -(\mathsf{I}+2\widehat{\mathsf{A}})\mathbf{e}_1$ where $\mathbf{e}_1 = (1,0,\ldots,0)^T$. Therefore $\lambda = -1$ is an eigenvalue of $\mathsf{M}$. Similarly, we can prove that $\mathsf{M}\mathbf{e}_m = -\mathbf{e}_m$.

Now, suppose that there exists an eigenvalue $\lambda < -1$, with

$$\mathsf{A}^{-1}\mathsf{B}\boldsymbol{v} = \lambda\boldsymbol{v} \quad \text{and} \quad \|\boldsymbol{v}\|_2 = 1. \tag{4.24}$$

The matrices $\mathsf{A}$ and $\mathsf{B}$ for $\sigma^n = 2$ are given, respectively, as follows

$$\mathsf{A} = \begin{bmatrix} 3 & -1 & 0 & 0 & \ldots \\ -1 & 3 & -1 & 0 & \ldots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \ldots & 0 & -1 & 3 & -1 \\ \ldots & 0 & 0 & -1 & 3 \end{bmatrix}_{m\times m} \quad \text{and} \quad \mathsf{B} = \begin{bmatrix} -3 & 1 & 0 & 0 & \ldots \\ 1 & -1 & 1 & 0 & \ldots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \ldots & 0 & 1 & -1 & 1 \\ \ldots & 0 & 0 & 1 & -3 \end{bmatrix}_{m\times m}. \tag{4.25}$$

From (4.24) we can write

$$\boldsymbol{v}^T\mathsf{B}\boldsymbol{v} = \lambda\boldsymbol{v}^T\mathsf{A}\boldsymbol{v}. \tag{4.26}$$

Note from (4.25) that

$$\mathsf{B} = -\mathsf{I}+\mathsf{E} \quad \text{and} \quad \mathsf{A} = \mathsf{I}+\mathsf{F}, \tag{4.27}$$

where

$$\mathsf{E} = \begin{bmatrix} -2 & 1 & 0 & 0 & \ldots \\ 1 & 0 & 1 & 0 & \ldots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \ldots & 0 & 1 & 0 & 1 \\ \ldots & 0 & 0 & 1 & -2 \end{bmatrix}_{m\times m} \quad \text{and} \quad \mathsf{F} = \begin{bmatrix} 2 & -1 & 0 & 0 & \ldots \\ -1 & 2 & -1 & 0 & \ldots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \ldots & 0 & -1 & 2 & -1 \\ \ldots & 0 & 0 & -1 & 2 \end{bmatrix}_{m\times m}. \tag{4.28}$$

It follows from (4.26) and (4.27) that

$$-1 + \boldsymbol{v}^T\mathsf{E}\boldsymbol{v} = \lambda + \lambda\boldsymbol{v}^T\mathsf{F}\boldsymbol{v}. \tag{4.29}$$

The matrix $\mathsf{F}$ satisfies the assumptions of Theorem 3.1, with $\alpha = \beta = 0, a = c = 1$ and $b = 2$. So its eigenvalues can be computed from (3.7) giving $\lambda_i^{\mathsf{F}} = 2+2\cos(\frac{i\pi}{m+1})$ for $i = 1,\ldots,m$, i.e., $\lambda_i^{\mathsf{F}} > 0$. Thus, as $\mathsf{F}$ is symmetric positive definite and $\lambda < -1$, we can write

$$-1 + \boldsymbol{v}^T \mathsf{E}\boldsymbol{v} < -1 - \boldsymbol{v}^T \mathsf{F}\boldsymbol{v} \Rightarrow \boldsymbol{v}^T(\mathsf{E}+\mathsf{F})\boldsymbol{v} < 0. \tag{4.30}$$

Note that

$$\mathsf{E}+\mathsf{F} = \begin{bmatrix} 0 & 0 & 0 & 0 & \dots \\ 0 & 2 & 0 & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \dots & 0 & 0 & 2 & 0 \\ \dots & 0 & 0 & 0 & 0 \end{bmatrix}_{m\times m}. \tag{4.31}$$

Hence $\boldsymbol{v}^T(\mathsf{E}+\mathsf{F})\boldsymbol{v} = 2\sum_{i=2}^{m-1} v_i^2$ and cannot be negative, so there cannot exist an eigenvalue $\lambda < -1$. Therefore, for $\sigma^n = 2$, $\lambda_{min}^{\mathsf{M}} = -1$. □

In addition, in order to demonstrate our result on the eigenvalues of the iteration matrix for the Crank–Nicolson scheme with explicit boundary conditions we also need the following results.

**Theorem 4.3** (Rayleigh–Ritz). *Let* $\mathsf{A} \in \mathbb{R}^{m\times m}$ *be symmetric, and let the eigenvalues of* $\mathsf{A}$ *be ordered as*

$$\lambda_{min}^{\mathsf{A}} = \lambda_1^{\mathsf{A}} \le \lambda_2^{\mathsf{A}} \le \dots \lambda_{m-1}^{\mathsf{A}} \le \lambda_m^{\mathsf{A}} = \lambda_{max}^{\mathsf{A}}.$$

*Then*

$$\lambda_1^{\mathsf{A}} \boldsymbol{v}^T\boldsymbol{v} \le \boldsymbol{v}^T\mathsf{A}\boldsymbol{v} \le \lambda_m^{\mathsf{A}}\boldsymbol{v}^T\boldsymbol{v} \quad \textit{for all} \quad \boldsymbol{v} \in \mathbb{R}^m,$$

$$\lambda_{min}^{\mathsf{A}} = \lambda_1^{\mathsf{A}} = \min_{\boldsymbol{v}\neq 0} \frac{\boldsymbol{v}^T\mathsf{A}\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}}, \tag{4.32}$$

$$\lambda_{max}^{\mathsf{A}} = \lambda_m^{\mathsf{A}} = \max_{\boldsymbol{v}\neq 0} \frac{\boldsymbol{v}^T\mathsf{A}\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}}. \tag{4.33}$$

The proof of this theorem can be found in Horn and Johnson [8].

**Theorem 4.4** (Danskin). *Suppose that* $f : X \times Y \to \mathbb{R}$ *is a continuous function, where* $X \subset \mathbb{R}^n$ *is an open set,* $Y$ *is a compact set of a topological space* $F$*, and the gradient* $\nabla_x f(x, y)$ *exists and is continuous. Then the marginal function*

$$\phi(x) = \max_{y\in Y} f(x, y)$$

*is continuous and has directional derivative in every direction* $h$*, which is given by the formula*

$$D_h\phi(x) = \max_{y\in Y(x)} \nabla_x f(x, y)^T h,$$

*where* $Y(x) = \{y \in Y \mid \phi(x) = f(x, y)\}$ *is the set of maximizers in the definition of* $\phi(x)$.

More details about this theorem can be found in [5,6].

We are now ready to state one of the main results of this Section.

**Theorem 4.5** *The eigenvalues of the matrix* $\mathsf{M}(\sigma^j)$ *in (4.22) for a generic index* $j$ *satisfy:*

1. $|\lambda_i^{\mathsf{M}(\sigma^j)}| < 1$, $i = 1, \ldots, m$, *where* $m$ *is the dimension of* $\mathsf{M}(\sigma^j)$, *if* $\sigma^j < 2$;
2. $|\lambda_i^{\mathsf{M}(\sigma^j)}| \geq 1$, *for some* $i$ *if* $\sigma^j \geq 2$, *for any* $j$.

