Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Positivity preservation of implicit discretizations of the advection equation

Yiannis Hadjimichael[1], David I. Ketcheson[2], Lajos Lóczi[3]

submitted: May 31, 2021

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: yiannis.hadjimichael@wias-berlin.de

[2] Computer, Electrical and Mathematical Science
and Engineering Division
King Abdullah University of Science
and Technology (KAUST)
Thuwal 23955-6900
Saudi Arabia
E-Mail: david.ketcheson@kaust.edu.sa

[3] Department of Numerical Analysis
Eötvös Loránd University, and
Department of Differential Equations
Budapest University of Technology and Economics
Hungary
E-Mail: LLoczi@inf.elte.hu

No. 2846
Berlin 2021

2020 *Mathematics Subject Classification.* 65M12, 65L07, 65M06, 35P15, 35B09.

Key words and phrases. Positivity preservation, implicit time-discretization, finite difference, spectral collocation, linear partial differential equations.

The Application-domain Specific Highly Reliable IT Solutions project has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Thematic Excellence Programme TKP2020-NKA-06 (National Challenges Subprogramme) funding scheme. This work was supported by the King Abdullah University of Science and Technology (KAUST), 4700 Thuwal, 23955-6900, Saudi Arabia, and by the Leibniz Competition.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Positivity preservation of implicit discretizations of the advection equation

Yiannis Hadjimichael, David I. Ketcheson, Lajos Lóczi

Abstract

We analyze, from the viewpoint of positivity preservation, certain discretizations of a fundamental partial differential equation, the one-dimensional advection equation with periodic boundary condition. The full discretization is obtained by coupling a finite difference spatial semi-discretization (the second- and some higher-order centered difference schemes, or the Fourier spectral collocation method) with an arbitrary θ-method in time (including the forward and backward Euler methods, and a second-order method by choosing $\theta \in [0,1]$ suitably). The full discretization generates a two-parameter family of circulant matrices $M \in \mathbb{R}^{m\times m}$, where each matrix entry is a rational function in θ and ν. Here, ν denotes the CFL number, being proportional to the ratio between the temporal and spatial discretization step sizes. The entrywise non-negativity of the matrix M—which is equivalent to the positivity preservation of the fully discrete scheme—is investigated via discrete Fourier analysis and also by solving some low-order parametric linear recursions. We find that positivity preservation of the fully discrete system is impossible if the number of spatial grid points m is even. However, it turns out that positivity preservation of the fully discrete system is recovered for *odd* values of m provided that $\theta \geq 1/2$ and ν are chosen suitably. These results are interesting since the systems of ordinary differential equations obtained via the spatial semi-discretizations studied are *not* positivity preserving.

1 Background and motivation

In this work, we investigate the positivity of some discretizations of the advection equation with periodic boundary condition

$$\begin{aligned} U_t(x,t) &= aU_x(x,t), \qquad x\in[0,1], t>0,\\ U(x,0) &= U_0(x),\\ U(0,t) &= U(1,t), \end{aligned} \tag{1}$$

where $U : \mathbb{R}\times[0,+\infty) \to \mathbb{R}$ is the unknown function, $U_0 : \mathbb{R} \to \mathbb{R}$ is a given differentiable initial function, and $a > 0$ is a constant. The exact solution of the Cauchy problem (1), given by $U(x,t) = U_0(\{x+at\})$ (where $\{\cdot\}$ denotes the fractional part), is positivity preserving; i.e.

$$\forall x \in [0,1], \forall t > 0 \qquad U_0(x) \geq 0 \implies U(x,t) \geq 0. \tag{2}$$

Positivity is often important in this context, since U may represent a concentration or density that cannot be negative.

Remark 1. *Herein the term* positivity *is always meant in the weak sense; i.e. it means* non-negativity.

DOI 10.20347/WIAS.PREPRINT.2846

Finite difference spatial semi-discretization of (1) on a uniform grid $\{\Delta x, 2\Delta x, \ldots, m\Delta x\} \subset [0,1]$ with mesh spacing $\Delta x > 0$ and $m\Delta x = 1$ yields a system of ordinary differential equations

$$u'(t) = \frac{a}{\Delta x} L u(t), \tag{3}$$

where $u : \mathbb{R} \to \mathbb{R}^m$, $L \in \mathbb{R}^{m\times m}$ is a circulant matrix [2, Section 5.16], and $m \in \mathbb{N}^+$ is the number of grid points (the points $x = 0$ and $x = 1$ are identified due to the periodic boundary condition). If one uses an upwind spatial discretization

$$L = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & -1 & 1 \\ 1 & 0 & \cdots & 0 & 0 & -1 \end{pmatrix}, \tag{4}$$

then the exact solution of (3) is also positivity preserving. Moreover, a corresponding full discretization will be positivity preserving too, under an appropriate time step size restriction $0 < \Delta t \leq \Delta t_0$ if, for example, the forward (explicit) Euler method or any strong stability preserving method [5] is used in time; see, e.g. [3].

Positivity-preserving methods for transport equations are typically based on low-order upwind-biased spatial discretizations like that above, or involve nonlinear limiters (or both). Here we instead consider the positivity of linear higher-order centered discretizations. A second-order scheme is obtained with the centered difference discretization

$$L = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \cdots & 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 & \frac{1}{2} & \cdots & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \cdots & 0 & -\frac{1}{2} & 0 \end{pmatrix}. \tag{5}$$

However, this spatial semi-discretization is not positivity preserving, since the matrix L has at least one negative off-diagonal entry [7, Chapter I, Theorem 7.2]. This implies that any consistent full discretization based on (5) must fail to preserve positivity under sufficiently small step sizes $\Delta t > 0$. Indeed, a full discretization based on the scheme (5) and forward Euler in time is not positivity preserving for any step size $\Delta t > 0$.

On the other hand, interestingly, using (5) with backward (implicit) Euler time integration, one observes positivity preservation under *large* enough time step sizes provided that the parity of the number of spatial grid points is *odd*. To investigate the differences between the behavior of the forward and backward Euler methods, we will study the θ-method [6, Chapter IV.3] as time discretization applied to (3)

$$u^{n+1} = u^n + \frac{a\Delta t}{\Delta x}((1-\theta)Lu^n + \theta L u^{n+1}). \tag{6}$$

For $\theta \in [0,1]$, the θ-family includes both Euler methods as limiting cases: the forward Euler method for $\theta = 0$, the backward Euler method for $\theta = 1$, and the only second-order θ-method for $\theta = 1/2$; any θ-method with $\theta \in (0,1]$ is implicit.

Remark 2. *As is customary in the context of space-time discretizations of partial differential equations, superscripts of u in* (6) *(and later in this work) are not exponents but denote time discretization steps.*

Remark 3. *Similarly to* (2)*, the semi-discrete system* (3) *and the fully discrete system* (6) *are said to be* positivity preserving*, if for* ***any*** *componentwise non-negative vector of initial condition*

- $u(0)$*, the solution* $u(t)$ *of* (3) *stays componentwise non-negative* $\forall t > 0$
- u^0*, the solution* u^n *of* (6) *stays componentwise non-negative* $\forall n \in \mathbb{N}^+$*,*

respectively.

The motivation for this work is not to develop new positivity preserving methods, but to study the positivity of some of fundamental discretizations such as the second-order centered difference method (5) and the θ-method (6). As we will see, the combination of these methods does not preserve positivity in general, nor in the limit of small time step size, so it is not typically recommended in practice. Nevertheless, this study may both shed light on the behavior of more complicated methods used in practice and provide tools that can be used to study the positivity of those methods. In the later sections of the paper, we combine higher-order methods in space with the θ-method in time, as a next step in this direction.

1.1 Structure of the paper and notation

In Section 2, we first characterize positivity preservation of full discretizations of (1) resulting from finite difference spatial and one-step time discretizations. Then, in Section 2.1, we study positivity of the second-order centered differences in space combined with the θ-method in time, using discrete Fourier analysis. We also point out some connections with structured non-negative inverse eigenvalue problems. In Section 3, we study this particular full discretization in more detail. In Section 3.1, by setting up and solving certain parametric linear recursions, we derive explicit, non-trigonometric formulae for the entries of the full discretization matrix M. Then, in Section 3.2, we use these formulae to provide precise results on the non-negativity of M in terms of roots of some sparse polynomials. In Section 4, we discuss some observations and results regarding higher-order spatial discretizations, including high-order centered differences (in Section 4.1) and spectral collocation methods (in Section 4.2). We summarize our findings in Section 5.

Throughout the paper, the set of positive integers is denoted by $\mathbb{N}^+$, the complex imaginary unit is $\imath$, the identity matrix is $I \in \mathbb{R}^{m\times m}$, and to emphasize the dimensions of a matrix, we will sometimes write, for example, $L_{m\times m}$. The symbol $M \geq 0$ means that $M_{i,j} \geq 0$ for every entry $1 \leq i, j \leq m$ of the matrix $M \in \mathbb{R}^{m\times m}$. The positive integer m is the number of spatial grid points within the interval $[0, 1]$, and the matrices L and M are the matrices corresponding to the spatial and the full discretizations, respectively. The three key parameters in our investigations will be m, $\theta \in [0, 1]$ and $\nu > 0$ (see (6) and (12)).

The computations in this work have been carried out by using Wolfram *Mathematica* version 11.

2 Discrete Fourier analysis

From here on, we consider the problem (1) on the domain $x \in [0,1]$ with periodic boundary condition $U(0,t) = U(1,t)$. Finite difference discretization in space with step size Δx leads to (3) with $u_j \approx u(j\Delta x)$ for $1 \le j \le m$. The circulant matrix $L \in \mathbb{R}^{m\times m}$ has the eigendecomposition

$$L = \mathcal{F}\Lambda\mathcal{F}^*, \tag{7}$$

where the (unitary) matrix of normalized eigenvectors $\mathcal{F}$ has entries

$$f_{j,\ell} := \frac{1}{\sqrt{m}} \exp(\imath(j-1)\xi_\ell) \qquad (1 \le j, \ell \le m), \tag{8}$$

and Λ is the diagonal matrix of eigenvalues λ_ℓ, which depends on the particular finite difference method chosen. Here ξ_ℓ are evenly spaced angles

$$\xi_\ell := \frac{2\pi(\ell-1)}{m} \qquad (1 \le \ell \le m), \tag{9}$$

such that $\exp(\imath\xi_\ell)$ are the m^{th} roots of unity. Applying a one-step time discretization with step size Δt and stability function $R : \mathbb{C} \to \mathbb{C}$ to (3) leads to the iteration

$$u^{n+1} = Mu^n, \tag{10}$$

where

$$M := R(\nu L) = \mathcal{F}R(\nu\Lambda)\mathcal{F}^*, \tag{11}$$

and

$$\nu := a\frac{\Delta t}{\Delta x} > 0 \tag{12}$$

is the CFL number. Then it is easily seen that

$$\boxed{\text{positivity preservation of the fully discrete numerical solution} \quad \Longleftrightarrow \quad M \ge 0.}$$

For one-step methods, R is a rational function, and products and inverses of circulant matrices are also circulant [2, Fact 5.16.7], so M is also a real, circulant matrix. Thus it is defined completely by the entries of its first row, which are given by

$$M_{1,j} = \frac{1}{m}\sum_{\ell=1}^{m} R(\nu\lambda_\ell)\exp(-\imath(j-1)\xi_\ell) \qquad (1 \le j \le m). \tag{13}$$

