Journal of Computational Physics 404 (2020) 109121

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

A roadmap for discretely energy-stable schemes for dissipative systems based on a generalized auxiliary variable with guaranteed positivity

Zhiguo Yang, Suchuan Dong *

Center for Computational and Applied Mathematics, Department of Mathematics, Purdue University, USA

ARTICLE INFO

Article history:
Received 30 March 2019
Received in revised form 22 August 2019
Accepted 10 November 2019
Available online 15 November 2019

Keywords:
Energy stability
Unconditional stability
Dissipative systems
Conservative systems
Auxiliary variables
Positivity

ABSTRACT

We present a framework for devising discretely energy-stable schemes for general dissipative systems based on a generalized auxiliary variable. The auxiliary variable, a scalar number, can be defined in terms of the energy functional by a general class of functions, not limited to the square root function adopted in previous approaches. The current method has another remarkable property: the computed values for the generalized auxiliary variable are guaranteed to be positive on the discrete level, regardless of the time step sizes or the external forces. This property of guaranteed positivity is not available in previous approaches. A unified procedure for treating the dissipative governing equations and the generalized auxiliary variable on the discrete level has been presented. The discrete energy stability of the proposed numerical scheme and the positivity of the computed auxiliary variable have been proved for general dissipative systems. The current method, termed gPAV (generalized Positive Auxiliary Variable), requires only the solution of linear algebraic equations within a time step. With appropriate choice of the operator in the algorithm, the resultant linear algebraic systems upon discretization involve only constant and time-independent coefficient matrices, which only need to be computed once and can be pre-computed. Several specific dissipative systems are studied in relative detail using the gPAV framework. Ample numerical experiments are presented to demonstrate the performance of the method, and the robustness of the scheme at large time step sizes.

1. Introduction

Dissipative systems are of immense interest to science and engineering. Physical systems encountered in the real world are dissipative, thanks to the second law of thermodynamics. In dissipative systems there exists a storage function that is bounded from below [40]. We will refer to this function as the energy in the current work. Dissipative systems are distinguished from general dynamical systems by the dissipation inequality, which basically states that the increase in storage of the system over a time interval cannot exceed the supply to the system during that interval [40,41]. The governing partial differential equations (PDE) describing dissipative systems are typically nonlinear, and they satisfy a balance equation for the energy (or entropy) as an embodiment of the dissipation inequality [11,35,2,32,1,15].

* Corresponding author.
E-mail address: sdong@purdue.edu (S. Dong).

https://doi.org/10.1016/j.jcp.2019.109121
0021-9991/

A highly desirable property for numerical algorithms for dissipative systems is the preservation of the energy dissipation (or conservation) on the discrete level. This not only preserves one important aspect of the underlying structure of the continuous system [23], but more practically also provides a control on the numerical stability in actual computer simulations. The history for such strategies is long and they can be traced to at least the work of [9] on discrete energy conservation for finite difference approximations in the 1920s. While energy-stable schemes for specific domains of science and engineering have been under intensive studies and these efforts have borne invaluable fruits, the schemes and methods developed usually have only limited applicability across domains. The energy-stable schemes for one area are hardly transferable to a different field, and they can hardly shed light on the development of such types of schemes in new unexplored domains. Unified techniques that can be broadly applied to treat different PDEs from different domains for devising energy-stable schemes are generally lacking. The metaphor used in [24] (page 139) to compare the motley collection of PDEs to a huge unhappy family (each unhappy in its own way; Tolstoy, "Anna Karenina") seems fitting in describing this situation (see also [6]).

Occasionally, certain methods appear and seem to be broadly applicable to a wide class of problems spanning different areas. The average vector field (AVF) method [6,36] and the discrete variational derivative method (DVDM) [19], both of which can be traced to the idea of discrete gradients [22,33], are two such examples. For gradient systems that can be expressed into the form $\frac{\partial \boldsymbol{u}}{\partial t} = \boldsymbol{L} \cdot \frac{\delta H}{\delta \boldsymbol{u}}$, where $\boldsymbol{L}$ is an anti-symmetric or negative semi-definite matrix, $\boldsymbol{u}$ is the field variable, $H(\boldsymbol{u})$ is the energy functional and $\frac{\delta H}{\delta \boldsymbol{u}}$ denotes the variational derivative, the AVF and DVDM methods can preserve the energy conservation (resp. energy dissipation) discretely. We refer the reader to e.g. [18,10,34,5,17] (among others) for related and variants of these methods. A potential drawback of these methods is their computational cost. Because these are fully implicit schemes and the governing PDEs are in general nonlinear, these methods will entail the solution of nonlinear algebraic equations on the discrete level. Consequently, some nonlinear algebraic solver (e.g. Newton's method) will be required for computing the field functions, and the associated computational cost can be substantial.

In the current work we present a framework for devising energy-stable schemes for general dissipative systems that can potentially be useful and applicable to different domains. Our method does not require the governing PDEs to be in any particular form, as long as they are dissipative (or conserving). When devising the energy-stable numerical schemes, we are particularly mindful of the computational cost involved therein. The resultant energy-stable schemes from our method involve only the solution of linear algebraic equations when computing the field functions within a time step, and no nonlinear algebraic solver is needed. Furthermore, with appropriate choice of the operator in the scheme, the resultant linear algebraic systems upon discretization can involve only constant and time-independent coefficient matrices, which only need to be computed once and can be pre-computed during pre-processing. Thanks to these properties, the presented method and the resultant energy-stable schemes are computationally very competitive and attractive. In terms of the computational cost the presented method enjoys a notable advantage when compared with the aforementioned methods.

The key to achieving the above useful properties for general dissipative systems in the presented method lies in the introduction of a generalized auxiliary variable. The generalized auxiliary variable introduced here is inspired by the scalar auxiliary variable (SAV) approach proposed by [38], and to a lesser extent, by the invariant energy quadratization (IEQ) method [44], both of which are devised for gradient flows; see also e.g. [37,20,8,49,27,28,47,45] (among others) for extensions and applications of these techniques. In SAV a scalar-valued auxiliary variable is defined, as the square root of the shifted potential energy integral. In IEQ an auxiliary field variable is defined, as the square root of the shifted potential energy density function. With these auxiliary variables, energy-stable schemes can be devised for gradient flows and their discrete energy stability can be proven in the SAV and IEQ methods. In both SAV and IEQ, the use of the square root function is critical to the proof of the discrete energy stability of the resultant numerical schemes, due to the interesting property that the square root is the only function form that satisfies the relation

$$2 f(x) f'(x) = 1.$$

In the current work we will show that the square root function is not essential to devising energy-stable schemes. In the generalized auxiliary variable method developed here, the auxiliary variable (a scalar number) can be defined by a rather general class of functions (conditions specifically given in Section 2.1) in terms of the energy functional, which is why the method is termed "generalized", and the resultant numerical schemes can be proven to be discretely energy stable.

The method presented here is applicable to general dissipative systems, which is another key difference from previous auxiliary-variable approaches. The ability to deal with general dissipative systems hinges on how the governing PDEs are treated based on the generalized auxiliary variable and how the generalized auxiliary variable is numerically treated on the discrete level. A unified procedure for treating discretely the dissipative governing equations and the generalized auxiliary variable has been presented. These numerical treatments have drawn inspirations from the recent developments in [29, 46] for incompressible Navier-Stokes equations and for the incompressible two-phase flows, which are not gradient-type systems.

The generalized auxiliary variable method proposed herein has another remarkable property: The computed values for the auxiliary variable are guaranteed to be positive on the discrete level. Such a property is not available in the SAV (or IEQ) method. In both SAV and IEQ, as well as in the current method, the auxiliary variable is computed discretely by solving an associated dynamic equation, which is derived based on the definition of the auxiliary variable in terms of the square root function in SAV and IEQ or a general function in the current method. The auxiliary variable physically should be positive

according to its definition. However, this positivity property is in general not guaranteed in the computed values for the auxiliary variable, because they are obtained by numerically solving a differential equation. Indeed, in numerical experiments we have observed negative values for the computed auxiliary variable using the previous methods, especially at large time step sizes. With the current method, on the other hand, we can prove that the computed values for the generalized auxiliary variable are guaranteed to be positive, regardless of the time step sizes or the external forces. The guaranteed positivity of the auxiliary variable in the current method is intimately related to and is critical to the proof of discrete energy stability of the proposed numerical schemes.

Because of these crucial properties, we will refer to the framework proposed herein as "gPAV", which stands for the generalized Positive Auxiliary Variable method.

In this paper we consider general dissipative systems and outline the gPAV procedure for devising discretely energy-stable schemes. The discrete energy stability of the proposed numerical scheme and the positivity property of the computed auxiliary variable will be proven for general dissipative systems. As already mentioned, the gPAV method requires only the solution of linear algebraic equations within a time step, and with appropriate choice of the operator in the algorithm, the resultant linear algebraic systems involve only constant and time-independent coefficient matrices that can be pre-computed. We demonstrate the gPAV procedure by looking into three specific dissipative systems: a chemo-repulsion model [21], the Cahn-Hilliard equation [4] with constant and variable mobility, and the nonlinear Klein-Gordon equation [39]. Ample numerical experiments are provided for each system to demonstrate the performance of the algorithm and the effects of the parameters.

The current work contains several new aspects: (i) the framework for developing discretely energy-stable schemes for general dissipative systems; (ii) the generalized auxiliary variable introduced herein; and (iii) the guaranteed positivity of the computed auxiliary variable on the discrete level. Some other aspects, such as the generalization of the numerical algorithm as discussed in Remarks 2.5 and 2.6, are also potentially useful to other researchers and the community.

The remainder of this paper is structured as follows. In Section 2 we introduce a generalized auxiliary variable and present the gPAV framework for devising discretely energy-stable schemes for general dissipative systems. The discrete energy stability of the presented algorithm and the positivity of the computed auxiliary variable will be proven. The solution algorithm for implementing the proposed energy-stable scheme will be presented. An alternative formulation for the energy-stable scheme will also be discussed in this section. Then in the three subsequent sections (Sections 3–5) we apply the gPAV framework to three specific dissipative systems (a chemo-repulsion model, Cahn-Hilliard equation with constant and variable mobility, and Klein-Gordon equation). Ample numerical experiments are provided to demonstrate the performance of the method for each system, and numerical results with large time step sizes are presented to show the robustness of the proposed scheme. Section 6 concludes the discussions with some closing remarks. In Appendix A we provide a method for approximating the variables for the first time step, which guarantees the positivity of the computed auxiliary variable to start off. This startup procedure is important for the proof of discrete energy stability of the presented numerical scheme.

2. The gPAV framework for energy-stable schemes for dissipative systems

Consider a domain Ω in two or three dimensions and a dissipative system on this domain, whose dynamics is described by,

$$\frac{\partial \boldsymbol{u}}{\partial t} = \boldsymbol{F}(\boldsymbol{u}) + \boldsymbol{f}(\boldsymbol{x}, t) \tag{2.1}$$

where $\boldsymbol{x}$ and t denote the spatial coordinate and time, $\boldsymbol{u}(\boldsymbol{x}, t)$ denotes the state variables of the system and can be a scalar- or vector-valued field function, and $\boldsymbol{f}(\boldsymbol{x}, t)$ is an external source term (hereafter referred to as the external force). $\boldsymbol{F}(\boldsymbol{u})$ is an operator that gives rise to the dissipative dynamics of the system and can be nonlinear in general. Equation (2.1) is supplemented by the boundary condition

$$\boldsymbol{B}(\boldsymbol{u}) = \boldsymbol{f}_b, \quad \text{on } \Gamma \tag{2.2}$$

where Γ denotes the domain boundary, $\boldsymbol{f}_b$ is an external source term on the boundary, which will be referred to as the external boundary force hereafter, and $\boldsymbol{B}$ is assumed to be a linear operator for the sake of simplicity. The initial condition is

$$\boldsymbol{u}(\boldsymbol{x}, t = 0) = \boldsymbol{u}_{in}(\boldsymbol{x}) \tag{2.3}$$

where $\boldsymbol{u}_{in}(\boldsymbol{x})$ is the initial distribution of the state variable.

Because the system is dissipative, there exists a storage function that is bounded from below [40], which hereafter will be referred to as the energy,

$$E_{tot}(t) = E_{tot}[\boldsymbol{u}] = \int_{\Omega} e(\boldsymbol{u}) d\Omega, \tag{2.4}$$

where $e(\boldsymbol{u})$ is the energy density function. The evolution of the energy is described by

$$\frac{dE_{tot}}{dt} = \int_\Omega e'(\boldsymbol{u}) \cdot \frac{\partial \boldsymbol{u}}{\partial t} d\Omega = \int_\Omega e'(\boldsymbol{u}) \cdot [\boldsymbol{F}(\boldsymbol{u}) + \boldsymbol{f}]\, d\Omega, \tag{2.5}$$

where $e'(\boldsymbol{u})$ is the partial derivative of $e(\boldsymbol{u})$ with respect to $\boldsymbol{u}$ ($e'(\boldsymbol{u}) = \frac{\partial e}{\partial \boldsymbol{u}}$) and we have used equation (2.1). With integration by part, the right-hand-side (RHS) of equation (2.5) can be transformed into

$$\int_\Omega e'(\boldsymbol{u}) \cdot [\boldsymbol{F}(\boldsymbol{u}) + \boldsymbol{f}] d\Omega = -\int_\Omega V(\boldsymbol{u}) d\Omega + \int_\Omega V_s(\boldsymbol{f}, \boldsymbol{u}) d\Omega + \int_\Gamma B_s(\boldsymbol{f}_b, \boldsymbol{u}) d\Gamma, \tag{2.6}$$

where $V_s(\boldsymbol{f}, \boldsymbol{u}) = e'(\boldsymbol{u}) \cdot \boldsymbol{f}$ denotes the volume terms involving the external force $\boldsymbol{f}$, which satisfies the property

$$V_s(\boldsymbol{f}, \boldsymbol{u}) = 0, \quad \text{if } \boldsymbol{f} = 0. \tag{2.7}$$

The rest of the volume terms are denoted by $-V(\boldsymbol{u})$, not involving $\boldsymbol{f}$. $B_s(\boldsymbol{f}_b, \boldsymbol{u})$ denotes the boundary terms, which may involve the boundary source term ($\boldsymbol{f}_b$) through the boundary conditions.