*Proof* In Lemma 4.3 we proved that $\lambda_{min}^{\mathsf{M}} = -1$ when $\sigma^j = 2$. Now, we are going to show that the eigenvalues $\lambda_{min}^{\mathsf{M}(\sigma^j)}$ and $\lambda_{max}^{\mathsf{M}(\sigma^j)}$ are monotonically decreasing functions of $\sigma^j$. For this, we follow the idea proposed by Oishi et al. [11].

Note that the matrix $\mathsf{M}(\sigma^j)$ is similar to a symmetric matrix $\tilde{\mathsf{M}}(\sigma^j)$. From Lemma 4.2 we have that $\lambda_i^{\mathsf{A}(\sigma^j)} > 0$ for $\sigma^j > -\frac{1}{2}$ and consequently the symmetric matrix $\mathsf{A}(\sigma^j)$ is positive definite. Then there exists symmetric and positive-definite $\mathsf{A}^{\frac{1}{2}}(\sigma^j)$ such that $\mathsf{A}^{\frac{1}{2}}(\sigma^j)\mathsf{A}^{\frac{1}{2}}(\sigma^j) = \mathsf{A}(\sigma^j)$. Hence

$$\mathsf{A}^{\frac{1}{2}}(\sigma^j)\mathsf{M}(\sigma^j)\mathsf{A}^{-\frac{1}{2}}(\sigma^j) = \mathsf{A}^{-\frac{1}{2}}(\sigma^j)\mathsf{B}(\sigma^j)\mathsf{A}^{-\frac{1}{2}}(\sigma^j) = \widehat{\mathsf{M}}(\sigma^j), \tag{4.34}$$

which is symmetric, and consequently has only real eigenvalues.

From Theorem 4.3, we have

$$\lambda_{min}^{\widehat{\mathsf{M}}(\sigma^j)} = \min_{\boldsymbol{v}\neq 0} \frac{\boldsymbol{v}^T \widehat{\mathsf{M}}(\sigma^j)\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}} = \min_{\boldsymbol{w}\neq 0} \frac{\boldsymbol{w}^T(\mathsf{I} + \sigma^j\widehat{\mathsf{B}})\boldsymbol{w}}{\boldsymbol{w}^T(\mathsf{I} + \sigma^j\widehat{\mathsf{A}})\boldsymbol{w}}, \tag{4.35}$$

where use has been made of the positive definite linear transformation $\boldsymbol{w} = \mathsf{A}^{-\frac{1}{2}}(\sigma^j)\boldsymbol{v}$ from $\mathbb{R}^n$ to $\mathbb{R}^n$.

We shall now prove that $\lambda_{min}^{\mathsf{M}(\sigma^j)}$ is a monotonically decreasing function of $\sigma^j$, for all $\sigma^j \geq 0$.

Expression (4.35) can be written as

$$\lambda_{min}^{\widehat{\mathsf{M}}(\sigma^j)} = \min_{\|\boldsymbol{w}\|=1} \frac{\boldsymbol{w}^T(\mathsf{I} + \sigma^j\widehat{\mathsf{B}})\boldsymbol{w}}{\boldsymbol{w}^T(\mathsf{I} + \sigma^j\widehat{\mathsf{A}})\boldsymbol{w}}.$$

Define the compact set $C = \{x \in \mathbb{R}^n \mid \|\boldsymbol{w}\| = 1\}$ and the function

$$\sigma^j > 0, \boldsymbol{w} \in C \mapsto g((\sigma^j), \boldsymbol{w}) = \frac{\boldsymbol{w}^T(\mathsf{I} + \sigma^j\widehat{\mathsf{B}})\boldsymbol{w}}{\boldsymbol{w}^T(\mathsf{I} + \sigma^j\widehat{\mathsf{A}})\boldsymbol{w}}.$$

This function is continuous and continuously differentiable with respect to $\sigma^j$, satisfying the hypotheses of Danskin's theorem (see Theorem 4.4). We can then compute its directional derivative from the right,

$$\frac{\mathrm{d}\lambda_{min}^{\widehat{\mathsf{M}}(\sigma^j)}}{\mathrm{d}(\sigma^j)^+} = \min_{\boldsymbol{w}\in W(\sigma^j)} \frac{\mathrm{d}}{\mathrm{d}(\sigma^j)^+} g(\sigma^j, \boldsymbol{w}),$$

where $W(\sigma^j) = \{\boldsymbol{w} \in C \mid \phi(\sigma^j) = g(\sigma^j, \boldsymbol{w})\}$ is a compact set.

The computation of this derivative is straightforward:

$$\frac{\mathrm{d}}{\mathrm{d}(\sigma^j)^+} g(\sigma^j, \boldsymbol{w}) = \frac{\mathrm{d}}{\mathrm{d}(\sigma^j)} g(\sigma^j, \boldsymbol{w}) = \frac{\boldsymbol{w}^T(\widehat{\mathsf{B}} - \widehat{\mathsf{A}})\boldsymbol{w}\ \boldsymbol{w}^T\boldsymbol{w}}{[\boldsymbol{w}^T(\mathsf{I} + \sigma^j\widehat{\mathsf{A}})\boldsymbol{w}]^2} = \frac{\boldsymbol{w}^T(\widehat{\mathsf{B}} - \widehat{\mathsf{A}})\boldsymbol{w}}{[\boldsymbol{w}^T(\mathsf{I} + \sigma^j\widehat{\mathsf{A}})\boldsymbol{w}]^2}.$$

Notice that the resulting matrix $(\widehat{\mathsf{B}} - \widehat{\mathsf{A}})$ is given by

$$\widehat{\mathsf{B}} - \widehat{\mathsf{A}} = \begin{bmatrix} -3 & 1 & 0 & 0 & \dots \\ 1 & -2 & 1 & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \dots & 0 & 1 & -2 & 1 \\ \dots & 0 & 0 & 1 & -3 \end{bmatrix}_{m\times m}. \tag{4.36}$$

From Theorem 3.1, we can calculate its eigenvalues as

$$\lambda_i^{\widehat{\mathsf{B}}-\widehat{\mathsf{A}}} = -2 + 2\cos\left(\frac{i\pi}{m}\right) \quad \text{for} \quad i = 1, \dots, m, \tag{4.37}$$

and so $\boldsymbol{w}^T(\widehat{\mathsf{B}} - \widehat{\mathsf{A}})\boldsymbol{w} < 0$ for all $\boldsymbol{w} \in C$. In particular, for any $\boldsymbol{w} \in W(\sigma^j)$,

$$\frac{\mathrm{d}}{\mathrm{d}(\sigma^j)^+} g(\sigma^j, \boldsymbol{w}) = \frac{\boldsymbol{w}^T(\widehat{\mathsf{B}} - \widehat{\mathsf{A}})\boldsymbol{w}}{[\boldsymbol{w}^T(\mathsf{I} + \sigma^j\widehat{\mathsf{A}})\boldsymbol{w}]^2} < 0,$$

as $\mathsf{I} + \sigma^j\widehat{\mathsf{A}}$ is positive definite. Since $W(\sigma^j)$ is compact, we conclude that

$$\frac{\mathrm{d}\lambda_{min}^{\widehat{\mathsf{M}}(\sigma^j)}}{\mathrm{d}(\sigma^j)^+} < 0.$$

Therefore, using the fact that $\mathsf{M}(\sigma^j)$ and $\tilde{\mathsf{M}}(\sigma^j)$ are similar, we have proved that

$$\frac{\mathrm{d}\lambda_{min}^{\widehat{\mathsf{M}}(\sigma^j)}}{\mathrm{d}(\sigma^j)} = \frac{\mathrm{d}\lambda_{min}^{\mathsf{M}(\sigma^j)}}{\mathrm{d}(\sigma^j)} < 0. \tag{4.38}$$

For the maximum eigenvalue, the same argument applies, simply by replacing the minimum by the maximum in Theorem 4.3, so

$$\frac{\mathrm{d}\lambda_{max}^{\widehat{\mathsf{M}}(\sigma^j)}}{\mathrm{d}(\sigma^j)} = \frac{\mathrm{d}\lambda_{max}^{\mathsf{M}(\sigma^j)}}{\mathrm{d}(\sigma^j)} < 0, \tag{4.39}$$

that is, $\lambda_{max}^{\mathsf{M}(\sigma^j)}$ is also a monotonically decreasing function of $\sigma^j$.

