Computers and Mathematics with Applications 109 (2022) 191–203

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

www.elsevier.com/locate/camwa

Stable rotational symmetric schemes for nonlinear reaction-diffusion equations

Philku Lee [a,b,*], George V. Popescu [a,c], Seongjai Kim [d]

[a] Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, MS 39762 USA
[b] Korea Automotive Technology Institute, Bussiness growth support center 218(216), Changeop-ro 42, Sujeong-gu, Seongnam-si, Gyeonggi-do 13449, Republic of Korea
[c] The National Institute for Laser, Plasma & Radiation Physics, 077126 Măgurele, Ilfov, Romania
[d] Department of Mathematics and Statistics, Mississippi State University, Mississippi State, MS 39762 USA

ARTICLE INFO

Keywords:
Reaction-diffusion equations
Biological pattern formation
Time-stepping procedure
Variable-θ method
Rotational symmetry
Averaging scheme

ABSTRACT

In the simulation of biological pattern forming, it has been observed that the numerical solution is more sensitive to the spatial mesh resolution than the temporal one. Such a higher sensitivity to the spatial resolution is mainly originated from an inaccurate approximation of diffusion differential operators, which might violate the rotational symmetry to be seriously erroneous in low spatial resolutions. Also, it has been known that the second-order Crank-Nicolson time-stepping procedure may introduce spurious oscillations when the initial data or the source term is nonsmooth and the temporal step size is set relatively large. This article studies 9-point finite difference schemes for the diffusion operator to enhance the rotational symmetry, employs the variable-θ method to achieve a nonoscillatory second-order time-stepping procedure, and adopts an effective relaxation linear solver to solve the algebraic systems efficiently. The variable-θ method is proved to satisfy the maximum principle, which guarantees that the time-stepping procedure is unconditionally stable. When the successive over-relaxation method with an optimal relaxation parameter is adopted for the algebraic solver, the iteration converges in 2-4 iterations in most time steps. The overall algorithm is second-order in accuracy and scalable in efficiency. Various examples are given to show the accuracy and efficiency of the proposed algorithm for the numerical solution of the system of nonlinear reaction-diffusion equations.

1. Introduction

Let Ω be a connected, bounded open domain in $\mathbb{R}^d$ with a piecewise smooth boundary $\Gamma = \partial\Omega$. Let $J = (0, T]$ for $T > 0$. Consider the following system of *reaction-diffusion* (RD) equations for $\mathbf{u} = [u_1, u_2, \dots, u_m]^T$, $m \geq 2$:

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} - D\Delta\mathbf{u} &= \mathbf{f}(\mathbf{u}), && \Omega \times J, \\ \frac{\partial \mathbf{u}}{\partial \nu} &= 0, && \Gamma \times J, \\ \mathbf{u}(\mathbf{x}, 0) &= \mathbf{u}^0, && \Omega \times \{t = 0\}, \end{aligned} \tag{1.1}$$

where u_l are real-valued functions, $D = \mathrm{diag}[D_1, D_2, \dots, D_m]$ is the diffusion tensor whose elements D_l's are strictly positive constants, Δ denotes the Laplace operator, $\partial/\partial\nu$ is the outward normal derivative on the boundary Γ, and $\mathbf{f}(\mathbf{u})$ is the reaction kinetics of the system given as

$$\mathbf{f}(\mathbf{u}) = [f_1(\mathbf{u}), f_2(\mathbf{u}), \dots, f_m(\mathbf{u})]^T, \tag{1.2}$$

which is often nonlinear with respect to $\mathbf{u}$. In this article, we consider the RD systems with $d = 2$ to investigate accuracy issues, concerning approximations of spatial derivatives in high dimensions.

Since Turing [19] proposed an RD problem to explain biological pattern formation in 1952, there have been many efforts to solve RD problems numerically. In [21], Zegeling and Kok proposed an adaptive moving mesh method and its application to RD models, and Madzvamuse [14] integrated it with a special form of linearization of the reaction terms and a second-order semi-implicit backward differentiation formula. McCourt *et el.* [15] provided numerical results for an RD problem (the Geirer-Meinhardt model) by employing a high-order spectral collocation method. Then the spectral method was collaborated with finite volume technique by Shakeri and Dehghan [17]. Recently, Fernandes and Fairweather applied the orthogonal spline collocation method to solve RD problems and also introduced a time-stepping procedure integrated with alternating direction implicit (ADI) methods; see [3–5].

* Corresponding author.
E-mail addresses: pklee@katech.re.kr (P. Lee), popescu@igbb.msstate.edu (G.V. Popescu), skim@math.msstate.edu (S. Kim).

https://doi.org/10.1016/j.camwa.2022.01.026
Received 30 June 2021; Received in revised form 10 January 2022; Accepted 19 January 2022
Available online 28 January 2022

In 2020, the authors [10] introduced a nonoscillatory second-order time-stepping procedure called the *variable-θ method*, as a perturbation of the *Crank-Nicolson* (CN) method, for the numerical solution of parabolic problems of nonsmooth data. Then, in [11], we performed a sensitivity analysis for the numerical solution of one-dimensional (1D) biological pattern formation problems and concluded that the accuracy of the numerical solution might be much more sensitive to the spatial mesh resolution than the temporal one. Also, it was experimentally verified that the sensitivity to the spatial resolution might introduce undesirable numerical solutions in low spatial resolutions and deteriorate biological patterns.

For 1D cases, this sensitivity issue can be well-explained via grid effect. That is, the accuracy of the numerical solution degenerates when the spatial spacing becomes large (low resolution) compared with a desired characteristic length of reaction and diffusion. However, in two and higher dimensions, biological patterns can be affected not only by the grid effect but also by whether or not the approximation scheme enforces the rotational invariance. The property of rotational invariance is often translated into rotational symmetry in the *finite difference* (FD) discrete domain [13]. It has been numerically verified that the standard FD approximation of the Laplace diffusion operator may fail to hold rotational symmetry in biological pattern formation.

In this article, to study the effect of rotational symmetry, we perform a sensitivity analysis of two-dimensional (2D) RD problems with various FD approximations for the Laplace operator. In order to investigate the spatial sensitivity issue, we study the averaging scheme $\mathcal{A}_\alpha$ for the approximation of the negative Laplacian $(-\Delta)$, which is defined as an average of the standard 5-point scheme $\mathcal{A}_+$ and the skewed 5-point scheme $\mathcal{A}_\times$:

$$\mathcal{A}_\alpha = \alpha\mathcal{A}_+ + (1-\alpha)\mathcal{A}_\times. \tag{1.3}$$

It has been *numerically* verified that such an averaging scheme can effectively suppress certain deterioration in biological patterns; the averaging scheme lets the numerical solution evolve in desired biological patterns. An effective strategy is considered to optimize the averaging parameter α, which minimizes the leading truncation error of the Laplacian approximation.

In addition to incorporating the averaging scheme, the resulting algorithm for solving the nonlinear RD problem is equipped with an effective extrapolation for the linearization of nonlinear source terms and the variable-θ method for time-stepping, which is effective particularly when a larger time step or a lower spatial mesh resolution is desirable. The variable-θ method is a variant of the CN method ($\theta = 1/2$), in which $\theta = 1$ at grid points where the numerical solution shows a certain portent of oscillations. This article proves that the variable-θ method satisfies the maximum principle *unconditionally*, i.e. for all choices of spatial and temporal grid sizes.

The article is organized as follows. The next section presents a brief review on state-of-the-art FD methods for nonlinear RD systems, and their accuracy issue concerning rotational symmetry is discussed by exemplifying a biological pattern problem in 2D. In Section 3, an averaging scheme for the Laplace operator is suggested to enhance the rotational symmetry of the numerical solution. Then, an effective time-stepping procedure is formulated the averaging scheme and the variable-θ method. Section 4 states and proves the maximum principle for the variable-θ method incorporated with the averaging scheme. Section 5 discusses an optimization procedure for the averaging scheme, which minimizes the leading truncation error of the Laplacian approximation. In Section 6, numerical examples are included to verify the effectiveness of the suggested methods. Section 7 concludes the article summarizing our experiments and findings.

For error analysis, the *accumulated* L^2*-error* $E_2[0,T]$ is measured over the whole time period ($t \in [0,T]$) and the eventual L^∞-error $E_\infty[T]$ is measured at the last moment ($t = T$):

$$E_2[0,T] := \tau \sum_{n=1}^{n_t} \|\mathbf{u}^n - \hat{\mathbf{u}}^n\|_2 \quad \text{and} \quad E_\infty[T] := \|\mathbf{u}^{n_t} - \hat{\mathbf{u}}^{n_t}\|_\infty, \tag{1.4}$$

where $\hat{\mathbf{u}}$ is the exact solution (or a desirable solution),

$$\|\mathbf{u}\|_2 := \Big(h_x h_y \sum_{ij} |u_{ij}|^2\Big)^{1/2}, \quad \text{and} \quad \|\mathbf{u}\|_\infty := \max_{ij} |u_{ij}|.$$

Here h_x and h_y are respectively the x- and y-directional spatial step sizes and τ is the temporal step size.

2. Preliminaries

This section presents a brief review on state-of-the-art FD methods for nonlinear RD systems and certain accuracy issues related to spatial approximation.

2.1. FD schemes for the second-order spatial derivatives

We begin with FD schemes for the negative Laplace operator $-\Delta$. Let Ω be a rectangular domain in $\mathbb{R}^2$: $\Omega = (a_x, b_x) \times (a_y, b_y)$. By partitioning $\Omega \times J$, we obtain the space-time grid points

$$\begin{aligned} &(\mathbf{x}_{ij}, t^n) := (x_i, y_j, t^n); \\ &i = 0, 1, \cdots, n_x, \quad j = 0, 1, \cdots, n_y, \quad n = 0, 1, \cdots, n_t, \end{aligned} \tag{2.1}$$

where n_x, n_y, and n_t are prescribed positive integers and

$$\begin{aligned} &x_i = a_x + i \cdot h_x, \quad y_j = a_y + j \cdot h_y, \quad t^n = n \cdot \tau; \\ &h_x = \frac{b_x - a_x}{n_x}, \quad h_y = \frac{b_y - a_y}{n_y}, \quad \tau = \frac{T}{n_t}. \end{aligned} \tag{2.2}$$

Define the discrete domain, the set of the spatial grid points, by

$$\Omega_d = \{(x_i, y_j) : \ 0 \le i \le n_x, \ \ 0 \le j \le n_y\}, \tag{2.3}$$

and denote the set of boundary grid points by $\Gamma_d = \Omega_d \cap \Gamma$ and the set of interior grid points by $\Omega_d^0 = \Omega_d \setminus \Gamma_d$. Moreover, we define $g_{ij}^n := g(\mathbf{x}_{ij}, t^n)$ for all functions g defined in $(\mathbf{x}, t)$.