Substituting equation (2.6) into equation (2.5), we arrive at the following energy balance equation for the system,

$$\frac{dE_{tot}}{dt} = -\int_\Omega V(\boldsymbol{u}) d\Omega + \int_\Omega V_s(\boldsymbol{f}, \boldsymbol{u}) d\Omega + \int_\Gamma B_s(\boldsymbol{f}_b, \boldsymbol{u}) d\Gamma. \tag{2.8}$$

We assume that the boundary conditions (2.2) satisfy the following property,

$$B_s(\boldsymbol{f}_b, \boldsymbol{u}) = 0 \ \text{ if } \boldsymbol{f}_b = 0, \quad \text{on } \Gamma. \tag{2.9}$$

The dissipative nature of the system ensures that $\frac{dE_{tot}}{dt} \leqslant 0$ in the absence of the external forces (i.e. $\boldsymbol{f} = 0$ and $\boldsymbol{f}_b = 0$). Because the domain Ω can be arbitrary, it follows that $V(\boldsymbol{u})$ must be non-negative, i.e.

$$V(\boldsymbol{u}) \geqslant 0. \tag{2.10}$$

2.1. Reformulated equivalent system

To facilitate energy-stable numerical approximations of the system (2.1), we define a shifted energy of the following form

$$E(t) = E[\boldsymbol{u}] = \int_\Omega e(\boldsymbol{u}) d\Omega + C_0, \tag{2.11}$$

where C_0 is a chosen energy constant such that $E(t) > 0$ for $0 \leqslant t \leqslant T$, and T is the time interval on which the computation is to be carried out. Note that for a physical system the energy is bounded from below, and thus C_0 can always be found.

Let $\mathscr{F}$ denote a one-to-one increasing differentiable function, with its inverse $\mathscr{F}^{-1} = \mathscr{G}$, satisfying the property

$$\begin{cases} \mathscr{F}(\chi) > 0, & \text{for } \chi > 0; \\ \mathscr{G}(\chi) > 0, & \text{for } \chi > 0. \end{cases} \tag{2.12}$$

We define a scalar variable $R(t)$ by

$$R(t) = \mathscr{G}(E), \tag{2.13a}$$

$$E(t) = \mathscr{F}(R), \tag{2.13b}$$

where $E(t)$ is the shifted energy given by (2.11). $R(t)$ then satisfies the following evolution equation,

$$\mathscr{F}'(R) \frac{dR}{dt} = \int_\Omega e'(\boldsymbol{u}) \cdot \frac{\partial \boldsymbol{u}}{\partial t} d\Omega \tag{2.14}$$

which is obtained by taking the time derivative of equation (2.13b) and using equation (2.11).

Remark 2.1. The choice for $\mathscr{F}$ and $\mathscr{G}$ is rather general. Some examples are,

$$\mathscr{F}(\chi) = \chi^m, \quad \mathscr{G}(\chi) = \chi^{1/m}, \quad m \in \mathcal{Z}^+ = \{1, 2, 3, \ldots\}; \tag{2.15}$$

or

$$\mathscr{F}(\chi)=\frac{e_0}{2}\ln\left(\frac{\kappa_0+\chi}{\kappa_0-\chi}\right),\quad \mathscr{G}(\chi)=\kappa_0\tanh\left(\frac{\chi}{e_0}\right), \tag{2.16}$$

where κ_0 and e_0 are positive constants. It is important to notice that a function like $\mathscr{F}(\chi)=\chi^{2m+1}$ (with an integer $m\geqslant 0$) or $\mathscr{F}(\chi)=\ln(1+\chi)$ does not automatically guarantee that $\mathscr{F}(\chi)>0$ with arbitrary χ. However, if one can ensure that the argument satisfies $\chi>0$, the property $\mathscr{F}(\chi)>0$ can be guaranteed with such choices of functions when defining $R(t)$. This point is critical in the subsequent development of the numerical algorithm.

Noting that $\frac{\mathscr{F}(R)}{E}=1$, we rewrite equation (2.1) into an equivalent form

$$\frac{\partial \boldsymbol{u}}{\partial t}=\boldsymbol{F}_L(\boldsymbol{u})+\frac{\mathscr{F}(R)}{E}\Big(\boldsymbol{F}(\boldsymbol{u})-\boldsymbol{F}_L(\boldsymbol{u})\Big)+\boldsymbol{f}, \tag{2.17}$$

where $\boldsymbol{F}_L(\boldsymbol{u})$ is a chosen linear operator about $\boldsymbol{u}$. $\boldsymbol{F}_L(\boldsymbol{u})$ should be of the same spatial order as $\boldsymbol{F}(\boldsymbol{u})$. For improved accuracy $\boldsymbol{F}_L(\boldsymbol{u})$ should be an approximation of $\boldsymbol{F}(\boldsymbol{u})$ in some way, such as the linear component of $\boldsymbol{F}(\boldsymbol{u})$ or a linearized approximation of $\boldsymbol{F}(\boldsymbol{u})$. For improved numerical efficiency $\boldsymbol{F}_L(\boldsymbol{u})$ should be easy to compute and implement.

Remark 2.2. $\boldsymbol{F}(\boldsymbol{u})$ often consists of linear components and nonlinear components for many systems, and oftentimes one can choose the linear components as the $\boldsymbol{F}_L$ operator. One can also add/subtract certain linear operators, and treat one part freely and the other part together with $\frac{\mathscr{F}(R)}{E}$ as in equation (2.17). By choosing an $\boldsymbol{F}_L$ operator that involves only time-independent (or constant) coefficients, the resultant method will become computationally very efficient, because the coefficient matrices for the linear algebraic systems upon discretization will be time-independent and therefore can be pre-computed when solving the field variables. This point will become clearer from later discussions.

We reformulate equation (2.14) as follows,

$$\begin{aligned}\mathscr{F}'(R)\frac{dR}{dt}=&\int_\Omega e'(\boldsymbol{u})\cdot\frac{\partial \boldsymbol{u}}{\partial t}d\Omega+\left[\frac{\mathscr{F}(R)}{E}-1\right]\int_\Omega e'(\boldsymbol{u})\cdot[\boldsymbol{F}_L(\boldsymbol{u})+\boldsymbol{f}]d\Omega\\&+\frac{\mathscr{F}(R)}{E}\left(\int_\Omega e'(\boldsymbol{u})\cdot[\boldsymbol{F}(\boldsymbol{u})-\boldsymbol{F}_L(\boldsymbol{u})]d\Omega-\int_\Omega e'(\boldsymbol{u})\cdot[\boldsymbol{F}(\boldsymbol{u})-\boldsymbol{F}_L(\boldsymbol{u})]d\Omega\right)\\&+\left[1-\frac{\mathscr{F}(R)}{E}\right]\left|\int_\Omega V_s(\boldsymbol{f},\boldsymbol{u})d\Omega+\int_\Gamma B_s(\boldsymbol{f}_b,\boldsymbol{u})d\Gamma\right|\\=&\int_\Omega e'(\boldsymbol{u})\cdot\frac{\partial \boldsymbol{u}}{\partial t}d\Omega-\int_\Omega e'(\boldsymbol{u})\cdot\left(\boldsymbol{F}_L(\boldsymbol{u})+\frac{\mathscr{F}(R)}{E}[\boldsymbol{F}(\boldsymbol{u})-\boldsymbol{F}_L(\boldsymbol{u})]+\boldsymbol{f}\right)d\Omega\\&+\frac{\mathscr{F}(R)}{E}\int_\Omega e'(\boldsymbol{u})\cdot[\boldsymbol{F}(\boldsymbol{u})+\boldsymbol{f}]d\Omega+\left[1-\frac{\mathscr{F}(R)}{E}\right]\left|\int_\Omega V_s(\boldsymbol{f},\boldsymbol{u})d\Omega+\int_\Gamma B_s(\boldsymbol{f}_b,\boldsymbol{u})d\Gamma\right|\end{aligned} \tag{2.18}$$

where it can be noted that a number of zero terms have been incorporated. In the above equation $|(\cdot)|$ denotes the absolute value of $(\cdot)$. In light of (2.6), we transform equation (2.18) into the final reformulated equivalent form

$$\begin{aligned}\mathscr{F}'(R)\frac{dR}{dt}=&\int_\Omega e'(\boldsymbol{u})\cdot\frac{\partial \boldsymbol{u}}{\partial t}d\Omega-\int_\Omega e'(\boldsymbol{u})\cdot\left(\boldsymbol{F}_L(\boldsymbol{u})+\frac{\mathscr{F}(R)}{E}[\boldsymbol{F}(\boldsymbol{u})-\boldsymbol{F}_L(\boldsymbol{u})]+\boldsymbol{f}\right)d\Omega\\&+\frac{\mathscr{F}(R)}{E}\left[-\int_\Omega V(\boldsymbol{u})d\Omega+\int_\Omega V_s(\boldsymbol{f},\boldsymbol{u})d\Omega+\int_\Gamma B_s(\boldsymbol{f}_b,\boldsymbol{u})d\Gamma\right]\\&+\left[1-\frac{\mathscr{F}(R)}{E}\right]\left|\int_\Omega V_s(\boldsymbol{f},\boldsymbol{u})d\Omega+\int_\Gamma B_s(\boldsymbol{f}_b,\boldsymbol{u})d\Gamma\right|.\end{aligned} \tag{2.19}$$

The reformulated system consists of equations (2.17) and (2.19), the boundary conditions (2.2), the initial condition (2.3) for $\boldsymbol{u}$, and the following initial condition for $R(t)$,

$$R(0)=\mathscr{G}(E(0)),\quad \text{where } E(0)=\int_\Omega e(\boldsymbol{u}_{in})d\Omega+C_0. \tag{2.20}$$

In the reformulated system, the dynamic variables are $\boldsymbol{u}$ and $R(t)$, which are coupled in the equations (2.17) and (2.19). $E(t)$ is given by equation (2.11). Note that in this system $R(t)$ is determined by solving the coupled system of equations, not by using the equation (2.13a).

2.2. An energy-stable scheme

We next present an energy-stable scheme for the reformulated system consisting of (2.17) and (2.19), together with the boundary condition (2.2) and the initial conditions (2.3) and (2.20).

Let $n \geqslant 0$ denote the time step index, and $(\cdot)^n$ represent the variable $(\cdot)$ at time step n, corresponding to the time $t = n\Delta t$, where Δt is the time step size. If a real-valued parameter θ is involved, $(\cdot)^{n+\theta}$ represents the variable $(\cdot)$ at time step $(n+\theta)$, corresponding to the time $(n+\theta)\Delta t$.

Let χ denote a generic scalar or vector-valued variable. We consider the following second-order approximations:

$$\chi^{n+\frac{3}{2}} = \frac{3}{2}\chi^{n+1} - \frac{1}{2}\chi^{n}, \quad \chi^{n+\frac{1}{2}} = \frac{3}{2}\chi^{n} - \frac{1}{2}\chi^{n-1}, \tag{2.21a}$$

$$\frac{\partial \chi}{\partial t}\Big|^{n+1} = \frac{\chi^{n+\frac{3}{2}} - \chi^{n+\frac{1}{2}}}{\Delta t} = \frac{1}{\Delta t}\Big(\frac{3}{2}\chi^{n+1} - 2\chi^{n} + \frac{1}{2}\chi^{n-1}\Big), \tag{2.21b}$$

$$\bar{\chi}^{n+1} = 2\chi^{n} - \chi^{n-1}, \tag{2.21c}$$

where (2.21b) is the second-order backward differentiation formula (BDF) and $\bar{\chi}^{n+1}$ is an explicit approximation of χ^{n+1}. We also consider the following second-order approximation of $\frac{d\mathscr{F}(\chi)}{d\chi}\Big|^{n+1} = \mathscr{F}'(\chi)\Big|^{n+1}$ based on the discrete directional derivative [22]

$$D_{\mathscr{F}}(\chi)\Big|^{n+1} = \frac{\mathscr{F}(\chi^{n+\frac{3}{2}}) - \mathscr{F}(\chi^{n+\frac{1}{2}}) - \mathscr{F}'(\chi^{n+1})\cdot(\chi^{n+\frac{3}{2}} - \chi^{n+\frac{1}{2}})}{\|\chi^{n+\frac{3}{2}} - \chi^{n+\frac{1}{2}}\|^2}(\chi^{n+\frac{3}{2}} - \chi^{n+\frac{1}{2}}) + \mathscr{F}'(\chi^{n+1}), \tag{2.22}$$

which satisfies the property

$$D_{\mathscr{F}}(\chi)\Big|^{n+1}\cdot\Big(\frac{3}{2}\chi^{n+1} - 2\chi^{n} + \frac{1}{2}\chi^{n-1}\Big) = D_{\mathscr{F}}(\chi)\Big|^{n+1}\cdot\Big(\chi^{n+\frac{3}{2}} - \chi^{n+\frac{1}{2}}\Big) = \mathscr{F}(\chi^{n+\frac{3}{2}}) - \mathscr{F}(\chi^{n+\frac{1}{2}}). \tag{2.23}$$

Note that in these equations $\chi^{n+3/2}$ and $\chi^{n+1/2}$ are given by (2.21a). If χ represents a scalar-valued variable, one can also approximate $\mathscr{F}'(\chi)\Big|^{n+1}$ by

$$D_{\mathscr{F}}(\chi)\Big|^{n+1} = \frac{\mathscr{F}(\chi^{n+\frac{3}{2}}) - \mathscr{F}(\chi^{n+\frac{1}{2}})}{\chi^{n+\frac{3}{2}} - \chi^{n+\frac{1}{2}}} = \frac{\mathscr{F}(\chi^{n+\frac{3}{2}}) - \mathscr{F}(\chi^{n+\frac{1}{2}})}{\frac{3}{2}\chi^{n+1} - 2\chi^{n} + \frac{1}{2}\chi^{n-1}}, \tag{2.24}$$

which satisfies the same property (2.23).