Therefore, for all $\sigma^j > 0$, we have that $\lambda_{min}^{\mathsf{M}(\sigma^j)}$ and $\lambda_{max}^{\mathsf{M}(\sigma^j)}$ are monotonically decreasing functions of $\sigma^j$.

Now $\lambda_i^{\mathsf{M}(0)} = 1$, for $i = 1, \ldots, m$, which means that $\lambda_{min}^{\mathsf{M}(0)} = \lambda_{max}^{\mathsf{M}(0)} = 1$. Since $\mathrm{d}\lambda_{max}^{\mathsf{M}(\sigma^j)}/\mathrm{d}(\sigma^j) < 0$, $\lambda_{max}^{\mathsf{M}(\sigma^j)}$ is always less than 1 for all $\sigma^j > 0$.

Therefore, for $\sigma^j < 2$ we have that $|\lambda_i^{\mathsf{M}(\sigma^j)}| < 1, i = 1, \ldots, m$, while $|\lambda_i^{\mathsf{M}(\sigma^j)}| \geq 1$, for some $i$ if $\sigma^j \in [2, \infty)$. □

Having studied the eigenvalues of each iteration matrix $\mathsf{M}(\sigma^j)$ we are now going to analyse the following problem: under what conditions can we guarantee that the product

$$\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1}) \cdots \mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1) \to 0$$

as $n \to \infty$? Our main result in this section is presented below.

**Theorem 4.6** *If* $\gamma = \sum_{j=1}^{\infty} \left|\sigma^{j+1} - \sigma^j\right| < \infty$, $0 \leq \sigma^j \leq 2$ *and there is an* $\varepsilon > 0$ *where* $\varepsilon \leq \sigma^{(j)} \leq 2 - \varepsilon$ *for all* $j$, *then* $\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1}) \cdots \mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1) \to 0$ *as* $n \to \infty$.

*Proof* Note that in the case of the Crank–Nicolson scheme with explicit boundary conditions, from Eq. (4.18) the matrix $\widehat{\mathsf{B}}$ can be written in the form

$$\widehat{\mathsf{B}} = -\widehat{\mathsf{A}} - (\mathbf{e}_1\mathbf{e}_1^T + \mathbf{e}_m\mathbf{e}_m^T), \tag{4.40}$$

resulting in Eqs.(4.16)–(4.17) and, for a generic index $j$,

$$\mathsf{M}(\sigma^j) = \mathsf{A}(\sigma^j)^{-1}\mathsf{B}(\sigma^j).$$

Since $\widehat{\mathsf{A}}$ is symmetric, let $\mathsf{Q}$ be an orthogonal matrix such that $\mathsf{Q}^T\widehat{\mathsf{A}}\mathsf{Q} = \mathsf{D}$ ($\mathsf{D}$ diagonal). The eigenvalues of $\widehat{\mathsf{A}}$ [defined in (4.18)] are all in the interval $[0, 2]$ as can be checked using Theorem 3.1. Thus the diagonal entries of $\mathsf{D}$ are the eigenvalues of $\mathsf{A}$, and so are all in the interval $[0, 2]$. It is easily checked from Theorem 3.1 that the eigenvalues of $\widehat{\mathsf{A}}$ are real and lie in the interval $[0, 2]$. Then $\mathsf{Q}^T\widehat{\mathsf{B}}\mathsf{Q} = -\mathsf{Q}^T\widehat{\mathsf{A}}\mathsf{Q} - (\mathsf{Q}^T\mathbf{e}_1\mathbf{e}_1^T\mathsf{Q} + \mathsf{Q}^T\mathbf{e}_m\mathbf{e}_m^T\mathsf{Q}) = -\mathsf{D} - (\mathbf{q}_1\mathbf{q}_1^T + \mathbf{q}_m\mathbf{q}_m^T)$. Note that both $\widehat{\mathsf{A}}$ and $\widehat{\mathsf{B}}$ are symmetric, so that $\mathsf{A}(\sigma^j)$ and $\mathsf{B}(\sigma^j)$ are also symmetric, although $\mathsf{M}(\sigma^j)$ is not unless $\sigma^j = 0$.

Note that

$$\begin{aligned}\mathsf{M}(\sigma^j) &= \left(\mathsf{I} + \sigma^j\mathsf{Q}\mathsf{D}\mathsf{Q}^T\right)^{-1}\left(\mathsf{I} - \sigma^j\mathsf{Q}\mathsf{D}\mathsf{Q}^T - \sigma^j\mathsf{Q}(\mathbf{q}_1\mathbf{q}_1^T + \mathbf{q}_m\mathbf{q}_m^T)\mathsf{Q}^T\right)\\ &= \mathsf{Q}\left(\mathsf{I} + \sigma^j\mathsf{D}\right)^{-1}\left(\mathsf{I} - \sigma^j\mathsf{D} - \sigma^j\mathbf{q}_1\mathbf{q}_1^T - \sigma^j\mathbf{q}_m\mathbf{q}_m^T\right)\mathsf{Q}^T\end{aligned}$$

$$
\begin{aligned}
&= \mathsf{Q}\left(\mathsf{I}+\sigma^j\mathsf{D}\right)^{-1/2}\left(\mathsf{I}+\sigma^j\mathsf{D}\right)^{-1/2}\left(\mathsf{I}-\sigma^j\mathsf{D}-\sigma^j\mathbf{q}_1\mathbf{q}_1^T-\sigma^j\mathbf{q}_m\mathbf{q}_m^T\right)\\
&\qquad\times\left(\mathsf{I}+\sigma^j\mathsf{D}\right)^{-1/2}\left(\mathsf{I}+\sigma^j\mathsf{D}\right)^{1/2}\mathsf{Q}^T\\
&= \mathsf{Q}\left(\mathsf{I}+\sigma^j\mathsf{D}\right)^{-1/2}\mathsf{N}(\sigma^j)\left(\mathsf{I}+\sigma^j\mathsf{D}\right)^{1/2}\mathsf{Q}^T,
\end{aligned}
$$

where

$$
\mathsf{N}(\sigma^j)=\left(\mathsf{I}+\sigma^j\mathsf{D}\right)^{-1/2}\left(\mathsf{I}-\sigma^j\mathsf{D}-\sigma^j\mathbf{q}_1\mathbf{q}_1^T-\sigma^j\mathbf{q}_m\mathbf{q}_m^T\right)\left(\mathsf{I}+\sigma^j\mathsf{D}\right)^{-1/2}.
$$

Thus $\mathsf{M}(\sigma^j)$ is similar to $\mathsf{N}(\sigma^j)$, and $\mathsf{N}(\sigma^j)$ is symmetric, yielding $\rho(\mathsf{M}(\sigma^j)) = \rho(\mathsf{N}(\sigma^j)) = \left\|\mathsf{N}(\sigma^j)\right\|_2$. From Theorem 4.5 we have already seen that $\rho(\mathsf{M}(\sigma^j)) < 1$ for $0 < \sigma^j < 2$.