For convenience, we assume the uniform grid $h_x = h_y = h$. Then, the Taylor series gives us the following FD approximations for each grid point $\mathbf{x}_{ij} = (x_i, y_j)$. The standard 5-point FD approximation $\mathcal{A}_+$ of $-\Delta$ $(= -\partial_x^2 - \partial_y^2)$ reads

$$\mathcal{A}_+ u_{ij} = \frac{-u_{i-1,j} - u_{i,j+1} + 4u_{ij} - u_{i+1,j} - u_{i,j-1}}{h^2}, \tag{2.4}$$

where the truncation error $\mathcal{E}_{+,ij}$ at the grid point $\mathbf{x}_{ij} = (x_i, y_j)$ becomes

$$\mathcal{E}_{+,ij} = \frac{h^2}{12}(u_{xxxx} + u_{yyyy}) + \mathcal{O}(h^4).$$

On the other hand, applying a 45°-rotated FD approximation to the negative Laplacian operator, we can obtain the following skewed 5-point scheme:

$$\mathcal{A}_\times u_{ij} = \frac{-u_{i-1,j-1} - u_{i-1,j+1} + 4u_{ij} - u_{i+1,j+1} - u_{i+1,j-1}}{2h^2}, \tag{2.5}$$

where the truncation error $\mathcal{E}_{\times,ij}$ is

$$\mathcal{E}_{\times,ij} = \frac{h^2}{12}(u_{xxxx} + 6u_{xxyy} + u_{yyyy}) + \mathcal{O}(h^4).$$

When averaging $\mathcal{A}_+$ and $\mathcal{A}_\times$ with the weight $\alpha = 2/3$, we obtain the 9-point FD approximation of the Laplacian known as the *Mehrstellen* discretization [1]:

$$\begin{aligned} -\Delta u_{ij} \ &= \ \frac{2}{3}\mathcal{A}_+ u_{ij} + \frac{1}{3}\mathcal{A}_\times u_{ij} + \mathcal{E}_{M,ij} \\ &= \ \frac{1}{6h^2}\begin{array}{|c|c|c|}\hline -1 & -4 & -1 \\ \hline -4 & 20 & -4 \\ \hline -1 & -4 & -1 \\ \hline \end{array} u_{ij} + \mathcal{E}_{M,ij}, \end{aligned} \tag{2.6}$$

where the truncation error $\mathcal{E}_{M,ij}$ reads

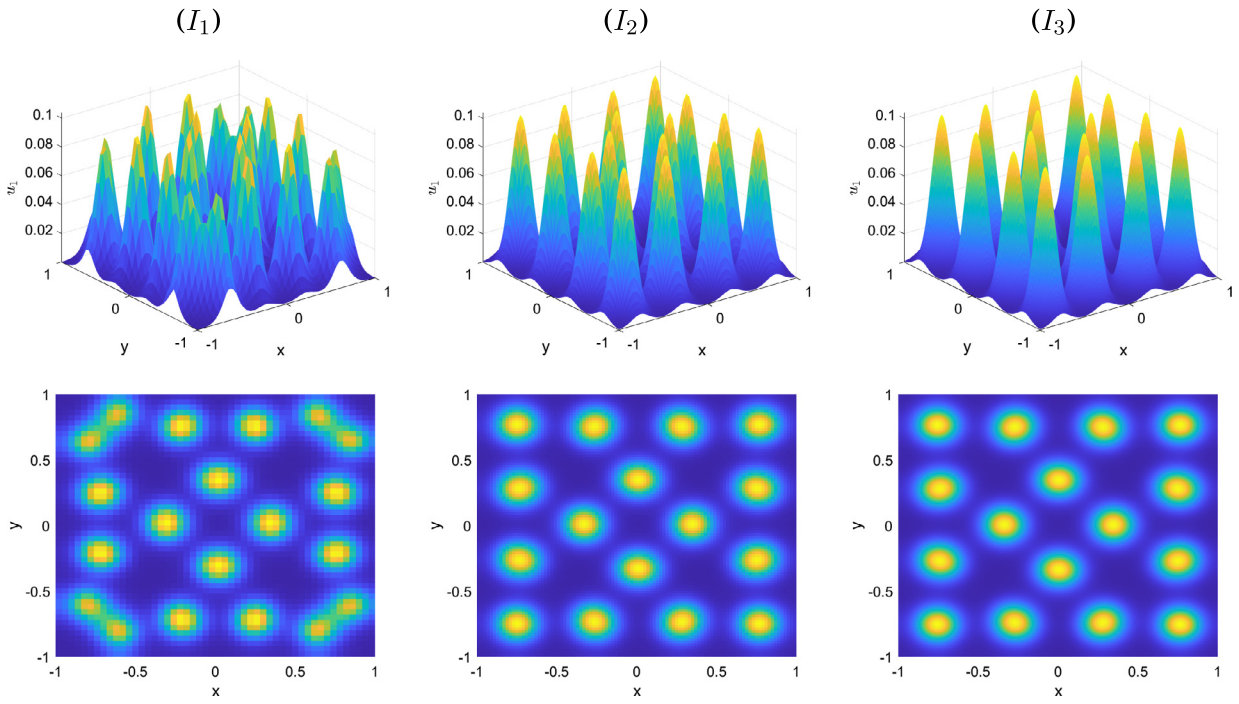

Fig. 1. Numerical solutions for Gierer-Meinhardt model (2.8)–(2.9) of the steady-state ($T = 500$) approximated by the standard 5-point scheme at the fixed time step $\tau = 0.01$ and various spatial resolutions. Each image column I_ℓ represents the numerical solution and its aerial view obtained with the spatial resolution $n_x = n_y = 50 \cdot 2^{\ell-1}$.

$$\mathcal{E}_{M,ij} = \frac{h^2}{12}(u_{xxxx} + 2u_{xxyy} + u_{yyyy}) + \mathcal{O}(h^4).$$

Note that the leading truncation error of the Mehrstellen discretization coincides with a scaled biharmonic operator:

$$\frac{h^2}{12}(u_{xxxx} + 2u_{xxyy} + u_{yyyy}) = \frac{h^2}{12}\Delta^2 u = \frac{h^2}{12}\Delta(\Delta u). \tag{2.7}$$

Thus, for harmonic ($\Delta u = 0$) and biharmonic ($\Delta^2 u = 0$) solutions, the Mehrstellen discretization can achieve the fourth-order truncation error. Moreover, it is known that Mehrstellen discretization can give the best approximation of rotational invariant when it applies to the heat equation; see [13] for details.

2.2. Accuracy issues on the spatial resolution

As pointed out in one of the authors' earlier publications [11], the accuracy of the numerical solution is much more sensitive to the spatial mesh resolution than to the temporal one. This sensitivity phenomenon might be significant in low spatial resolution (of large h), in which the RD process does not have enough time to grow to reach the margins of the spatial mesh. In this case, the RD pattern deteriorates and neither evolves in an appropriate speed nor reaches a condition to replicate itself on time; see [7, §4.2] for similar observations. To investigate this issue, we carried out a sensitivity analysis with the Gierer-Meinhardt model in 2D, applying two different numerical methods: the standard 5-point scheme and the ADI extrapolated *Crank-Nicolson orthogonal spline collocation* (CNOSC) method [4]. The CNOSC method is a state-of-the-art algorithm for the numerical solution of various scalar transient problems [2–5,12].

The Gierer-Meinhardt model [6] is a two-component RD system defined in $\Omega = (-1, 1)^2$ with the following reaction kinetics:

$$D = [\epsilon^2, \kappa/\mu]^T, \quad \mathbf{f}(\mathbf{u}) = \left[\frac{u_1^2}{u_2} - u_1, \frac{1}{\mu}\left(\frac{u_1^2}{\epsilon} - u_2\right)\right]^T. \tag{2.8}$$

We employ the following parameters and initial conditions used in [16]:

$$\begin{aligned} &\epsilon = 0.04, \quad \mu = 0.1, \quad \kappa = 0.0128, \\ &u_1(x, y, 0) = \frac{1}{2}\left[1 + 0.001\sum_{k=1}^{20}\cos\left(\frac{k\pi y}{2}\right)\right]\operatorname{sech}^2\left(\frac{\sqrt{x^2+y^2}}{2\epsilon}\right), \\ &u_2(x, y, 0) = \frac{\cosh\left(1 - \sqrt{x^2+y^2}\right)}{3\cosh(1)}. \end{aligned} \tag{2.9}$$

Fig. 1 and Fig. 2 present the numerical solutions associated with the Gierer-Meinhardt model (2.8)–(2.9) at the steady-state at $T = 500$, varying spatial resolutions, respectively for the standard 5-point scheme and the CNOSC method with $r = 3$. Here, for simplicity, we restrict our attention to the dynamics of u_1 of the model. We set the time step size $\tau = 0.01$ and choose the comparable numbers of spatial grid points for the two methods: $n_x = n_y = 50 \cdot 2^{\ell-1}$, $\ell = 1, 2, 3$. In both figures, one can observe that the numerical solutions in the lowest resolution (Figs. 1(I_1), 2(J_1)) show quite different steady-state patterns from the patterns of the higher resolutions (Figs. 1(I_2, I_3), 2(J_2, J_3)). It should be noticed that the CNOSC method has produced an unreliable steady-state pattern in the low spatial resolution, although it is of fourth-order accuracy in spatial direction. We can see from the example that higher-order spatial schemes may not be advantageous over the second-order scheme, when the spatial resolution is low.

For RD problems in 1D, the sensitivity to the spatial resolution can be explained by the grid effect. For an accurate solution, it requires the spatial spacing to be small (high resolution) compared with a desired characteristic length of physical evolution. Thus, as aforementioned, the RD patterns can be deteriorated mostly by the grid effect, in low spatial resolutions. However, in two and higher dimensions, the RD patterns can be deteriorated by not only the grid effect but also the asymmetry of the approximation schemes. The Laplacian is rotationally invariant in two and higher dimensions; however, its numerical approximation may not guarantee the rotational invariance, depending on the point stencil utilized in the scheme. Effective schemes for the Laplace operator should be designed to enhance the rotational symmetry as much as possible. To enhance the symmetry, we consider the averaging scheme (1.3), with the parameter α being selected adaptively; see Section 5 for details.

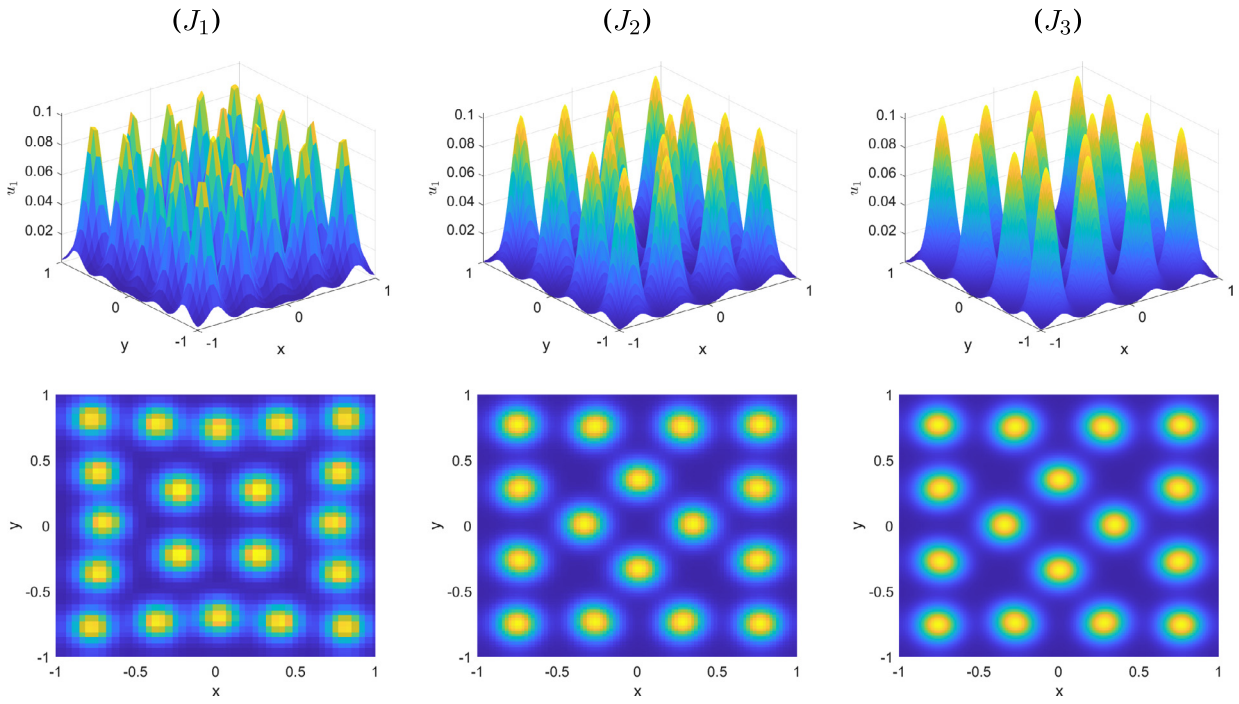

Fig. 2. Numerical solutions for Gierer-Meinhardt model (2.8)–(2.9) of the steady-state ($T = 500$) approximated by the CNOSC method computed with $r = 3$ at the fixed time step $\tau = 0.01$ and various spatial resolutions. Each image column J_ℓ represents the numerical solution and its aerial view obtained with the mesh resolution $n_x = n_y = 50 \cdot 2^{\ell-1}$.

3. The time-stepping procedure

In this section, we introduce an effective time-stepping procedure for the nonlinear RD system (1.1). The resulting algorithm involves an averaging scheme for the Laplace operator to enhance the rotational symmetry, takes the linearization through extrapolation as introduced in [4], employs the variable-θ method in [10] to achieve a nonoscillatory time-stepping procedure, and adopts an effective relaxation linear solver to solve the algebraic systems efficiently.