We propose the following scheme to approximate the reformulated system:

$$\frac{\partial \boldsymbol{u}}{\partial t}\Big|^{n+1} = \boldsymbol{F}_L(\boldsymbol{u}^{n+1}) + \xi\Big[\boldsymbol{F}(\bar{\boldsymbol{u}}^{n+1}) - \boldsymbol{F}_L(\bar{\boldsymbol{u}}^{n+1})\Big] + \boldsymbol{f}^{n+1}, \tag{2.25a}$$

$$\xi = \frac{\mathscr{F}(R^{n+3/2})}{E[\tilde{\boldsymbol{u}}^{n+3/2}]}, \tag{2.25b}$$

$$E[\tilde{\boldsymbol{u}}^{n+3/2}] = \int_{\Omega} e(\tilde{\boldsymbol{u}}^{n+3/2})d\Omega + C_0, \tag{2.25c}$$

$$\boldsymbol{B}(\boldsymbol{u}^{n+1}) = \boldsymbol{f}_b^{n+1}, \quad \text{on } \Gamma, \tag{2.25d}$$

$$
\begin{aligned}
D_{\mathscr{F}}(R)\Big|^{n+1}\ \frac{dR}{dt}\Big|^{n+1} &= \int_\Omega e'(\boldsymbol{u}^{n+1})\cdot\frac{\partial \boldsymbol{u}}{\partial t}\Big|^{n+1} d\Omega \\
&- \int_\Omega e'(\boldsymbol{u}^{n+1})\cdot\Big(\boldsymbol{F}_L(\boldsymbol{u}^{n+1}) + \xi\Big[\boldsymbol{F}(\bar{\boldsymbol{u}}^{n+1}) - \boldsymbol{F}_L(\bar{\boldsymbol{u}}^{n+1})\Big] + \boldsymbol{f}^{n+1}\Big) d\Omega \\
&+ \xi\left[-\int_\Omega V(\tilde{\boldsymbol{u}}^{n+1})d\Omega + \int_\Omega V_s(\boldsymbol{f}^{n+1},\tilde{\boldsymbol{u}}^{n+1})d\Omega + \int_\Gamma B_s(\boldsymbol{f}_b^{n+1},\tilde{\boldsymbol{u}}^{n+1})d\Gamma\right] \\
&+ (1-\xi)\left|\int_\Omega V_s(\boldsymbol{f}^{n+1},\tilde{\boldsymbol{u}}^{n+1})d\Omega + \int_\Gamma B_s(\boldsymbol{f}_b^{n+1},\tilde{\boldsymbol{u}}^{n+1})d\Gamma\right|.
\end{aligned}
\tag{2.25e}
$$

In the above equations, $\frac{\partial \boldsymbol{u}}{\partial t}\big|^{n+1}$ and $\frac{dR}{dt}\Big|^{n+1}$ are defined by (2.21b), $D_{\mathscr{F}}(R)|^{n+1}$ is defined by (2.22) (or (2.24)), $\tilde{\boldsymbol{u}}^{n+1}$ is defined by (2.21c), and $R^{n+3/2}$ is defined by (2.21a). $\tilde{\boldsymbol{u}}^{n+1}$ and $\tilde{\boldsymbol{u}}^{n+3/2}$ are second-order approximations of $\boldsymbol{u}^{n+1}$ and $\boldsymbol{u}^{n+3/2}$, respectively, to be specifically defined later in (2.42).

Remark 2.3. It is critical to note that in the scheme (2.25a)–(2.25e), $\frac{\mathscr{F}(R)}{E[\boldsymbol{u}]}$ is approximated at step $(n+\frac{3}{2})$ while the other variables are approximated at step $(n+1)$. This feature, together with the approximation (2.24), allows R^{n+1} to be computed from a linear algebraic equation (no nonlinear algebraic solver), and endows the scheme with the property that the computed R^{n+1} and $\mathscr{F}(R^{n+1})$ (resp. $R^{n+3/2}$ and $\mathscr{F}(R^{n+3/2})$, for all $n\geqslant 0$) are guaranteed to be positive. These points will become clear from later discussions. It should be noted that the approximation $\frac{\mathscr{F}(R^{n+3/2})}{E[\tilde{\boldsymbol{u}}^{n+3/2}]}$ at step $(n+3/2)$ is a second-order approximation of $\frac{\mathscr{F}(R)}{E}=1$. In fact, the approximation involving any real parameter θ,

$$
\frac{\mathscr{F}(R^{n+\theta})}{E[\tilde{\boldsymbol{u}}^{n+\theta}]} = 1+\mathcal{O}(\Delta t)^2,
\tag{2.26}
$$

is a second-order approximation of $\frac{\mathscr{F}(R)}{E}=1$, as long as $R^{n+\theta}$ and $\tilde{\boldsymbol{u}}^{n+\theta}$ are second-order approximations of $R(t)$ and $\boldsymbol{u}(t)$ at time $(n+\theta)\Delta t$. Therefore, the approximation in (2.25b) does not affect the second-order accuracy of the scheme.

The scheme given by (2.25a)–(2.25e) has the following property.

Theorem 2.1. *In the absence of the external force and external boundary force (i.e. $\boldsymbol{f}=\boldsymbol{0}$ and $\boldsymbol{f}_b=0$), the following relation holds with the scheme (2.25):*

$$
\mathscr{F}(R^{n+\frac{3}{2}}) - \mathscr{F}(R^{n+\frac{1}{2}}) = -\Delta t\frac{\mathscr{F}(R^{n+\frac{3}{2}})}{E[\tilde{\boldsymbol{u}}^{n+3/2}]}\int_\Omega V(\tilde{\boldsymbol{u}}^{n+1}) \leqslant 0, \quad \text{for } n\geqslant 0,
\tag{2.27}
$$

if the approximation of $R(t)$ at time step $\frac{1}{2}$ is positive, i.e. $Y_0 = R^{n+1/2}\big|_{n=0} > 0$.

Proof. By equations (2.21b) and (2.22), we have

$$
D_{\mathscr{F}}(R)\Big|^{n+1}\ \frac{dR}{dt}\Big|^{n+1} = \frac{\mathscr{F}(R^{n+\frac{3}{2}}) - \mathscr{F}(R^{n+\frac{1}{2}})}{\Delta t}.
\tag{2.28}
$$

Taking the L^2 inner product between equation (2.25a) and $e'(\boldsymbol{u}^{n+1})$, and adding the resultant equation to equation (2.25e) and noting equation (2.28), we arrive at

$$
\mathscr{F}(R^{n+\frac{3}{2}}) - \mathscr{F}(R^{n+\frac{1}{2}}) = -\Delta t\frac{\mathscr{F}(R^{n+\frac{3}{2}})}{E[\tilde{\boldsymbol{u}}^{n+3/2}]}\int_\Omega V(\tilde{\boldsymbol{u}}^{n+1}) + \left(1-\frac{\mathscr{F}(R^{n+\frac{3}{2}})}{E[\tilde{\boldsymbol{u}}^{n+3/2}]}\right)|S_0|\,\Delta t + \frac{\mathscr{F}(R^{n+\frac{3}{2}})}{E[\tilde{\boldsymbol{u}}^{n+3/2}]}S_0\Delta t,
\tag{2.29}
$$

where we have used equation (2.25b), and S_0 is defined by

$$
S_0 = \int_\Omega V_s(\boldsymbol{f}^{n+1},\tilde{\boldsymbol{u}}^{n+1})d\Omega + \int_\Gamma B_s(\boldsymbol{f}_b^{n+1},\tilde{\boldsymbol{u}}^{n+1})d\Gamma.
\tag{2.30}
$$

Then it follows that, if $\boldsymbol{f}=0$ and $\boldsymbol{f}_b=0$,

$$\mathscr{F}(R^{n+3/2})=\frac{\mathscr{F}(R^{n+\frac{1}{2}})}{1+\frac{\Delta t}{E[\tilde{\boldsymbol{u}}^{n+3/2}]}\int_\Omega V(\tilde{\boldsymbol{u}}^{n+1})d\Omega} \tag{2.31}$$

where we have used the relations (2.7) and (2.9).

Note that $E[\tilde{\boldsymbol{u}}^{n+3/2}]>0$ and $V(\tilde{\boldsymbol{u}}^{n+1})\geqslant 0$, in light of (2.11) and (2.10). If $Y_0=R^{n+1/2}|_{n=0}>0$, then $\mathscr{F}(Y_0)>0$ based on the property (2.12). By induction, we can conclude from equation (2.31) that $\mathscr{F}(R^{n+3/2})>0$ for all $n\geqslant 0$. The inequality in (2.27) then holds. We therefore conclude that, if $R^{n+1/2}|_{n=0}>0$,

$$0<\mathscr{F}(R^{n+\frac{3}{2}})\leqslant\mathscr{F}(R^{n+\frac{1}{2}}),\quad \text{for } n\geqslant 0. \tag{2.32}$$

Thus, the scheme is unconditionally energy stable with respect to the modified energy $\mathscr{F}(R)$, if the approximation of $R(t)$ at time step $\frac{1}{2}$ is positive. □

There are many ways to approximate $R(t)$ to ensure that it is positive at time step $\frac{1}{2}$ and that the overall scheme is second-order accurate in time. One such method is given in the Appendix A. Therefore we have the following result:

Theorem 2.2. *With $\boldsymbol{u}^1$ and R^1 approximated using the method from Appendix A, in the absence of external forces ($\boldsymbol{f}=0$ and $\boldsymbol{f}_b=0$), the scheme represented by (2.25a)–(2.25e) is unconditionally energy-stable in the sense of the relation (2.32).*

Remark 2.4. If the functional form of $\mathscr{F}(\chi)$ is such that $\mathscr{F}(\chi)\geqslant 0$ for all $\chi\in(-\infty,\infty)$, e.g. $\mathscr{F}(\chi)=\chi^{2m}$ (with an integer $m\geqslant 1$), then the scheme given by (2.25a)–(2.25e) is unconditionally energy stable regardless of the approximation of $R(t)$ at the time step $\frac{1}{2}$.

Remark 2.5. The scheme (2.25) is devised by enforcing the system of equations consisting of (2.17), (2.19) and (2.2) at time step $(n+1)$, approximating $\frac{\mathscr{F}(R)}{E}$ at time step $(n+\frac{3}{2})$, and employing the approximations (2.21a)–(2.22). Inspired by the recent work [47], we can generalize this scheme by enforcing the system of equations at time step $(n+\theta)$, where θ is a real-valued parameter, to arrive at a family of energy-stable schemes.

In brief, let us consider the following second-order approximations at time step $(n+\theta)$ with $\theta\geqslant\frac{1}{2}$: ($\chi$ denoting a generic variable, and $\beta\geqslant 0$ denoting a real parameter below)

$$\chi^{n+\theta+\frac{1}{2}}=\left(\theta+\frac{1}{2}\right)\chi^{n+1}-\left(\theta-\frac{1}{2}\right)\chi^{n},\qquad \chi^{n+\theta-\frac{1}{2}}=\left(\theta+\frac{1}{2}\right)\chi^{n}-\left(\theta-\frac{1}{2}\right)\chi^{n-1}; \tag{2.33a}$$

$$\begin{aligned}\chi^{n+\theta}=&\frac{1}{2}(\chi^{n+\theta+\frac{1}{2}}+\chi^{n+\theta-\frac{1}{2}})+\beta(\chi^{n+1}-2\chi^{n}+\chi^{n-1})\\ =&\left(\beta+\frac{\theta}{2}+\frac{1}{4}\right)\chi^{n+1}+\left(\frac{1}{2}-2\beta\right)\chi^{n}+\left(\beta-\frac{\theta}{2}+\frac{1}{4}\right)\chi^{n-1};\quad \text{(implicit approximation)}\end{aligned} \tag{2.33b}$$

$$\tilde{\chi}^{n+\theta}=(1+\theta)\chi^{n}-\theta\chi^{n-1};\quad \text{(explicit approximation)} \tag{2.33c}$$

$$\left.\frac{\partial\chi}{\partial t}\right|^{n+\theta}=\frac{\chi^{n+\theta+\frac{1}{2}}-\chi^{n+\theta-\frac{1}{2}}}{\Delta t}=\frac{1}{\Delta t}\left[\left(\theta+\frac{1}{2}\right)\chi^{n+1}-2\theta\chi^{n}+\left(\theta-\frac{1}{2}\right)\chi^{n-1}\right]; \tag{2.33d}$$

and the following approximation of $\left.\frac{d\mathscr{F}(\chi)}{d\chi}\right|^{n+\theta}=\mathscr{F}'(\chi)|^{n+\theta}$ based on discrete directional derivative,

$$\begin{aligned}D_{\mathscr{F}}(\chi)|^{n+\theta}=&\frac{\mathscr{F}(\chi^{n+\theta+\frac{1}{2}})-\mathscr{F}(\chi^{n+\theta-\frac{1}{2}})-\mathscr{F}'(\chi^{n+\theta})\cdot(\chi^{n+\theta+\frac{1}{2}}-\chi^{n+\theta-\frac{1}{2}})}{\|\chi^{n+\theta+\frac{1}{2}}-\chi^{n+\theta-\frac{1}{2}}\|^2}(\chi^{n+\theta+\frac{1}{2}}-\chi^{n+\theta-\frac{1}{2}})\\ &+\mathscr{F}'(\chi^{n+\theta}).\end{aligned} \tag{2.34}$$

These approximations satisfy the following properties:

$$\begin{aligned}\chi^{n+\theta}\left[\left(\theta+\frac{1}{2}\right)\chi^{n+1}-2\theta\chi^{n}+\left(\theta-\frac{1}{2}\right)\chi^{n-1}\right]=&\frac{1}{2}\left(\left|\chi^{n+\theta+\frac{1}{2}}\right|^2-\left|\chi^{n+\theta-\frac{1}{2}}\right|^2\right)\\ &+\frac{\beta}{2}\left(\left|\chi^{n+1}-\chi^{n}\right|^2-\left|\chi^{n}-\chi^{n-1}\right|^2\right)+\theta\beta\left|\chi^{n+1}-2\chi^{n}+\chi^{n-1}\right|^2;\end{aligned} \tag{2.35a}$$

$$D_{\mathscr{F}}(\chi)|^{n+\theta}\left[\left(\theta+\frac{1}{2}\right)\chi^{n+1}-2\theta\chi^{n}+\left(\theta-\frac{1}{2}\right)\chi^{n-1}\right]=\mathscr{F}(\chi^{n+\theta+\frac{1}{2}})-\mathscr{F}(\chi^{n+\theta-\frac{1}{2}}). \tag{2.35b}$$

Note that the parameter $\beta \geqslant 0$ in (2.33b) can often be used to control the numerical dissipation of the approximations, which will be useful for approximating energy-conserving systems. An example will be given with the Klein-Gordon equation in a later section. The scheme given in (2.25) corresponds to $\theta = 1$ and $\beta = \frac{1}{4}$.

By approximating the terms in equations (2.17), (2.19) and (2.2) at time step $(n+\theta)$, except for the term $\frac{\mathscr{F}(R)}{E}$, which will be approximated at time step $(n+\theta+\frac{1}{2})$, and employing the approximations (2.33a)–(2.34), one can prove that the resultant family of schemes (with θ and β as parameters) is unconditionally energy-stable. The details will not be provided here.