We can use these results to bound the product:

$$
\begin{aligned}
&\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1})\cdots\mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1)\\
&\quad= \mathsf{Q}\left(\mathsf{I}+\sigma^n\mathsf{D}\right)^{-1/2}\mathsf{N}(\sigma^n)\left(\mathsf{I}+\sigma^n\mathsf{D}\right)^{1/2}\mathsf{Q}^T\\
&\qquad\mathsf{Q}\left(\mathsf{I}+\sigma^{n-1}\mathsf{D}\right)^{-1/2}\mathsf{N}(\sigma^{n-1})\left(\mathsf{I}+\sigma^{n-1}\mathsf{D}\right)^{1/2}\mathsf{Q}^T\\
&\qquad\quad\vdots\\
&\qquad\mathsf{Q}\left(\mathsf{I}+\sigma^1\mathsf{D}\right)^{-1/2}\mathsf{N}(\sigma^1)\left(\mathsf{I}+\sigma^1\mathsf{D}\right)^{1/2}\mathsf{Q}^T\\
&\quad= \mathsf{Q}\left(\mathsf{I}+\sigma^n\mathsf{D}\right)^{-1/2}\mathsf{N}(\sigma^n)\left(\mathsf{I}+\sigma^n\mathsf{D}\right)^{1/2}\\
&\qquad\left(\mathsf{I}+\sigma^{n-1}\mathsf{D}\right)^{-1/2}\mathsf{N}(\sigma^{n-1})\left(\mathsf{I}+\sigma^{n-1}\mathsf{D}\right)^{1/2}\cdots\\
&\qquad\left(\mathsf{I}+\sigma^2\mathsf{D}\right)^{1/2}\left(\mathsf{I}+\sigma^1\mathsf{D}\right)^{-1/2}\mathsf{N}(\sigma^1)\left(\mathsf{I}+\sigma^1\mathsf{D}\right)^{1/2}\mathsf{Q}^T.
\end{aligned}
$$

Then

$$
\begin{aligned}
&\left\|\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1})\cdots\mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1)\right\|_2\\
&\quad\le\left\|\left(\mathsf{I}+\sigma^n\mathsf{D}\right)^{-1/2}\right\|_2\left\|\mathsf{N}(\sigma^n)\right\|_2\left\|\left(\mathsf{I}+\sigma^n\mathsf{D}\right)^{1/2}\left(\mathsf{I}+\sigma^{n-1}\mathsf{D}\right)^{-1/2}\right\|_2\\
&\qquad\left\|\mathsf{N}(\sigma^{n-1})\right\|_2\left\|\left(\mathsf{I}+\sigma^{n-1}\mathsf{D}\right)^{1/2}\left(\mathsf{I}+\sigma^{n-2}\mathsf{D}\right)^{-1/2}\right\|_2\cdots\\
&\qquad\left\|\mathsf{N}(\sigma^2)\right\|_2\left\|\left(\mathsf{I}+\sigma^2\mathsf{D}\right)^{1/2}\left(\mathsf{I}+\sigma^1\mathsf{D}\right)^{-1/2}\right\|_2\left\|\mathsf{N}(\sigma^1)\right\|_2\left\|\left(\mathsf{I}+\sigma^1\mathsf{D}\right)^{1/2}\right\|_2\\
&\quad=\prod_{j=1}^{n}\left\|\mathsf{N}(\sigma^j)\right\|_2\prod_{j=1}^{n-1}\left\|\left(\mathsf{I}+\sigma^{j+1}\mathsf{D}\right)^{1/2}\left(\mathsf{I}+\sigma^j\mathsf{D}\right)^{-1/2}\right\|_2\\
&\qquad\left\|\left(\mathsf{I}+\sigma^n\mathsf{D}\right)^{-1/2}\right\|_2\left\|\left(\mathsf{I}+\sigma^1\mathsf{D}\right)^{1/2}\right\|_2.
\end{aligned}
$$

The following bound is easily obtained

$$\begin{aligned}
&\left\| \left(\mathsf{I}+\sigma^{j+1}\mathsf{D}\right)^{1/2}\left(\mathsf{I}+\sigma^{j}\mathsf{D}\right)^{-1/2}\right\|_2 = \max_{k=1,\ldots,m}\left|\frac{1+\lambda_k\sigma^{j+1}}{1+\lambda_k\sigma^{j}}\right|^{1/2}\\
&\leq \max_{0\leq\lambda\leq 2}\left|\frac{1+\lambda\sigma^{j+1}}{1+\lambda\sigma^{j}}\right|^{1/2} = \max\left(1,\ \left|\frac{1+2\sigma^{j+1}}{1+2\sigma^{j}}\right|^{1/2}\right)\\
&= \exp\left(\frac{1}{2}\max(0,\ \log\left(1+2\sigma^{j+1}\right)-\log\left(1+2\sigma^{j}\right))\right)\\
&\leq \exp\left(\frac{1}{2}\left|\log\left(1+2\sigma^{j+1}\right)-\log\left(1+2\sigma^{j}\right)\right|\right).
\end{aligned}$$

Thus

$$\begin{aligned}
&\prod_{j=1}^{n-1}\left\| \left(\mathsf{I}+\sigma^{j+1}\mathsf{D}\right)^{1/2}\left(\mathsf{I}+\sigma^{j}\mathsf{D}\right)^{-1/2}\right\|_2\\
&\quad\leq \exp\left(\frac{1}{2}\sum_{j=1}^{n-1}\left|\log\left(1+2\sigma^{j+1}\right)-\log\left(1+2\sigma^{j}\right)\right|\right).
\end{aligned}$$

Now the function $f:\ [0,2]\to\left[0,\log 5\right]$ given by $f(\sigma)=\log(1+2\sigma)$ is a Lipschitz function with Lipschitz constant 2. Thus

$$\prod_{j=1}^{n-1}\left\| \left(\mathsf{I}+\sigma^{j+1}\mathsf{D}\right)^{1/2}\left(\mathsf{I}+\sigma^{j}\mathsf{D}\right)^{-1/2}\right\|_2 \leq \exp\left(\sum_{j=1}^{n-1}\left|\sigma^{j+1}-\sigma^{j}\right|\right).$$

Using the assumption that

$$\sum_{j=1}^{n-1}\left|\sigma^{j+1}-\sigma^{j}\right|\leq\gamma,$$

we have

$$\begin{aligned}
\left\|\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1})\cdots\mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1)\right\|_2 &\leq \sqrt{5}\,e^{\gamma}\prod_{j=1}^{n}\left\|\mathsf{N}(\sigma^j)\right\|_2\\
&= \sqrt{5}\,e^{\gamma}\prod_{j=1}^{n}\rho(\mathsf{M}(\sigma^j)).
\end{aligned}$$