2.1 Second-order centered discretization in space, θ-method in time

In what follows we assume $m \ge 3$. Consider the case of a 3-point centered difference approximation in space (having order 2):

$$U_x\Big|_{x=x_j} \approx \frac{u_{j+1} - u_{j-1}}{2\Delta x},$$

so that $L \in \mathbb{R}^{m\times m}$ is a circulant matrix with entries $(-1/2, 0, 1/2)$ on the central three diagonals:

$$L := \begin{pmatrix} 0 & \frac{1}{2} & 0 & \cdots & 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 & \frac{1}{2} & \cdots & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \cdots & 0 & -\frac{1}{2} & 0 \end{pmatrix}, \tag{14}$$

that is,

$$L_{i,i-1} := -\frac{1}{2}, \qquad L_{i,i+1} := \frac{1}{2} \tag{15a}$$

$$L_{1,m} := -\frac{1}{2}, \qquad L_{m,1} := \frac{1}{2}. \tag{15b}$$

It is known that the eigenvalues of L are

$$\lambda_\ell = \imath \sin(\xi_\ell) \quad (1 \le \ell \le m). \tag{16}$$

Now we consider the θ-method [6, Chapter IV.3] in time, whose stability function is

$$R(z) := \frac{1+(1-\theta)z}{1-\theta z}, \tag{17}$$

so with the second-order centered difference in space we get from (13) for $1 \le j \le m$ that the entries of the full discretization matrix are

$$\begin{aligned} M_{1,j} &= \frac{1}{m}\sum_{\ell=1}^{m} \frac{1+(1-\theta)\nu\imath\sin(\xi_\ell)}{1-\theta\nu\imath\sin(\xi_\ell)} \exp\left(-\imath(j-1)\xi_\ell\right) \\ &= \frac{1}{m}\sum_{\ell=1}^{m} \frac{\left(1-\theta(1-\theta)\nu^2\sin^2(\xi_\ell)\right)\cos((j-1)\xi_\ell)+\nu\sin(\xi_\ell)\sin((j-1)\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)} \\ &\quad - \frac{\imath}{m}\sum_{\ell=1}^{m} \frac{\left(1-\theta(1-\theta)\nu^2\sin^2(\xi_\ell)\right)\sin((j-1)\xi_\ell)-\nu\sin(\xi_\ell)\cos((j-1)\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)}. \end{aligned} \tag{18}$$

Note that the angles $\{(j-1)\xi_\ell\}_{\ell=1}^m$ are symmetric about the x-axis if m is odd, and also if m is even and j is odd. If both m and j are even, then the angles are symmetric about the origin. Therefore, we have that

$$\sum_{\ell=1}^{m} \sin((j-1)\xi_\ell) = 0 \qquad \text{and} \qquad \sum_{\ell=1}^{m} \sin(\xi_\ell)\cos((j-1)\xi_\ell) = 0,$$

for all $1 \le j \le m$ and for any value of m. Moreover the factors $\frac{1-\theta(1-\theta)\nu^2\sin^2(\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)}$ and $\frac{\nu}{1+\theta^2\nu^2\sin^2(\xi_\ell)}$ in (18) keep this symmetry. Thus, the imaginary part of (18) vanishes, yielding $M_{1,j} \in \mathbb{R}$ for all j, as expected. So for $1 \le j \le m$ we get

$$M_{1,j} = \frac{1}{m}\sum_{\ell=1}^{m} \frac{\left(1-\theta(1-\theta)\nu^2\sin^2(\xi_\ell)\right)\cos((j-1)\xi_\ell)+\nu\sin(\xi_\ell)\sin((j-1)\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)}.$$

We will also make use of the identities

$$\cos((m-1)\xi_\ell) = \cos(\xi_\ell), \qquad\qquad \sin((m-1)\xi_\ell) = -\sin(\xi_\ell). \tag{19}$$

This leads to the following expressions for the first, second and last entries of the first row of M:

$$\begin{aligned} M_{1,1} &= \frac{1}{m}\sum_{\ell=1}^{m} \frac{1-\theta(1-\theta)\nu^2\sin^2(\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)}, \\ M_{1,2} &= \frac{1}{m}\sum_{\ell=1}^{m} \frac{\left(1-\theta(1-\theta)\nu^2\sin^2(\xi_\ell)\right)\cos(\xi_\ell)+\nu\sin^2(\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)}, \\ M_{1,m} &= \frac{1}{m}\sum_{\ell=1}^{m} \frac{\left(1-\theta(1-\theta)\nu^2\sin^2(\xi_\ell)\right)\cos(\xi_\ell)-\nu\sin^2(\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)}. \end{aligned}$$

These entries will have a special role in the forthcoming analysis. We distinguish two cases.

Case 1: m is even.

By using similar symmetry arguments as before, we conclude that for even $m = 2k \geq 4$ the entries of matrix M are given by

$$M_{1,j} = \begin{cases} \dfrac{1}{m} \displaystyle\sum_{\ell=1}^{m} \dfrac{\left(1-\theta(1-\theta)\nu^2 \sin^2(\xi_\ell)\right)\cos((j-1)\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)}, & \text{if } j \text{ is odd,} \\ \dfrac{1}{m} \displaystyle\sum_{\ell=1}^{m} \dfrac{\nu \sin(\xi_\ell)\sin((j-1)\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)}, & \text{if } j \text{ is even.} \end{cases}$$

Considering the above expression for $j = m$ we have

$$M_{1,m} = \frac{1}{m}\sum_{\ell=1}^{m} \frac{-\nu\sin^2(\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)} < 0.$$

Thus the discretization using second-order centered differences in space and the θ-method in time cannot preserve positivity when m is even, regardless of the values of $\theta \in [0,1]$ and $\nu > 0$. We can arrive at the same conclusion by observing that for any $m = 2k \geq 4$ we have $M_{1,2} = -M_{1,m}$, so that one of these (non-zero) entries must always be negative.

Case 2: m is odd.

Let us now consider the case of odd $m = 2k+1 \geq 3$. Then $\sin(\xi_\ell) \neq 0$ for $2 \leq \ell \leq m$. Writing

$$M_{1,1} = \frac{1}{m}\left(1 + \sum_{\ell=2}^{m} \frac{1-\theta(1-\theta)\nu^2\sin^2(\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)}\right)$$

$$M_{1,j} = \frac{1}{m}\left(1 + \sum_{\ell=2}^{m} \frac{(1-\theta(1-\theta)\nu^2\sin^2(\xi_\ell))\cos((j-1)\xi_\ell) + \nu\sin(\xi_\ell)\sin((j-1)\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)}\right) \quad (j \geq 2).$$

and taking $\nu \to +\infty$ with m and $\theta \in (0,1]$ fixed, we find that

$$M_{1,1}^{\infty} := \lim_{\nu\to+\infty} M_{1,1} = \frac{1}{m}\left(1 - \sum_{\ell=2}^{m}\frac{1-\theta}{\theta}\right) = 1 - \frac{m-1}{m\theta}$$

$$M_{1,j}^{\infty} := \lim_{\nu\to+\infty} M_{1,j} = \frac{1}{m}\left(1 - \sum_{\ell=2}^{m}\frac{1-\theta}{\theta}\cos((j-1)\xi_\ell)\right) = \frac{1}{m}\left(1+\frac{1-\theta}{\theta}\right) = \frac{1}{m\theta} \quad (j \geq 2).$$

We see that

$$M_{1,j}^{\infty} > 0 \text{ for all } 2 \leq j \leq m \text{ and } \theta \in (0,1],$$

while

$$M_{1,1}^{\infty} > 0 \quad \Longleftrightarrow \quad \theta > \frac{m-1}{m}.$$

Thus for fixed $m \geq 3$ and $\theta > \frac{m-1}{m}$, the matrix M is non-negative if $\nu > 0$ is large enough.

We now show that $M \geq 0$ also holds for $\theta = \frac{m-1}{m}$ with m fixed and for $\nu > 0$ large enough. Clearly, we only need to verify the non-negativity of entry $M_{1,1}$ for $\nu > 0$ large enough. In fact, for $\theta = \frac{m-1}{m}$ and for *any* $\nu > 0$ we have $M_{1,1} > 0$. To see this, consider a summand with $2 \leq \ell \leq m$ in $M_{1,1}$:

$$\left.\frac{1-\theta(1-\theta)\nu^2\sin^2(\xi_\ell)}{1+\theta^2\nu^2\sin^2(\xi_\ell)}\right|_{\theta=\frac{m-1}{m}} = \frac{m^2-(m-1)\nu^2\sin^2(\xi_\ell)}{m^2+(m-1)^2\nu^2\sin^2(\xi_\ell)} =: \varphi(\nu,\ell).$$

Its partial derivative with respect to ν is

$$\partial_\nu\varphi(\nu,\ell) = -\frac{2(m-1)m^3\nu\sin^2(\xi_\ell)}{\left(m^2+(m-1)^2\nu^2\sin^2(\xi_\ell)\right)^2} < 0,$$

and $\varphi(0,\ell) = 1$, hence the function

$$\nu \mapsto \left.M_{1,1}\right|_{\theta=\frac{m-1}{m}} = \frac{1}{m}\left(1+\sum_{\ell=2}^{m}\varphi(\nu,\ell)\right)$$

is positive at $\nu = 0$, strictly decreases, and its limit when $\nu \to +\infty$ is $\left.M_{1,1}^{\infty}\right|_{\theta=\frac{m-1}{m}} = 0$, completing the proof of the claim.

Summarizing the above, we have proved the following for any $\nu > 0$.

Theorem 1. *Consider the advection equation* (1) *with periodic boundary condition discretized using* 2^{nd}*-order centered differences in space and the* θ*-method in time with* $m \geq 3$ *spatial grid points and* $\theta \in [0,1]$*. The full discretization takes the form* (10)*, where*
(i) if m *is even, then* M *has at least one negative entry;*
(ii) if m *is odd and* $\theta \in \left[\frac{m-1}{m}, 1\right]$*, then for large enough* Δt *all entries of* M *are non-negative.*

A refinement of Theorem 1 for odd values of m will be given at the end of Section 3; see Theorem 2. As for the interval $\theta \in \left[\frac{m-1}{m}, 1\right]$ appearing in Theorem 1, see also Figures 3–4.

Remark 4. *In the formulae leading to Theorem 1 we used a trigonometric representation of the matrix* entries $M_{1,j}$*. Here we highlight a related approach to studying the non-negativity of* M *by relying only on the* eigenvalues σ_ℓ *(*$1 \leq \ell \leq m$*) of* M*. According to* (11)*,* (16) *and* (17)*, we have*

$$\sigma_\ell := R(\nu\lambda_\ell) = \frac{1+(1-\theta)\nu\imath\sin(\xi_\ell)}{1-\theta\nu\imath\sin(\xi_\ell)}.$$

The main question in the context of non-negative inverse eigenvalue problems *is to find (necessary or sufficient) conditions for a set* $\Sigma := \{\sigma_1, \ldots, \sigma_m\} \subset \mathbb{C}$ *to be the spectrum of* some *non-negative* $m \times m$ *matrix. One such condition is the following. It is known [1, Chapter 4] that if* Σ *is the spectrum of an* $m \times m$ *non-negative matrix, then*

$$\forall p, q \in \mathbb{N}^+ : \qquad 0 \leq \left(\sum_{j=1}^{m}\sigma_j^p\right)^q \leq m^{q-1}\sum_{j=1}^{m}\sigma_j^{pq}. \tag{20}$$

For example, for $m = 5$ *and* $\theta = 1$*,* (20) *with* $p \in \{1, \ldots, 9\}$ *and* $q \in \{2, 3\}$ *yields the lower bounds*

$$\nu \geq \nu_*(p,q), \tag{21}$$

where the approximate values of $\nu_*(p,q)$ are given below:

$\nu_*(p,q)$	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$	$p=7$	$p=8$	$p=9$
$q=2$	3.0074	1.462	0.9669	0.7219	0.5753	0.4778	0.4082	0.3563	0.3160
$q=3$	2.1497	1.0269	0.6694	0.4941	0.3907	0.3227	0.2749	0.2393	0.2119

As we see, the necessary condition (20)*—valid for* any *non-negative matrix—already implies that there are positive* lower *bounds on ν, although these bounds are not optimal.*

It is possible to sharpen the lower bounds in (21) *by making use of some more specific results. We know in addition that the matrix M is* circulant*, which leads us to the realm of* structured non-negative inverse eigenvalue problems*. For example, the spectra of non-negative circulant matrices have been characterized (with a necessary* and *sufficient condition) in [9, Theorem 10]. From this theorem we get (still for $m=5$ and $\theta=1$) the lower bound*

$$\nu \geq 3.9173.$$

As we will see, the precise lower bound for this matrix—according to our Theorem 2 with $k=2$ and $\theta=1$—is

$$\nu \geq \nu_R(2,1) \approx 4.4111.$$

Remark 5. *It is not restrictive to assume $a>0$ in* (1)*. If we assumed $a<0$ instead, then the results of Theorems 1 and 2 would remain valid (together with Figures 3–4, for example), with all the arguments in their proofs being essentially the same. For example, as we will see in Section 3, the non-negativity of (the first row of) matrix M is governed by the elements $M_{1,1}$ and $M_{1,m}$ for $a>0$ and m odd—this would change to elements $M_{1,1}$ and $M_{1,2}$ for $a<0$ and m odd.*

3 Second-order centered discretization in space and θ-method in time—algebraic characterization of the entries of the full discretization matrix

The results of Section 2 are based on the eigendecomposition of the full discretization matrix $M = \mathcal{F}R(\nu\Lambda)\mathcal{F}^*$. In this section, instead of using trigonometric functions, we give an algebraic description of the matrix entires by exploiting the relation $M = R(\nu L)$ in (11) with L defined in (14). Explicitly, this means

$$M(m,\theta,\nu) = (I-\theta\nu L)^{-1}(I+(1-\theta)\nu L) \in \mathbb{R}^{m\times m}, \tag{22}$$

but the dependence of M on its parameters will often be suppressed.