2.3. Solution algorithm

Let us now consider how to implement the algorithm represented by equations (2.25a)-(2.25e). We first introduce some notations (χ again denoting a generic variable):

$$\gamma_0 = \frac{3}{2}, \quad \hat{\chi} = 2\chi^n - \frac{1}{2}\chi^{n-1}. \tag{2.36}$$

Then the approximation in (2.21b) can be written as

$$\left.\frac{\partial \chi}{\partial t}\right|^{n+1} = \frac{\gamma_0 \chi^{n+1} - \hat{\chi}}{\Delta t}. \tag{2.37}$$

Inserting notation (2.37) into equation (2.25a), we have

$$\frac{\gamma_0}{\Delta t}\boldsymbol{u}^{n+1} - \boldsymbol{F}_L(\boldsymbol{u}^{n+1}) = \xi\Big[\boldsymbol{F}(\bar{\boldsymbol{u}}^{n+1}) - \boldsymbol{F}_L(\bar{\boldsymbol{u}}^{n+1})\Big] + \boldsymbol{f}^{n+1} + \frac{\hat{\boldsymbol{u}}}{\Delta t}. \tag{2.38}$$

Note that $\bar{\boldsymbol{u}}^{n+1}$ and $\hat{\boldsymbol{u}}$ are both explicitly known, and ξ is an unknown depending on $\boldsymbol{u}^{n+1}$. Taking advantage of the fact that ξ is a scalar number instead of a field function and the linearity of the operator $\boldsymbol{B}$ in the boundary condition (2.2), we introduce two field functions $(\boldsymbol{u}_1^{n+1}, \boldsymbol{u}_2^{n+1})$ as solutions to the following two linear systems:

$$\frac{\gamma_0}{\Delta t}\boldsymbol{u}_1^{n+1} - \boldsymbol{F}_L(\boldsymbol{u}_1^{n+1}) = \frac{\hat{\boldsymbol{u}}}{\Delta t} + \boldsymbol{f}^{n+1}, \tag{2.39a}$$

$$\boldsymbol{B}(u_1^{n+1}) = \boldsymbol{f}_b^{n+1}, \quad \text{on } \Gamma. \tag{2.39b}$$

$$\frac{\gamma_0}{\Delta t}\boldsymbol{u}_2^{n+1} - \boldsymbol{F}_L(\boldsymbol{u}_2^{n+1}) = \boldsymbol{F}(\bar{\boldsymbol{u}}^{n+1}) - \boldsymbol{F}_L(\bar{\boldsymbol{u}}^{n+1}). \tag{2.40a}$$

$$\boldsymbol{B}(\boldsymbol{u}_2^{n+1}) = 0, \quad \text{on } \Gamma. \tag{2.40b}$$

Since the operator $\boldsymbol{F}_L$ is chosen to be a linear operator and relatively easy to compute, $\boldsymbol{u}_1^{n+1}$ and $\boldsymbol{u}_2^{n+1}$ can be solved efficiently from these equations. Then we have the following result.

Theorem 2.3. *Given scalar value ξ, the following function solves the system consisting of equations (2.25a) and (2.25d):*

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}_1^{n+1} + \xi \boldsymbol{u}_2^{n+1}, \tag{2.41}$$

where $\boldsymbol{u}_1^{n+1}$ and $\boldsymbol{u}_2^{n+1}$ are given by the equations (2.39a)-(2.40b).

The scalar value ξ still needs to be determined. Define

$$\begin{cases} \tilde{\boldsymbol{u}}^{n+1} = \boldsymbol{u}_1^{n+1} + \boldsymbol{u}_2^{n+1}, \\ \tilde{\boldsymbol{u}}^{n+3/2} = \dfrac{3}{2}\tilde{\boldsymbol{u}}^{n+1} - \dfrac{1}{2}\boldsymbol{u}^n, \end{cases} \tag{2.42}$$

which are second-order approximations of $\boldsymbol{u}^{n+1}$ and $\boldsymbol{u}^{n+3/2}$. These field variables can be explicitly computed after $\boldsymbol{u}_1^{n+1}$ and $\boldsymbol{u}_2^{n+1}$ are obtained. By equation (2.25b), we have

$$\mathscr{F}(R^{n+\frac{3}{2}}) = \xi E[\tilde{\boldsymbol{u}}^{n+\frac{3}{2}}]. \tag{2.43}$$

Note that equation (2.25e) can be transformed into equation (2.29). Inserting equation (2.43) into equation (2.29) leads to the solution for ξ,

$$\xi = \frac{\mathscr{F}(R^{n+1/2}) + \Delta t|S_0|}{E[\tilde{\boldsymbol{u}}^{n+\frac{3}{2}}] + \Delta t \int_\Omega V(\tilde{\boldsymbol{u}}^{n+1}) + \Delta t(|S_0| - S_0)}, \tag{2.44}$$

where $\tilde{\boldsymbol{u}}^{n+1}$ and $\tilde{\boldsymbol{u}}^{n+3/2}$ are given by (2.42), S_0 is given by equation (2.30), and $E[\tilde{\boldsymbol{u}}^{n+3/2}]$ is computed by equation (2.25c).

In light of equations (2.43) and (2.21a), we can then compute R^{n+1} by

$$\begin{cases} R^{n+3/2} = \mathscr{G}\left(\xi E[\tilde{\boldsymbol{u}}^{n+3/2}]\right), \quad n \geqslant 0; \\ R^{n+1} = \dfrac{2}{3}R^{n+3/2} + \dfrac{1}{3}R^{n}, \quad n \geqslant 0. \end{cases} \tag{2.45}$$

The following result holds.

Theorem 2.4. *The scalar value ξ computed by equation (2.44) and the variable R^{n+1} ($n \geqslant 0$) computed by equation (2.45) are always positive, if the approximation of $R(t)$ at time step $\frac{1}{2}$ is positive, i.e. $Y_0 = R^{n+1/2}|_{n=0} > 0$.*

Proof. If $Y_0 = R^{n+1/2}|_{n=0} > 0$, then $\mathscr{F}(Y_0) > 0$ based on (2.12). Since $E(\boldsymbol{u})$ is a positive function, $V(\boldsymbol{u}) \geqslant 0$ and $|S_0| - S_0 \geqslant 0$, we conclude by induction ξ computed from (2.44) is always positive.

Note that $R^0 = R(0) > 0$ according to equation (2.20). In light of the property (2.12), we conclude that $R^{n+3/2}$ and R^{n+1} computed from equation (2.45) are both positive. □

Using the method from the Appendix A can ensure the positiveness of the approximation of $R(t)$ at the time step $\frac{1}{2}$. We have the following result.

Theorem 2.5. *With $\boldsymbol{u}^1$ and R^1 computed based on the method from Appendix A, the ξ given by (2.44) and R^{n+1} and $R^{n+3/2}$ given by (2.45) satisfy the property*

$$\xi > 0, \quad R^{n+1} > 0, \quad \text{and} \quad R^{n+3/2} > 0, \tag{2.46}$$

for all $n \geqslant 0$, regardless of the external forces $\boldsymbol{f}$ and $\boldsymbol{f}_b$ and the time step size Δt.

Combining the above discussions, we arrive at the solution procedure for solving the system consisting of equations (2.25a)-(2.25e). Given $(\boldsymbol{u}^n, R^n)$, we compute $(\boldsymbol{u}^{n+1}, R^{n+1})$ through the following steps:

1. Solve equations (2.39a)–(2.39b) for $\boldsymbol{u}_1^{n+1}$;
   Solve equations (2.40a)–(2.40b) for $\boldsymbol{u}_2^{n+1}$.
2. Compute $\tilde{\boldsymbol{u}}^{n+1}$ and $\tilde{\boldsymbol{u}}^{n+3/2}$ based on equation (2.42);
   Compute $E[\tilde{\boldsymbol{u}}^{n+\frac{3}{2}}]$, $\int_\Omega V(\tilde{\boldsymbol{u}}^{n+1})$ and S_0 based on equations (2.11), (2.6) and (2.30).
3. Compute ξ based on equation (2.44).
4. Compute $\boldsymbol{u}^{n+1}$ based on equation (2.41). Compute R^{n+1} based on equation (2.45).

It can be noted that the numerical scheme and the solution algorithm developed in this section has several attractive properties: (i) Only linear systems need to be solved for the field variables $\boldsymbol{u}$ within a time step. Moreover, with appropriate choice for the $\boldsymbol{F}_L$ operator, the system can involve only constant and time-independent coefficient matrices, which can be pre-computed. Therefore, the solution for $\boldsymbol{u}$ will be computationally very efficient. (ii) The auxiliary variables R and ξ can be computed by a well-defined explicit formula, and no nonlinear algebraic solver is involved. Their computed values are guaranteed to be positive. (iii) The auxiliary variable R can be defined by a rather general class of functions ($\mathscr{F}$ and $\mathscr{G}$) using the method developed here. (iv) The scheme is unconditionally energy-stable for general dissipative systems.

2.4. An alternative formulation and energy-stable scheme

The numerical formulation presented in the previous subsections is not the only way to devise energy-stable schemes for dissipative systems. In this subsection we outline an alternative formulation and associated energy-stable scheme. The process is analogous to the developments in the sections 2.1–2.3. So many details will be omitted in the following discussions.

The main idea with the alternative formulation is to realize that $\frac{R(t)}{\mathscr{G}(E)} = 1$ with the auxiliary variable $R(t)$ defined in (2.13a). Therefore, one can potentially employ $\frac{R}{\mathscr{G}(E)}$, instead of $\frac{\mathscr{F}(R)}{E}$, in the numerical formulations. With appropriate reformulation and treatments of different terms, it turns out that a discretely energy-stable scheme can be obtained with similar attractive properties, such as the guaranteed positiveness of the computed values for the variable $R(t)$.

Note that $R(t)$ is defined by (2.13a), where $\mathscr{G}$ is a one-to-one increasing differentiable function with $\mathscr{G}(\chi) > 0$ and $\mathscr{G}'(\chi) > 0$ for $\chi > 0$. $R(t)$ satisfies the following dynamic equation

$$\frac{dR}{dt} = \mathscr{G}'(E) \int_\Omega e'(\boldsymbol{u}) \cdot \frac{\partial \boldsymbol{u}}{\partial t} d\Omega, \tag{2.47}$$

where $E(t)$ is defined by (2.11).

We reformulate equation (2.1) into

$$\frac{\partial \boldsymbol{u}}{\partial t} = \boldsymbol{F}_L(\boldsymbol{u}) + \frac{R}{\mathscr{G}(E)}\Big(\boldsymbol{F}(\boldsymbol{u}) - \boldsymbol{F}_L(\boldsymbol{u})\Big) + \boldsymbol{f}, \tag{2.48}$$

where the notations follow those defined in previous subsections. Analogously, by incorporating appropriate zero terms we can transform (2.47) into

$$\begin{aligned}\frac{dR}{dt} =& \mathscr{G}'(E)\int_\Omega e'(\boldsymbol{u})\cdot\frac{\partial \boldsymbol{u}}{\partial t}d\Omega - \mathscr{G}'(E)\int_\Omega e'(\boldsymbol{u})\cdot\left(\boldsymbol{F}_L(\boldsymbol{u}) + \frac{R}{\mathscr{G}(E)}\left[\boldsymbol{F}(\boldsymbol{u}) - \boldsymbol{F}_L(\boldsymbol{u})\right] + \boldsymbol{f}\right)d\Omega \\ &+ \frac{R}{\mathscr{G}(E)}\mathscr{G}'(E)\left[-\int_\Omega V(\boldsymbol{u})d\Omega + \int_\Omega V_s(\boldsymbol{f},\boldsymbol{u})d\Omega + \int_\Gamma B_s(\boldsymbol{f}_b,\boldsymbol{u})d\Gamma\right] \\ &+ \left[1 - \frac{R}{\mathscr{G}(E)}\right]\mathscr{G}'(E)\left|\int_\Omega V_s(\boldsymbol{f},\boldsymbol{u})d\Omega + \int_\Gamma B_s(\boldsymbol{f}_b,\boldsymbol{u})d\Gamma\right|.\end{aligned} \tag{2.49}$$

The reformulated system now consists of equations (2.48) and (2.49), the boundary condition (2.2), and the initial conditions (2.3) and (2.20).

We discretize the reformulated system as follows:

$$\frac{1}{\Delta t}\left(\frac{3}{2}\boldsymbol{u}^{n+1} - 2\boldsymbol{u}^n + \frac{1}{2}\boldsymbol{u}^{n-1}\right) = \boldsymbol{F}_L(\boldsymbol{u}^{n+1}) + \xi\Big[\boldsymbol{F}(\bar{\boldsymbol{u}}^{n+1}) - \boldsymbol{F}_L(\bar{\boldsymbol{u}}^{n+1})\Big] + \boldsymbol{f}^{n+1}, \tag{2.50a}$$

$$\xi = \frac{R^{n+3/2}}{\mathscr{G}(E[\tilde{\boldsymbol{u}}^{n+3/2}])}, \tag{2.50b}$$

$$E[\tilde{\boldsymbol{u}}^{n+3/2}] = \int_\Omega e(\tilde{\boldsymbol{u}}^{n+3/2})d\Omega + C_0, \tag{2.50c}$$

$$\boldsymbol{B}(\boldsymbol{u}^{n+1}) = \boldsymbol{f}_b^{n+1}, \quad \text{on } \Gamma, \tag{2.50d}$$

$$\begin{aligned}\frac{R^{n+3/2} - R^{n+1/2}}{\Delta t} =& \mathscr{G}'(E[\tilde{u}^{n+1}])\left\{\int_\Omega e'(\boldsymbol{u}^{n+1})\cdot\frac{\frac{3}{2}\boldsymbol{u}^{n+1} - 2\boldsymbol{u}^n + \frac{1}{2}\boldsymbol{u}^{n-1}}{\Delta t}d\Omega\right. \\ &- \int_\Omega e'(\boldsymbol{u}^{n+1})\cdot\left(\boldsymbol{F}_L(\boldsymbol{u}^{n+1}) + \xi\Big[\boldsymbol{F}(\bar{\boldsymbol{u}}^{n+1}) - \boldsymbol{F}_L(\bar{\boldsymbol{u}}^{n+1})\Big] + \boldsymbol{f}^{n+1}\right)d\Omega \\ &+ \xi\left[-\int_\Omega V(\tilde{\boldsymbol{u}}^{n+1})d\Omega + \int_\Omega V_s(\boldsymbol{f}^{n+1},\tilde{\boldsymbol{u}}^{n+1})d\Omega + \int_\Gamma B_s(\boldsymbol{f}_b^{n+1},\tilde{\boldsymbol{u}}^{n+1})d\Gamma\right] \\ &\left.+ (1-\xi)\left|\int_\Omega V_s(\boldsymbol{f}^{n+1},\tilde{\boldsymbol{u}}^{n+1})d\Omega + \int_\Gamma B_s(\boldsymbol{f}_b^{n+1},\tilde{\boldsymbol{u}}^{n+1})d\Gamma\right|\right\}.\end{aligned} \tag{2.50e}$$

In these equations $\bar{\boldsymbol{u}}^{n+1}$ is defined by (2.21c), $R^{n+3/2}$ and $R^{n+1/2}$ are defined by (2.21a), and $\tilde{\boldsymbol{u}}^{n+1}$ and $\tilde{\boldsymbol{u}}^{n+3/2}$ are second-order approximations of $\boldsymbol{u}^{n+1}$ and $\boldsymbol{u}^{n+3/2}$ respectively to be specified later.