Choose $\varepsilon>0$ so that $\varepsilon\leq\sigma^j\leq 2-\varepsilon$ for infinitely many $j$. Let $\mu=\max_{\varepsilon\leq\sigma\leq 2-\varepsilon}\rho(\mathsf{M}(\sigma))$. Note that $\mu<1$ since the maximum exists and $\rho(\mathsf{M}(\sigma))<1$

for all $0 < \sigma < 2$. Also let $k(n) = \left|\left\{ j \mid \varepsilon \le \sigma^j \le 2-\varepsilon,\ 1 \le j \le n \right\}\right|$. Note that $k(n) \to \infty$ as $n \to \infty$. Then $\prod_{j=1}^{n} \rho(\mathsf{M}(\sigma^j)) \le \mu^{k(n)} \to 0$ as $n \to \infty$. Thus $\|\mathsf{M}\|_2 \to 0$ as $n \to \infty$. □

*Remarks*. The assumption that $\sum_{j=1}^{n-1} \left|\sigma^{j+1} - \sigma^j\right| \le \gamma$ is usually quite a mild one. In terms of the diffusion problem (2.1), this will be true provided the time step $\delta t$ is fixed and $d(\cdot)$ is a function of bounded variation. It is less clear that this would be satisfied for *adaptive time step* methods where the time step is $\delta t_i$ and depends on the step number. In this case, this condition may fail to hold for certain adaptive time stepping methods. Even if the condition that $\sum_{j=1}^{n-1} \left|\sigma^{j+1} - \sigma^j\right| \le \gamma$ fails to hold, there is numerical evidence that the product $\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1}) \cdots \mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1)$ will still go to zero.

Finally, we can summarize all the results above in the following theorem.

**Theorem 4.7** *The Crank–Nicolson method with explicit boundary conditions (4.16)–(4.17) applied for solving problem (2.1)–(2.3) on a staggered grid is stable if*

$$0 < \sigma^j < 2, \quad j = 1, \ldots, n, \tag{4.41}$$

*and $\sigma^j$ are bounded away from zero and two infinitely often.*

*Proof* The proof is immediate combining Theorems 4.5 and 4.6. □

### 4.3 Stability study for the Euler schemes

As the stability analysis of the Euler schemes follows very closely that for the Crank–Nicolson method, we shall present it in this Section skipping the details already dealt with for the Crank–Nicolson study.

For the implicit Euler method with $\sigma^n = (d^{n+1})\,\delta t/(\delta x)^2$, we have

$$u_0^{n+1} = -u_1^{n+1} \quad \text{and} \quad u_{m+1}^{n+1} = -u_m^{n+1}, \tag{4.42}$$

for the implicit boundary conditions, and

$$u_0^{n+1} = -u_1^{n} \quad \text{and} \quad u_{m+1}^{n+1} = -u_m^{n}, \tag{4.43}$$

in the case of explicit boundary conditions.

For the explicit Euler method, the explicit boundary conditions in (2.7) are

$$u_0^{n} = -u_1^{n} \quad \text{and} \quad u_{m+1}^{n} = -u_m^{n}. \tag{4.44}$$

In this scheme, we set $\sigma^n = (d^n)\,\delta t/(\delta x)^2$.

*4.3.1 Implicit Euler with implicit boundary conditions*

For this case, using the implicit boundary conditions (4.42), $\mathsf{A}(\sigma^n)$ is given by

$$\mathsf{A}(\sigma^n) = \mathsf{I} - \sigma^n \check{\mathsf{A}}, \tag{4.45}$$

where

$$\check{\mathsf{A}} = \begin{bmatrix} -3 & 1 & 0 & 0 & \dots \\ 1 & -2 & 1 & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \dots & 0 & 1 & -2 & 1 \\ \dots & 0 & 0 & 1 & -3 \end{bmatrix}_{m\times m} \tag{4.46}$$

and

$$\mathsf{B}(\sigma^n) = \mathsf{I}. \tag{4.47}$$

Thus, we have that

$$\mathsf{M}(\sigma^n) = (\mathsf{I} - \sigma^n \check{\mathsf{A}})^{-1}. \tag{4.48}$$

**Theorem 4.8** *Let $\sigma^j > 0$ for all $j$ and bounded away from zero infinitely often. Then*

$$\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1})\cdots\mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1) \to 0$$

*as $n \to \infty$.*

*Proof* By using Eqs. (4.48) and (4.6), and following the ideas of the proof in Theorem 4.1, we have

$$\mathsf{M} = \prod_{j=1}^{n} \mathsf{M}(\sigma^j) = \prod_{j=1}^{n} \mathsf{Q}\left(\mathsf{I} - \sigma^j \mathsf{D}\right)^{-1} \mathsf{Q}^T = \mathsf{Q}\left[\prod_{j=1}^{n} \left(\mathsf{I} - \sigma^j \mathsf{D}\right)^{-1}\right] \mathsf{Q}^T, \tag{4.49}$$

where $\mathsf{D}$ is the diagonal matrix with the eigenvalues of $\check{\mathsf{A}}$ which is given by (4.46).

Note that $\mathsf{M}$ is symmetric and that

$$\lambda_i^{\mathsf{M}} = \prod_{j=1}^{n} \frac{1}{1 - \sigma^j \lambda_i^{\check{\mathsf{A}}}} = \prod_{j=1}^{n} \frac{1}{1 - \sigma^j (-2 + 2\cos(\frac{i\pi}{m}))}, \quad i = 1, \dots, m, \tag{4.50}$$

where in the last term we obtain the eigenvalues of $\check{\mathsf{A}}$ from Theorem 3.1.

Choose $\varepsilon > 0$ so that $\sigma^j \geq \varepsilon$ infinitely often, and $k(n) = \left|\left\{ j \mid \sigma^j \geq \varepsilon \right\}\right|$. Then $k(n) \to \infty$ when $n \to \infty$. Also put $\mu = 1/(1 + 4\varepsilon(\sin^2(\frac{\pi}{2m}))) < 1$. We can bound

$$\lambda_i^{\mathsf{M}} = \prod_{j=1}^{n} \frac{1}{1 + 4\sigma^j(\sin^2(\frac{i\pi}{2m}))} \leq \mu^{k(n)} < 1, \quad i = 1, \ldots, m. \tag{4.51}$$

Finally, we have that

$$\left\| \mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1}) \cdots \mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1) \right\|_2 \leq \mu^{k(n)} \to 0$$

as $n \to \infty$. That is, $\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1}) \cdots \mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1) \to 0$ as $n \to \infty$. □

*4.3.2 Implicit Euler with explicit boundary conditions*

For the case of the implicit Euler method with explicit boundary conditions [see Eqs. (4.43)] the matrices $\mathsf{A}(\sigma^n)$ and $\mathsf{B}(\sigma^n)$ of (3.1) become:

$$\mathsf{A}(\sigma^n) = \mathsf{I} + \sigma^n \overline{\mathsf{A}}, \tag{4.52}$$

and

$$\mathsf{B}(\sigma^n) = diag(1 - \sigma^n, 1, \ldots, 1, 1 - \sigma^n) = \mathsf{I} - \sigma^n(\mathbf{e}_1\mathbf{e}_1^T + \mathbf{e}_m\mathbf{e}_m^T), \tag{4.53}$$

where

$$\overline{\mathsf{A}} = \begin{bmatrix} 2 & -1 & 0 & 0 & \ldots \\ -1 & 2 & -1 & 0 & \ldots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \ldots & 0 & -1 & 2 & -1 \\ \ldots & 0 & 0 & -1 & 2 \end{bmatrix}_{m \times m} \tag{4.54}$$