3.1. The averaging scheme for the Laplace operator

First, we would recall an approximation $\mathcal{A}_\alpha$ for the negative Laplace operator, which is derived from averaging $\mathcal{A}_+$ and $\mathcal{A}_\times$ as follows: for $0 < \alpha < 1$,

$$\mathcal{A}_\alpha u_{ij} = \alpha \mathcal{A}_+ u_{ij} + (1-\alpha)\mathcal{A}_\times u_{ij} = \frac{\alpha}{h^2}\begin{array}{|c|c|c|}\hline & -1 & \\ \hline -1 & 4 & -1 \\ \hline & -1 & \\ \hline\end{array}\, u_{ij} + \frac{1-\alpha}{2h^2}\begin{array}{|c|c|c|}\hline -1 & & -1 \\ \hline & 4 & \\ \hline -1 & & -1 \\ \hline\end{array}\, u_{ij}. \tag{3.1}$$

Then the negative Laplacian at the grid point $\mathbf{x}_{ij}$ can be written as

$$-\Delta u_{ij} = \frac{1}{2h^2}\begin{array}{|c|c|c|}\hline -1+\alpha & -2\alpha & -1+\alpha \\ \hline -2\alpha & 4(1+\alpha) & -2\alpha \\ \hline -1+\alpha & -2\alpha & -1+\alpha \\ \hline\end{array}\, u_{ij} + \frac{h^2}{12}\left[u_{xxxx} + 6(1-\alpha)u_{xxyy} + u_{yyyy}\right] + \mathcal{O}(h^4). \tag{3.2}$$

Here, it is noticeable that the averaging scheme becomes Mehrstellen discretization when $\alpha = 2/3$.

In [8], Jo, Shin, and Suh proposed the above averaging scheme for the numerical solution of the Helmholtz wave equation. They gave an optimized parameter by minimizing the numerical dispersion error of the phase velocity. Their optimal averaging scheme with 5 points per wavelength could achieve the same accuracy as the standard 5-point scheme with 10 points per wavelength. Also one of the authors introduced a fourth-order 9-point compact scheme for the Helmholtz wave equation by employing the method of modified equations [9]. However, for nonlinear RD systems, it is difficult to derive mathematical formulas for such optimal averaging (or fourth-order compact) schemes. We will discuss an effective numerical strategy in Section 5 for an optimal averaging parameter for the nonlinear RD system (1.1).

3.2. The extrapolated relaxation algorithm

Let $\mathbf{u}^n$ be the numerical solution at the n-th time level, $n \geq 0$. For the numerical solution in the $(n+1)$-th level, we first extrapolate numerical solutions in the two previous levels to approximate the solution at an intermediate point $t^{n+\theta} = (1+\theta)t^n - \theta t^{n+1}$:

$$\widetilde{\mathbf{u}}^{n+\theta} := (1+\theta)\mathbf{u}^n - \theta \mathbf{u}^{n-1}, \tag{3.3}$$

where $u^{-1} = u_0$. For example, for $\theta = 1/2$,

$$\widetilde{\mathbf{u}}^{n+1/2} := \frac{3\mathbf{u}^n - \mathbf{u}^{n-1}}{2}. \tag{3.4}$$

See [4] for details of second-order extrapolation schemes.

Recall the averaging scheme:

$$\mathcal{A}_a u_{ij} = \frac{1}{2h^2}\begin{array}{|c|c|c|}\hline -1+\alpha & -2\alpha & -1+\alpha \\ \hline -2\alpha & 4(1+\alpha) & -2\alpha \\ \hline -1+\alpha & -2\alpha & -1+\alpha \\ \hline\end{array}\, u_{ij} \approx -\Delta u_{ij}. \tag{3.5}$$

Then the time-stepping procedure for the system of RD equations (1.1), incorporating the linearization through extrapolation (3.3) and the proposed averaging scheme $\mathcal{A}_\alpha$, simply reads:

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau} + D\mathcal{A}_\alpha[\theta \mathbf{u}^{n+1} + (I-\theta)\mathbf{u}^n] = \mathbf{f}(\widetilde{\mathbf{u}}^{n+\theta}), \tag{3.6}$$

where $\mathbf{u}^n = [u_1^n, u_2^n, \ldots, u_m^n]^T$, $D = \mathrm{diag}[D_1, D_2, \ldots, D_m]$, and $\theta = \mathrm{diag}[\theta_1, \theta_2, \ldots, \theta_m]$.

The linearized problem (3.6) can be resolved by solving for m separate components: $u_1^{n+1}, u_2^{n+1}, \dots, u_m^{n+1}$. Each component in (3.6) is formulated as follows:

$$\frac{u_{ij}^{n+1} - u_{ij}^n}{\tau} + D\mathcal{A}_\alpha[\theta u_{ij}^{n+1} + (1-\theta)u_{ij}^n] = f(\widetilde{\mathbf{u}}^{n+\theta})_{ij}, \tag{3.7}$$

where u, D, θ, and f denote respectively u_l, D_l, θ_l, and f_l for $l = 1, 2, \dots, m$. We rewrite (3.7) in a vector form as

$$(I + \theta\tau D\mathcal{A}_\alpha)u^{n+1} = [I - (1-\theta)\tau D\mathcal{A}_\alpha]u^n + \tau f(\widetilde{\mathbf{u}}^{n+\theta}). \tag{3.8}$$

Define

$$\begin{aligned} \mathcal{B}_\alpha &= I + \theta\tau D\mathcal{A}_\alpha, \\ \mathcal{R}_\alpha &= I - (1-\theta)\tau D\mathcal{A}_\alpha, \\ \mathbf{r}^n &= \mathcal{R}_\alpha u^n + \tau f(\widetilde{\mathbf{u}}^{n+\theta}). \end{aligned} \tag{3.9}$$

Then (3.8) reads

$$\mathcal{B}_\alpha u^{n+1} = \mathbf{r}^n \tag{3.10}$$

It is often the case that relaxation methods solving an algebraic system begin with a regular splitting of the coefficient matrix $\mathcal{B}_\alpha$. Given an initialization

$$u^{n+1,0} = 2u^n - u^{n-1}, \tag{3.11}$$

and a regular splitting

$$\mathcal{B}_\alpha = \mathcal{M}_\alpha - \mathcal{N}_\alpha, \tag{3.12}$$

a relaxation algorithm for (3.10) can be formulated as

$$\begin{array}{l} \text{for } k = 1, 2, \cdots \\ \quad \text{for } \mathbf{x}_{ij} \in \Omega_d^0 \\ \qquad u_{ij}^{n+1,k} = \left[u^{n+1,k-1} + \mathcal{M}_\alpha^{-1}(\mathbf{r}^n - \mathcal{B}_\alpha u^{n+1,k-1})\right]_{ij}; \qquad (\mathcal{RA}) \\ \quad \text{end} \\ \text{end} \end{array} \tag{3.13}$$

Remark 3.1. It has been well known [20] that if $\mathcal{B} = \mathcal{M} - \mathcal{N}$ is a regular splitting and $\mathcal{B}^{-1} \geq 0$, then the spectral radius of the iteration matrix $(\mathcal{M}^{-1}\mathcal{N} = I - \mathcal{M}^{-1}\mathcal{B})$ is strictly less than 1. That is,

$$\rho(\mathcal{M}^{-1}\mathcal{N}) = \frac{\rho(\mathcal{B}^{-1}\mathcal{N})}{1 + \rho(\mathcal{B}^{-1}\mathcal{N})} < 1. \tag{3.14}$$

Thus relaxation methods of regular splittings (such as the Jacobi, the Gauss-Seidel (GS), and the successive over-relaxation (SOR) iterations) are all convergent. In this article, we will utilize the SOR with an optimal relaxation parameter, because of the following three reasons. It is simple to implement, not difficult to find an optimal parameter, and convergent faster than most of modern sophisticated algebraic solvers, particularly for such an evolutionary problem (1.1).

4. Maximum principle for the variable-θ method

In this section, we analyze the maximum principle for the variable-θ method. In the absence of sources and sinks, it is known mathematically and physically that the extreme values of the solution appear either in the initial data or on the boundary. This property is called the *maximum principle*. Once a numerical algorithm satisfies the maximum principle, its numerical solution will never introduce interior local extrema. Thus the maximum principle guarantees the stability of the algorithm.

4.1. The variable-θ method

Prior to proving the maximum principle for the variable-θ method, we briefly describe the variable-θ method presented by the authors in [10]. The method takes the advantage of the CN method ($\theta = 1/2$; a second-order accuracy in time) and the implicit method ($\theta = 1$; the immunity to spurious oscillations). It is well-known that the CN method of a second-order accuracy may introduce spurious oscillations near non-smooth data points. In order to suppress the undesirable oscillations, the authors simply allow the parameter θ to become a variable; $\theta = 1$ in the vicinity of nonsmoothness, while θ remains $1/2$ at other grid points. It is claimed that spurious non-physical oscillations of the CN method arise from its explicit half step.

The *wobble set* is defined as the collection of grid points showing non-physical oscillations so that the grid points would be treated by the implicit method ($\theta = 1$) to resolve the oscillations. In order to determine it, an effective strategy is introduced as follows.

For simplicity, we begin with a linear heat equation in 1D and its θ-method with an appropriate FD approximation $\mathcal{A}$ of $-\partial_{xx}$:

$$\partial_t u - \partial_{xx} u = f, \tag{4.1}$$

$$(I + \theta\tau\mathcal{A})u^{n+1} = [I - (1-\theta)\tau\mathcal{A}]u^n + f^{n+\theta} \tag{4.2}$$

Recall the explicit half step of the CN method and denote it as

$$u^{n+1,*} \equiv \left(I - \frac{\tau}{2}\mathcal{A}\right)u^n. \tag{4.3}$$

Define an index function for local extrema (`idxt`) as

$$\texttt{idxt}(a,b,c) = \begin{cases} 0, & \text{if } \min(a,c) < b < \max(a,c), \\ 1, & \text{if } b = \max(a,c), \\ -1, & \text{if } b = \min(a,c), \\ 2, & \text{if } \max(a,c) < b, \\ -2, & \text{if } b < \min(a,c). \end{cases} \tag{4.4}$$

Then, the wobble set of u^n is defined, to be used for the computation of u^{n+1}, as

$$\begin{aligned} \mathcal{W}_{1D}^n = \Big\{ x_i \in (-1,1) \,|\, & \texttt{idxt}(u_{i-1}^{n+1,*}, u_i^{n+1,*}, u_{i+1}^{n+1,*}) \neq 0 \text{ and} \\ & \left|\texttt{idxt}(u_{i-1}^{n+1,*}, u_i^{n+1,*}, u_{i+1}^{n+1,*}) + \texttt{idxt}(u_{i-1}^n, u_i^n, u_{i+1}^n)\right| < 4 \Big\}, \end{aligned} \tag{4.5}$$

where $u^{n+1,*}$ is the result of the explicit half step of the CN method given in (4.3). Thus the wobble set in 1D is a collection of interior points x_i where $u_i^{n+1,*}$ becomes a local extremum while u_i^n is either a non-extreme value or an extremum in the opposite sense. The wobble set in (4.5) excludes cases where a *strict* extremum in u^n becomes a *strict* extremum in the same sense as in $u^{n+1,*}$; that is,

$$\left|\texttt{idxt}(u_{i-1}^{n+1,*}, u_i^{n+1,*}, u_{i+1}^{n+1,*}) + \texttt{idxt}(u_{i-1}^n, u_i^n, u_{i+1}^n)\right| = 4. \tag{4.6}$$

The above 1D wobble set can be easily expanded to the 2D case by considering the four partial directions as in Fig. 3. Applying the 1D wobble scheme (4.5) to the four partial directions; if at least one of the directional lines wobbles, then we regard the point $\mathbf{x}_{ij}$ as a wobble point. Let P, Q, and R be point indices and define

$$\begin{aligned} &\texttt{iswb}(P,Q,R,n) \\ &= \begin{cases} 1, & \text{if } \texttt{idxt}(u_P^{n+1,*}, u_Q^{n+1,*}, u_R^{n+1,*}) \neq 0 \text{ and} \\ & \quad \left|\texttt{idxt}(u_P^{n+1,*}, u_Q^{n+1,*}, u_R^{n+1,*}) + \texttt{idxt}(u_P^n, u_Q^n, u_R^n)\right| < 4, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{4.7}$$

Then, the wobble set (for the computation of u^{n+1}) is defined as

$$\begin{aligned} \mathcal{W}_{2D}^n = \Big\{ \mathbf{x}_{ij} \in \Omega_d^0 \,|\, & \texttt{iswb}[(i,j-1),(i,j),(i,j+1),n] = 1 \\ & \text{or } \texttt{iswb}[(i-1,j-1),(i,j),(i+1,j+1),n] = 1 \\ & \text{or } \texttt{iswb}[(i-1,j),(i,j),(i+1,j),n] = 1 \\ & \text{or } \texttt{iswb}[(i-1,j+1),(i,j),(i+1,j-1),n] = 1 \Big\}. \end{aligned} \tag{4.8}$$

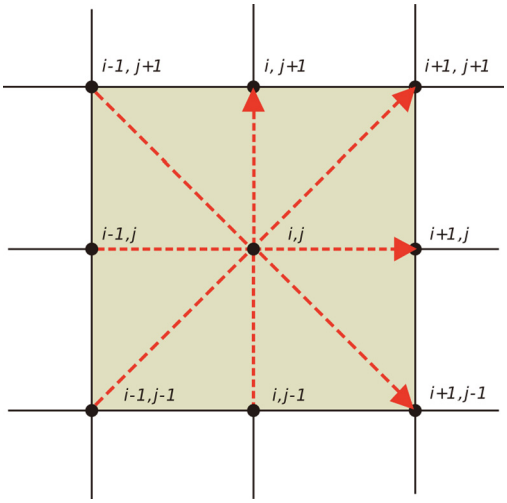

Fig. 3. The eight vicinal points of $\mathbf{x}_{ij}$ and four partial directions.