It is trivial that for $\theta=0$ we have $M(m,0,\nu)=I+\nu L$, hence $M\geq 0$ cannot hold for any $\nu>0$. The case $m=2k$ has been discussed in Section 2.1. Thus, throughout the rest of this section, we can assume that

$$\boxed{m=2k+1 \quad (k\in\mathbb{N}^+), \quad \nu>0 \text{ and } 0<\theta\leq 1.} \tag{23}$$

3.1 Explicit description of the matrix entries for odd values of m

To illustrate the structure of M, we present its first row (as a vector, and with the common denominator of the entries in front of it) for the smallest values of m.

Example 1. *For $m = 3$ the first row of* (22) *is*

$$\frac{1}{3\theta^2\nu^2/4+1}\left(\frac{3\theta^2\nu^2}{4}-\frac{\theta\nu^2}{2}+1, \frac{\theta\nu^2}{4}+\frac{\nu}{2}, \frac{\theta\nu^2}{4}-\frac{\nu}{2}\right),$$

while for $m = 5$ we have

$$\frac{1}{5\theta^4\nu^4/16+5\theta^2\nu^2/4+1}\left(\frac{5\theta^4\nu^4}{16}-\frac{\theta^3\nu^4}{4}+\frac{5\theta^2\nu^2}{4}-\frac{\theta\nu^2}{2}+1,\right.$$

$$\left.\frac{\theta^3\nu^4}{16}+\frac{\theta^2\nu^3}{4}+\frac{\nu}{2}, \frac{\theta^3\nu^4}{16}-\frac{\theta^2\nu^3}{8}+\frac{\theta\nu^2}{4}, \frac{\theta^3\nu^4}{16}+\frac{\theta^2\nu^3}{8}+\frac{\theta\nu^2}{4}, \frac{\theta^3\nu^4}{16}-\frac{\theta^2\nu^3}{4}-\frac{\nu}{2}\right).$$

Each element of M is a rational function in the variables θ and ν. From (22) it is clear that

$$M_{1,j} = \frac{\mathcal{P}_{j,k}(\theta,\nu)}{\mathcal{D}_k(\theta,\nu)} \qquad (j = 1, 2, \dots, 2k+1), \tag{24}$$

where $\mathcal{P}_{j,k}$ and $\mathcal{D}_k$ are certain bivariate polynomials in θ and ν, and

$$\mathcal{D}_k := \det\left(I_{(2k+1)\times(2k+1)} - \theta\nu L_{(2k+1)\times(2k+1)}\right). \tag{25}$$

Remark 6. *The subscripts of $\mathcal{P}_{j,k}$ thus refer to the position of the polynomial within the first row of M, and the size of $M \in \mathbb{R}^{(2k+1)\times(2k+1)}$, respectively.*

The key to describing M algebraically is the observation that the polynomials $\mathcal{P}_{j,k}$ and $\mathcal{D}_k$ satisfy certain low-order linear recursions with constant coefficients. As already indicated by Section 2.1, the leftmost entry ($j = 1$) behaves differently than the rest ($2 \le j \le 2k+1$).

Remark 7. *Mathematica's* `FindLinearRecurrence` *command proved to be an efficient tool for discovering these linear recursions.*

First, let us introduce some new variables. On the one hand, as suggested by Example 1, it seems convenient to set

$$\mu := \theta^2\nu^2 > 0.$$

Then, due to the sign assumptions, $\sqrt{\mu} = \theta\nu$. On the other hand, as we will soon see, the polynomial

$$\kappa^2 - \kappa\left(1+\frac{\mu}{2}\right) + \frac{\mu^2}{16}$$

will appear as a (factor of a) characteristic polynomial, and its roots are

$$\kappa_{1,2} = \frac{2+\mu \pm 2\sqrt{\mu+1}}{4} = \left(\frac{\sqrt{1+\mu}\pm 1}{2}\right)^2. \tag{26}$$

This motivates us to introduce yet another variable, which will further simplify our exposition. We set

$$y := \frac{\sqrt{1+\mu}-1}{\sqrt{\mu}} = \frac{\sqrt{1+\theta^2\nu^2}-1}{\theta\nu} \in (0,1). \tag{27}$$

It is seen that the transformation

$$(0,+\infty) \ni \mu \longleftrightarrow y \in (0,1)$$

is a bijection. Moreover, the following (inverse) relations

$$\mu = \left(\frac{2y}{1-y^2}\right)^2,$$

$$\mu y^2 = 2+\mu-2\sqrt{1+\mu},$$
$$\mu/y^2 = 2+\mu+2\sqrt{1+\mu},$$

and

$$\nu = \frac{2y}{1-y^2}\cdot\frac{1}{\theta} \tag{28}$$

are easily verified. We can now start describing the entries of the first row of M.

Remark 8. *Although the expressions $\mathcal{P}_{j,k}$ and $\mathcal{D}_k$ will become in general rational functions in the variable y, we still call them polynomials (referring to their structure in the original variables θ and ν).*

- The polynomials $\mathcal{D}_k$. By carrying out some determinant expansions, we see that the determinants (25) obey the second-order parametric recursion

$$\mathcal{D}_{k+2} = \left(1+\frac{\mu}{2}\right)\mathcal{D}_{k+1} - \frac{\mu^2}{16}\mathcal{D}_k \tag{29}$$

with initial conditions

$$\mathcal{D}_1 = 1+\frac{3\mu}{4}, \quad \mathcal{D}_2 = 1+\frac{5\mu}{4}+\frac{5\mu^2}{16}$$

(cf. Example 1). After solving this recursion, we obtain

$$\mathcal{D}_k = \left(\frac{\sqrt{1+\mu}+1}{2}\right)^{2k+1} - \left(\frac{\sqrt{1+\mu}-1}{2}\right)^{2k+1},$$

which, in terms of the variable y, becomes

$$\mathcal{D}_k = \frac{1-y^{4k+2}}{(1-y^2)^{2k+1}}. \tag{30}$$

- The polynomials $\mathcal{P}_{1,k}$. They satisfy the recursion

$$\mathcal{P}_{1,k+2} = \left(1+\frac{\mu}{2}\right)\mathcal{P}_{1,k+1} - \frac{\mu^2}{16}\mathcal{P}_{1,k},$$

that is, with coefficients being the same as in (29), but with initial conditions

$$\mathcal{P}_{1,1} = 1+\frac{3\mu}{4}-\frac{\mu/\theta}{2}, \quad \mathcal{P}_{1,2} = 1+\frac{5\mu}{4}+\frac{5\mu^2}{16}-\frac{\mu/\theta}{2}-\frac{\mu^2/\theta}{4}$$

(cf. Example 1). By solving this recursion, we derive that

$$\mathcal{P}_{1,k} = \frac{P_{L,k,\theta}(y)}{(1+y^2)(1-y^2)^{2k+1}\theta}, \tag{31}$$

where the numerator is

$$P_{L,k,\theta}(y) := -\theta y^{4k+4} - (\theta-2)y^{4k+2} + (\theta-2)y^2 + \theta. \tag{32}$$

Remark 9. *Here, the subscript L stands for* leftmost*. This polynomial will play a special role in the next section.*

• The polynomials $\mathcal{P}_{2,k}$. They satisfy a third-order recursion in the variable k,

$$\mathcal{P}_{2,k+3} = \left(1 + \frac{3\mu}{4}\right)\mathcal{P}_{2,k+2} - \left(\frac{\mu}{4} + \frac{3\mu^2}{16}\right)\mathcal{P}_{2,k+1} + \frac{\mu^3}{64}\mathcal{P}_{2,k}, \tag{33}$$

with initial conditions

$$\mathcal{P}_{2,1} = \left(\frac{1}{2} + \frac{\sqrt{\mu}}{4}\right)\nu, \quad \mathcal{P}_{2,2} = \left(\frac{1}{2} + \frac{\mu}{4} + \frac{\mu^{3/2}}{16}\right)\nu,$$

$$\mathcal{P}_{2,3} = \left(\frac{1}{2} + \frac{\mu}{2} + \frac{3\mu^2}{32} + \frac{\mu^{5/2}}{64}\right)\nu.$$

The characteristic polynomial of recursion (33) is

$$\kappa^3 - \kappa^2\left(1 + \frac{3\mu}{4}\right) + \kappa\left(\frac{\mu}{4} + \frac{3\mu^2}{16}\right) - \frac{\mu^3}{64} = \left(\kappa - \frac{\mu}{4}\right)\left(\kappa^2 - \kappa\left(1 + \frac{\mu}{2}\right) + \frac{\mu^2}{16}\right),$$

hence the characteristic roots are $\kappa_{1,2}$ as in (26), and $\kappa_3 = \mu/4$. Based on this, one easily obtains the explicit solution as

$$\mathcal{P}_{2,k} = \frac{\nu\left(1-y^2\right)^{1-2k}\left(1 + y^{2k-1} + y^{2k+1} - y^{4k}\right)}{2\left(1+y^2\right)}. \tag{34}$$

• The polynomials $\mathcal{P}_{3,k}$. They satisfy the same third-order recursion in the variable k as (33),

$$\mathcal{P}_{3,k+3} = \left(1 + \frac{3\mu}{4}\right)\mathcal{P}_{3,k+2} - \left(\frac{\mu}{4} + \frac{3\mu^2}{16}\right)\mathcal{P}_{3,k+1} + \frac{\mu^3}{64}\mathcal{P}_{3,k},$$

but with initial conditions

$$\mathcal{P}_{3,1} = \left(-\frac{1}{2} + \frac{\sqrt{\mu}}{4}\right)\nu, \quad \mathcal{P}_{3,2} = \left(\frac{\sqrt{\mu}}{4} - \frac{\mu}{8} + \frac{\mu^{3/2}}{16}\right)\nu,$$

$$\mathcal{P}_{3,3} = \left(\frac{\sqrt{\mu}}{4} - \frac{\mu^2}{32} + \frac{3\mu^{3/2}}{16} + \frac{\mu^{5/2}}{64}\right)\nu.$$

The explicit solution of this recursion is

$$\mathcal{P}_{3,k} = \frac{\nu\left(1-y^2\right)^{1-2k}\left(y - y^{2k-2} + y^{2k+2} + y^{4k-1}\right)}{2\left(1+y^2\right)}. \tag{35}$$

Remark 10. *We note that, for any* fixed $j \geq 2$*, the polynomials* $\mathcal{P}_{j,k}$ *satisfy the same third-order recursion* (33) *in the variable* k*, with triplets of initial conditions depending on* j*. However, we cannot use this approach to proceed, since setting up the initial conditions would require, among others, the knowledge of the polynomials* $\mathcal{P}_{j,1}$ *(for* $j = 2, 3$*),* $\mathcal{P}_{j,2}$ *(for* $j = 4, 5$*),* $\mathcal{P}_{j,3}$ *(for* $j = 6, 7$*), and so on.*

- The polynomials $\mathcal{P}_{j,k}$ $(4 \leq j \leq 2k+1,\ k \geq 2)$. They satisfy the following second-order recursion in the variable j when k is *fixed* (hence having only finitely many terms for a particular k):

$$\mathcal{P}_{j+2,k} = -\frac{2}{\sqrt{\mu}}\mathcal{P}_{j+1,k} + \mathcal{P}_{j,k}.$$

For the initial conditions of this final recursion, we use the general forms of $\mathcal{P}_{2,k}$ and $\mathcal{P}_{3,k}$ in (34) and (35) to get for any $k \geq 1$ and $2 \leq j \leq 2k+1$ that

$$\mathcal{P}_{j,k} = \frac{\nu\left(1-y^2\right)^{1-2k}}{2\left(1+y^2\right)} P_{j,k}(y), \tag{36}$$

where the polynomials $P_{j,k}$ are defined as

$$P_{j,k}(y) := (-1)^{j-1} y^{4k+2-j} + y^{2k-1+j} + (-1)^j y^{2k+1-j} + y^{j-2}. \tag{37}$$

As a special case, we set

$$P_{R,k}(y) := P_{2k+1,k}(y),$$

in other words we have

$$P_{R,k}(y) = y^{4k} + y^{2k+1} + y^{2k-1} - 1, \tag{38}$$

where the subscript R stands for *rightmost*.