In this case the iteration matrix is given by

$$\mathsf{M}(\sigma^n) = (\mathsf{I} + \sigma^n \overline{\mathsf{A}})^{-1} \mathsf{B}(\sigma^n). \tag{4.55}$$

Now we can enunciate the main result on this Section:

**Theorem 4.9** *Suppose that $\sigma^j > 0$ are bounded and bounded away from zero for all $j$, and assume that $\sum_{j=1}^{n-1} |\sigma^{j+1} - \sigma^j| \leq \gamma$. Then*

$$\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1}) \cdots \mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1) \to 0$$

*as $n \to \infty$.*

*Proof* The proof is similar to the proof of Theorem 4.6. For $\sigma^j > 0$, $\mathsf{A}(\sigma^j) = \mathsf{I} + \sigma^j \overline{\mathsf{A}}$ is symmetric and positive definite, so $\mathsf{A}(\sigma^j)^{-1}$ is also symmetric and positive definite. Now $\mathsf{M}(\sigma^j) = \mathsf{A}(\sigma^j)^{-1}\mathsf{B}(\sigma^j)$ is similar to $\mathsf{N}(\sigma^j) = \mathsf{A}(\sigma^j)^{-1/2}\mathsf{B}(\sigma^j)\mathsf{A}(\sigma^j)^{-1/2}$: $\mathsf{M}(\sigma^j) = \mathsf{A}(\sigma^j)^{-1/2}\mathsf{N}(\sigma^j)\mathsf{A}(\sigma^j)^{1/2}$. While $\mathsf{M}(\sigma^j)$ is generally not symmetric, $\mathsf{N}(\sigma^j)$ is. Thus $\|\mathsf{N}(\sigma^j)\|_2 = \rho(\mathsf{N}(\sigma^j)) = \rho(\mathsf{M}(\sigma^j)) < 1$. To deal with the product we use

$$\begin{aligned}&\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1})\cdots\mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1)\\ &\quad= \mathsf{A}(\sigma^n)^{-1/2}\mathsf{N}(\sigma^n)\mathsf{A}(\sigma^n)^{1/2}\mathsf{A}(\sigma^{n-1})^{-1/2}\mathsf{N}(\sigma^{n-1})\cdots\\ &\qquad\cdots\mathsf{N}(\sigma^2)\mathsf{A}(\sigma^2)^{1/2}\mathsf{A}(\sigma^1)^{-1/2}\mathsf{N}(\sigma^1)\mathsf{A}(\sigma^1)^{1/2}.\end{aligned}$$

Thus

$$\begin{aligned}&\|\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1})\cdots\mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1)\|_2\\ &\quad\le \|\mathsf{A}(\sigma^n)^{-1/2}\|_2\|\mathsf{A}(\sigma^1)^{1/2}\|_2\prod_{j=1}^{n}\|\mathsf{N}(\sigma^j)\|_2\prod_{j=1}^{n-1}\|\mathsf{A}(\sigma^{j+1})^{1/2}\mathsf{A}(\sigma^j)^{-1/2}\|_2\end{aligned}$$

The eigenvalues of $\overline{\mathsf{A}}$, $\lambda_i^{\overline{\mathsf{A}}}$ $(i = 1, \ldots, n)$, are positive real numbers as $\overline{\mathsf{A}}$ is symmetric positive definite. From Theorem 3.1, $0 < \lambda_i^{\overline{\mathsf{A}}} \le 4$. Thus the eigenvalues of $\mathsf{A}(\sigma^j)$ are

$$\lambda_i^{\mathsf{A}(\sigma^j)} = 1 + \sigma^j\lambda_i^{\overline{\mathsf{A}}}.$$

Since $\mathsf{A}(\sigma^j)$ is symmetric, so are the matrices $\mathsf{A}(\sigma^j)^{\pm 1/2}$, and the 2-norms of these matrices are the same as the maximum magnitude eigenvalues. So $\|\mathsf{A}(\sigma^j)^{1/2}\|_2 = \rho(\mathsf{A}(\sigma^j)^{1/2}) \le (1 + 4\sigma^j)^{1/2}$, and $\|\mathsf{A}(\sigma^j)^{-1/2}\|_2 = \rho(\mathsf{A}(\sigma^j)^{-1/2}) \le 1$. Also, since $\mathsf{A}(\sigma^j)$ commutes with $\mathsf{A}(\sigma^k)$ for every $\sigma^j$, $\sigma^k$, we have

$$\begin{aligned}\|\mathsf{A}(\sigma^{j+1})^{1/2}\mathsf{A}(\sigma^j)^{-1/2}\|_2 &= \max_i\left(\frac{1+\sigma^{j+1}\lambda_i^{\overline{\mathsf{A}}}}{1+\sigma^j\lambda_i^{\overline{\mathsf{A}}}}\right)^{1/2}\\ &\le \max\left(1,\ \frac{1+4\sigma^{j+1}}{1+4\sigma^j}\right)^{1/2}\\ &\le \exp\left(\left|\log(1+4\sigma^{j+1}) - \log(1+4\sigma^j)\right|\right)\\ &\le \exp\left(4\left|\sigma^{j+1}-\sigma^j\right|\right).\end{aligned}$$

Then

$$\begin{aligned}&\|\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1})\cdots\mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1)\|_2\\ &\quad\le (1+4\sigma^1)^{1/2}\exp\left(4\sum_{i=1}^{n-1}\left|\sigma^{j+1}-\sigma^j\right|\right)\prod_{j=1}^{n}\|\mathsf{N}(\sigma^j)\|_2.\end{aligned}$$

Provided $\sigma^j$ is bounded and bounded away from zero, $\|\mathsf{N}(\sigma^j)\|_2 < 1$ and bounded away from one. Under the assumption that $\sum_{j=1}^{n-1} \left|\sigma^{j+1} - \sigma^j\right| \leq \gamma$, we then have

$$\begin{aligned}&\|\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1})\cdots\mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1)\|_2\\ &\quad\leq (1+4\sigma^1)^{1/2}\, e^{4\gamma} \prod_{j=1}^{n} \|\mathsf{N}(\sigma^j)\|_2 \to 0\end{aligned}$$

as $n \to \infty$, as we wanted. □

Therefore, in the context of implicit Euler scheme, we have the following theorem:

**Theorem 4.10** *The implicit Euler method applied for solving problem (2.1)–(2.3) on a staggered grid is unconditionally stable.*

*Proof* From Theorem 4.8 we verify that the implicit Euler method with implicit boundary conditions (4.45)–(4.47) satisfy condition (3.6). On the other hand, theorem 4.9 ensures that implicit Euler with explicit boundary conditions (4.52)–(4.53) is unconditionally stable. □

*4.3.3 Explicit method with explicit boundary conditions*

In this case, by using Eq. (4.44), we obtain $\mathsf{A}(\sigma^n) = \mathsf{I}$ and

$$\mathsf{B}(\sigma^n) = \mathsf{I} + \sigma^n \tilde{\mathsf{B}}, \tag{4.56}$$

where

$$\tilde{\mathsf{B}} = \begin{bmatrix} -3 & 1 & 0 & 0 & \dots \\ 1 & -2 & 1 & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \dots & 0 & 1 & -2 & 1 \\ \dots & 0 & 0 & 1 & -3 \end{bmatrix}_{m\times m} \tag{4.57}$$

Thus, we have that $\mathsf{M}(\sigma^n) = \mathsf{B}(\sigma^n)$.