Once the wobble set is determined, the parameter θ_{ij}^{n+1} for the computation of u^{n+1} can be assigned pointwisely

$$\theta_{ij}^{n+1} := \theta(\mathbf{x}_{ij}, t^{n+1}) = \begin{cases} 1, & \text{if } \mathbf{x}_{ij} \in \mathcal{W}^n, \\ 1/2, & \text{otherwise.} \end{cases} \tag{4.9}$$

Then the *variable-θ method* is formulated as

$$\frac{u_{ij}^{n+1} - u_{ij}^n}{\tau} + \mathcal{A}[\theta_{ij}^{n+1} u_{ij}^{n+1} + (1-\theta_{ij}^{n+1}) u_{ij}^n] = f_{ij}^{n+\theta_{ij}^n}. \tag{4.10}$$

4.2. The maximum principle

Now, we analyze the maximum principle for the variable-θ method.

Theorem 4.1. *The numerical solution of the variable-θ method (4.10) with the standard 5-point scheme $\mathcal{A}_+$ to the heat equation satisfies the maximum principle* unconditionally *(i.e., for all choices of spatial and temporal grid sizes).*

Proof. For simplicity, we first consider the 1D heat equation, without the source term:

$$\partial_t u - \partial_{xx} u = 0. \tag{4.11}$$

Define grid points as in (2.2) and let $\mu = \tau/h^2 > 0$. Then, for fixed $0 \le i \le n_x$, $0 \le n \le n_t$, the variable-θ method with the central spatial scheme for (4.11) can be expressed as

$$\begin{aligned}(1+2\theta_i^{n+1}\mu)u_i^{n+1} = {} & \theta_i^{n+1}\mu(u_{i-1}^{n+1} + u_{i+1}^{n+1}) + (1-\theta_i^{n+1})\mu(u_{i-1}^n + u_{i+1}^n) \\ & + [1 - 2(1-\theta_i^{n+1})\mu]u_i^n,\end{aligned} \tag{4.12}$$

where θ_i^{n+1} is either 1 or 1/2.

Case A: $\theta_i^{n+1} = 1$. In this case, (4.12) becomes

$$(1+2\mu)u_i^{n+1} = \mu(u_{i-1}^{n+1} + u_{i+1}^{n+1}) + u_i^n. \tag{4.13}$$

Since u_i^{n+1} is an average of its neighboring values $\{u_{i-1}^{n+1}, u_{i+1}^{n+1}, u_i^n\}$ with positive weights, u_i^{n+1} can be a local maximum or minimum only if all three neighboring points have the same maximum or minimum value. That is, the implicit time-stepping method ($\theta = 1$) does not introduce strict local extrema to the numerical solution for all $\mu > 0$.

Case B: $\theta_i^{n+1} = 1/2$. In this case, (4.12) reads

$$(1+\mu)u_i^{n+1} = \frac{\mu}{2}(u_{i-1}^{n+1} + u_{i+1}^{n+1}) + u_i^{n+1,*}, \tag{4.14}$$

where

$$u_i^{n+1,*} = \frac{\mu}{2}(u_{i-1}^n + u_{i+1}^n) + (1-\mu)u_i^n. \tag{4.15}$$

If $\theta_i^{n+1} = 1/2$, then $x_i \notin \mathcal{W}_{1D}^n$, where $\mathcal{W}_{1D}^n$ is the wobble set defined in (4.5). Thus we have the following two possible cases: either (4.6) is satisfied or

$$\min\{u_{i-1}^{n+1,*}, u_{i+1}^{n+1,*}\} < u_i^{n+1,*} < \max\{u_{i-1}^{n+1,*}, u_{i+1}^{n+1,*}\}. \tag{4.16}$$

B-1. Assume that (4.6) is satisfied. Suppose that u_i^n is a strict local maximum. Then $u_{i\pm1}^n < u_i^n$ and therefore

$$u_i^{n+1,*} = \frac{\mu}{2}(u_{i-1}^n + u_{i+1}^n) + (1-\mu)u_i^n < \mu u_i^n + (1-\mu)u_i^n = u_i^n. \tag{4.17}$$

Thus, utilizing (4.14), we have

$$u_i^{n+1} \le \max\{u_{i-1}^{n+1}, u_{i+1}^{n+1}, u_i^{n+1,*}\} < \max\{u_{i-1}^{n+1}, u_{i+1}^{n+1}, u_i^n\}. \tag{4.18}$$

Suppose that u_i^n is a strict local minimum. Then, using the same arguments, we have

$$u_i^{n+1} \ge \min\{u_{i-1}^{n+1}, u_{i+1}^{n+1}, u_i^{n+1,*}\} > \min\{u_{i-1}^{n+1}, u_{i+1}^{n+1}, u_i^n\}. \tag{4.19}$$

Now, one should notice that u_i^{n+1} is obtained by the implicit half step of the CN method (4.14), which does not introduce strict local extrema as shown in Case A. Thus, for both (4.18) and (4.19), u_i^{n+1} cannot have a value outside the interval

$$[\min\{u_{i-1}^{n+1}, u_{i+1}^{n+1}, u_i^{n+1,*}\}, \max\{u_{i-1}^{n+1}, u_{i+1}^{n+1}, u_i^{n+1,*}\}].$$

Thus, it follows from the second inequalities of (4.18) and (4.19) that u_i^{n+1} cannot be outside the range of its neighboring values $\{u_{i-1}^{n+1}, u_{i+1}^{n+1}, u_i^n\}$, when u_i^n and $u_i^{n+1,*}$ are strict local extrema in the same sense.

B-2. Assume that (4.16) holds. This case represents the largest set of grid points for locally-smooth nonconstant solutions, where the solution is most likely monotone locally. Here the main task is to prove that

$$u_i^{n+1} \textit{ is not a strict local extremum, when (4.16) holds.} \tag{4.20}$$

When x_i is an interior point of the set, it is clear to see it, because the implicit half step of the CN method (4.14) does not introduce local extrema to the numerical solution.

Let x_i be an edge point of the set; that is, at least one of x_{i-1} and x_{i+1} is in the wobble set. In this case, a mathematical analysis for the task (4.20) is hard to be carried out explicitly due to the nature of implicit equations. For example, let x_{i-1} be in the wobble set ($\theta_{i-1}^{n+1} = 1$). Then it follows from (4.13), (4.14), and (4.15) that

$$\begin{aligned}(1+2\mu)u_{i-1}^{n+1} &= \mu(u_{i-2}^{n+1} + u_i^{n+1}) + u_{i-1}^n, \\ (1+\mu)u_i^{n+1} &= \frac{\mu}{2}(u_{i-1}^{n+1} + u_{i+1}^{n+1}) + \frac{\mu}{2}(u_{i-1}^n + u_{i+1}^n) + (1-\mu)u_i^n.\end{aligned} \tag{4.21}$$

Thus, with one more implicit equation defined at x_{i+1}, u_i^{n+1} is related to 9 neighboring values: $\{u_{i+j}^{n+1} \mid j = -2, -1, 1, 2\}$, $\{u_{i+k}^n \mid k = -2, \cdots, 2\}$, each of which again related to its neighboring values. Thus we decided to prove (4.20) *numerically*.

As a numerical test, we consider the following 1D parabolic equation of a discontinuous initial condition on $[-1, 1]$:

$$\begin{aligned}&\partial_t u - \partial_{xx} u = 0, \quad (x,t) \in (-1,1) \times [0,T], \\ &u(x,0) = u_0(x) = \begin{cases} 1 & \text{if } |x| < 0.5, \\ 0.5 & \text{if } |x| = 0.5, \\ 0 & \text{if } |x| > 0.5. \end{cases}\end{aligned} \tag{4.22}$$

The Dirichlet boundary condition is set to satisfy the analytic solution given in [18]:

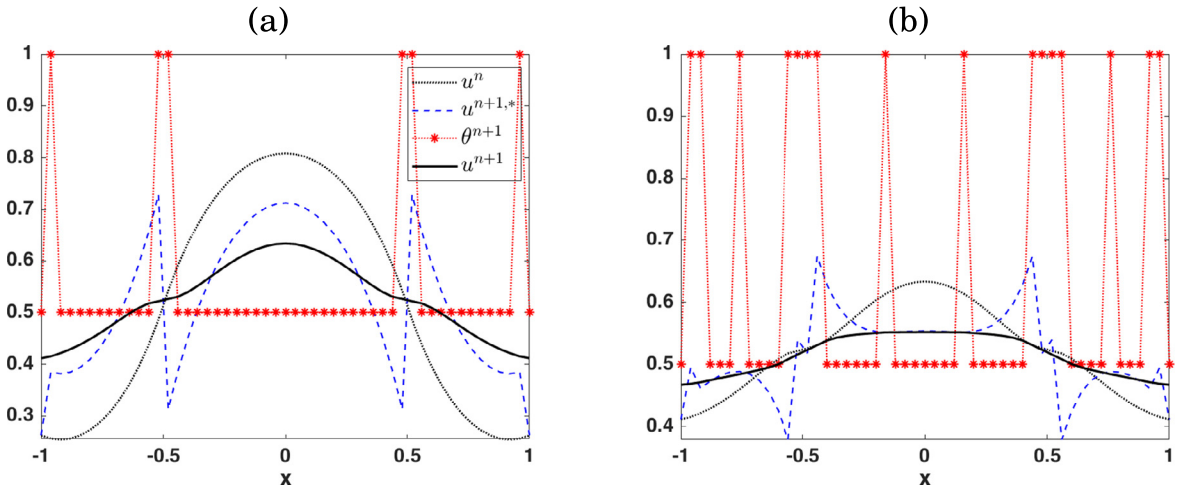

Fig. 4. The solutions: u^n (black dotted) and u^{n+1} (black solid), intermediate solution $u^{n+1,*}$ (blue dashed), and variable θ^{n+1} (red asterisk marker) to the problem (4.22) at the time step: (a) $n=2$ and (b) $n=3$.

$$u(x,t)=\frac{1}{2}+2\sum_{k=0}^{\infty}(-1)^k\frac{\cos\pi(2k+1)x}{\pi(2k+1)}e^{-\pi^2(2k+1)^2t},\quad (x,t)\in[-1,1]\times[0,T]. \tag{4.23}$$

Fig. 4 exhibits numerical solutions, u^n and u^{n+1}, the intermediate solution $u^{n+1,*}$, and θ^{n+1} at the time steps $n=2$ and $n=3$, for $h=0.04$ and $\tau=0.1$. One can check from the figures that u_i^{n+1} never involves a strict local extremum at points where (4.16) holds, which proves (4.20) experimentally. It has been verified from various numerical tests that the claim (4.20) is true.

The above proves (partially experimentally, though) that the variable-θ method does not introduce an interior local extremum to the numerical solution of 1D heat equation (4.11) for all choices of $\mu>0$. Thus the variable-θ method satisfies the maximum principle *unconditionally*. One can apply the above arguments for the 2D case. □

It is hoped that readers having advanced mathematical insights can prove it *mathematically*.