Taking the L^2 inner product between $\mathscr{G}'(E[\tilde{u}^{n+1}])e'(\boldsymbol{u}^{n+1})$ and equation (2.50a), and summing up the resultant equation and equation (2.50e), we get

$$R^{n+3/2} - R^{n+1/2} = \Delta t\mathscr{G}'(E[\tilde{u}^{n+1}])\left[-\xi\int_\Omega V(\tilde{u}^{n+1})d\Omega + (1-\xi)|S_0| + \xi S_0\right] \tag{2.51}$$

where S_0 is given by the equation (2.30). In the absence of external forces ($\boldsymbol{f} = 0$ and $\boldsymbol{f}_b = 0$), $S_0 = 0$ and equation (2.51) leads to

$$R^{n+3/2} = \frac{R^{n+1/2}}{1 + \Delta t\frac{\mathscr{G}'(E[\tilde{u}^{n+1}])}{\mathscr{G}(E[\tilde{u}^{n+3/2}])}\int_\Omega V(\tilde{\boldsymbol{u}}^{n+1})d\Omega} \tag{2.52}$$

where we have used (2.50b). Note that $E[\tilde{\boldsymbol{u}}^{n+3/2}] > 0$, $E[\tilde{\boldsymbol{u}}^{n+1}] > 0$, $V(\tilde{\boldsymbol{u}}^{n+1}) \geqslant 0$, and that $\mathscr{G}(\chi) > 0$ and $\mathscr{G}'(\chi) > 0$ for $\chi > 0$. By induction we can conclude from (2.52) that $R^{n+3/2} \geqslant 0$ (for all $n \geqslant 0$) if the approximation of $R(t)$ at time step $\frac{1}{2}$ is non-negative. Equation (2.51) then leads to the following result.

Theorem 2.6. *In the absence of external forces ($\boldsymbol{f} = 0$ and $\boldsymbol{f}_b = 0$), if the approximation of $R(t)$ at time step $\frac{1}{2}$ is non-negative, the scheme given by (2.50a)–(2.50e) is unconditionally energy-stable in the sense that*

$$0 \leqslant R^{n+3/2} \leqslant R^{n+1/2}, \quad \text{for all } n \geqslant 0. \tag{2.53}$$

In the Appendix A, we have presented a method for computing the first time step, which can ensure that the approximation of $R(t)$ at step $\frac{1}{2}$ is positive. This leads to the following result.

Theorem 2.7. *In the absence of external forces ($\boldsymbol{f} = 0$ and $\boldsymbol{f}_b = 0$), when the first time step is approximated using the method from Appendix A, the numerical scheme given by (2.50a)–(2.50e) is unconditionally energy-stable in the sense of equation (2.53).*

The scheme represented by (2.50a)–(2.50e) can be implemented in a similar way to that of Section 2.3, with the following steps:

- Compute $\boldsymbol{u}_1^{n+1}$ and $\boldsymbol{u}_2^{n+1}$ by solving equations (2.39a)–(2.40b).
- Define $\tilde{\boldsymbol{u}}^{n+1}$ and $\tilde{\boldsymbol{u}}^{n+3/2}$ again by equations (2.42). These variables can be computed.
- Compute ξ based on equation (2.51), specifically by

$$\xi = \frac{R^{n+1/2} + \Delta t |S_0| \mathscr{G}'(E[\tilde{\boldsymbol{u}}^{n+1}])}{\mathscr{G}(E[\tilde{\boldsymbol{u}}^{n+3/2}]) + \Delta t \mathscr{G}'(E[\tilde{\boldsymbol{u}}^{n+1}]) \left[\int_\Omega V(\tilde{\boldsymbol{u}}^{n+1}) d\Omega + (|S_0| - S_0) \right]} \tag{2.54}$$

  where S_0 is given by (2.30).
- Compute $\boldsymbol{u}^{n+1}$ by equation (2.41). Compute R^{n+1} by

$$\begin{cases} R^{n+3/2} = \xi \mathscr{G}(E[\tilde{\boldsymbol{u}}^{n+3/2}]), \\ R^{n+1} = \dfrac{2}{3} R^{n+3/2} + \dfrac{1}{3} R^n, \end{cases} \tag{2.55}$$

  where we have used equations (2.50b) and (2.21a).

Noting the positiveness of energy $E(t)$ and the other functions involved in equations (2.54) and (2.55), we have the following result.

Theorem 2.8. *If the first time step is approximated using the method from Appendix A, regardless of the external forces $\boldsymbol{f}$ and $\boldsymbol{f}_b$ and the time step size Δt, the computed values for ξ and R^{n+1} with the scheme (2.50a)–(2.50e) satisfy the property,*

$$\xi > 0, \quad \text{and} \quad R^{n+1} > 0 \tag{2.56}$$

for all time steps.

Remark 2.6. In the current paper we have used the total energy (shifted) $E_{tot}(t)$ (see equation (2.11)) to define the auxiliary variable $R(t)$. One can also define an auxiliary variable based on a part of the total energy. Suppose the total energy of the system can be written as

$$E_{tot}(t) = E_1(t) + E_2(t), \quad \text{with } E_1(t) = E_1[\boldsymbol{u}] = \int_\Omega e_1(\boldsymbol{u}) d\Omega, \quad E_2(t) = E_2[\boldsymbol{u}] = \int_\Omega e_2(\boldsymbol{u}) d\Omega \tag{2.57}$$

where each of the energy components $E_1[\boldsymbol{u}]$ and $E_2[\boldsymbol{u}]$ is bounded from below. One can define an auxiliary variable $R(t)$ based on e.g. $E_2(t)$ (shifted appropriately),

$$\begin{cases} \mathscr{F}(R) = E_s(t) = E_2(t) + C_0 = \displaystyle\int_\Omega e_2(\boldsymbol{u}) d\Omega + C_0, \\ R(t) = \mathscr{G}(E_s), \end{cases} \tag{2.58}$$

where the chosen energy constant C_0 is to ensure that $E_s(t) > 0$. By appropriate reformulation of the system one can devise energy-stable schemes in an analogous way. We refer the reader to [46] for such an energy-stable scheme for incompressible

two-phase flows with different densities and viscosities for the two fluids, which corresponds to a specific mapping function $\mathscr{F}(R) = R^2$. A drawback with this lies in that one needs to solve a nonlinear algebraic equation (or a quadratic equation), albeit about a scalar number, when computing the auxiliary variable, and that the property for guaranteed positiveness of the computed auxiliary-variable values will be lost.

In the subsequent sections, we consider three dissipative (or conserving) systems (a chemotaxis model, Cahn-Hilliard equation, and Klein-Gordon equation) as specific applications and demonstrations of the gPAV method developed in this section.

3. A chemo-repulsion model

3.1. Model and numerical scheme

Consider the following repulsive-productive chemotaxis model with a quadratic production term (see e.g. [21]) in a domain Ω (with boundary Γ):

$$\frac{\partial u}{\partial t} = \nabla^2 u + \nabla \cdot (u \nabla v) + f_1(\boldsymbol{x}, t), \tag{3.1a}$$

$$\frac{\partial v}{\partial t} = \nabla^2 v - v + p(u) + f_2(\boldsymbol{x}, t), \tag{3.1b}$$

$$\boldsymbol{n} \cdot \nabla u = d_a(\boldsymbol{x}, t), \quad \boldsymbol{n} \cdot \nabla v = d_b(\boldsymbol{x}, t), \quad \text{on } \Gamma, \tag{3.1c}$$

$$u(\boldsymbol{x}, 0) = u_{in}(\boldsymbol{x}), \quad v(\boldsymbol{x}, 0) = v_{in}(\boldsymbol{x}), \tag{3.1d}$$

where $p(u) = u^2$ is the quadratic production term, $u(\boldsymbol{x}, t) \geq 0$ is the cell density, and $v(\boldsymbol{x}, t) \geq 0$ is the chemical concentration. f_1, f_2, d_a and d_b denote the volume and boundary source terms, respectively. u_{in} and v_{in} are the initial distributions of the field variables. This system is dissipative in the absence of the source terms, with the total energy given by (see [21])

$$E_{\text{tot}} = \int_\Omega \Big(\frac{1}{2}|u|^2 + \frac{1}{4}|\nabla v|^2\Big) d\Omega. \tag{3.2}$$

By taking the L^2 inner products between (3.1a) and u, and between (3.1b) and $-\frac{1}{2}\nabla^2 v$, summing them up and performing integration by part and imposing boundary conditions in (3.1c), we can obtain the following energy balance equation:

$$\begin{aligned}\int_\Omega \frac{\partial}{\partial t}\Big(\frac{1}{2}|u|^2 + \frac{1}{4}|\nabla v|^2\Big) d\Omega = &- \int_\Omega \Big(|\nabla u|^2 + \frac{1}{2}|\nabla^2 v|^2 + \frac{1}{2}|\nabla v|^2\Big) d\Omega + \int_\Omega \Big(f_1 u + \frac{1}{2}\nabla f_2 \cdot \nabla v\Big) d\Omega \\ &+ \int_\Gamma \Big(d_a u + \frac{1}{2} d_b u^2 + \frac{1}{2} d_b \frac{\partial v}{\partial t} + \frac{1}{2} d_b v - \frac{1}{2} d_b f_2\Big) d\Gamma.\end{aligned} \tag{3.3}$$

Following the gPAV procedure from section 2, we define a shifted energy according to equation (2.11)

$$E(t) = E[u, v] = \int_\Omega \Big(\frac{1}{2}|u|^2 + \frac{1}{4}|\nabla v|^2\Big) d\Omega + C_0, \tag{3.4}$$

where C_0 is a chosen energy constant such that $E(t) > 0$. Define a scalar auxiliary variable $R(t)$ according to equation (2.13a). Thus, equation (2.14) becomes

$$\mathscr{F}'(R)\frac{dR}{dt} = \int_\Omega \Big(u\frac{\partial u}{\partial t} + \frac{1}{2}\nabla v \cdot \nabla \frac{\partial v}{\partial t}\Big) d\Omega = \int_\Omega \Big(u\frac{\partial u}{\partial t} - \frac{1}{2}\nabla^2 v \frac{\partial v}{\partial t}\Big) d\Omega + \frac{1}{2}\int_\Gamma (\boldsymbol{n} \cdot \nabla v)\frac{\partial v}{\partial t} d\Gamma. \tag{3.5}$$

Following equations (2.17)-(2.19), we reformulate equations (3.1a)-(3.1b) into the following equivalent form:

$$\frac{\partial u}{\partial t} = \nabla^2 u + \frac{\mathscr{F}(R)}{E}\nabla \cdot (u \nabla v) + f_1, \tag{3.6a}$$

$$\frac{\partial v}{\partial t} = \nabla^2 v - v + \frac{\mathscr{F}(R)}{E} p(u) + f_2. \tag{3.6b}$$

By incorporating the following zero terms into the right hand side of equation (3.5),

$$\begin{aligned}
&\left(\frac{\mathscr{F}(R)}{E}-1\right)\int_\Omega u(\nabla^2 u+f_1)d\Omega+\frac{\mathscr{F}(R)}{E}\left[\int_\Omega u\nabla\cdot(u\nabla v)d\Omega-\int_\Omega u\nabla\cdot(u\nabla v)d\Omega\right]\\
&-\left(\frac{\mathscr{F}(R)}{E}-1\right)\int_\Omega \frac{1}{2}\nabla^2 v(\nabla^2 v-v+f_2)d\Omega+\frac{\mathscr{F}(R)}{E}\left[\int_\Omega \frac{1}{2}(\nabla^2 v)p(u)d\Omega-\int_\Omega \frac{1}{2}(\nabla^2 v)p(u)d\Omega\right]\\
&+\left(\frac{\mathscr{F}(R)}{E}-1\right)\int_\Gamma \frac{1}{2}(\boldsymbol{n}\cdot\nabla v)\frac{\partial v}{\partial t}d\Gamma\\
&+\left(1-\frac{\mathscr{F}(R)}{E}\right)\left|\int_\Omega f_1 u d\Omega+\frac{1}{2}\int_\Omega \nabla f_2\cdot\nabla v d\Omega+\int_\Gamma d_a u d\Gamma+\int_\Gamma \frac{d_b}{2}\left(\frac{\partial v}{\partial t}+v+u^2-f_2\right)d\Gamma\right|,
\end{aligned}$$

we can transform this equation into

$$\begin{aligned}
\mathscr{F}'(R)\frac{dR}{dt}=&\int_\Omega\left(u\frac{\partial u}{\partial t}-\frac{1}{2}\nabla^2 v\frac{\partial v}{\partial t}\right)d\Omega\\
&+\frac{\mathscr{F}(R)}{E}\left[-\int_\Omega\left(|\nabla u|^2+\frac{1}{2}|\nabla^2 v|^2+\frac{1}{2}|\nabla v|^2\right)d\Omega+\int_\Omega\left(f_1 u+\frac{1}{2}\nabla f_2\cdot\nabla v\right)d\Omega\right.\\
&\qquad\left.+\int_\Gamma\left(d_a u+\frac{1}{2}d_b u^2+\frac{1}{2}d_b\frac{\partial v}{\partial t}+\frac{1}{2}d_b v-\frac{1}{2}d_b f_2\right)d\Gamma\right]\\
&+\left(1-\frac{\mathscr{F}(R)}{E}\right)\left|\int_\Omega\left(f_1 u+\frac{1}{2}\nabla f_2\cdot\nabla v\right)d\Omega+\int_\Gamma\left(d_a u+\frac{1}{2}d_b u^2+\frac{1}{2}d_b\frac{\partial v}{\partial t}+\frac{1}{2}d_b v-\frac{1}{2}d_b f_2\right)d\Gamma\right|\\
&-\int_\Omega u\left(\nabla^2 u+\frac{\mathscr{F}(R)}{E}\nabla\cdot(u\nabla v)+f_1\right)d\Omega+\int_\Omega\frac{1}{2}\nabla^2 v\left(\nabla^2 v-v+\frac{\mathscr{F}(R)}{E}p(u)+f_2\right)d\Omega,
\end{aligned}\tag{3.7}$$

where we have used the fact $\frac{\mathscr{F}(R)}{E}=1$ and the boundary conditions (3.1c).