**Theorem 4.11** *Let $0 < \sigma^j < \frac{1}{2}$ for all $j$, and $\sigma^j$ is bounded away from both zero and $1/2$. Then*

$$\mathsf{M}(\sigma^n)\mathsf{M}(\sigma^{n-1})\cdots\mathsf{M}(\sigma^2)\mathsf{M}(\sigma^1) \to 0$$

*as $n \to \infty$.*

*Proof* Let $\varepsilon > 0$ be chosen so that $\varepsilon \leq \sigma^j \leq 1/2 - \varepsilon$ for all $j$.

By using the fact that $\mathsf{M}(\sigma^j) = \mathsf{I} + \sigma^j \tilde{\mathsf{B}}$, for a generic index $j$, and that $\tilde{\mathsf{B}}$ is symmetric we can define

$$\begin{aligned}\mathsf{M} &= \prod_{j=1}^{n} \mathsf{M}(\sigma^j) = \prod_{j=1}^{n} \mathsf{Q}\left(\mathsf{I} + \sigma^j \mathsf{D}\right)\mathsf{Q}^T \\ &= \mathsf{Q}\left[\prod_{j=1}^{n}\left(\mathsf{I} + \sigma^j \mathsf{D}\right)\right]\mathsf{Q}^T. \end{aligned} \tag{4.58}$$

The eigenvalues of $\mathsf{M}$ are given by

$$\lambda_i^{\mathsf{M}} = \prod_{j=1}^{n}(1 + \sigma^j \lambda_i^{\tilde{\mathsf{B}}}) = \prod_{j=1}^{n}(1 - 4\sigma^j \sin^2(\frac{i\pi}{2m})), \quad i = 1, \ldots, m, \tag{4.59}$$

where in the last term we obtain the eigenvalues of $\tilde{\mathsf{B}}$ from Theorem 3.1. Since $\mathsf{M}$ is symmetric, $\|\mathsf{M}\|_2$ is the maximum magnitude eigenvalue of $\mathsf{M}$; that is,

$$\|\mathsf{M}\|_2 = \max_{i=1,2,\ldots,m} \prod_{j=1}^{n} \left|1 - 4\sigma^j \sin^2\left(\frac{i\pi}{2m}\right)\right|. \tag{4.60}$$

To bound this product we use the well-known inequality that $2\theta/\pi \le \sin\theta \le \min(\theta, 1)$ for $0 \le \theta \le \pi/2$. Note that $i = m$ gives $i\pi/(2m) = \pi/2$ and $\sin(i\pi/(2m)) = 1$, while $1 \le i \le m-1$ gives $1/m \le i/m \le \sin(i\pi/(2m)) < 1$. Then for all $j$ and $1 \le i \le m$,

$$\begin{aligned}\left|1 - 4\sigma^j \sin^2\left(\frac{i\pi}{2m}\right)\right| &\le \max\left(\left|1 - 4\sigma^j/m\right|, \left|1 - 4\sigma^j\right|\right) \\ &\le \max\left(\left|1 - 4\sigma^j/m\right|, \left|1 - 4\sigma^j\right|\right) \\ &\le \max\left(\left|1 - 4\varepsilon/m\right|, \left|1 - 4(\frac{1}{2} - \varepsilon)\right|\right) \\ &= \left|1 - 4\varepsilon/m\right|. \end{aligned}$$

Thus

$$\|\mathsf{M}\|_2 \le |1 - 4\varepsilon/m|^n \to 0, \tag{4.61}$$

as $n \to \infty$, as we wanted. □

In order to describe a result for the explicit Euler scheme, we propose the following theorem.

**Theorem 4.12** *The explicit Euler method with explicit boundary conditions (4.56) applied for solving problem (2.1)–(2.3) on a staggered grid is stable if*

$$0 < \sigma^j < 1/2, \quad j = 1, \ldots, n. \tag{4.62}$$

*Proof* The proof is immediate from Theorem 4.11. □

## 5 Numerical results

In order to illustrate the significance of the stability results presented in this paper, we perform numerical tests for the problem (2.1)–(2.3) setting $u(x, 0) = \sin(\pi x)$, $d(t) = 4 - t$ and $q(x, t) = 0$. The exact solution in this case is a time dependent function given by the following expression

$$u(x, t) = \exp(-\pi^2(4t - t^2/2)) \sin(\pi x). \tag{5.1}$$

The numerical solution is compared to the exact solution by calculating the maximum error $E(u)$ while the rate of convergence is calculated from the formula

$$Rate = \frac{\log\left(\frac{E(u)_{M_{i+1}}}{E(u)_{M_i}}\right)}{\log\left(\frac{1}{2}\right)}, \ i = 1, 2. \tag{5.2}$$

Firstly, we verify the theoretical results presented in Theorems 4.2 and 4.10 investigating the numerical solutions obtained from Crank–Nicolson and implicit Euler methods. In this test we used three meshes in space/time: $M1(\delta x = 2.5 \times 10^{-3}, \delta t = 5 \times 10^{-4})$, $M2(\delta x = 1.25 \times 10^{-3}, \delta t = 2.5 \times 10^{-4})$, and $M3(\delta x = 6.25 \times 10^{-4}, \delta t = 1.25 \times 10^{-4})$.

Table 2 shows the values of $\sigma$, errors and the rate of convergence in the numerical solution calculated at $t = 0.01$. One can see from this Table that the Crank–Nicolson method with implicit boundary conditions and the implicit Euler method are stable for a large value of $\sigma$. These results are in agreement with the study presented in Sect. 4 which guarantees the unconditionally stability of these schemes independent of $\sigma$. Moreover, mesh refinement reveals second order convergence for the Crank–Nicolson scheme while the implicit Euler scheme converges at first order.

In order to confirm that the Crank–Nicolson method is conditionally stable in the case of explicit boundary conditions, as predicted by Theorem 4.7, we have monitored the maximum errors as a time function at different $\delta t$ for $\delta x = 2.5 \times 10^{-3}$. Results are presented in Table 3 which also describes the values of $\sigma$ at different times. The growth of the errors as a function of the time for $\delta t = 5 \times 10^{-4}$, $\delta t = 5 \times 10^{-5}$, $\delta t = 5 \times 10^{-6}$, and $\delta t = 3.15 \times 10^{-6}$ confirms the instability of this method when explicit boundary conditions are employed. In this case, for stability, we need to reduce the time-step $\delta t$. For example, when the time-step is reduced to $\delta t = 2.5 \times 10^{-6}$ (or $\delta t = 5 \times 10^{-7}$), the scheme is stable, as shown in Table 3.