Remark 11. *As a by-product, we have obtained the following set of identities by comparing the trigonometric and algebraic representations presented so far. They are also interesting from a structural point of view: although the number of terms in the trigonometric sums increases as k gets larger, the polynomials in y are sparse polynomials (also known as lacunary polynomials or fewnomials)—the number of terms does not increase as the polynomial degree increases.*

Corollary 1. *With M defined in* (22), $\theta > 0$, $\nu > 0$, $k \in \mathbb{N}^+$, $y = \frac{\sqrt{1+\theta^2\nu^2}-1}{\theta\nu}$, *and* $\xi_\ell = \frac{2\pi(\ell-1)}{2k+1}$, *we have that*

$$\frac{1}{2k+1}\sum_{\ell=1}^{2k+1} \frac{1+\imath(1-\theta)\nu\sin(\xi_\ell)}{1-\imath\theta\nu\sin(\xi_\ell)} =$$

$$M_{1,1} = \frac{\mathcal{P}_{1,k}}{\mathcal{D}_k} =$$

$$\frac{-\theta y^{4k+4} - (\theta-2)y^{4k+2} + (\theta-2)y^2 + \theta}{\left(1+y^2\right)\left(1-y^{4k+2}\right)\theta}.$$

Moreover, for $j = 2, 3, \ldots, 2k+1$ we have that

$$\frac{1}{2k+1}\sum_{\ell=1}^{2k+1} \frac{1+\imath(1-\theta)\nu\sin(\xi_\ell)}{1-\imath\theta\nu\sin(\xi_\ell)} \exp\left(-\imath(j-1)\xi_\ell\right) =$$

$$M_{1,j} = \frac{\mathcal{P}_{j,k}}{\mathcal{D}_k} =$$

$$\frac{\nu\left(1-y^2\right)^2}{2\left(1+y^2\right)\left(1-y^{4k+2}\right)}\Big((-1)^{j-1} y^{4k+2-j} + y^{2k-1+j} + (-1)^j y^{2k+1-j} + y^{j-2}\Big).$$

In particular,

$$\prod_{\ell=1}^{2k+1}\left(1-\imath\theta\nu\sin(\xi_\ell)\right) = \mathcal{D}_k =$$

$$\left(\frac{\sqrt{1+\theta^2\nu^2}+1}{2}\right)^{2k+1} - \left(\frac{\sqrt{1+\theta^2\nu^2}-1}{2}\right)^{2k+1} = \frac{1-y^{4k+2}}{\left(1-y^2\right)^{2k+1}}.$$

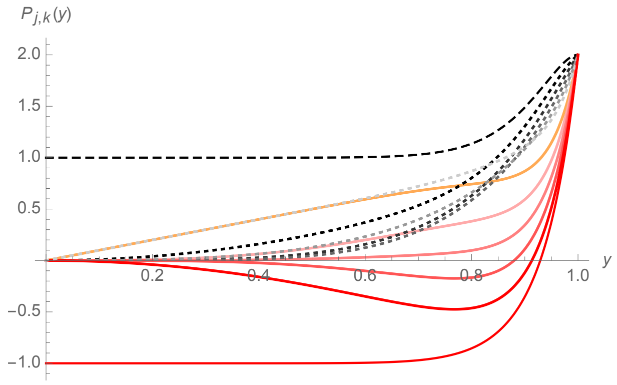

Figure 1: The typical behavior of the polynomials $P_{j,k}$ appearing in Corollary 2 for $2 \le j \le 2k+1$ and k fixed: curves in shades of gray (or black) correspond to even j, while curves in shades of red (or orange) correspond to odd j indices. Based on this figure, one can make the following observations. On the one hand, for each fixed and even j, $P_{j,k}$ is strictly increasing in y; however, for any fixed $y \in (0,1)$, $P_{j,k}$ is in general not monotone in its even index j. On the other hand, for each fixed and odd j, $P_{j,k}$ is in general not monotone in y; however, for any fixed $y \in (0,1)$, $P_{j,k}$ is strictly decreasing in its odd index j.

3.2 Non-negativity of the matrix entries for odd values of m

In this section we present a detailed description of the non-negativity properties of the matrix M, thanks to the explicit forms for the entries $M_{1,j}$ obtained in Section 3.1. Throughout this section we still assume (23).

By taking into account (24), (30), (31), (32), (36), (37), and the fact that now $y \in (0,1)$ (see (27)), the following corollary is evident.

Corollary 2. *For a given pair* (θ, ν)

$$M_{1,1}(2k+1,\theta,\nu) \ge 0 \quad \Longleftrightarrow \quad P_{L,k,\theta}(y) \ge 0 \quad \textit{(see (32))},$$

and for any $2 \le j \le 2k+1$

$$M_{1,j}(2k+1,\theta,\nu) \ge 0 \quad \Longleftrightarrow \quad P_{j,k}(y) \ge 0 \quad \textit{(see (37))}.$$

The following lemma proves some of the observations about the polynomials $P_{j,k}$ suggested by Figure 1 for even and odd indices $2 \le j \le 2k+1$.

Lemma 1. *Let us fix* $y \in (0,1)$ *arbitrarily. Then*
- *for any* $1 \le \ell \le k$, $P_{2\ell,k}(y) > 0$;
- *for any* $2 \le \ell \le k$, $P_{2\ell+1,k}(y) < P_{2\ell-1,k}(y)$.

Proof. For the even indices, we have

$$P_{2\ell,k}(y) = y^{2k-2l+1}(1-y^{2k+1}) + y^{2k+2l-1} + y^{2l-2} > 0,$$

while for the odd indices,

$$P_{2\ell+1,k}(y) - P_{2\ell-1,k}(y) = -(1-y^2)\Big(y^{2k+2l-2} + y^{2l-3} + y^{2k-2l}(1-y^{2k+1})\Big) < 0.$$

□

By combining Corollary 2 and Lemma 1, we have obtained the following result, expressing the fact that the non-negativity of $M(2k+1,\theta,\nu)$ is determined only by the polynomials appearing in the numerators of its top left and top right entries.

Corollary 3. *For a given pair* (θ,ν)

$$M_{1,1}(2k+1,\theta,\nu)\geq 0 \quad\Longleftrightarrow\quad P_{L,k,\theta}(y)\geq 0 \quad \textit{(see (32))},$$

and

$$M_{1,j}(2k+1,\theta,\nu)\geq 0 \textit{ for each } 2\leq j\leq 2k+1 \quad\Longleftrightarrow\quad P_{R,k}(y)\geq 0 \quad \textit{(see (38))}.$$

The non-negativity of $M(2k+1,\theta,\nu)$ has therefore been reduced to studying the simultaneous non-negativity of two parametric polynomials, $P_{L,k,\theta}$ and $P_{R,k}$, over the y-interval $(0,1)$. The content of Lemmas 2 and 3 is illustrated by Figure 2.

Lemma 2 (about the sign of $P_{R,k}(y)$)**.** *Let us fix k arbitrarily, and recall that by definition $P_{R,k}(y)=y^{4k}+y^{2k+1}+y^{2k-1}-1$. Then there is a unique $y\in(0,1)$ such that $P_{R,k}(y)=0$.*
Let

$$y_R(k) \textit{ denote this root.} \tag{39}$$

Then $P_{R,k}(y)<0$ for $y\in(0,y_R(k))$, and $P_{R,k}(y)>0$ for $y\in(y_R(k),1)$.
Moreover, $y_R(k)<y_R(k+1)$, $\lim_{k\to+\infty} y_R(k)=1$, and

$$\left(\sqrt{2}-1\right)^{\frac{1}{2k-1}}<y_R(k)<\left(\sqrt{2}-1\right)^{\frac{1}{2k+1}}. \tag{40}$$

Proof. For fixed k, the continuous function $y\mapsto P_{R,k}(y)=y^{4k}+y^{2k+1}+y^{2k-1}-1$ is strictly increasing, $P_{R,k}(0)<0$ and $P_{R,k}(1)>0$, hence there is a unique root. This root is strictly increasing in k, because the function $k\mapsto P_{R,k}(y)$ is strictly decreasing for fixed $y\in(0,1)$. Finally notice that $y^{4k}+y^{2k+1}+y^{2k-1}-1=0$ is equivalent to $\left(y^{2k-1}+1\right)\left(y^{2k+1}+1\right)=2$, and for any $y\in(0,1)$ one has

$$\left(y^{2k+1}+1\right)^2<\left(y^{2k-1}+1\right)\left(y^{2k+1}+1\right)<\left(y^{2k-1}+1\right)^2.$$

From this we easily get (40) and also the limit of $y_R(k)$ ($k\to+\infty$). □

Remark 12. *The asymptotic series (as $k\to+\infty$) of both bounds in* (40) *has the form*

$$1+\frac{\ln\left(\sqrt{2}-1\right)}{2k}+\mathcal{O}\left(\frac{1}{k^2}\right)\approx 1-\frac{0.44069}{k}+\mathcal{O}\left(\frac{1}{k^2}\right).$$

Lemma 3 (about the sign of $P_{L,k,\theta}(y)$)**.** *Let us fix k arbitrarily, and recall that by definition $P_{L,k,\theta}(y)=-\theta y^{4k+4}-(\theta-2)y^{4k+2}+(\theta-2)y^2+\theta$.*
(i) Suppose that $\frac{2k}{2k+1}\leq\theta\leq 1$. Then, for any $y\in(0,1)$, $P_{L,k,\theta}(y)>0$.
(ii) Suppose now that $0<\theta<\frac{2k}{2k+1}$. Then there is a unique $y\in(0,1)$ such that $P_{L,k,\theta}(y)=0$.
Let

$$y_L(k,\theta) \textit{ denote this root.} \tag{41}$$

Then $P_{L,k,\theta}(y)>0$ for $y\in(0,y_L(k,\theta))$, and $P_{L,k,\theta}(y)<0$ for $y\in(y_L(k,\theta),1)$.
Moreover, on the one hand, for fixed $0<\theta<\frac{2k}{2k+1}$, the function $k\mapsto y_L(k,\theta)$ is strictly decreasing, and $\lim_{k\to+\infty} y_L(k,\theta)=\sqrt{\frac{\theta}{2-\theta}}\in(0,1)$.
On the other hand, for fixed $k\in\mathbb{N}^+$, the function $\left(0,\frac{2k}{2k+1}\right)\ni\theta\mapsto y_L(k,\theta)$ is strictly increasing, and we have the one-sided limits $\lim_{\theta\to 0+0} y_L(k,\theta)=0$ and $\lim_{\theta\to\frac{2k}{2k+1}-0} y_L(k,\theta)=1$.

Proof. We notice that the expression $P_{L,k,\theta}(y)$ is linear in θ, so by setting

$$\Theta(y,k) := \frac{2y^2\left(1-y^{4k}\right)}{\left(1+y^2\right)\left(1-y^{4k+2}\right)},$$

we easily get for any $y \in (0,1)$ that

$$P_{L,k,\theta}(y) \lesseqqgtr 0 \quad \Longleftrightarrow \quad \theta \lesseqqgtr \Theta(y,k), \tag{42}$$

where the symbol $\lesseqqgtr$ denotes either $<$, or $=$, or $>$ on both sides of the equivalence. It is seen that for fixed k we have the one-sided limits

$$\lim_{y\to 0+0}\Theta(y,k) = 0 \quad \text{and} \quad \lim_{y\to 1-0}\Theta(y,k) = \frac{2k}{2k+1}. \tag{43}$$

Now we show that the function

$$(0,1) \ni y \mapsto \Theta(y,k) \quad \text{is strictly increasing.} \tag{44}$$

The partial derivative

$$\partial_y\Theta(y,k) = \frac{4y\left(1-(2k+1)y^{4k}+(2k+1)y^{4k+4}-y^{8k+4}\right)}{\left(1+y^2\right)^2\left(1-y^{4k+2}\right)^2}$$

is positive, if $\widetilde{P}(y,k) := 1-(2k+1)y^{4k}+(2k+1)y^{4k+4}-y^{8k+4} > 0$. But

$$\widetilde{P}(0,k) = 1 \quad \text{and} \quad \widetilde{P}(1,k) = 0,$$

so the positivity of $\widetilde{P}(y,k)$ will follow if we show that $y \mapsto \widetilde{P}(y,k)$ is strictly decreasing. Indeed,

$$\partial_y\widetilde{P}(y,k) = -4(2k+1)y^{4k-1}\widetilde{Q}(y,k),$$

where

$$\widetilde{Q}(y,k) := y^{4k+4}-(k+1)y^4+k,$$

hence it is enough to verify $\widetilde{Q}(y,k) > 0$. And this is true, since $\widetilde{Q}(0,k) = k$, $\widetilde{Q}(1,k) = 0$ and

$$\partial_y\widetilde{Q}(y,k) = -4(k+1)y^3\left(1-y^{4k}\right) < 0.$$

Now, as (44) has been checked, it is obvious that continuity, (42), (43) and (44) imply statement (*i*) of the lemma, and, at the same time, regarding statement (*ii*) of the lemma, the existence of a unique root $y_L(k,\theta) \in (0,1)$, the positivity of $P_{L,k,\theta}$ on $(0, y_L(k,\theta))$, and the negativity of $P_{L,k,\theta}$ on $(y_L(k,\theta),1)$.