Remark 4.2. The wobble set $\mathcal{W}_{1D}^n$ in (4.5) can be defined as

$$\begin{aligned}\widehat{\mathcal{W}}_{1D}^n \;=\; \Big\{ & x_i\in(-1,1)\,|\,[\mathtt{idxt}(u_{i-1}^{n+1,*},u_i^{n+1,*},u_{i+1}^{n+1,*})\neq 0\\ & \text{or }\ \mathtt{idxt}(u_{i-1}^{n},u_i^{n},u_{i+1}^{n})\neq 0]\\ & \text{and }\ \Big|\mathtt{idxt}(u_{i-1}^{n+1,*},u_i^{n+1,*},u_{i+1}^{n+1,*})+\mathtt{idxt}(u_{i-1}^{n},u_i^{n},u_{i+1}^{n})\Big|<4\Big\}.\end{aligned} \tag{4.24}$$

When $x_i\notin\widehat{\mathcal{W}}_{1D}^n$, we have

$$\begin{aligned}&[\min\{u_{i-1}^{n+1,*},u_{i+1}^{n+1,*}\}<u_i^{n+1,*}<\max\{u_{i-1}^{n+1,*},u_{i+1}^{n+1,*}\}\\ &\quad\text{and }\ \min\{u_{i-1}^{n},u_{i+1}^{n}\}<u_i^{n}<\max\{u_{i-1}^{n+1,*},u_{i+1}^{n+1,*}\}],\ \text{or}\\ &\Big|\mathtt{idxt}(u_{i-1}^{n+1,*},u_i^{n+1,*},u_{i+1}^{n+1,*})+\mathtt{idxt}(u_{i-1}^{n},u_i^{n},u_{i+1}^{n})\Big|=4.\end{aligned} \tag{4.25}$$

It is clear to see that $\mathcal{W}_{1D}^n\subset\widehat{\mathcal{W}}_{1D}^n$. However, their performances are not observably different in practice, because it is occasional for $\widehat{\mathcal{W}}_{1D}^n$ to include more points than $\mathcal{W}_{1D}^n$; the extra points are quite few.

Like the standard 5-point scheme, the averaging scheme $\mathcal{A}_\alpha$ is an approximation of the negative Laplacian by using a weighted sum of the standard 5-point scheme and the skewed 5-point scheme. Hence, the same arguments in the proof of Theorem 4.1 can be extended for the variable-θ method with the averaging scheme $\mathcal{A}_\alpha$ for $0\le\alpha\le 1$.

Corollary 4.3. *The numerical solution of the variable-θ method (4.10) with the averaging scheme $\mathcal{A}_\alpha$ for $0\le\alpha\le 1$ to the heat equation satisfies the maximum principle* unconditionally.

5. The optimal averaging parameter $\widetilde{\alpha}$

In this section, we will try to derive an optimal averaging parameter which minimizes the leading truncation error. Since the variable-θ method is a variant of the CN method, to focus on the effect of the optimal parameter, we restrict our interest to the truncation error of the CN method ($\theta\equiv 1/2$) with the averaging scheme. Then the leading truncation error simply reads

$$\begin{aligned}&\frac{h^2}{12}\left[u_{xxxx}+6(1-\alpha)u_{xxyy}+u_{yyyy}\right]-\frac{\tau^2}{24}u_{ttt}\\ &\quad=\frac{h^2}{12}\left[\Delta^2u+(4-6\alpha)u_{xxyy}\right]-\frac{\tau^2}{24}u_{ttt},\end{aligned} \tag{5.1}$$

where we have utilized the identity $\Delta^2u=u_{xxxx}+2u_{xxyy}+u_{yyyy}$. To choose α which make vanish the leading error, we consider the following equation: for $\gamma=\tau/h$,

$$\Delta^2u+(4-6\alpha)u_{xxyy}-\frac{\gamma^2}{2}u_{ttt}=0. \tag{5.2}$$

Solving (5.2) for α, we obtain

$$\alpha=\frac{2\Delta^2u-\gamma^2u_{ttt}}{12u_{xxyy}}+\frac{2}{3}. \tag{5.3}$$

Let

$$\overline{\partial}_tu^n:=\frac{u^n-u^{n-1}}{\tau}.$$

Then, since $\overline{\partial}_tu^n\approx u_t+\frac{\tau^2}{24}u_{ttt}$, we have

$$u_{ttt}\approx\frac{24}{\tau^2}(\overline{\partial}_tu^n-u_t^n). \tag{5.4}$$

Thus it follows from (5.3) and (5.4) that the parameter α in the n-th level becomes

$$\begin{aligned}\alpha^n&\approx\frac{h^2\Delta^2u-12(\overline{\partial}_tu^n-u_t^n)}{6h^2u_{xxyy}}+\frac{2}{3}\\ &=\frac{h^2\Delta^2u-12(\overline{\partial}_tu^n-D\Delta u^n-f(\mathbf{u}^n))}{6h^2u_{xxyy}}+\frac{2}{3}.\end{aligned} \tag{5.5}$$

Define discrete operators of a second-order accuracy

$$\mathcal{A}_xu\approx-u_{xx}\ \text{ and }\ \mathcal{A}_yu\approx-u_{yy}.$$

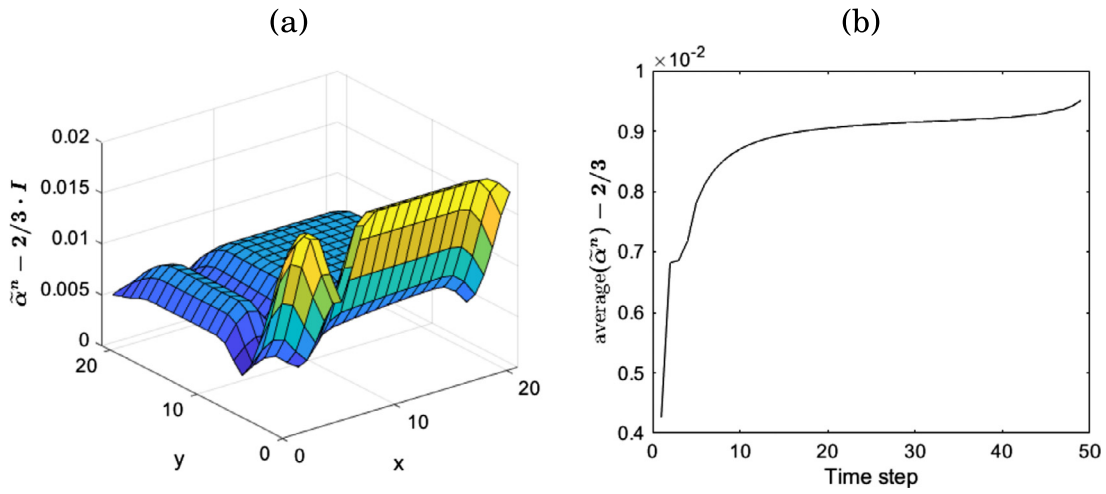

Fig. 5. (a) $\widetilde{\alpha}^n - 2/3 \cdot I$ at $t = 0.1$ ($n = 10$) and (b) average($\widetilde{\alpha}^n$) $- 2/3$ on the 50 time steps.

Table 1
Accumulated L^2-errors $E[0,T]$ and CPU-time for the problem associated with (5.8) and (5.9) at $T = 0.5$ and the mesh resolution $(\tau, h) = (0.01, 0.05)$ for $\alpha = 1, 2/3$, and $\widetilde{\alpha}^n$.

$\varepsilon = 10^{-8}$	$\alpha = 1$	$\alpha = 2/3$	$\widetilde{\alpha}^n$
L^2-error	$8.65 \cdot 10^{-7}$	$6.36 \cdot 10^{-7}$	$6.07 \cdot 10^{-7}$
CPU	1.09	1.52	1.60

Then, by approximating (5.5), we obtain the optimal parameter matrix $\boldsymbol{\alpha}^n$ at the n-th time step as follows.

$$\boldsymbol{\alpha}^n \approx \frac{h^2 \mathcal{A}_+^2 u^n - 12(\overline{\partial}_t u^n - D\mathcal{A}_+ u^n - f(\mathbf{u}^n))}{6h^2 \mathcal{A}_x \mathcal{A}_y u^n} + \frac{2}{3}. \tag{5.6}$$

In practice, we must restrict entries of $\boldsymbol{\alpha}^n$ between 0 and 1,

$$[\boldsymbol{\alpha}^n]_{ij} = \min(\max([\boldsymbol{\alpha}^n]_{ij}, 0), 1),$$

and apply an appropriate smoothing operator S (e.g., Gaussian 5×5 filter with $\sigma = 1.0$) to attain a reliable smooth parameter matrix:

$$\widetilde{\boldsymbol{\alpha}}^n \approx S\left(\frac{h^2 \mathcal{A}_+^2 u^n - 12(\overline{\partial}_t u^n - D\mathcal{A}_+ u^n - f(\mathbf{u}^n))}{6h^2 \mathcal{A}_x \mathcal{A}_y u^n} \right) + \frac{2}{3}. \tag{5.7}$$

Now, we will verify the effectiveness of the optimal parameter (5.7) with the numerical solution of the heat equation:

$$\partial_t u - \Delta u = f(x, y, t), \tag{5.8}$$

where the source term f, and the initial and boundary condition are set corresponding to the analytic solution given by

$$u(x, y, t) = e^{-2\pi^2 t} \sin \pi x \sin \pi y, \quad (x, y, t) \in [-0.25, 0.75]^2 \times [0, T]. \tag{5.9}$$

Here we have set an asymmetric domain $[-0.25, 0.75]^2$ to add an asymmetry to the numerical solution; we have selected the heat equation (a linear problem) to focus on the discretization error of the averaging scheme, without being mixed by the error from nonlinear terms.

For three choices of $\alpha = 1, 2/3$, and $\widetilde{\alpha}^n$, Table 1 summarizes the L^2-error $E_2[0,T]$ and the CPU-time for the problem associated with (5.8)-(5.9) at $T = 0.5$, when grid sizes $(\tau, h) = (0.01, 0.05)$. One can see from Table 1 that the error of Mehrstellen discretization ($\alpha = 2/3$) is smaller than that of the standard 5-point scheme ($\alpha = 1$); the optimal parameter $\widetilde{\alpha}^n$ shows a slightly smaller error than the error of Mehrstellen discretization. The optimal procedure consumes more CPU-time compared with the other two fixed parameters, due to the extra computation of the optimal parameter matrix.

Fig. 5 shows $\widetilde{\alpha}^n - 2/3 \cdot I$, the difference between optimal parameter matrix $\widetilde{\alpha}^n$ and 2/3 (Mehrstellen) at $t = 0.1$, and the difference between the averages of $\widetilde{\alpha}^n$ and 2/3 on the 50 time steps. One can see from the figure that all the differences are pretty close to zero.

We have found from various experiments that the *fixed* averaging parameter employed in the Mehrstellen scheme ($\alpha = 2/3$) is effective enough to represent the variable optimal parameter $\widetilde{\alpha}^n$. In the rest of the article, the fixed parameter $\alpha = 2/3$ will be utilized as the optimal averaging parameter, unless otherwise indicated.

6. Numerical experiments

In this section, we present numerical experiments to show the effectiveness of the proposed method, the averaging scheme incorporated with the variable-θ time-stepping procedure, in both accuracy and efficiency. The algorithm is implemented in Matlab and carried out on a Desktop computer of AMD Ryzen 7 PRO 4750U 1.7GHz (4.1GHZ) processor with 16.0GB RAM. For a comparison purpose, we also implement the CNOSC method [4]. For the algebraic solver, we employ the SOR with the near-optimal parameter studied in [11, §4]. The SOR iteration is stopped when the maximum difference of consecutive iterates becomes smaller than a prescribed tolerance,

$$\|u^{n,k} - u^{n,k-1}\|_\infty < \varepsilon, \tag{6.1}$$

where we set $\varepsilon = 10^{-8}$. We set $\alpha = 2/3$ for the averaging scheme. The elapsed time is measured in second and denoted by *CPU*.

6.1. Convergence analysis

We begin with a convergence analysis for the proposed method. Consider a heat equation in 2D.

$$\begin{aligned} \partial_t u - \Delta u &= f(x,t), \quad (x, y, t) \in (-1, 1)^2 \times (0, T], \\ u(x, y, 0) &= \sin \pi x \sin \pi y, \quad (x, y) \in [-1, 1]^2, \end{aligned} \tag{6.2}$$

where the boundary condition is set corresponding to the exact solution given by $u(x, y, t) = e^{-2\pi^2 t} \sin \pi x \sin \pi y$, with which the source term vanishes, i.e., $f \equiv 0$.

Table 2 summarizes the accumulated L^2-error $E[0,T]$ with $T = 0.5$, the CPU-time, and the convergence order, for the numerical solution of the heat equation (6.2) in various resolutions. The convergence order is measured along the diagonal entries of the table, which is slightly higher than the second-order. The proposed method, the averaging scheme incorporated with the variable-θ time-stepping procedure, results in a desirable accuracy. It can achieve a near second-order accuracy in the temporal direction by the variable-θ method and a slightly

Table 2
Accumulated L^2-error $E[0,T]$ with $T=0.5$, the CPU-time, and the convergence order, for the proposed method solving the heat equation (6.2). The convergence order is measured along the diagonal entries of the table.