The reformulated equivalent system consists of equations (3.6a)-(3.7) and (3.1c)-(3.1d). The energy-stable scheme for this system is as follows:

$$\left.\frac{\partial u}{\partial t}\right|^{n+1}=\nabla^2 u^{n+1}+\xi\nabla\cdot\left(\bar{u}^{n+1}\nabla\bar{v}^{n+1}\right)+f_1^{n+1};\tag{3.8a}$$

$$\left.\frac{\partial v}{\partial t}\right|^{n+1}=\nabla^2 v^{n+1}-v^{n+1}+\xi\, p(\bar{u}^{n+1})+f_2^{n+1};\tag{3.8b}$$

$$\xi=\frac{\mathscr{F}(R^{n+\frac{3}{2}})}{E[\tilde{u}^{n+3/2},\tilde{v}^{n+3/2}]};\tag{3.8c}$$

$$E[\tilde{u}^{n+3/2},\tilde{v}^{n+3/2}]=\int_\Omega\left(\frac{1}{2}|\tilde{u}^{n+3/2}|^2+\frac{1}{4}|\nabla\tilde{v}^{n+3/2}|^2\right)d\Omega+C_0;\tag{3.8d}$$

$$\boldsymbol{n}\cdot\nabla u^{n+1}=d_a^{n+1},\ \ \boldsymbol{n}\cdot\nabla v^{n+1}=d_b^{n+1};\tag{3.8e}$$

and

$$\begin{aligned}
&D_{\mathscr{F}}(R)\Big|^{n+1}\left.\frac{dR}{dt}\right|^{n+1}=\int_\Omega\left(u^{n+1}\left.\frac{\partial u}{\partial t}\right|^{n+1}-\frac{1}{2}\nabla^2 v^{n+1}\left.\frac{\partial v}{\partial t}\right|^{n+1}\right)d\Omega\\
&-\xi\int_\Omega\left(|\nabla\tilde{u}^{n+1}|^2+\frac{1}{2}|\nabla^2\tilde{v}^{n+1}|^2+\frac{1}{2}|\nabla\tilde{v}^{n+1}|^2\right)d\Omega+\xi S_0+(1-\xi)|S_0|\\
&-\int_\Omega u^{n+1}\left(\nabla^2 u^{n+1}+\xi\nabla\cdot(\bar{u}^{n+1}\nabla\bar{v}^{n+1})+f_1^{n+1}\right)d\Omega\\
&+\int_\Omega\frac{1}{2}\nabla^2 v^{n+1}\left(\nabla^2 v^{n+1}-v^{n+1}+\xi\, p(\bar{u}^{n+1})+f_2^{n+1}\right)d\Omega.
\end{aligned}\tag{3.9}$$

In these equations, $\frac{\partial u}{\partial t}\Big|^{n+1}$, $\frac{\partial v}{\partial t}\Big|^{n+1}$ and $\frac{dR}{dt}\Big|^{n+1}$ are defined by equation (2.21b). $\bar{u}^{n+1}$ and $\bar{v}^{n+1}$ are defined by (2.21c). $\tilde{u}^{n+1}$ and $\tilde{v}^{n+1}$ are second-order approximations of u^{n+1} and v^{n+1} to be specified later in (3.21). $\tilde{u}^{n+3/2}$ and $\tilde{v}^{n+3/2}$ are second-order approximations of $u^{n+3/2}$ and $v^{n+3/2}$ to be specified later in (3.22). S_0 in equation (3.9) is given by

$$\begin{aligned} S_0 = &\int_\Omega \Big(f_1^{n+1}\tilde{u}^{n+1} + \frac{1}{2}\nabla f_2^{n+1}\cdot\nabla\tilde{v}^{n+1}\Big)d\Omega \\ &+ \int_\Gamma \Big(d_a^{n+1}\tilde{u}^{n+1} + \frac{1}{2}d_b^{n+1}(\tilde{u}^{n+1})^2 + \frac{1}{2}d_b^{n+1}\frac{\partial v}{\partial t}\Big|^{*,n+1} + \frac{1}{2}d_b^{n+1}\tilde{v}^{n+1} - \frac{1}{2}d_b^{n+1}f_2^{n+1}\Big)d\Gamma, \end{aligned} \tag{3.10}$$

where

$$\frac{\partial v}{\partial t}\Big|^{*,n+1} = \frac{\frac{3}{2}\tilde{v}^{n+1} - 2v^n + \frac{1}{2}v^{n-1}}{\Delta t}. \tag{3.11}$$

These equations are supplemented by the following initial conditions

$$u^0 = u_{\text{in}}(\boldsymbol{x}), \quad v^0 = v_{\text{in}}(\boldsymbol{x}), \quad R^0 = \mathscr{G}(E^0), \text{ with } E^0 = \int_\Omega \Big(\frac{1}{2}|u_{\text{in}}|^2 + \frac{1}{4}|\nabla v_{\text{in}}|^2\Big)d\Omega + C_0. \tag{3.12}$$

Theorem 3.1. *In the absence of the external force $f_1 = f_2 = 0$, and with homogeneous boundary conditions $d_a = d_b = 0$, the scheme consisting of (3.8a)-(3.9) is unconditionally energy stable in the sense that:*

$$\mathscr{F}(R^{n+\frac{3}{2}}) - \mathscr{F}(R^{n+\frac{1}{2}}) = -\xi\Delta t\int_\Omega \Big(|\nabla\tilde{u}^{n+1}|^2 + \frac{1}{2}|\nabla^2\tilde{v}^{n+1}|^2 + \frac{1}{2}|\nabla\tilde{v}^{n+1}|^2\Big)d\Omega \leqslant 0, \tag{3.13}$$

if the approximation of $R(t)$ at the time step $\frac{1}{2}$ is non-negative.

This theorem can be proved in a way analogous to Theorem 2.1. We can apply the method from Appendix A to this chemo-repulsion model for the first time step, and this ensures that $R^{n+1/2}|_{n=0} > 0$.

3.2. Solution algorithm and implementation

Using the notation (2.37), we rewrite equations (3.8a)-(3.8b) into

$$\frac{\gamma_0}{\Delta t}u^{n+1} - \nabla^2 u^{n+1} = \frac{\hat{u}}{\Delta t} + f_1^{n+1} + \xi\nabla\cdot\big(\bar{u}^{n+1}\nabla\bar{v}^{n+1}\big), \tag{3.14}$$

$$\Big(\frac{\gamma_0}{\Delta t} + 1\Big)v^{n+1} - \nabla^2 v^{n+1} = \frac{\hat{v}}{\Delta t} + f_2^{n+1} + \xi p(\bar{u}^{n+1}). \tag{3.15}$$

Barring the unknown scalar ξ, (3.14) and (3.15) are two decoupled Helmholtz-type equations about u^{n+1} and v^{n+1}, respectively.

Note that ξ is a scalar number instead of a field function, we define two sets of variables (u_i^{n+1}, v_i^{n+1}) $(i = 1, 2)$ as the solutions to the following equations:

$$\frac{\gamma_0}{\Delta t}u_1^{n+1} - \nabla^2 u_1^{n+1} = \frac{\hat{u}}{\Delta t} + f_1^{n+1}, \quad \boldsymbol{n}\cdot\nabla u_1^{n+1} = d_a^{n+1}; \tag{3.16}$$

$$\frac{\gamma_0}{\Delta t}u_2^{n+1} - \nabla^2 u_2^{n+1} = \nabla\cdot\big(\bar{u}^{n+1}\nabla\bar{v}^{n+1}\big), \quad \boldsymbol{n}\cdot\nabla u_2^{n+1} = 0; \tag{3.17}$$

$$\Big(\frac{\gamma_0}{\Delta t} + 1\Big)v_1^{n+1} - \nabla^2 v_1^{n+1} = \frac{\hat{v}}{\Delta t} + f_2^{n+1}, \quad \boldsymbol{n}\cdot\nabla v_1^{n+1} = d_b^{n+1}; \tag{3.18}$$

$$\Big(\frac{\gamma_0}{\Delta t} + 1\Big)v_2^{n+1} - \nabla^2 v_2^{n+1} = p(\bar{u}^{n+1}), \quad \boldsymbol{n}\cdot\nabla v_2^{n+1} = 0. \tag{3.19}$$

Then we have the following result: Given the scalar number ξ, the following field functions solve the system consisting of equations (3.14)-(3.15):

$$u^{n+1} = u_1^{n+1} + \xi u_2^{n+1}, \quad v^{n+1} = v_1^{n+1} + \xi v_2^{n+1}, \tag{3.20}$$

where (u_i^{n+1}, v_i^{n+1}) $i = 1, 2$ is given by equations (3.16)-(3.19), respectively.

Once (u_i^{n+1}, v_i^{n+1}) $i=1,2$ are known, we determine $\tilde{u}^{n+1}$, $\tilde{v}^{n+1}$, $\tilde{u}^{n+3/2}$ and $\tilde{v}^{n+3/2}$ according to (2.42), specifically by

$$\tilde{u}^{n+1} = u_1^{n+1} + u_2^{n+1}, \quad \tilde{v}^{n+1} = v_1^{n+1} + v_2^{n+1}; \tag{3.21}$$

$$\tilde{u}^{n+3/2} = \frac{3}{2}\tilde{u}^{n+1} - \frac{1}{2}u^n, \quad \tilde{v}^{n+3/2} = \frac{3}{2}\tilde{v}^{n+1} - \frac{1}{2}v^n. \tag{3.22}$$

In light of equations (3.1b), (3.21) and (3.11), we compute $\nabla^2\tilde{v}^{n+1}$ in equation (3.9) by

$$\nabla^2\tilde{v}^{n+1} = \frac{\partial v}{\partial t}\Big|^{*,n+1} + \tilde{v}^{n+1} - p(\tilde{u}^{n+1}) - f_2^{n+1}, \tag{3.23}$$

where $\frac{\partial v}{\partial t}\big|^{*,n+1}$ is given by (3.11).

Combining equations (3.8a)–(3.8b) and (3.9), and using the property (2.23), we have

$$\frac{\mathscr{F}(R^{n+\frac{3}{2}}) - \mathscr{F}(R^{n+\frac{1}{2}})}{\Delta t} = -\xi \int_\Omega \Big(|\nabla\tilde{u}^{n+1}|^2 + \frac{1}{2}|\nabla^2\tilde{v}^{n+1}|^2 + \frac{1}{2}|\nabla\tilde{v}^{n+1}|^2\Big)d\Omega + \xi S_0 + (1-\xi)|S_0|. \tag{3.24}$$

This gives rise to

$$\xi = \frac{\mathscr{F}(R^{n+1/2}) + \Delta t|S_0|}{E[\tilde{u}^{n+3/2}, \tilde{v}^{n+3/2}] + \Delta t\Big[\int_\Omega \Big(|\nabla\tilde{u}^{n+1}|^2 + \frac{1}{2}|\nabla^2\tilde{v}^{n+1}|^2 + \frac{1}{2}|\nabla\tilde{v}^{n+1}|^2\Big)d\Omega + (|S_0| - S_0)\Big]}, \tag{3.25}$$

in which S_0 is given by (3.10), $\nabla^2\tilde{v}$ is to be computed by (3.23), and $E[\tilde{u}^{n+3/2}, \tilde{v}^{n+3/2}]$ is given by (3.8d). With ξ known, R^{n+1} and (u^{n+1}, v^{n+1}) can be evaluated directly by (2.45) and (3.20), respectively.

We employ C^0-continuous high-order spectral elements for spatial discretizations in our implementation. Note that equations (3.16)–(3.19) involve Helmholtz type equations with Neumann type boundary conditions. The weak formulations of these equations are: Find u_i^{n+1} and $v_i^{n+1} \in H^1(\Omega)$ for $i=1,2$, such that

$$\begin{aligned}
&\big(\nabla u_1^{n+1}, \nabla\varphi\big)_\Omega + \frac{\gamma_0}{\Delta t}\big(u_1^{n+1}, \varphi\big)_\Omega = \big(\frac{\hat{u}}{\Delta t} + f_1^{n+1}, \varphi\big)_\Omega + \big\langle d_a^{n+1}, \varphi\big\rangle_\Gamma,\\
&\big(\nabla u_2^{n+1}, \nabla\varphi\big)_\Omega + \frac{\gamma_0}{\Delta t}\big(u_2^{n+1}, \varphi\big)_\Omega = -\big(\bar{u}^{n+1}\nabla\bar{v}^{n+1}, \nabla\varphi\big)_\Omega + \big\langle \boldsymbol{n}\cdot\nabla\bar{v}^{n+1}\bar{u}^{n+1}, \varphi\big\rangle_\Gamma,\\
&\big(\nabla v_1^{n+1}, \nabla\varphi\big)_\Omega + \big(\frac{\gamma_0}{\Delta t} + 1\big)\big(v_1^{n+1}, \varphi\big)_\Omega = \big(\frac{\hat{v}}{\Delta t} + f_2^{n+1}, \varphi\big)_\Omega + \big\langle d_b^{n+1}, \varphi\big\rangle_\Gamma,\\
&\big(\nabla v_2^{n+1}, \nabla\varphi\big)_\Omega + \big(\frac{\gamma_0}{\Delta t} + 1\big)\big(v_2^{n+1}, \varphi\big)_\Omega = \big(p(\bar{u}^{n+1}), \varphi\big)_\Omega,
\end{aligned}$$

for $\forall\varphi \in H^1(\Omega)$, where

$$(f,g)_\Omega = \int_\Omega f(\boldsymbol{x})g(\boldsymbol{x})d\Omega, \quad \langle f,g\rangle_\Gamma = \int_\Gamma f(\boldsymbol{x})g(\boldsymbol{x})d\Gamma. \tag{3.26}$$

These weak forms can be discretized using C^0 spectral elements in the standard way [25].