We finally discuss the monotonicity and limit properties of the root $y_L(k,\theta)$. For fixed $y \in (0,1)$, the function $k \mapsto \Theta(y,k)$ is strictly increasing, since

$$\Theta(y,k+1)-\Theta(y,k) = \frac{2\left(1-y^2\right)^2 y^{4k+2}}{\left(1-y^{4k+2}\right)\left(1-y^{4k+6}\right)} > 0.$$

This implies that, for any fixed $\theta \in \left(0, \frac{2k}{2k+1}\right)$, the function $k \mapsto y_L(k,\theta)$ is strictly decreasing. Moreover, for fixed $y \in (0,1)$, we see from the definition that $\lim_{k\to+\infty}\Theta(y,k) = \frac{2y^2}{1+y^2}$, so, due to (42) with "equality", one has for fixed $\theta \in \left(0, \frac{2k}{2k+1}\right)$ that $y_\infty(\theta) := \lim_{k\to+\infty} y_L(k,\theta)$ solves $\theta = \frac{2y_\infty(\theta)^2}{1+y_\infty(\theta)^2}$; in other words, $y_\infty(\theta) = \sqrt{\frac{\theta}{2-\theta}} \in (0,1)$. To show the validity of the last sentence of the lemma, we fix $k \in \mathbb{N}^+$, and simply take into account again (42) with "equality", (43) and (44). □

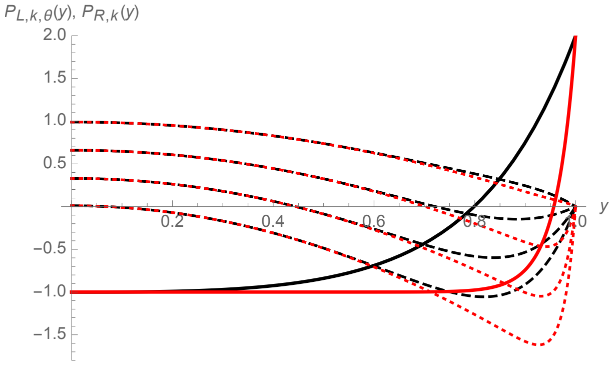

Figure 2: The two solid curves show the functions $y \mapsto P_{R,k}(y)$ for some $k = k_0$ (solid black) and $k = k_1$ (solid red) with $k_0 < k_1$. The dashed black curves show the functions $y \mapsto P_{L,k,\theta}(y)$ for $k = k_0$ and for various values of $\theta \in (0, 1]$. Finally, the dotted red curves show the functions $y \mapsto P_{L,k,\theta}(y)$ for $k = k_1$ and for the same values of $\theta \in (0, 1]$.

In order to return to the original variables (θ, ν) from the variable y—based on (28), (39) and (41)—we define

$$\nu_R(k, \theta) := \frac{2y_R(k)}{1 - y_R(k)^2} \cdot \frac{1}{\theta}, \tag{45}$$

and similarly,

$$\nu_L(k, \theta) := \begin{cases} \frac{2y_L(k,\theta)}{1-y_L(k,\theta)^2} \cdot \frac{1}{\theta} & \text{for } 0 < \theta < \frac{2k}{2k+1} \\ +\infty & \text{for } \frac{2k}{2k+1} \leq \theta \leq 1. \end{cases} \tag{46}$$

The value $+\infty$ is introduced here for convenience so as to make our descriptions shorter.

A reformulation of Corollary 3 in terms of the variables (θ, ν) is given below.

Corollary 4. *For any $k \in \mathbb{N}^+$ and $\theta \in (0, 1]$ we have*

$$M_{1,1}(2k + 1, \theta, \nu) \geq 0 \quad \Longleftrightarrow \quad \nu \leq \nu_L(k, \theta),$$

and

$$M_{1,j}(2k + 1, \theta, \nu) \geq 0 \text{ for each } 2 \leq j \leq 2k + 1 \quad \Longleftrightarrow \quad \nu \geq \nu_R(k, \theta).$$

In particular,

$$M_{1,j}(2k + 1, \theta, \nu) \geq 0 \text{ for each } 1 \leq j \leq 2k + 1 \quad \Longleftrightarrow \quad \nu_R(k, \theta) \leq \nu \leq \nu_L(k, \theta).$$

Proof. By taking into account Corollary 3, Lemmas 2 and 3, and the fact that the map in (28)

$$(0, 1) \ni y \mapsto \frac{2y}{1 - y^2} \in (0, +\infty) \text{ is a strictly increasing bijection,} \tag{47}$$

we get for fixed k and θ that $P_{R,k}(y) \geq 0$ is equivalent to $\nu \geq \nu_R(k, \theta)$, and $P_{L,k,\theta}(y) \geq 0$ is equivalent to $\nu \leq \nu_L(k, \theta)$. In particular, due to the definition of $\nu_L(k, \theta)$ in (46), this last inequality means that there is no upper bound on ν for $\frac{2k}{2k+1} \leq \theta \leq 1$. □

Some growth rates, monotonicity and limit properties of $\nu_R(k, \theta)$ and $\nu_L(k, \theta)$—defined in (45)–(46)—are collected below; see also Figures 3 and 4.

Corollary 5. *(i) For any $k \in \mathbb{N}^+$ and $\theta \in (0,1]$, we have $\nu_R(k,\theta) < \nu_R(k+1,\theta)$, and*

$$\frac{2\left(\sqrt{2}+1\right)^{\frac{1}{2k-1}}}{\left(\sqrt{2}+1\right)^{\frac{2}{2k-1}}-1}\cdot\frac{1}{\theta} < \nu_R(k,\theta) < \frac{2\left(\sqrt{2}-1\right)^{\frac{1}{2k+1}}}{1-\left(\sqrt{2}-1\right)^{\frac{2}{2k+1}}}\cdot\frac{1}{\theta}. \tag{48}$$

The asymptotic series for these lower and upper bounds have the form

$$\left(\frac{2}{\ln\left(\sqrt{2}+1\right)}k \mp \frac{1}{\ln\left(\sqrt{2}+1\right)} + \mathcal{O}\left(\frac{1}{k}\right)\right)\cdot\frac{1}{\theta},$$

being approximately $\left(2.26919k \mp 1.13459 + \mathcal{O}\left(\frac{1}{k}\right)\right)\cdot\frac{1}{\theta}$. In particular, $\lim_{k\to+\infty}\nu_R(k,\theta) = +\infty$.
(ii) For fixed $0 < \theta < \frac{2k}{2k+1}$, $\nu_L(k,\theta) > \nu_L(k+1,\theta)$ (and $\nu_L(k,\theta) = +\infty$ for $\frac{2k}{2k+1} \leq \theta \leq 1$). Finally, for fixed $\theta \in (0,1)$, we have the limit

$$\lim_{k\to+\infty}\nu_L(k,\theta) = \frac{1}{1-\theta}\sqrt{\frac{2-\theta}{\theta}}, \tag{49}$$

and, for fixed $k \in \mathbb{N}^+$, the one-sided limits

$$\lim_{\theta\to 0+0}\nu_L(k,\theta) = +\infty = \lim_{\theta\to\frac{2k}{2k+1}-0}\nu_L(k,\theta). \tag{50}$$

Proof. (*i*) The monotonicity of $\nu_R(k,\theta)$ in k for fixed θ follows from the monotonicity of $y_R(k)$ in Lemma 2 together with (47), and inequality (48) is just (40) under the transformation (47).
(*ii*) We similarly obtain the monotonicity of $\nu_L(k,\theta)$ in k for fixed θ, and the limit (49) from Lemma 3 via (47), by also noting that

$$\frac{2\sqrt{\frac{\theta}{2-\theta}}}{1-\left(\sqrt{\frac{\theta}{2-\theta}}\right)^2}\cdot\frac{1}{\theta} = \frac{1}{1-\theta}\sqrt{\frac{2-\theta}{\theta}}.$$

As for the $\theta \to \frac{2k}{2k+1} - 0$ limit in (50), we know from Lemma 3 that $y_L(k,\theta) \to 1$ (from below), and $\lim_{y\to 1-0}\frac{2y}{1-y^2} = +\infty$, hence $\nu_L(k,\theta) \to +\infty$ when $\theta \to \frac{2k}{2k+1} - 0$.

One needs to take care only when evaluating the $\theta \to 0+0$ limit in (50) for fixed $k \in \mathbb{N}^+$, since $\frac{2y_L(k,\theta)}{1-y_L(k,\theta)^2} \to 0$ and $\frac{1}{\theta} \to +\infty$ in (46) when $\theta \to 0+0$. But the monotonicity of $\nu_L(k,\theta)$ in k for fixed θ, and (49) imply for any k and $\theta \in (0,1)$ that

$$\nu_L(k,\theta) \geq \frac{1}{1-\theta}\sqrt{\frac{2-\theta}{\theta}},$$

and the right-hand side here tends to $+\infty$ as $\theta \to 0+0$. □

The following result explains why the "left half" of Figure 3 is "empty" (cf. Corollary 4)—the result is non-trivial, since for fixed k, $\lim_{\theta\to 0+0}\nu_L(k,\theta) = +\infty = \lim_{\theta\to 0+0}\nu_R(k,\theta)$ (cf. Figure 4).

Lemma 4. *For any $k \in \mathbb{N}^+$ there is a unique $\theta_k \in \left[\frac{1}{2}, \frac{2k}{2k+1}\right)$ such that*

$$\nu_R(k,\theta) = \nu_L(k,\theta) \quad (\text{for } \theta \in (0,1]) \quad \Longleftrightarrow \quad \theta = \theta_k.$$

This θ_k also satisfies

$$\nu_R(k,\theta) < \nu_L(k,\theta) \quad \Longleftrightarrow \quad \theta > \theta_k.$$

Moreover, the sequence θ_k is strictly increasing in k, and $\theta_1 = \frac{1}{2}$. In particular, for any $k \in \mathbb{N}^+$ and $\theta \in \left(0, \frac{1}{2}\right)$ we have

$$\nu_R(k,\theta) > \nu_L(k,\theta).$$

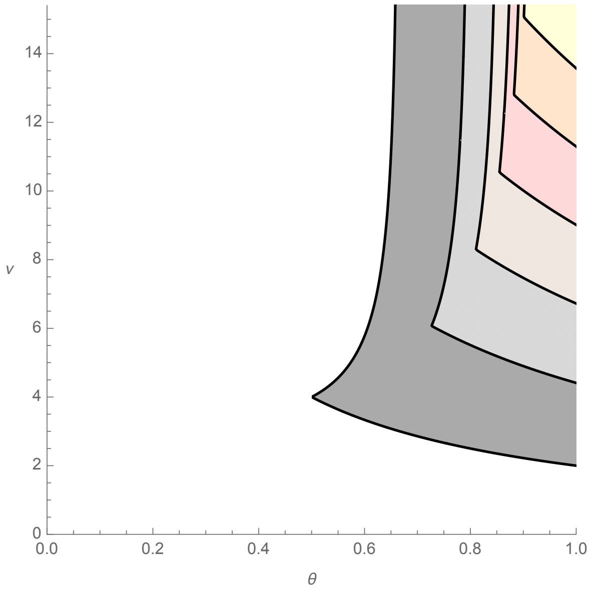

Figure 3: The parameter regions in the (θ, ν) parameter plane ensuring $M(2k+1, \theta, \nu) \geq 0$ for $k = 1, 2, \ldots, 6$ (different values of k are represented by different colors). The regions continue to extend to infinity "upward", but "shrink" in the horizontal direction as k is increased.