τ \ h	0.1 (CPU)	0.05 (CPU)	0.025 (CPU)	0.0125 (CPU)	order
0.1 (CPU)	$8.21\cdot10^{-4}$ (0.04)	$5.34\cdot10^{-4}$ (0.08)	$3.03\cdot10^{-4}$ (0.20)	$1.71\cdot10^{-4}$ (1.19)	
0.05 (CPU)	$5.80\cdot10^{-4}$ (0.08)	$1.75\cdot10^{-4}$ (0.13)	$5.94\cdot10^{-5}$ (0.35)	$3.00\cdot10^{-5}$ (1.80)	2.23
0.025 (CPU)	$1.90\cdot10^{-4}$ (0.15)	$8.87\cdot10^{-5}$ (0.23)	$4.14\cdot10^{-5}$ (0.62)	$1.15\cdot10^{-5}$ (2.71)	2.08
0.0125 (CPU)	$1.24\cdot10^{-4}$ (0.23)	$1.26\cdot10^{-5}$ (0.43)	$1.05\cdot10^{-5}$ (1.19)	$8.29\cdot10^{-6}$ (4.40)	2.32

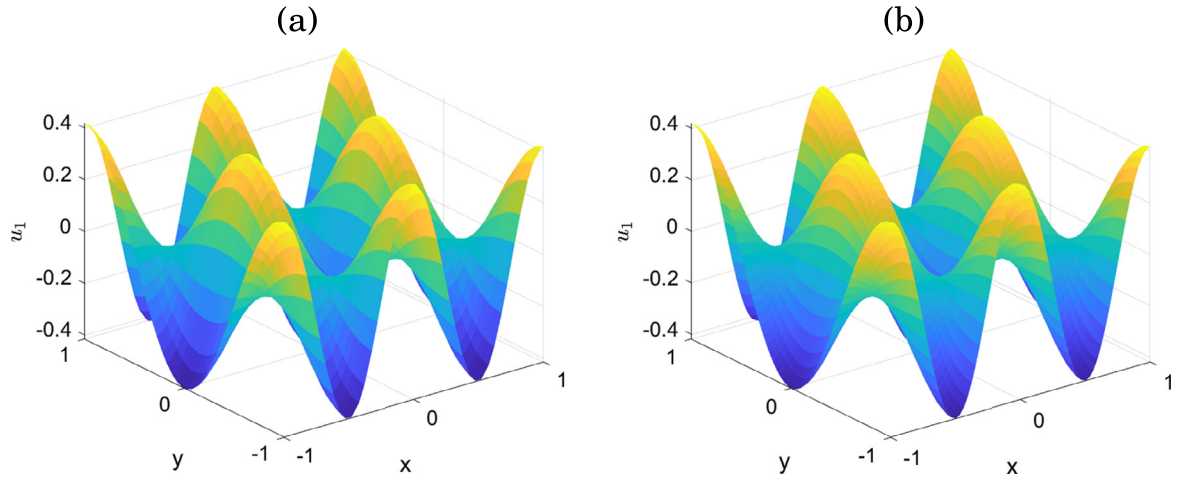

Fig. 6. Numerical solutions u_1 for Gray-Scott model (6.3)-(6.7) at $T=1$, approximated by (a) the CNOSC method and (b) the proposed method. Set $(\tau,h)=(0.001,0.1)$.

Table 3
The accumulated L^2-error $E[0,T]$ and the CPU-time for the numerical solution of the Gray-Scott model associated with (6.3)-(6.6) in various resolutions, when $T=1.0$.

$(\tau=h^3,h)$	CNOSC (CPU)	$\mathcal{A}_+$ (CPU)	$\mathcal{A}_{2/3}$ (CPU)	$\mathcal{A}_{\widetilde{\alpha}^n}$ (CPU)
$(0.2^3,0.2)$	$1.66\cdot10^{-2}$ (0.35s)	$1.41\cdot10^{-3}$ (0.14s)	$1.18\cdot10^{-3}$ (0.24s)	$1.13\cdot10^{-3}$ (0.89s)
$(0.1^3,0.1)$	$4.00\cdot10^{-3}$ (8.01s)	$9.21\cdot10^{-4}$ (3.09s)	$8.67\cdot10^{-4}$ (5.18s)	$8.44\cdot10^{-4}$ (24.23s)
$(0.05^3,0.05)$	$1.75\cdot10^{-4}$ (221.32s)	$6.88\cdot10^{-5}$ (88.33s)	$6.79\cdot10^{-5}$ (126.45s)	$6.77\cdot10^{-5}$ (707.56s)

higher accuracy than second-order in the spatial direction due to the averaging scheme.

It should be noticed that the proposed method is *scalable* when the mesh sizes are set as in practice. For example, for diagonal entries in the table ($\tau=h$), the problem size in the current level becomes eight times the previous one. However the CPU-time increases only by factors of 3.25, 4.77, and 7.10, respectively for $\tau=h=0.05, 0.025$, and 0.0125. The algebraic solver, the SOR, converges faster for smaller τ, while it converges slower for finer spatial resolutions. With the near-optimal parameter [11], the SOR can maintain the same efficiency when the workload grows.

From the above example, we can conclude that the proposed method is second-order in accuracy and scalable in efficiency, for the numerical solution of the heat equation. The claim can be applied for the numerical solution of the nonlinear RD system (1.1), provided that the nonlinear reaction term is approximated and treated accurately enough.

6.2. The Gray-Scott model

First, we will verify the accuracy and efficiency of the averaging schemes quantitatively. Consider the following two-component Gray-Scott model formulated as in (1.1) with the reaction kinetics $\mathbf{f}(\mathbf{u})$ given as

$$\mathbf{f}(\mathbf{u})=[F(1-u_1)-u_1u_2^2, u_1u_2^2-(F+k)u_2]^T. \tag{6.3}$$

We choose model coefficients as follows:

$$\Omega=(-1,1)^2,\quad D=[0.001,0.001]^T,\quad F=1,\quad k=0. \tag{6.4}$$

For a purpose of error analysis, we select a smooth solution $\widehat{\mathbf{u}}=[\widehat{u}_1,\widehat{u}_2]$ defined as

$$\begin{aligned}\widehat{u}_1(x,y,t)&=\cos(2t)\cos(2\pi x)\cos(\pi y),\\ \widehat{u}_2(x,y,t)&=\cos(2t)\cos(\pi x)\cos(2\pi y),\end{aligned} \tag{6.5}$$

and replace the reaction kinetics $\mathbf{f}(\mathbf{u})$ with $\mathbf{f}_{\widehat{\mathbf{u}}}(\mathbf{u})$:

$$\mathbf{f}_{\widehat{\mathbf{u}}}(\mathbf{u}):=\mathbf{f}(\mathbf{u})+\frac{\partial\widehat{\mathbf{u}}}{\partial t}-D\Delta\widehat{\mathbf{u}}-\mathbf{f}(\widehat{\mathbf{u}}). \tag{6.6}$$

Then $\widehat{\mathbf{u}}=[\widehat{u}_1,\widehat{u}_2]$ in (6.5) would be the exact solution of

$$\frac{\partial\mathbf{u}}{\partial t}-D\Delta\mathbf{u}-\mathbf{f}(\mathbf{u})=\mathbf{f}_{\widehat{\mathbf{u}}}(\mathbf{u}), \tag{6.7}$$

with the initial condition $\mathbf{u}^0=\widehat{\mathbf{u}}(x,y,0)$.

Fig. 6 presents u_1 of the numerical solutions approximated by the CNOSC method and the proposed method. For the CNOSC method, we select the same parameters as for Example 3 in [4]: $n_x=n_y=20$ ($h=0.1$), $\tau=h^3$, and $r=3$. We could check that Figs. 6 (a) and (b) show the same numerical solutions displayed as in Figure 2 of [4].

In Table 3, we present the accumulated L^2-error and the CPU-time for the numerical solution of the Gray-Scott model associated with (6.3)-(6.6) in various resolutions, when $T=1.0$. We compare performances of four different algorithms: the CNOSC method, the standard 5-point scheme ($\mathcal{A}_+=\mathcal{A}_1$), the averaging scheme with $\alpha=2/3$ ($\mathcal{A}_{2/3}$), and the averaging scheme with $\widetilde{\boldsymbol{\alpha}}^n$ ($\mathcal{A}_{\widetilde{\alpha}^n}$). The CNOSC method is implemented with the splines of degree $r=3$, while the FD schemes are

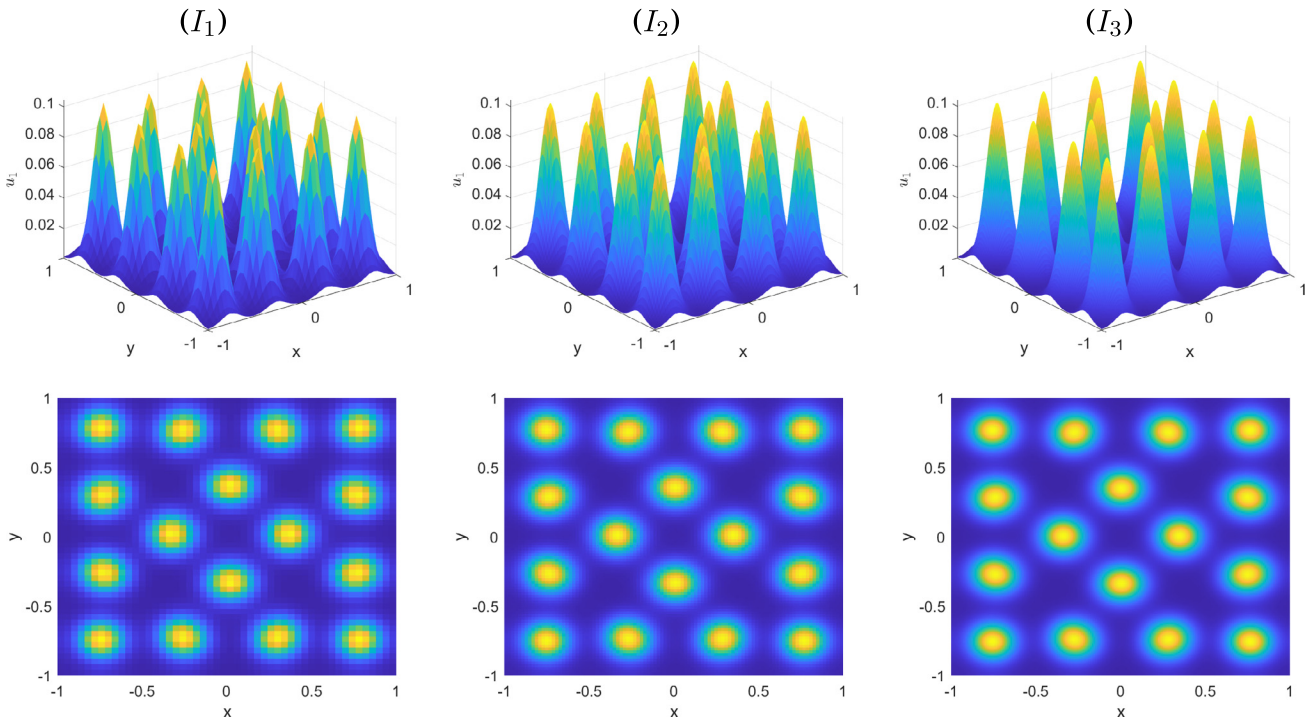

Fig. 7. Numerical solutions for the Gierer-Meinhardt model (2.8)-(2.9) in the steady-state ($T = 500$), approximated by the averaging scheme $\mathcal{A}_{2/3}$-VT. at the fixed time step $\tau = 0.01$ and various spatial resolutions. Each image column I_ℓ presents the numerical solution and its aerial view obtained with the mesh resolution $(\tau, n_x = n_y) = (0.01, 50 \cdot 2^{\ell-1})$.

not incorporated with the variable-θ method. In order to give an asymmetry to the numerical solution, we shift the domain of (6.4) by 0.5, i.e. from $(-1,1)^2$ to $(-0.5,1.5)^2$. Moreover, finer temporal resolutions ($\tau = h^{(r+1)/2}$) are chosen for the CNOSC method to achieve $\mathcal{O}(\tau^2 + h^{r+1})$ accuracy in the L^2-norm [4].