3.3. Numerical results

3.3.1. Convergence rate

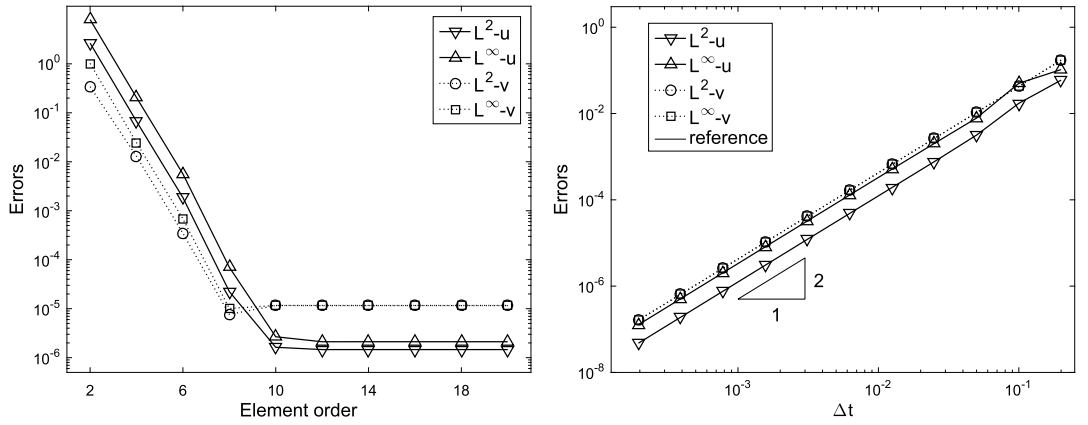
We first employ a manufactured analytical solution to the chemo-repulsion model to demonstrate the spatial and temporal convergence rates of the proposed algorithm.

Consider the computational domain $\Omega = [0,1]^2$ and the following contrived solution to the system (3.1) on this domain

$$u = \exp(-t)\big(\cos(2\pi x)\cos(2\pi y) + 2\big), \quad v = \big(1 + \sin(t)\big)\big(\cos(2\pi x)\cos(2\pi y) + 2\big). \tag{3.27}$$

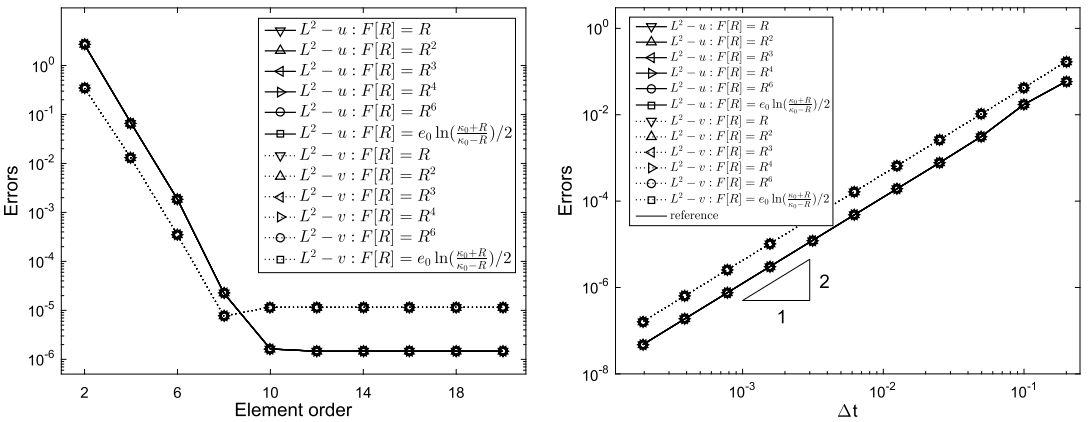
The external forces $f_1(\boldsymbol{x},t)$, $f_2(\boldsymbol{x},t)$ and boundary forces $d_a(\boldsymbol{x},t)$, $d_b(\boldsymbol{x},t)$ therein are chosen such that the expressions in (3.27) satisfy (3.1).

The domain is discretized with four equal-sized quadrilateral elements. The initial cell density u_{in} and initial chemical concentration v_{in} are given according to the analytic expressions in (3.27) by setting $t=0$. We simulate this problem from $t=0$ to $t=t_f$. Then we compare the numerical solutions of u and v at $t=t_f$ with the analytic solutions in (3.27) and various norms of the errors are computed. The element order and time step sizes are varied systematically in order to



(a) Errors vs Element order (b) Errors vs Δt

Fig. 3.1. Spatial/temporal convergence tests for chemo-repulsion model: L^2 and L^∞ errors of u and v versus (a) element order (fixed $\Delta t = 0.001$ and $t_f = 0.1$), and (b) Δt (fixed element order 18 and $t_f = 1$).



(a) Spatial convergence: L^2-errors vs. element order (b) Temporal convergence: L^2-errors vs. Δt

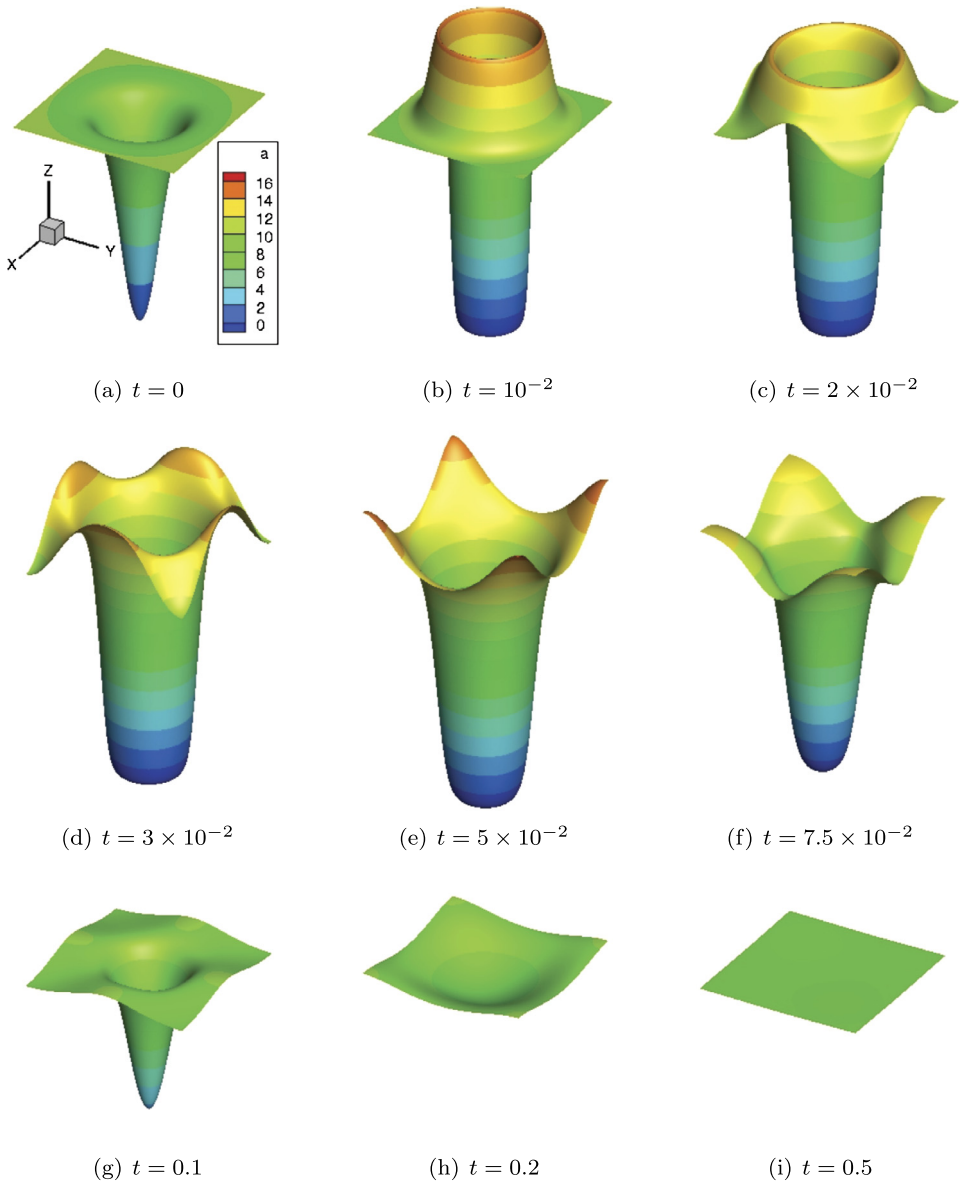
Fig. 3.2. Spatial/temporal convergence tests for the chemo-repulsion model obtained using several mapping functions $\mathscr{F}(R)$ as given in the legend: L^2 errors versus element order (a) (fixed $\Delta t = 0.001$ and $t_f = 0.1$), and L^2 errors versus Δt (b) (fixed element order 18 and $t_f = 1$).

investigate their effects on the numerical errors. We employ the function $\mathscr{F}(R) = R$ for defining the auxiliary variable $R(t)$ and the energy constant $C_0 = 1$ in the following convergence tests.

We first study the spatial convergence rate. A fixed $t_f = 0.1$ and $\Delta t = 0.001$ is employed and the element order is varied systematically between 2 and 20. We record the errors at $t = t_f$ between the numerical solution and the contrived solution (3.27) in both L^∞ and L^2 norms with respect to the element orders. Fig. 3.1(a) shows these numerical errors as a function of the element order. We observe an exponential decrease of the numerical errors with increasing element order, and a level-off of the error curves beyond element order 10 and 8, respectively for u and v, due to the saturation of temporal errors.

The study of the temporal convergence rate is summarized by the results in Fig. 3.1(b). Here we fix the integration time $t_f = 1.0$ and the element order at a large value 18, and vary Δt systematically between 0.2 and 1.953125×10^{-4}. This figure demonstrates the L^∞ and L^2 errors of u and v as a function of Δt. It is evident that the proposed scheme has a second-order convergence rate in time.

Note that a general mapping function $\mathscr{F}(R)$ can be employed for defining $R(t)$ with the gPAV method. Fig. 3.2 shows the spatial and temporal convergence behaviors of the method in terms of the L^2 errors of u and v corresponding to several mapping functions: $\mathscr{F}(R) = R^m$ $(m = 1, 2, 3, 4, 6)$ and $\mathscr{F}(R) = \frac{e_0}{2}\ln(\frac{\kappa_0 + R}{\kappa_0 - R})$ with $e_0 = \kappa_0 = 10$. It is evident that the method exhibits a spatial exponential convergence rate and a second-order temporal convergence rate with various mapping functions $\mathscr{F}(R)$. We also observe that the difference among the errors corresponding to different $\mathscr{F}(R)$ is very small and basically negligible. The choice for the specific mapping $\mathscr{F}$ appears to have very little or essentially no influence on the simulation results using the current method.



(a) $t = 0$ (b) $t = 10^{-2}$ (c) $t = 2 \times 10^{-2}$

(d) $t = 3 \times 10^{-2}$ (e) $t = 5 \times 10^{-2}$ (f) $t = 7.5 \times 10^{-2}$

(g) $t = 0.1$ (h) $t = 0.2$ (i) $t = 0.5$

Fig. 3.3. Chemo-repulsion model: Temporal sequence of snapshots of the cell density u distribution visualized by its contours. The color map in (a) applies to all the plots. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

3.3.2. Study of unconditional stability and effect of algorithmic parameters

We next consider the test problem used in [21], and show the efficiency and unconditional stability of the method proposed here. Consider the domain $\Omega = [0, 2]^2$ and the initial distributions for the cell density u and chemical concentration v in this domain given by

$$u_{in}(\boldsymbol{x}) = -10xy(2-x)(2-y)\exp(-10(y-1)^2 - 10(x-1)^2) + 10.0001, \tag{3.28a}$$

$$v_{in}(\boldsymbol{x}) = 100xy(2-x)(2-y)\exp(-30(y-1)^2 - 30(x-1)^2) + 0.0001. \tag{3.28b}$$

The external forces and boundary forces in (3.1) are set to $f_1 = f_2 = d_a = d_b = 0$. The computational domain is discretized with 400 equal-sized quadrilateral elements, and the element order is fixed to be 10.

Figs. 3.3 and 3.4 demonstrate the dynamics of the system. These results are obtained with $\Delta t = 10^{-5}$, $\mathscr{F}(R) = R$ and $C_0 = 1$ in the numerical algorithm. Fig. 3.3 shows the evolution of the cell density $u(\boldsymbol{x}, t)$ with a temporal sequence of snapshots of the distribution visualized by the contour plots. The z coordinate corresponds to u in these plots. The system exhibits a very rapid dynamics. The initial cell density has a Gaussian type distribution, taking a minimal value 0.0001 at the domain center $\boldsymbol{x}_0 = (1, 1)$ and gradually approaching the maximal value 10.0001 near the domain boundary. In a very

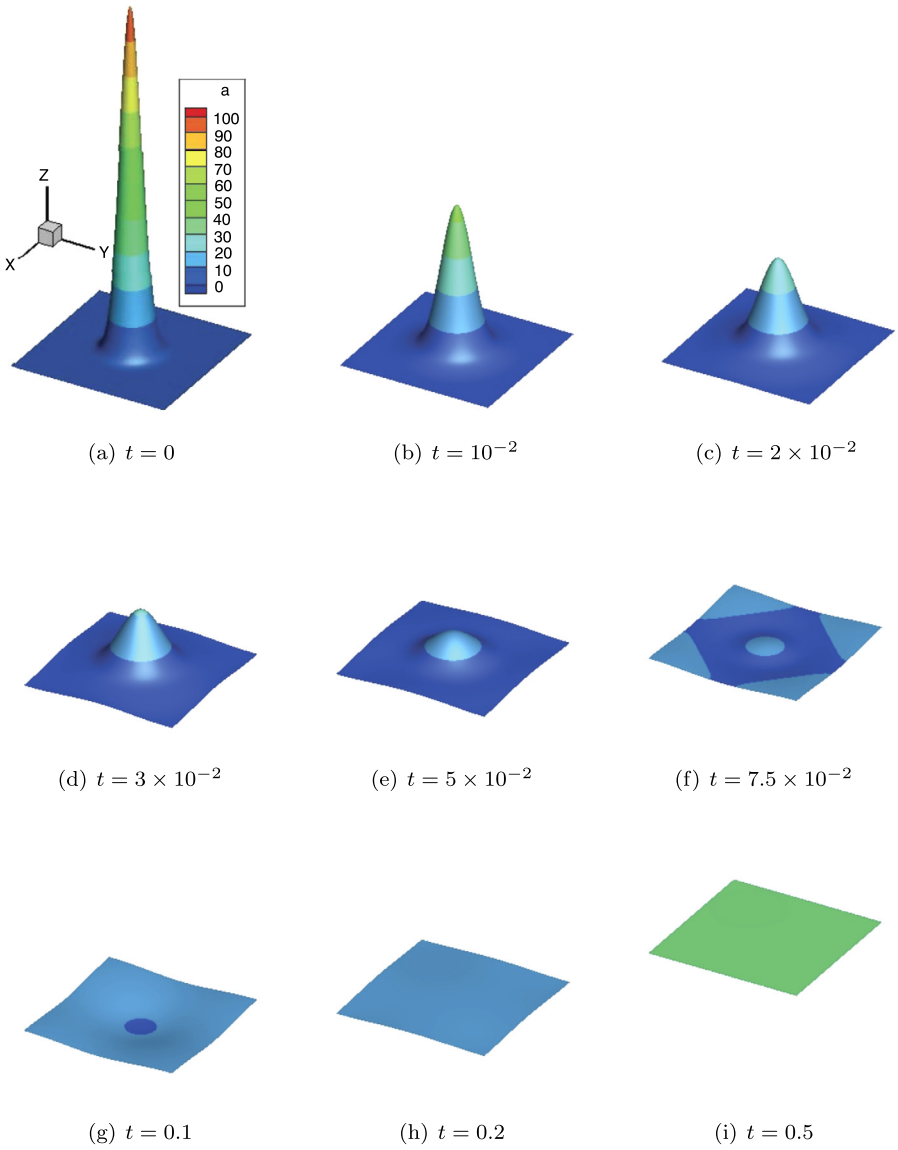


Fig. 3.4. Chemo-repulsion model: Temporal sequence of snapshots of the chemical concentration v visualized by its contours. The color map in (a) applies to all the plots herein.