**Table 2** Values of $\sigma$, errors and rate of convergence for the Crank–Nicolson method with implicit boundary conditions and implicit Euler scheme with explicit and implicit boundary conditions for different meshes

| Method | $M1$ $\sigma \approx 3.19e+02$ | $M2$ $\sigma \approx 6.39e+02$ | $M3$ $\sigma \approx 1.28e+03$ |
|---|---|---|---|
| Crank–Nicolson with implicit boundary conditions | | | |
| Error | $7.243e-06$ | $1.810e-06$ | $4.526e-07$ |
| Rate | – | 2.000 | 1.999 |
| Implicit Euler with explicit boundary conditions | | | |
| Error | $2.603e-03$ | $1.310e-03$ | $6.571e-04$ |
| Rate | – | 0.990 | 0.995 |
| Implicit Euler with implicit boundary conditions | | | |
| Error | $2.608e-03$ | $1.311e-03$ | $6.574e-04$ |
| Rate | – | 0.992 | 0.996 |

Results are computed at $t = 0.01$

**Table 3** Errors at different times for the Crank–Nicolson method with explicit boundary conditions using different values of $\delta t$

| | $t = 0.001$ | $t = 0.0025$ | $t = 0.005$ | $t = 0.01$ | $t = 0.015$ |
|---|---|---|---|---|---|
| $\delta t = 5 \times 10^{-4}$ | | | | | |
| $\sigma$ | $3.199e+02$ | $3.198e+02$ | $3.196e+02$ | $3.192e+02$ | $3.18e+02$ |
| Error | $1.26e-04$ | $3.14e-03$ | $5.21e-01$ | $1.45e+04$ | $4.03e+08$ |
| $\delta t = 5 \times 10^{-5}$ | | | | | |
| $\sigma$ | $3.199e+01$ | $3.198e+01$ | $3.196e+01$ | $3.192e+01$ | $3.18e+01$ |
| Error | $4.76e+01$ | $2.83e+12$ | $2.50e+31$ | $1.48e+68$ | $4.81e+105$ |
| $\delta t = 5 \times 10^{-6}$ | | | | | |
| $\sigma$ | 3.199 | 3.198 | 3.196 | 3.192 | 3.187 |
| Error | $1.08e+16$ | $4.25e+50$ | $8.46e+107$ | $4.72e+222$ | $\infty$ |
| $\delta t = 3.15 \times 10^{-6}$ | | | | | |
| $\sigma$ | 2.015 | 2.014 | 2.013 | 2.01 | 2.00 |
| Error | $5.40e-07$ | $4.71e-06$ | $1.88e-04$ | $1.17e-01$ | $1.94e+01$ |
| $\delta t = 2.5 \times 10^{-6}$ | | | | | |
| $\sigma$ | 1.599 | 1.599 | 1.598 | 1.596 | 1.594 |
| Error | $1.95e-07$ | $4.59e-07$ | $8.30e-07$ | $1.35e-06$ | $1.65e-06$ |
| $\delta t = 5 \times 10^{-7}$ | | | | | |
| $\sigma$ | 0.319 | 0.319 | 0.319 | 0.319 | 0.318 |
| Error | $1.95e-07$ | $4.59e-07$ | $8.32e-07$ | $1.36e-06$ | $1.67e-06$ |

In all computations we have used $\delta x = 2.5 \times 10^{-3}$

## 6 Conclusion

Although many of the results, presented in this paper, concerning the stability of the Crank–Nicolson and Euler methods are well known, their detailed proofs, fully taking

into account the boundary conditions, in the case of a time-dependent diffusion coefficient, do not seem to be available in the literature. Of special notice is the result that the Crank–Nicolson method can become restrictly stable if the boundary conditions are not correctly set. This result seems not to have been reported before the paper [11], that dealt with the constant diffusion coefficient case, appeared. In this paper we have extended the results of [11] to the more general time-dependent diffusion coefficient problem with the discretization parameters, space and time steps, constants. The proofs presented in this paper, despite leaning on the results of [11] are not straightforward applications of them as the coefficient matrices defining the numerical methods are now time-dependent and the study of their eigenvalues a much more complex task. In addition, whereas for the constant coefficient case the stability depends on the powers of a fixed matrix, in the time-dependent case it depends on boundedness of an infinite product of matrices each having eigenvalues in the unit circle.

The use of a staggered grid for discretizing the one dimension heat Eq. (2.1), does not seem reasonable at first, as it would be easier to use a conforming grid with end points coinciding with the boundaries. Nonetheless, in the numerical solution of the Navier–Stoke equations by finite differences it is advisable the use of a staggered grid to avoid instabilities (called the odd-even decoupling or spurious pressure modes) due to the relationship between the pressure and velocity.

The case of a space dependent diffusion coefficient has not been addressed in this paper as its finite difference discretization leads to completely generic tridiagonal coefficient matrices (with non-constant diagonal and sub-diagonals) for which there are no closed form formulae for the eigenvalues. The use of Yueh's theorem (3.1) was crucial for the proofs of all the results presented in the paper. We believe that distinct approach and tools need to be employed in this case.

In practice time-dependent diffusion coefficient problems do crop up in the simulation non-Newtonian flows, for instance when viscosity is time-dependent as in thixotropic fluids [2].

Finally, numerical results confirm that the Crank–Nicolson scheme with implicit boundary conditions is an unconditionally stable second-order accurate finite difference method.

**Acknowledgments** We gratefully acknowledge the financial support given by FAPESP(Fundação de Amparo a Pesquisa do Estado de São Paulo), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

## References

1. Axelson, O.: Error estimates over infinite intervals of some discretizations of evolution equations. BIT **24**, 413–424 (1984)
2. Barnes, H.A.: Thixotropy—a review. J. Non-Newton Fluid **70**, 1–33 (1997)
3. Chua, T.S., Dew, P.M.: The design of a variable-step integrator for the simulation of gas transmisson network. Int. J. Numer. Meth. Eng. **20**, 1797–1813 (1984)
4. Crank, J.: The Mathematics of Diffusion. Clarendon Press, Oxford (1975)
5. Danskin, J.M.: The theory of max–min with applications. SIAM J. Appl. Math. **14**, 641–664 (1966)
6. Guler, O.: Foundations of Optimization in Finite Dimensions. Springer-Verlag, Berlin (2010)
7. Huang, K.M., Lin, Z., Yang, X.Q.: Numerical simulation of microwave heating on chemical reaction in dilute solution. Prog. Eletromagn. Res. **49**, 273–289 (2004)

8. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1996)
9. Hunsdorfer, W., Verwer, J.G.: Numerical Solution of Time-Dependent Advection–Diffusion–Reaction Equations. Springer Series in Computational Mathematics, vol. 33. Springer, Berlin (2003)
10. McCartin, B.J., Labadie, S.M.: Accurate and efficient pricing of vanilla stock options via the Crandall–Douglas scheme. Appl. Math. Comput. **143**, 39–60 (2003)
11. Oishi, C.M., Cuminato, J.A., Yuan, J.Y., McKee, S.: Stability of numerical schemes on staggered grids. Numer. Linear. Algebr. **15**, 945–967 (2008)
12. Pozar, D.M.: Microwave Engineering. John-Wiley, New york (2005)
13. Sousa, E.: On the edge of stability analysis. Appl. Numer. Math. **59**, 1322–1336 (2009)
14. Sucec, J.: Practical stability analysis of finite difference equations by the matrix method. Int. J. Numer. Meth. Eng. **24**, 679–687 (1987)
15. Tadjeran, C.: Stability analysis of the Crank–Nicholson method for variable coefficient diffusion equation. Commun. Numer. Meth. En. **23**, 29–34 (2007)
16. Trefethen, L.N., Embree, M.: Spectra and Pseudospectra. Princeton University Press, New Jersey (2005)
17. Wilmott, P., Howison, S., Dewynne, J.: The Mathematics of Financial Derivatives: A Student Introduction. Cambridge University Press, Cambridge (1995)
18. Yueh, W.C.: Eigenvalues of several tridiagonal matrices. Appl. Math. E-Notes **5**, 66–74 (2005)