We can point out from the table that the L^2-errors of the averaging schemes are smaller than those of other methods. For example, in the low spatial resolution $(\tau, h) = (0.2^3, 0.2)$, the error of the CNOSC method mounts up about 15 times those of the averaging schemes. The CNOSC method may introduce imperceptible oscillations spreading out to all over the domain, which might be originated from its rough orthogonal basis taking fewer collocation points of the low spatial resolution. Furthermore, even though the CNOSC method employs the ADI method to accelerate its computation, the CPU-time is longer than the two FD methods, $\mathcal{A}_+$ and $\mathcal{A}_{2/3}$. It is partially due to the efficiency of the SOR method; also it is because of an intrinsic complexity of the CNOSC method, in which the calculation has to deal with large coefficient matrices. For example, when the spatial mesh is set with 40×40 grid points, the CNOSC method with the splines of degree $r = 3$ produces coefficient matrices in 82×82 dimensions ($82 = 2 \cdot 40 + 2$).

The errors of the three FD schemes ($\mathcal{A}_+$, $\mathcal{A}_{2/3}$, and $\mathcal{A}_{\widetilde{\alpha}^n}$) are different, but not significantly. However, the differences in their CPU-time are quite varied. As we examined before, the averaging scheme with $\widetilde{\alpha}^n$ showed much more CPU-time than the other two FD methods; $\mathcal{A}_{2/3}$ has achieved a good accuracy and efficiency with comparatively smaller CPU-time.

Table 4 shows the accumulated L^2-errors $E[0,T]$, $T = 1.0$, and the CPU-time of the averaging scheme $\mathcal{A}_{2/3}$ incorporated with the variable-θ method ($\mathcal{A}_\alpha$-VT), for the numerical solution of the Gray-Scott model with parameters set the same as in Table 3. We consider various temporal resolutions with the spatial grid size being fixed, $h = 0.1$. Since the overall error is dominated by the spatial error, we cannot witness a dramatic error change as the time step size varies. However, one can see from the table that even with $(\tau, h) = (0.1, 0.1)$, the accuracy of the $\mathcal{A}_{2/3}$-VT has surpassed that of the CNOSC method with $(\tau, h) = (0.1^3, 0.1)$; see the second row in the first column in Table 3. The CNOSC method requires at least 100 times more time steps to achieve the same level of accuracy as the $\mathcal{A}_{2/3}$-VT. We further stress out the efficiency of the proposed method, by comparing the error and the CPU-time of $\mathcal{A}_{2/3}$-VT ($1.15 \cdot 10^{-3}$, 0.09s) and the CNOSC method ($4.00 \cdot 10^{-3}$, 8.01s). The $\mathcal{A}_{2/3}$-VT is 100 times more efficient than the CNOSC method.

Table 4
The accumulated L^2-error $E[0,T]$, $T = 1.0$, and the CPU-time for the numerical solution of the Gray-Scott model associated with (6.3)-(6.6) by the $\mathcal{A}_{2/3}$-VT, in various temporal resolutions. The spatial grid size is fixed, $h = 0.1$.

(τ, h)	$(0.05, 0.1)$ (CPU)	$(0.1, 0.1)$ (CPU)	$(0.2, 0.1)$ (CPU)
$\mathcal{A}_{2/3}$-VT	$8.88 \cdot 10^{-4}$ (0.16s)	$1.15 \cdot 10^{-3}$ (0.09s)	$3.20 \cdot 10^{-3}$ (0.05s)

6.3. The Gierer-Meinhardt model

In Fig. 7, we depict the numerical solutions of the Gierer-Meinhardt model at the steady-state ($T = 500$) by the $\mathcal{A}_{2/3}$-VT, in order to investigate the rotational symmetry and the accuracy of the proposed method. Its mesh resolutions are the same as in Figs. 1 and 2. Unlike the low resolution cases of the standard 5-point scheme and the CNOSC method: Figs. 1 (I_1) and 2 (J_1), the proposed averaging scheme shows the same steady-state pattern as the high resolution cases, as illustrated in Fig. 7 (I_1). It is noticeable that the $\mathcal{A}_{2/3}$-VT can achieve a stable evolution even in low spatial resolutions.

To examine early pattern formations in low spatial resolutions, we depict the numerical solution of Gierer-Meinhardt model by three different schemes: the standard 5-point scheme, the CNOSC method ($r = 3$), and the averaging scheme $\mathcal{A}_{2/3}$-VT. Set $(\tau, n_x = n_y) = (0.01, 50)$. Figures (a), (b), and (c) in Fig. 8 correspond respectively to an early state at $t = 80$ of Figs. 1 (I_1), 2 (I_1), and 7 (I_1).

In Fig. 8 (a), one can see that the standard 5-point scheme shows the artifacts depending on variable directions (x and y), which is originated from the spatial approximation concerning only two directions. In Fig. 8 (b), the CNOSC method shows a rotationally invariant pat-

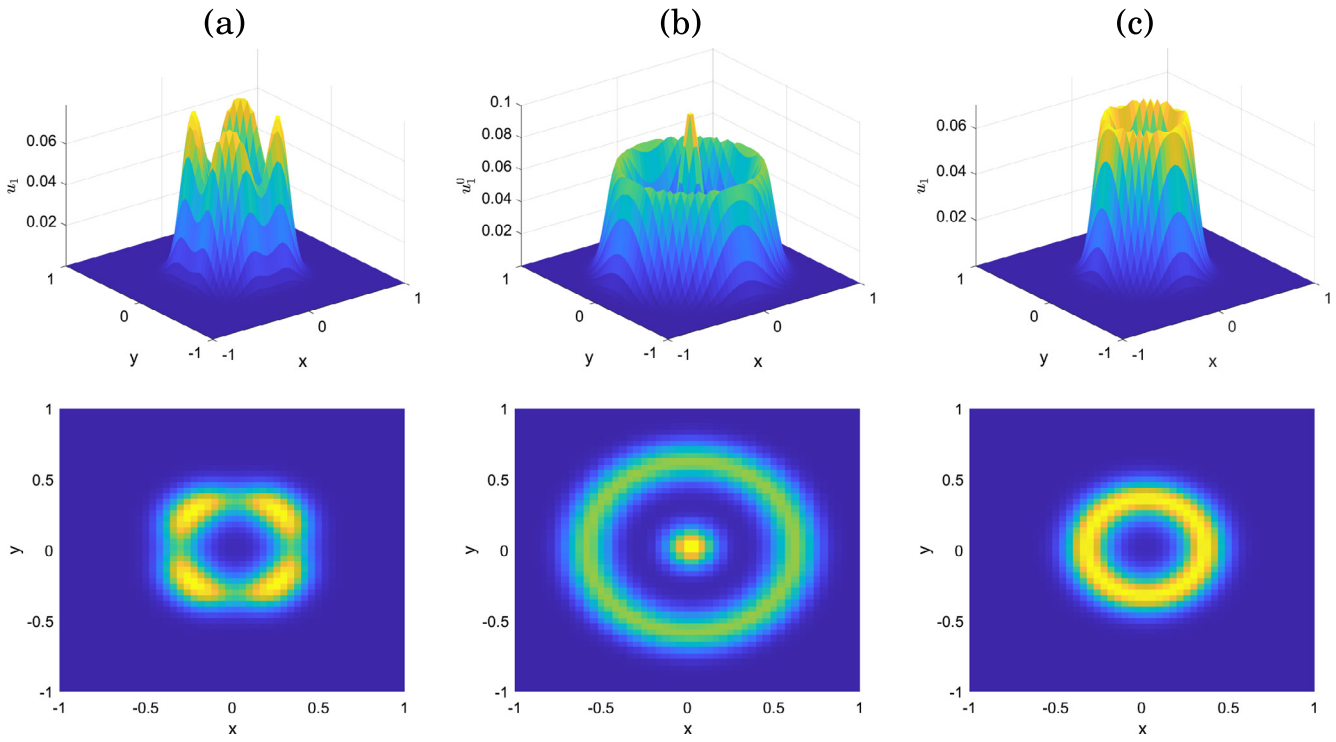

Fig. 8. Numerical solutions and their aerial views for Gierer-Meinhardt model (2.8)–(2.9) at $t = 80$, obtained by (a) the standard 5-point scheme, (b) the CNOSC method ($r = 3$), and (c) the $\mathcal{A}_{2/3}$-VT. Set $(\tau, n_x = n_y) = (0.01, 50)$.

tern formation. However it introduces a peak at the center of the figure, which results in a bigger circular pattern than in Fig. 8 (c). The peak is arisen from the rough spline polynomial generated by fewer collocation points and we can suppress the artifact by either adding more Gaussian quadrature points or employing higher-order polynomials; however the artifact may not be suppressed completely, although the CNOSC method becomes much more expensive computationally. These artifacts of the standard 5-point scheme and the CNOSC method would produce abnormal steady-state patterns in low spatial resolutions, which are different from their steady-state patterns in high spatial resolutions.

On the other hand, the $\mathcal{A}_{2/3}$-VT gives us a rotationally invariant pattern formation without any visible artifacts as shown in Fig. 8 (c); the averaging scheme $\mathcal{A}_{2/3}$-VT can achieve the same steady-state patterns in low spatial resolutions as in high spatial resolutions, since the scheme approximates the diffusion operator appropriately by considering all the possible directions.

To precisely examine the evolution of each method depending on spatial resolutions, Figs. 9, 10, and 11 present evolution patterns for the numerical solution of the Gierer-Meinhardt model (2.8)-(2.9) at $T = 80, 170, 270, 340$, approximated respectively by the standard 5-point scheme, the CNOSC method ($r = 3$), and the averaging scheme $\mathcal{A}_{2/3}$-VT. The algorithm parameters are set the same as for Fig. 8, with various spatial resolutions ($n_x = n_y = 50, 100, 200$). As you can see from the figures, the averaging scheme results in evolution patterns consistent over the spatial resolutions (Fig. 11), while other methods show evolution patterns quite different depending on the spatial resolutions (Figs. 9 and 10). Only the averaging scheme, $\mathcal{A}_{2/3}$-VT, can produce the correct evolution pattern in the low spatial resolution ($n_x = n_y = 50$). For the steady-state patterns ($T = 500$), one can refer to Figs. 1, 2, and 7.

We close the section with the following remark: we have *numerically* verified the rotational symmetry of the averaging scheme $\mathcal{A}_{2/3}$-VT.

7. Conclusions

The authors' previous publication [11] revealed that the spatial sensitivity of nonlinear reaction-diffusion problems might introduce undesirable artifacts into their numerical solutions, particularly in low spatial resolutions. The sensitivity issue in the one-dimensional space can be well-explained via grid effect; however in two and higher dimensions, the issue can be affected not only by the grid effect but also by the rotational symmetry of the approximation of the Laplacian diffusion operator. Moreover, most of the conventional Laplacian approximations fail to hold rotational symmetry. To investigate the effect of rotational symmetry, we have conducted a sensitivity analysis for two-dimensional reaction-diffusion problems approximated by various finite-difference methods. The averaging scheme $\mathcal{A}_\alpha$ for the approximation of the negative Laplacian $(-\Delta)$ is suggested as an average of the standard 5-point scheme and the skewed 5-point scheme. It has been found that the averaging scheme can effectively suppress artifacts arisen from an asymmetry of numerical approximation and eventually give us correct steady-state patterns, even in low spatial resolutions. An effective strategy is suggested to optimize the averaging parameter matrix $\widetilde{\boldsymbol{\alpha}}^n$, which can be replaced by the fixed parameter $2/3$ in practice. In addition to the averaging scheme, we have analyzed the maximum principle for the variable-θ method that is the time-stepping method employed to solve the reaction-diffusion systems. Various numerical examples have been considered to show the effectiveness of the proposed method.

Data availability

The data and figures used to support the findings of this study are included within the article. The codes in MATLAB for the experiments are available from the corresponding author upon request.

Acknowledgement

The authors deeply appreciate informative suggestions by two anonymous reviewers. This research was conducted as part of NSF-MCB 1714157 awarded to George V. Popescu.