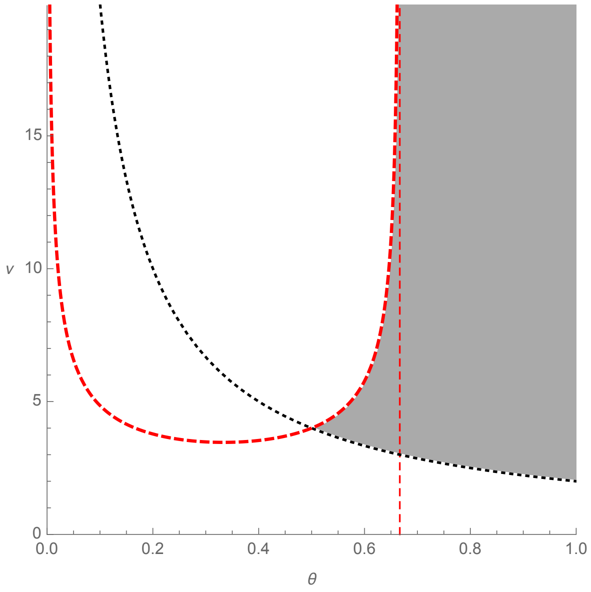

Figure 4: A typical shaded region in Figure 3 for which $M(2k+1, \theta, \nu) \geq 0$; in this particular case, for $k = 1$. The gray region is described by the inequalities $\nu_R(k, \theta) \leq \nu \leq \nu_L(k, \theta)$. The black dotted curve represents the function $\theta \mapsto \nu_R(k, \theta)$, while the red dashed curve is the function $\theta \mapsto \nu_L(k, \theta)$, having a vertical asymptote at $\theta = 2k/(2k+1)$.

Proof. Let us fix k. Due to (45)–(46), $\nu_R(k,\theta)$ is finite but $\nu_L(k,\theta)$ is infinite for any $\theta \in \left[\frac{2k}{2k+1}, 1\right]$, so $\nu_R(k,\theta) = \nu_L(k,\theta)$ cannot hold. For $\theta \in \left(0, \frac{2k}{2k+1}\right)$, by also using (47), we have

$$\nu_R(k,\theta) = \nu_L(k,\theta) \quad \Longleftrightarrow \quad \frac{2y_R(k)}{1-y_R(k)^2} = \frac{2y_L(k,\theta)}{1-y_L(k,\theta)^2} \quad \Longleftrightarrow \quad y_R(k) = y_L(k,\theta).$$

Here $y_R(k)$ is independent of θ, and k is fixed, so by definitions (39) and (41) this means that θ must be chosen in a way such that $P_{L,k,\theta}(y_R(k)) = 0$. By using the notation introduced in the proof of Lemma 3, this is equivalent to $\theta = \Theta(y_R(k), k) =: \theta_k$. Hence, if $\nu_R(k,\theta) = \nu_L(k,\theta)$ holds for some $\theta \in (0,1]$, then $\theta = \theta_k \in \left(0, \frac{2k}{2k+1}\right)$. Now, by Lemma 2, we have $y_R(k) < y_R(k+1)$, and Θ is strictly increasing in its first argument (see (44)), so the sequence θ_k is also strictly increasing.

The same monotonicity argument shows that $\nu_R(k,\theta) < \nu_L(k,\theta)$ holds for some $\theta \in (0,1]$ if and only if $\theta > \theta_k$ (see (42) and the characterization of $P_{L,k,\theta} > 0$ in Lemma 3).

For $k = 1$, one explicitly computes that

$$\nu_R(1,\theta) = \frac{2}{\theta} \quad \text{and} \quad \nu_L(1,\theta) = \frac{2}{\sqrt{\theta(2-3\theta)}}, \tag{51}$$

so $\nu_R(1,\theta) = \nu_L(1,\theta) \Longleftrightarrow \theta = \theta_1 = 1/2$. Therefore, we have $\theta_k \in \left(\frac{1}{2}, \frac{2k}{2k+1}\right)$ for $k \geq 2$, also implying that for any $k \in \mathbb{N}^+$ and $\theta \in \left(0, \frac{1}{2}\right)$, we have $\nu_R(k,\theta) > \nu_L(k,\theta)$. □

The following theorem summarizes the results of Section 3. In the theorem, we assume $k \in \mathbb{N}^+$, $\theta \in [0,1]$ and $\nu \in (0,+\infty)$.

Theorem 2 (About the full discretization matrix corresponding to the 2^{nd}-order centered discretization in space and θ-method in time)**.**

- *Fix $0 \leq \theta < 1/2$ arbitrarily. Then $M(2k+1,\theta,\nu) \geq 0$ can never hold, i.e. for any $k \in \mathbb{N}^+$ and $\nu > 0$ there is at least one strictly negative entry of the matrix M.*

- *Let $\theta = 1/2$. Then*

$$M\left(2k+1, \frac{1}{2}, \nu\right) \geq 0 \quad \Longleftrightarrow \quad k = 1 \text{ and } \nu = 4 \quad \text{(see (52))}.$$

- *Fix $1/2 < \theta < 1$ arbitrarily. Then there are finitely many values of k for which there exists $\nu > 0$ with $M(2k+1,\theta,\nu) \geq 0$. For any such value of k, the set of admissible values of ν has the form $\nu_R(k,\theta) \leq \nu \leq \nu_L(k,\theta)$, with suitable constants $0 < \nu_R(k,\theta) \leq \nu_L(k,\theta) \leq +\infty$ (the possible case $\nu_L(k,\theta) = +\infty$ means that there is no upper but only a lower bound on ν); see also Corollary 5.*

- *Let $\theta = 1$. Then for each $k \in \mathbb{N}^+$ there is a constant $\nu_R(k,1) > 0$ such that*

$$M(2k+1,1,\nu) \geq 0 \quad \Longleftrightarrow \quad \nu \geq \nu_R(k,1).$$

In addition, $\nu_R(k,1) < \nu_R(k+1,1)$ for any k, $\lim_{k\to+\infty} \nu_R(k,1) = +\infty$, and the two-sided estimates in (48) with $\theta = 1$ hold.

Proof. The case $\theta = 0$ has already been discussed at the beginning of Section 3. In general, for $\theta \in (0,1]$, we know from Corollary 4 that, for any k, the set of ν values for which $M(2k+1,\theta,\nu) \geq 0$ holds has the form $\nu_R(k,\theta) \leq \nu \leq \nu_L(k,\theta)$.

The range $\theta \in \left(0, \frac{1}{2}\right)$ is covered by Lemma 4.

For $\theta = 1/2$, (51) shows that for $k = 1$ one has $\nu_R(1,1/2) = \nu_L(1,1/2) = 4$. But for any $k \geq 2$ we know (see Corollary 5) that

$$\nu_L(k,1/2) < \nu_L(1,1/2) = \nu_R(1,1/2) < \nu_R(k,1/2),$$

hence $\nu_R(k,1/2) \leq \nu_L(k,1/2)$ cannot hold for any $k \geq 2$.

For fixed $1/2 < \theta < 1$, $\nu_L(k,\theta)$ becomes finite for all sufficiently large k (see (46)). But according to Corollary 5, $\nu_L(k,\theta)$ is decreasing in k for $\theta < \frac{2k}{2k+1}$, and $\lim_{k\to+\infty} \nu_R(k,\theta) = +\infty$, so the inequality $\nu_R(k,\theta) \leq \nu_L(k,\theta)$ can hold only for finitely many values of k.

Finally, for $\theta = 1$, $\nu_L(k,1) = +\infty$ and we can use Corollary 5 (*i*) with $\theta = 1$. □

Remark 13. *The "lower left corner point" of each shaded region in Figure 3 corresponds to a pair (θ,ν) for which $\nu = \nu_R(k,\theta) = \nu_L(k,\theta)$. This means that here the leftmost and the rightmost entries of the first row of $M(2k+1,\theta,\nu)$ simultaneously vanish (and the other entries are non-negative). For $k = 1$, this happens for $\theta = \theta_1 = 1/2$ and $\nu = 4$; the corresponding matrix is*

$$M\left(3, \frac{1}{2}, 4\right) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}. \tag{52}$$

4 Other spatial discretizations

The spatial semi-discretization considered in Sections 2.1–3 is not positivity preserving. The same will hold true for each spatial semi-discretization to be investigated in Sections 4.1–4.2 below: it is well-known [7, Chapter I, Theorem 7.2] that a linear constant-coefficient system of ordinary differential equations (3) is positivity preserving if and only if the matrix $\frac{a}{\Delta x}L$ has no negative off-diagonal entries. The violation of this last condition is clear for the matrix L in (14), for all matrices L in Section 4.1, and also for the ones in Section 4.2 (due to $L_{1,2} = -L_{2,1} \neq 0$).

However, as we will see, it is again possible to obtain a positivity-preserving full discretization scheme when the spatial discretizations covered in this section (higher-order centered spatial discretizations and Fourier spectral collocation methods) are suitably combined with the θ-method.

Remark 14. *For $\theta = 0$ (that is, when the explicit Euler time discretization is applied), the matrix M in* (11) *becomes $M = I + \nu L$, hence $M \geq 0$ cannot hold for any $\nu > 0$ due to the negative off-diagonal entries of L. Therefore, in what follows, we can assume $\theta \in (0,1]$.*

4.1 Higher-order centered discretizations in space, θ-method in time

The coefficients of the centered differences can be found, e.g., in [4], from which the corresponding circulant matrices L describing the spatial discretization can be constructed. Here, we examine the first few cases.

- When the stencil width is 3 (implying $m \geq 3$ and 2^{nd}-order accuracy), the entries on the central diagonals are $(-1/2, 0, 1/2)$. This is matrix L in (14) that has been considered in Sections 2.1–3.
- When the stencil width is 5 (implying $m \geq 5$ and 4^{th}-order accuracy), the entries on the central diagonals are $(1/12, -2/3, 0, 2/3, -1/12)$.
- When the stencil width is 7 (implying $m \geq 7$ and 6^{th}-order accuracy), the entries on the central diagonals are $(-1/60, 3/20, -3/4, 0, 3/4, -3/20, 1/60)$.

In all above central diagonals, the middle 0 corresponds to the main diagonal. We can obtain an eigendecomposition (7) for the matrix L, with eigenvectors given by (8) and eigenvalues $\lambda_\ell = \imath\psi(\xi_\ell)$, where

$$\psi(x) := 2 \sum_{k=1}^{(N-1)/2} C_k \sin(kx) \quad (x \in \mathbb{R}). \tag{53}$$

As before, ξ_ℓ is defined in (9), and C is a vector consisting of the last $(N-1)/2$ coefficients of the central diagonals of L, with N denoting the stencil width. For instance, the vector C is equal to $(1/2)$, $(2/3, -1/12)$, and $(3/4, -3/20, 1/60)$ for stencil widths 3, 5, and 7, respectively.

After the matrix L has been chosen, we couple this spatial discretization with the θ-method as time discretization, and the full discretization matrix M is obtained (see (11) and (17)). As seen in Section 2.1, the matrix M is a real, circulant matrix so it can be characterized by the entries of its first row, which take the form

$$M_{1,j} = \frac{1}{m} \sum_{\ell=1}^{m} \frac{(1-\theta(1-\theta)\nu^2\psi^2(\xi_\ell))\cos((j-1)\xi_\ell) + \nu\psi(\xi_\ell)\sin((j-1)\xi_\ell)}{1+\theta^2\nu^2\psi^2(\xi_\ell)} \quad (1 \leq j \leq m). \tag{54}$$

Our computations suggest that the non-negativity properties of the matrix family $M(m, \theta, \nu)$ again depend on the parity of m.

Case 1: m is even.

By using symbolic calculations, we have found that, for the 4^{th}-order scheme, $M(m, \theta, \nu) \geq 0$ cannot hold for any $\theta \in (0, 1]$ and $\nu > 0$ when $m \in \{6, 8, 10\}$. Similarly, for the 6^{th}-order scheme, we checked (again symbolically) that $M(m, \theta, \nu) \geq 0$ does not hold for any $\theta \in (0, 1]$ and $\nu > 0$ when $m \in \{8, 10\}$. Therefore, positivity preservation is impossible in these cases.

The following proposition extends the above observations for the 4^{th}-order scheme when m is a general even number—although only for sufficiently large values of ν.

Proposition 1. *Consider the iterative formula* (10) *applied to the advection equation* (1) *with periodic boundary condition. Let the matrix $M_{m\times m}$ result from the 4^{th}-order centered discretization in space with m spatial grid points and the θ-method in time. Also, let ν be the CFL number defined in* (12)*. If $m \geq 6$ is* even, *then there exists $\nu_0 > 0$ such that the matrix M has at least one negative entry for any $\nu > \nu_0$.*

Proof. We show that $M_{1,m} < 0$ for any $\theta \in (0, 1]$ and $\nu > \nu_0$, where $\nu_0 > 0$ is a constant depending on m and θ.

Let $j = m$ in (54), then by using (19) and $\frac{1}{m}\sum_{\ell=1}^{m}\cos(\xi_\ell) = 0$ we have

$$\begin{aligned} M_{1,m} &= \frac{1}{m}\sum_{\ell=1}^{m}\frac{(1-\theta(1-\theta)\nu^2\psi^2(\xi_\ell))\cos(\xi_\ell) - \nu\psi(\xi_\ell)\sin(\xi_\ell)}{1+\theta^2\nu^2\psi^2(\xi_\ell)} \\ &= \left(\frac{1}{m}\sum_{\ell=1}^{m}\frac{(1-\theta(1-\theta)\nu^2\psi^2(\xi_\ell))\cos(\xi_\ell) - \nu\psi(\xi_\ell)\sin(\xi_\ell)}{1+\theta^2\nu^2\psi^2(\xi_\ell)}\right) - \frac{1}{m}\sum_{\ell=1}^{m}\cos(\xi_\ell) \\ &= -\frac{\nu}{m}\sum_{\ell=1}^{m}\frac{\psi(\xi_\ell)\sin(\xi_\ell) + \theta\nu\psi^2(\xi_\ell)\cos(\xi_\ell)}{1+\theta^2\nu^2\psi^2(\xi_\ell)} \\ &= -\frac{2\nu}{m}\sum_{\ell=2}^{m/2}\frac{\psi(\xi_\ell)(\sin(\xi_\ell) + \theta\nu\psi(\xi_\ell)\cos(\xi_\ell))}{1+\theta^2\nu^2\psi^2(\xi_\ell)}, \end{aligned} \tag{55}$$

where in the last equality we used $\psi(\xi_1) = \psi(0) = 0$, $\psi(\xi_{m/2+1}) = 0$, and the symmetry of angles ξ_ℓ $(1 \le \ell \le m)$ about the x-axis when m is even (explicitly, the identities $\sin\left(k\frac{2\pi(\ell-1)}{m}\right) = -\sin\left(k\frac{2\pi(m-\ell+1)}{m}\right)$ and $\cos\left(k\frac{2\pi(\ell-1)}{m}\right) = \cos\left(k\frac{2\pi(m-\ell+1)}{m}\right)$ for positive integers k, ℓ and m). Define the function

$$f(x;\theta,\nu) := \frac{\psi(x)(\sin(x) + \theta\nu\psi(x)\cos(x))}{1+\theta^2\nu^2\psi^2(x)}.$$

Then, we can express (55) by summing only over indices ℓ for which $0 < \xi_\ell < \pi/2$ (and separating the case $\xi_\ell = \pi/2$ when m is divisible by 4), yielding

$$M_{1,m} = \begin{cases} \dfrac{-2\nu\psi(\pi/2)}{m\,(1+\theta^2\nu^2\psi^2(\pi/2))} - \dfrac{2\nu}{m}\displaystyle\sum_{\ell=2}^{m/4}\Big(f(\xi_\ell;\theta,\nu) + f(\pi-\xi_\ell;\theta,\nu)\Big), & \text{if } m \equiv 0 \pmod 4, \\ -\dfrac{2\nu}{m}\displaystyle\sum_{\ell=2}^{(m+2)/4}\Big(f(\xi_\ell;\theta,\nu) + f(\pi-\xi_\ell;\theta,\nu)\Big), & \text{if } m \equiv 2 \pmod 4. \end{cases} \tag{56}$$

We will also use the identity

$$f(\xi_\ell;\theta,\nu) + f(\pi-\xi_\ell;\theta,\nu) = \frac{\Big(\psi(\xi_\ell) + \psi(\pi-\xi_\ell)\Big)\Big(\sin(\xi_\ell) + \theta\nu\psi(\xi_\ell)\cos(\xi_\ell) + \theta\nu\psi(\pi-\xi_\ell)\big(\theta\nu\psi(\xi_\ell)\sin(\xi_\ell) - \cos(\xi_\ell)\big)\Big)}{\big(1+\theta^2\nu^2\psi^2(\xi_\ell)\big)\big(1+\theta^2\nu^2\psi^2(\pi-\xi_\ell)\big)}.$$

First, observe that $\sin(\xi_\ell)$ and $\cos(\xi_\ell)$ are positive for each index $2 \le \ell \le (m+2)/4$ in (56). Now for the 4^{th}-order centered spatial discretization, easy calculations show that $\psi(\pi/2) = 4/3$,

$$\psi(\xi_\ell) = \frac{1}{3}\sin(\xi_\ell)(4-\cos(\xi_\ell)), \quad \text{and} \quad \psi(\pi-\xi_\ell) = \frac{1}{3}\sin(\xi_\ell)(4+\cos(\xi_\ell)),$$

so they are all positive as well. Let us fix $\theta \in (0,1]$ arbitrarily, and notice that for each ℓ we can find $\nu_\ell > 0$ such that $\theta\nu\psi(\xi_\ell)\sin(\xi_\ell) - \cos(\xi_\ell) > 0$ for $\nu > \nu_\ell$. Let $\nu_0 := \max\nu_\ell$, then $f(\xi_\ell;\theta,\nu) + f(\pi-\xi_\ell;\theta,\nu) > 0$ for $\nu > \nu_0$. Therefore, $M_{1,m} < 0$ for any $\theta \in (0,1]$ and $\nu > \nu_0$.

□

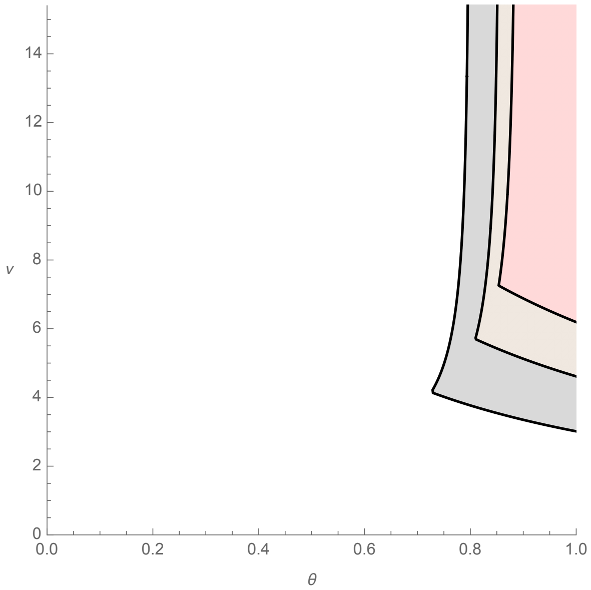

Figure 5: Non-negativity of the matrix family generated by the 4^{th}-order centered discretization in space and the θ-method in time; see Section 4.1. The parameter regions in the (θ, ν) parameter plane ensuring $M(m, \theta, \nu) \geq 0$ are highlighted for $m \in \{5, 7, 9\}$ (different values of m are represented by different colors; the regions "shrink" as m gets larger). For $m = 5$, $m = 7$, and $m = 9$, the "lower left corner" point of the region has $\theta \approx 0.726106$, $\theta \approx 0.809401$, and $\theta \approx 0.853562$, respectively.

Remark 15. *It is easily seen that the previous proof boils down to the fact that for the 4^{th}-order centered spatial discretization we have $\psi(x) > 0$ for $x \in (0, \pi)$, where*

$$\psi(x) = 2\left(\frac{2}{3}\sin(x) - \frac{1}{12}\sin(2x)\right) = \frac{1}{3}\sin(x)\big(4 - \cos(x)\big).$$

We know from (53) *that for the 6^{th}-order scheme*

$$\psi(x) = 2\left(\frac{3}{4}\sin(x) - \frac{3}{20}\sin(2x) + \frac{1}{60}\sin(3x)\right) = \frac{1}{15}\sin(x)\big(23 - 9\cos(x) + \cos(2x)\big),$$

so we again have $\psi(x) > 0$ for $x \in (0, \pi)$. Therefore, the analogue of Proposition 1 is true for the 6^{th}-order scheme as well (c.f. the proof of Proposition 2).

Remark 16. *Based on the observations above Proposition 1, we conjecture the following: given an arbitrary finite difference centered discretization in space coupled with the θ-method, then $M_{1,m} < 0$ for even m, and for* all *values $\nu > 0$ and $\theta \in (0, 1]$—that is, $\nu_0 = 0$ can be chosen in general.*

Case 2: m is odd. In this case we have found that positivity preservation is possible for a suitable set of $\theta \in (0, 1]$ and $\nu > 0$ values; see Figures 5 and 6.

Also, if we assume $\psi(\xi_\ell) \neq 0$ for each $2 \leq \ell \leq m$, then we can extend the asymptotic results of Section 2.1 for an arbitrary high-order centered discretization (c.f. the "odd m" case in Section 4.2).

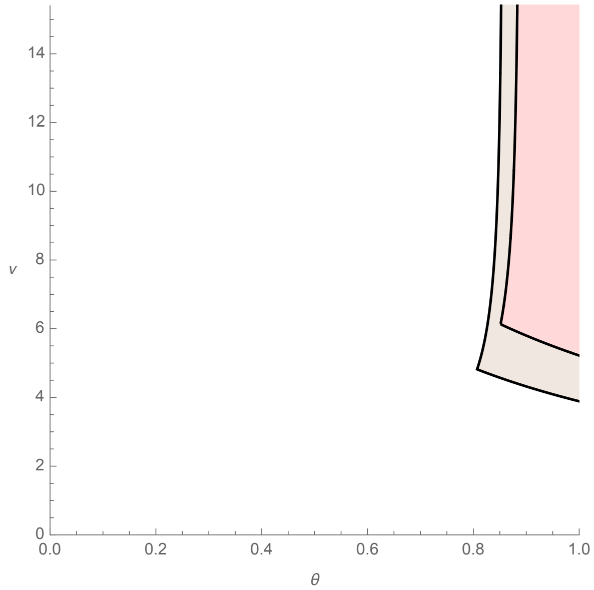

Figure 6: Non-negativity of the matrix family generated by the 6^{th}-order centered discretization in space and the θ-method in time; see Section 4.1. The parameter regions in the (θ, ν) parameter plane ensuring $M(m, \theta, \nu) \geq 0$ are highlighted for $m \in \{7, 9\}$ (different values of m are represented by different colors; the regions "shrink" as m gets larger). For $m = 7$ and $m = 9$, the "lower left corner" point of the region has $\theta \approx 0.807042$ and $\theta \approx 0.851437$, respectively.

4.2 Fourier spectral collocation in space, θ-method in time

Here, we consider the spectral method that results from extending the finite difference stencil to include the whole spatial grid. In this section, we assume $m \geq 4$. The resulting spatial semi-discretization can again be written in the form (3) where the matrix L takes the form [10, 8]

$$L_{i,j} = \begin{cases} 0 & \text{if } i = j, \\ \dfrac{\pi}{m}(-1)^{i+j} \cot\left(\frac{(i-j)\pi}{m}\right) & \text{if } i \neq j, \end{cases}$$

for even m, and

$$L_{i,j} = \begin{cases} 0 & \text{if } i = j, \\ \dfrac{\pi}{m}(-1)^{i+j} \csc\left(\frac{(i-j)\pi}{m}\right) & \text{if } i \neq j, \end{cases}$$

for odd m. As in Section 2, the spectral collocation matrices have an eigendecomposition given by (7), with eigenvectors (8), but now the entries of the diagonal matrix Λ are

$$\imath\lambda_\ell = \begin{cases} \imath\xi_\ell & 1 \leq \ell < \frac{m}{2} + 1, \\ 0 & \ell = \frac{m}{2} + 1 \text{ and } m \text{ is even}, \\ \imath(\xi_\ell - 2\pi) & \frac{m}{2} + 1 < \ell \leq m, \end{cases} \tag{57}$$

where ξ_ℓ has been defined in (9).