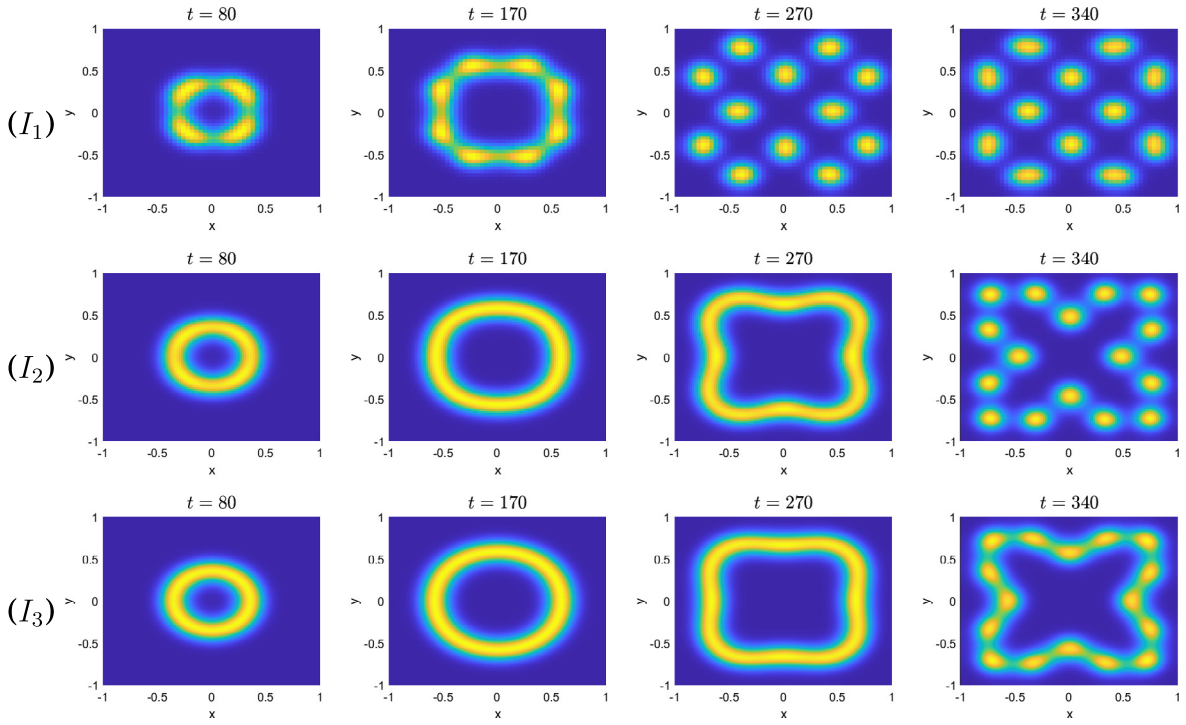

Fig. 9. Evolution of patterns for the Gierer-Meinhardt model (2.8)-(2.9) at $T = 80, 179, 270, 340$, approximated by the standard 5-point scheme. Each image row I_ℓ represents the numerical solution and its aerial view obtained with the mesh resolution $(\tau, n_x = n_y) = (0.01, 50 \cdot 2^{\ell-1})$.

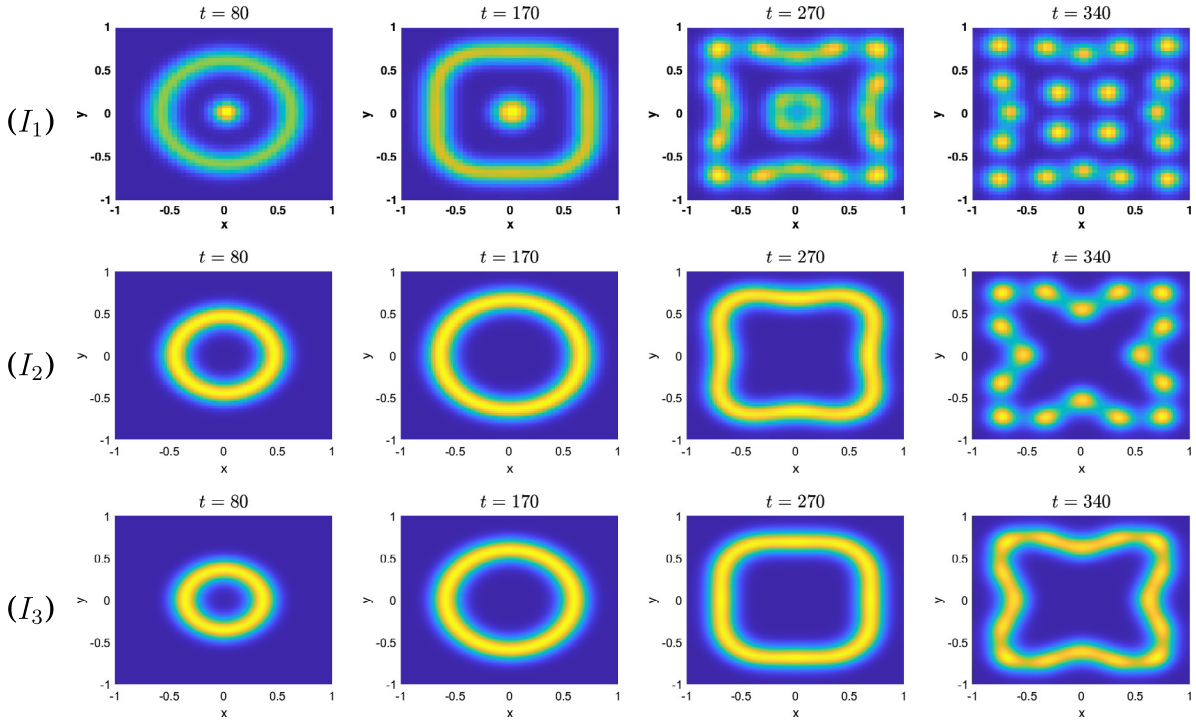

Fig. 10. Evolution of patterns for the Gierer-Meinhardt model (2.8)-(2.9) at $T = 80, 170, 270, 340$, approximated by the CNOSC method ($r = 3$). Each image row I_ℓ represents the numerical solution and its aerial view obtained with the mesh resolution $(\tau, n_x = n_y) = (0.01, 50 \cdot 2^{\ell-1})$.

References

[1] L. Collatz, The Numerical Treatment of Differential Equations, vol. 60, Springer Science & Business Media, 2012.

[2] G. Fairweather, X. Yang, D. Xu, H. Zhang, An adi Crank–Nicolson orthogonal spline collocation method for the two-dimensional fractional diffusion-wave equation, J. Sci. Comput. 65 (2015) 1217–1239.

[3] R.I. Fernandes, B. Bialecki, G. Fairweather, An adi extrapolated Crank–Nicolson orthogonal spline collocation method for nonlinear reaction–diffusion systems on evolving domains, J. Comput. Phys. 299 (2015) 561–580.

[4] R.I. Fernandes, G. Fairweather, An adi extrapolated Crank-Nicolson orthogonal spline collocation method for nonlinear reaction–diffusion systems, J. Comput. Phys. 231 (2012) 6248–6267.

[5] N. Fisher, B. Bialecki, Extrapolated adi Crank–Nicolson orthogonal spline collocation for coupled Burgers' equations, J. Differ. Equ. Appl. 26 (2020) 45–73.

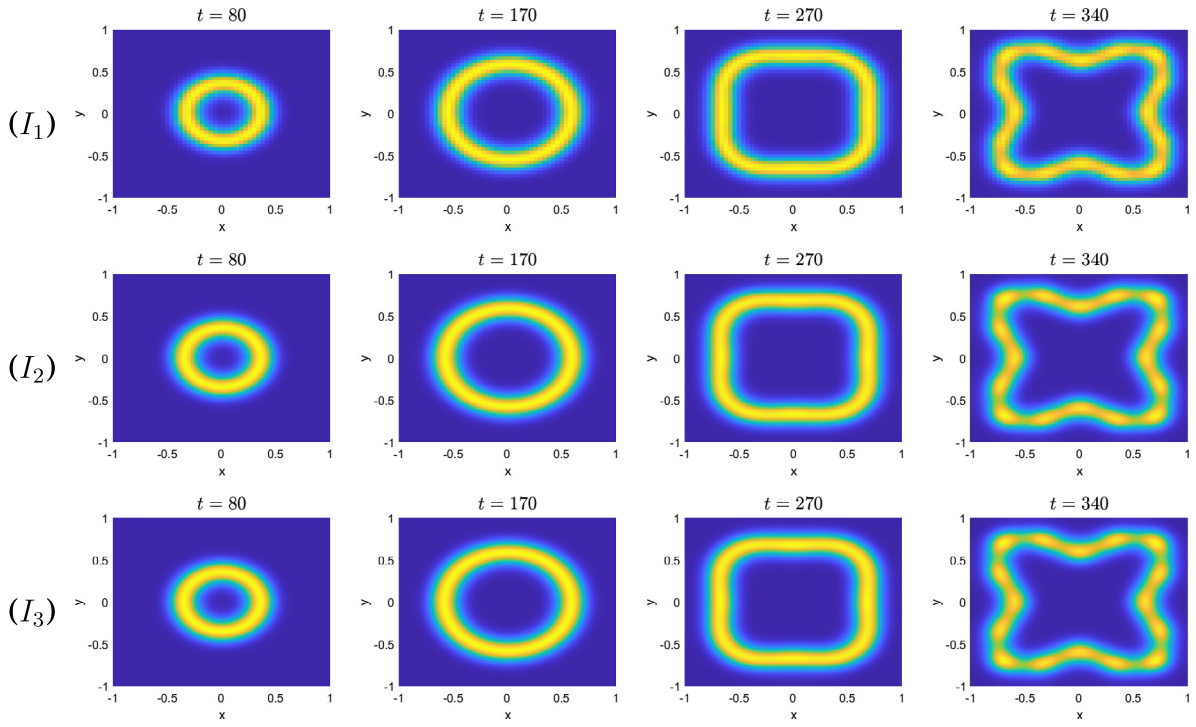

Fig. 11. Evolution of patterns for the Gierer-Meinhardt model (2.8)-(2.9) at $T = 80, 170, 270, 340$, approximated by the averaging scheme $\mathcal{A}_{2/3}$-VT. Each image row I_ℓ represents the numerical solution and its aerial view obtained with the mesh resolution $(\tau, n_x = n_y) = (0.01, 50 \cdot 2^{\ell-1})$.

[6] A. Gierer, H. Meinhardt, A theory of biological pattern formation, Kybernetik 12 (1972) 30–39.

[7] C. Gingras, P.G. Kry, Procedural modelling with reaction diffusion and growth of thin shells, in: Proceedings of the 45th Graphics Interface Conference on Proceedings of Graphics Interface 2019, Canadian Human-Computer Communications Society, 2019, pp. 1–7.

[8] C.-H. Jo, C. Shin, J.H. Suh, An optimal 9-point, finite-difference, frequency-space, 2-d scalar wave extrapolator, Geophysics 61 (1996) 529–537.

[9] S. Kim, Compact schemes for acoustics in the frequency domain, Math. Comput. Model. 37 (2003) 1335–1341.

[10] P. Lee, S. Kim, A variable-θ method for parabolic problems of nonsmooth data, Comput. Math. Appl. 79 (2020) 962–981.

[11] P. Lee, G.V. Popescu, S. Kim, A nonoscillatory second-order time-stepping procedure for reaction-diffusion equations, Complexity 2020 (2020) 1–15.

[12] F. Liao, L. Zhang, S. Wang, Numerical analysis of cubic orthogonal spline collocation methods for the coupled Schrödinger–Boussinesq equations, Appl. Numer. Math. 119 (2017) 194–212.

[13] T. Lindeberg, Scale-Space Theory in Computer Vision, vol. 256, Springer Science & Business Media, 2013.

[14] A. Madzvamuse, Time-stepping schemes for moving grid finite elements applied to reaction–diffusion systems on fixed and growing domains, J. Comput. Phys. 214 (2006) 239–263.

[15] M. McCourt, N. Dovidio, M. Gilbert, Spectral methods for resolving spike dynamics in the Gierer-Meinhardt model, Commun. Comput. Phys. 3 (2008) 659–678.

[16] Z. Qiao, Numerical investigations of the dynamical behaviors and instabilities for the Gierer-Meinhardt system, Commun. Comput. Phys. 3 (2008) 406–426.

[17] F. Shakeri, M. Dehghan, The finite volume spectral element method to solve Turing models in the biological pattern formation, Comput. Math. Appl. 62 (2011) 4322–4336.

[18] J.C. Strikwerda, Finite Difference Schemes and Partial Differential Equations, vol. 88, SIAM, 2004.

[19] A.M. Turing, The chemical basis of morphogenesis, Philos. Trans. R. Soc. Lond. B, Biol. Sci. 237 (1952) 37–72.

[20] R. Varga, Matrix Iterative Analysis, 2nd ed., Springer-Verlag, Berlin, Heidelberg, 2000.

[21] P.A. Zegeling, H. Kok, Adaptive moving mesh computations for reaction–diffusion systems, J. Comput. Appl. Math. 168 (2004) 519–528.