When this matrix L is coupled with the θ-method as time discretization, the full discretization matrix M in (11) is obtained. The entries of the first row of the circulant matrix M are again given by (13), and now they take the form

$$M_{1,j} = \frac{1}{m}\sum_{\ell=1}^{m} \frac{(1-\theta(1-\theta)\nu^2\lambda_\ell^2)\cos((j-1)\xi_\ell) + \nu\lambda_\ell \sin((j-1)\xi_\ell)}{1+\theta^2\nu^2\lambda_\ell^2} \quad (1 \le j \le m). \tag{58}$$

Similarly to the Sections 2 and 4.1, we distinguish between even and odd sizes of the discretization matrices M.

Case 1: m is even. From (58) we have (also using (19)) that

$$M_{1,m} = \frac{1}{m}\sum_{\ell=1}^{m} \frac{(1-\theta(1-\theta)\nu^2\lambda_\ell^2)\cos(\xi_\ell) - \nu\lambda_\ell \sin(\xi_\ell)}{1+\theta^2\nu^2\lambda_\ell^2}. \tag{59}$$

The following proposition proves that (59) is negative for sufficiently large CFL number ν.

Proposition 2. *Consider the iterative formula* (10) *applied to the advection equation* (1) *with periodic boundary condition. Let the matrix $M_{m\times m}$ result from a given even spactral collocation method in space (with m spatial grid points) and the θ-method in time. Let also ν be the CFL number defined in* (12)*. If $m \ge 4$ is* even*, then there exists $\nu_0 > 0$ such that the matrix M has at least one negative entry for any $\nu > \nu_0$.*

Proof. The proof is similar to that of Proposition 1: we show that $M_{1,m} < 0$ for all $\theta \in (0,1]$ and $\nu > \nu_0$, where $\nu_0 > 0$ depends on m and θ.

First, observe that by using $\frac{1}{m}\sum_{\ell=1}^{m}\cos(\xi_\ell) = 0$ we can rewrite (59) as

$$M_{1,m} = -\frac{\nu}{m}\sum_{\ell=1}^{m} \frac{\lambda_\ell \sin(\xi_\ell) + \theta\nu\lambda_\ell^2\cos(\xi_\ell)}{1+\theta^2\nu^2\lambda_\ell^2}.$$

Now, since $\lambda_1 = \lambda_{m/2+1} = 0$, we have

$$M_{1,m} = -\frac{\nu}{m}\sum_{\ell=2}^{m/2} \left(\frac{\lambda_\ell \sin(\xi_\ell) + \theta\nu\lambda_\ell^2\cos(\xi_\ell)}{1+\theta^2\nu^2\lambda_\ell^2}\right) - \frac{\nu}{m}\sum_{\ell=m/2+2}^{m} \left(\frac{\lambda_\ell \sin(\xi_\ell) + \theta\nu\lambda_\ell^2\cos(\xi_\ell)}{1+\theta^2\nu^2\lambda_\ell^2}\right).$$

One easily checks from (57) and (9) that for any $\frac{m}{2}+2 \le \ell \le m$

$$\lambda_\ell = \xi_\ell - 2\pi = -\xi_{m+2-\ell} = -\lambda_{m+2-\ell}.$$

This implies that

$$\sum_{\ell=2}^{m/2} \frac{\lambda_\ell \sin(\xi_\ell) + \theta\nu\lambda_\ell^2\cos(\xi_\ell)}{1+\theta^2\nu^2\lambda_\ell^2} = \sum_{\ell=m/2+2}^{m} \frac{\lambda_\ell \sin(\xi_\ell) + \theta\nu\lambda_\ell^2\cos(\xi_\ell)}{1+\theta^2\nu^2\lambda_\ell^2},$$

so

$$M_{1,m} = -\frac{2\nu}{m}\sum_{\ell=2}^{m/2} \frac{\lambda_\ell \sin(\xi_\ell) + \theta\nu\lambda_\ell^2\cos(\xi_\ell)}{1+\theta^2\nu^2\lambda_\ell^2}.$$

Now notice (due to $\lambda_\ell = \xi_\ell$ for $2 \le \ell \le \frac{m}{2}$) that we also have

$$M_{1,m} = -\frac{2\nu}{m}\sum_{\ell=2}^{m/2} \frac{\xi_\ell \sin(\xi_\ell) + \theta\nu\xi_\ell^2\cos(\xi_\ell)}{1+\theta^2\nu^2\xi_\ell^2},$$

thus

$$M_{1,m} = -\frac{2\nu}{m}\sum_{\ell=2}^{m/2}\frac{\psi(\xi_\ell)(\sin(\xi_\ell)+\theta\nu\psi(\xi_\ell)\cos(\xi_\ell))}{1+\theta^2\nu^2\psi^2(\xi_\ell)}$$

with $\psi(x) := x$. Since $\psi(x) > 0$ for $x \in (0, \pi)$, according to Remark 15, the proof is complete. □

As previously, we conjecture that $M_{1,m} < 0$ for *all* values of $\nu > 0$ and $\theta \in (0, 1]$. We have been able to verify this for $m \in \{4, 6, 8, 10\}$, as follows. The expression $M_{1,m}$ is a rational function in ν and θ, whose denominator is positive. By introducing a new variable $y := \pi\theta\nu \geq 0$ and dividing by $\nu > 0$, we can write the numerator as a univariate polynomial $p_m(y)$. For example,

$$p_8(y) = -3\left(8+3\sqrt{2}\right)y^4 + 64y^3 - 32\left(7+5\sqrt{2}\right)y^2 + 256y - 256\left(2+\sqrt{2}\right).$$

We have confirmed with symbolic calculations that $p_m(y) < 0$ for all $y \geq 0$ in the cases $m \in \{4, 6, 8, 10\}$.

Remark 17. *When generating the matrix M for the symbolic calculations for larger values of m for the actual full discretization, it is of course computationally more efficient to use the formula $\mathcal{F}R(\nu\Lambda)\mathcal{F}^*$ in* (11) *instead of $R(\nu L)$ (because in the latter form one would need to evaluate the inverse of a non-sparse matrix).*

Case 2: m is odd. This time we find a behavior similar to that observed in Section 2.1; see Figure 7. Moreover, since this time $\lambda_\ell \neq 0$ for $2 \leq \ell \leq m$, the same asymptotic results hold as in the case of the 2^{nd}-order finite difference scheme in Section 2.1. Indeed, for odd $m = 2k + 1 \geq 5$ we have that

$$\begin{aligned} M_{1,1} &= \frac{1}{m}\left(1+\sum_{\ell=2}^{m}\frac{1-\theta(1-\theta)\nu^2\lambda_\ell^2}{1+\theta^2\nu^2\lambda_\ell^2}\right)\\ M_{1,j} &= \frac{1}{m}\left(1+\sum_{\ell=2}^{m}\frac{(1-\theta(1-\theta)\nu^2\lambda_\ell^2)\cos((j-1)\xi_\ell)+\nu\lambda_\ell\sin((j-1)\xi_\ell)}{1+\theta^2\nu^2\lambda_\ell^2}\right) \quad (j \geq 2). \end{aligned} \tag{60}$$

Taking $\nu \to +\infty$ and $\theta \in (0, 1]$ fixed in (60), yields

$$\begin{aligned} M_{1,1}^\infty &:= \lim_{\nu\to+\infty} M_{1,1} = 1 - \frac{m-1}{m\theta}\\ M_{1,j}^\infty &:= \lim_{\nu\to+\infty} M_{1,j} = \frac{1}{m\theta} \quad (j \geq 2). \end{aligned}$$

As a result, we conclude that

$$M_{1,j}^\infty > 0 \text{ for all } 2 \leq j \leq m \text{ and } \theta \in (0, 1],$$

while

$$M_{1,1}^\infty > 0 \iff \theta > \frac{m-1}{m}.$$

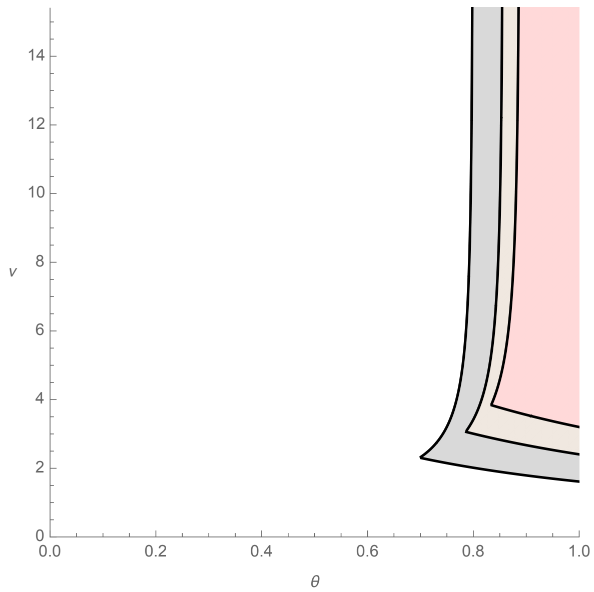

Figure 7: Non-negativity of the matrix family generated by the Fourier spectral collocation method in space and the θ-method in time; see Section 4.2. The parameter regions in the (θ, ν) parameter plane ensuring $M(m, \theta, \nu) \geq 0$ are highlighted for $m \in \{5, 7, 9\}$ (different values of m are represented by different colors; the regions "shrink" as m gets larger).

5 Conclusions

In this work, we have studied the positivity preservation of certain fundamental discretizations of the advection equation with periodic boundary condition (1). Our detailed investigations in Sections 2–3 were devoted to the full discretization obtained by coupling the second-order centered differences in space with the θ-method in time. Rather than using SSP theory [5], we have employed a direct approach, first based on discrete Fourier analysis and then on a polynomial representation of the entries of the full discretization matrix. The characterization of the matrix entries, along with the related trigonometric identities presented in Corollary 1, may be of independent interest. In Section 4, we considered higher-order centered differences or Fourier spectral collocation in space, and again the θ-method in time.

For all full discretizations constructed this way, we have found similar behavior. If the number of spatial grid points m is even, no method is positivity preserving, while if m is odd, some methods may be positivity preserving. Positivity is generally enhanced by taking larger values of the CFL number $\nu > 0$, larger values of the time-discretization parameter $\theta \in [0, 1]$, or smaller values of the spatial grid points $m \in \mathbb{N}^+$. These tendencies, and more specific results, are described in Theorem 2, and can be seen in Figures 3, 5, 6, and 7. Our positive results about the full discretizations are perhaps unexpected, since neither of the underlying spatial semi-discretizations preserves positivity.

Although some of the spatial discretizations considered above have high order, the θ-method as time discretization typically has order only 1 (order 2 occurs only for $\theta = 1/2$). Therefore, we emphasize that our goal in this work is not to provide efficient discretizations but rather to understand the behavior of these simple building blocks, as a means of gaining insight and understanding the positivity of more complicated discretizations that may not be amenable to a thorough analysis.

There are several possible future directions for research building on this work. Other finite difference spatial discretizations could be studied using similar techniques, and higher-order one-step time discretizations could easily be incorporated via (11). Similarly, finite difference discretizations of other linear partial differential equations could be analyzed with the same techniques. Further areas for extension might include other boundary conditions or multidimensional problems.

References

[1] BERMAN, A., AND PLEMMONS, R. J. *Nonnegative matrices in the mathematical sciences*, vol. 9 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994. Revised reprint of the 1979 original.

[2] BERNSTEIN, D. S. *Matrix mathematics*, second ed. Princeton University Press, Princeton, NJ, 2009. Theory, facts, and formulas.

[3] FEKETE, I., KETCHESON, D. I., AND LÓCZI, L. Positivity for convective semi-discretizations. *J. Sci. Comput. 74*, 1 (2018), 244–266.

[4] FORNBERG, B. Generation of finite difference formulas on arbitrarily spaced grids. *Math. Comp. 51*, 184 (1988), 699–706.

[5] GOTTLIEB, S., KETCHESON, D., AND SHU, C.-W. *Strong stability preserving Runge-Kutta and multistep time discretizations*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2011.

[6] HAIRER, E., AND WANNER, G. *Solving ordinary differential equations. II*, vol. 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2010. Stiff and differential-algebraic problems, Second revised edition.

[7] HUNDSDORFER, W., AND VERWER, J. *Numerical solution of time-dependent advection-diffusion-reaction equations*, vol. 33 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2003.

[8] PEYRET, R. *Spectral methods for incompressible viscous flow*, vol. 148 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2002.

[9] ROJO, O., AND SOTO, R. L. Existence and construction of nonnegative matrices with complex spectrum. *Linear Algebra Appl. 368* (2003), 53–69.

[10] TREFETHEN, L. N. *Spectral methods in MATLAB*, vol. 10 of *Software, Environments, and Tools*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.