short time $t = 10^{-2}$, the maximal density increases to around 16, attained near the boundary of a circular region with radius 0.6 and center at $\boldsymbol{x}_0$; see Fig. 3.3(b). Then the maximal density gradually moves from the circular boundary to the domain boundary between $t = 2 \times 10^{-2}$ and $t = 7.5 \times 10^{-2}$; see Fig. 3.3(c)-(f). The high density near the domain boundary then appears to diffuse to the region with low density near the center $\boldsymbol{x}_0$, and the system finally reaches an equilibrium state between $t = 0.1$ and $t = 0.5$ with a constant density level; see Fig. 3.3(g)-(i). Fig. 3.4 illustrates the evolution of the chemical concentration $v(\boldsymbol{x}, t)$. Fig. 3.4(a) shows the distribution of the initial chemical concentration. It has also a Gaussian type distribution, with a maximal value 100.0001 at the origin $\boldsymbol{x}_0$ and decreasing to 0.0001 gradually near the domain boundary. The concentration diffuses rapidly between $t = 0$ to $t = 5 \times 10^{-2}$ (Figs. 3.4(a)-(e)), and the maximal concentration decreases to around 10 at the origin. From $t = 7.5 \times 10^{-2}$ to $t = 0.2$, the contrast in the concentration levels in the domain becomes even smaller (Fig. 3.4(f)-(h)), and the concentration reaches its equilibrium with a constant level around 36.6 (Fig. 3.4(i)).

Fig. 3.5 shows time histories of three quantities: $E(t)$, $\mathscr{F}(R)$, and $\xi = \frac{\mathscr{F}(R)}{E(t)}$, corresponding to three time step sizes $\Delta t = 10^{-5}$, 10^{-4} and 10^{-3}. Note that $E(t)$ is computed based on equation (3.4), $\mathscr{F}(R)$ is computed based on the $R(t)$ obtained from the algorithm, and ξ is computed based on equation (3.25). These results are obtained with $\mathscr{F}(R) = R$ and $C_0 = 1$ in the algorithm. It is observed from Fig. 3.5(a) that both $E(t)$ and $\mathscr{F}(R)$ decrease over time and gradually level off at certain levels over time. A comparison of the $E(t)$ histories obtained using different Δt indicates that they are quite

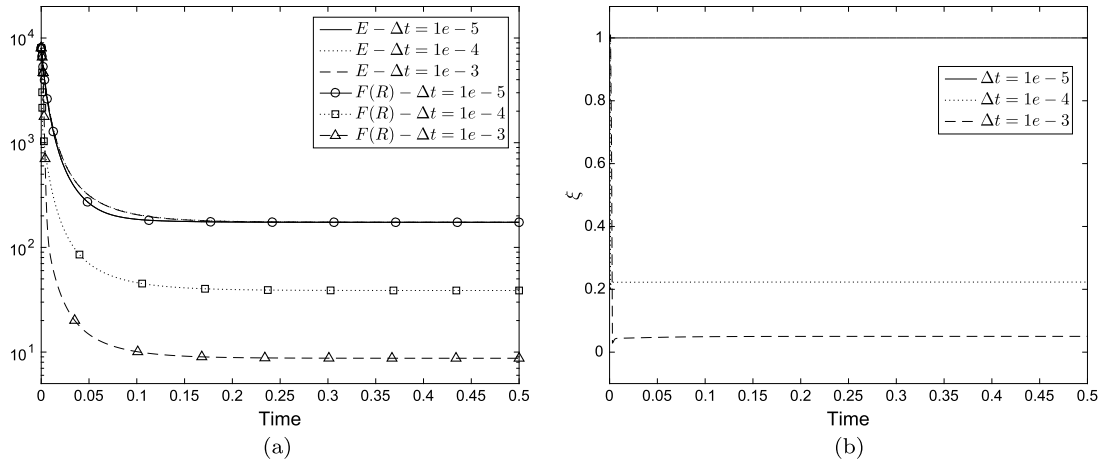


Fig. 3.5. Chemo-repulsion model: time histories of (a) $E(t)$ and $\mathscr{F}(R)$, and (b) $\xi = \mathscr{F}(R)/E(t)$, for several $\Delta t = 10^{-3}, 10^{-4}, 10^{-5}$.

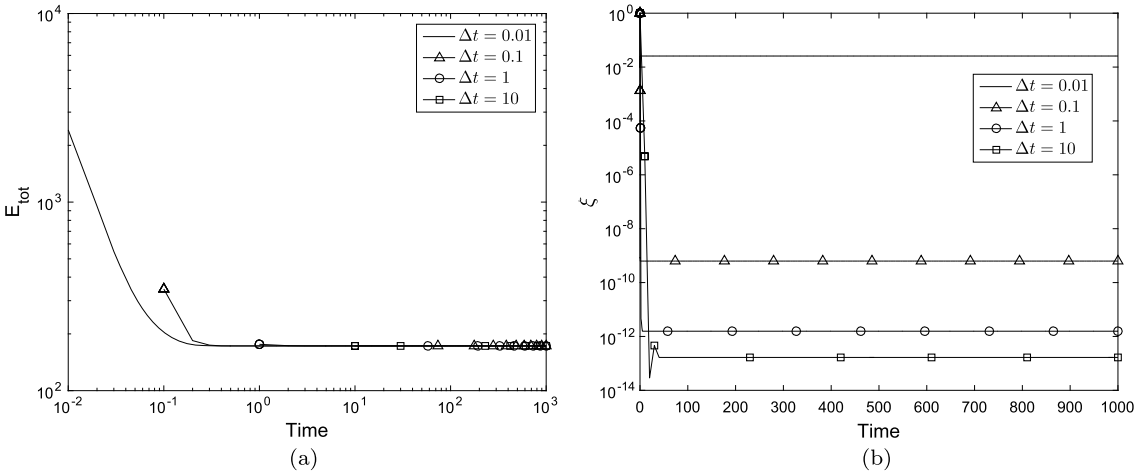


Fig. 3.6. Chemo-repulsion model: time histories of (a) $E_{tot}(t)$ and (b) $\xi = \mathscr{F}(R)/E(t)$ obtained with several large time step sizes $\Delta t = 0.01, 0.1, 1, 10$.

close, with only some slight difference on the interval between $t = 0.002$ and $t = 0.15$. Note that $\mathscr{F}(R)$ is an approximation of $E(t)$ in the current method, and the evolution equation for $R(t)$ stems from this relation; see equations (2.13a)–(2.14). Therefore, the difference between $E(t)$ and $\mathscr{F}(R)$, and also the quantity $\xi = \frac{\mathscr{F}(R)}{E(t)}$, can serve as an indicator of the accuracy of the simulations. If the difference between $E(t)$ and $\mathscr{F}(R)$ is small, or the deviation of ξ from the unit value is small, then the simulation tends to be more accurate. On the other hand, when the difference between $E(t)$ and $\mathscr{F}(R)$ is pronounced, or the deviation between ξ and the unit value is significant, it implies that $\mathscr{F}(R)$ is no longer an accurate approximation of $E(t)$ and the simulation will contain large numerical errors. Here it can be observed that $E(t)$ and $\mathscr{F}(R)$ computed with $\Delta t = 10^{-5}$ essentially overlap with each other, indicating $\mathscr{F}(R)$ approximates well the quantity $E(t)$. However, the time histories for $E(t)$ and $\mathscr{F}(R)$ obtained with $\Delta t = 10^{-4}$ and 10^{-3} exhibit noticeable discrepancies. This suggests that in these cases $\mathscr{F}(R)$ is no longer an accurate approximation of $E(t)$. We also observe from Fig. 3.5(b) that ξ computed by $\Delta t = 10^{-5}$ is essentially 1, while with larger values $\Delta t = 10^{-4}$ and $\Delta t = 10^{-3}$ the computed ξ attains values significantly smaller than 1. These results indicate that with the larger time step sizes $\Delta t = 10^{-4}$ and 10^{-3} the simulation results contain pronounced errors and they are not accurate anymore. Because this problem exhibits very rapid dynamics (see Figs. 3.3 and 3.4), to capture such dynamics accurately the requirement on Δt is very stringent.

Thanks to its energy-stable nature, our algorithm can produce stable simulation results even with very large Δt values. This is demonstrated by Fig. 3.6 with several large time step sizes, ranging from $\Delta t = 0.01$ to $\Delta t = 10$, with $\mathscr{F}(R) = R$ and $C_0 = 1$ in the algorithm. We show the time histories of the total energy $E_{tot}(t)$ (see equation (3.2)) and the ratio $\xi = \frac{\mathscr{F}(R)}{E}$ for a much longer simulation (up to $t = 1000$). The long time histories demonstrate that the computations with these large Δt values are indeed stable using the current algorithm. On the other hand, because these Δt values are very large, we cannot expect that the results will be accurate. This is evident from the values of ξ in Fig. 3.6(b). These time histories for ξ tend to level off at very small but positive values, with large deviations from the unit value. It is noted that the simulations are nonetheless stable, regardless of Δt.

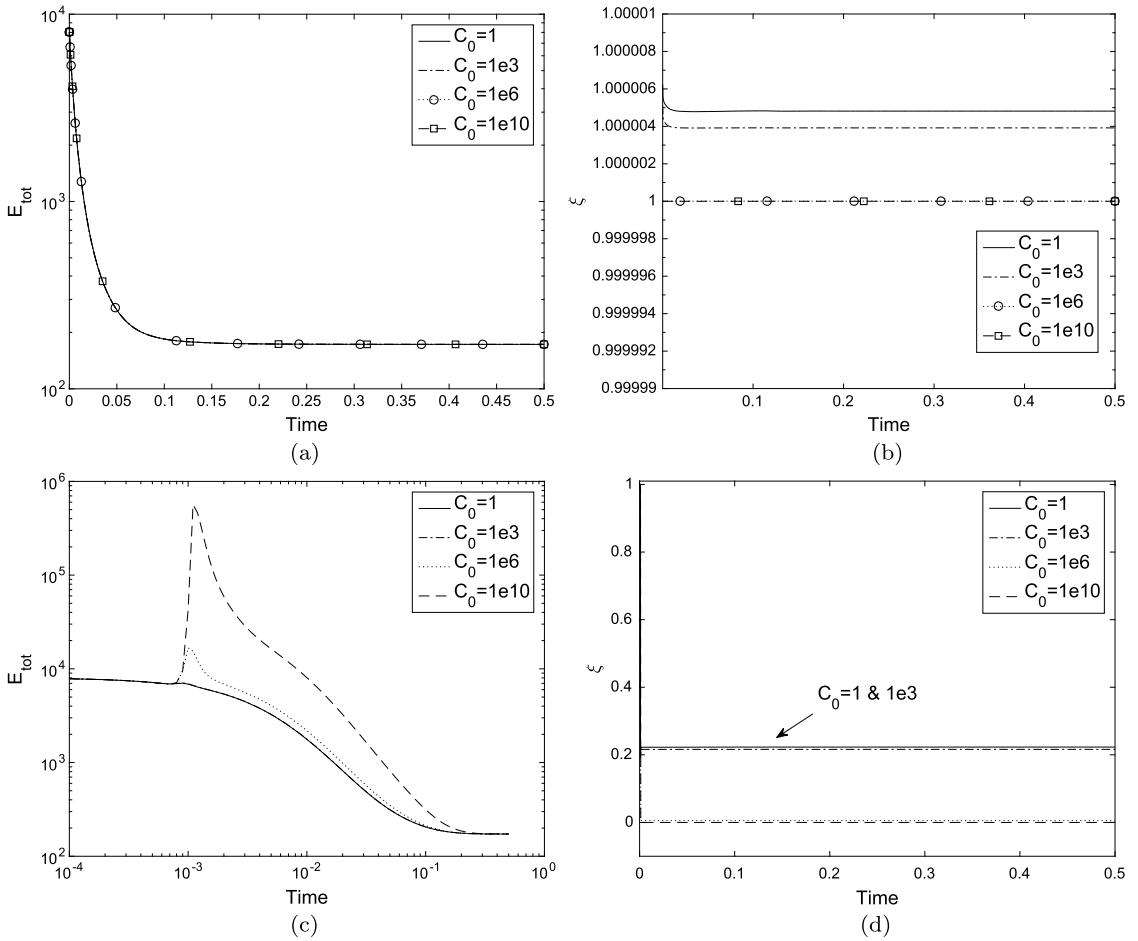


Fig. 3.7. Chemo-repulsion model: time histories of $E_{tot}(t)$ (plots (a) and (c)) and $\xi = \frac{\mathscr{F}(R)}{E(t)}$ (plots (b) and (d)) attained with various $C_0 = 1, 1e3, 1e6, 1e10$. The simulation results correspond to $\Delta t = 10^{-5}$ in (a) and (b), and $\Delta t = 10^{-4}$ in (c) and (d).

When defining the modified energy $E(t)$ (see equation (3.4)) we have incorporated an energy constant C_0. The goal of C_0 is to ensure that $E(t) > 0$ for all time, even in certain extreme cases such as when $E_{tot} = 0$, so that $\frac{1}{E(t)}$ (as in $\frac{\mathscr{F}(R)}{E(t)}$) is always well-defined. We observe that the choice of the C_0 value seems to have some influence on the numerical results. This effect is illustrated by Fig. 3.7. Here we employ $\mathscr{F}(R) = R$ and $\Delta t = 10^{-5}$ and 10^{-4}, and depict the time histories of $E_{tot}(t)$ and ξ obtained with several C_0 values ($C_0 = 1$, 10^3, 10^6 and 10^{10}). With the smaller $\Delta t = 10^{-5}$, the obtained E_{tot} histories corresponding to different C_0 values overlap with one another. The computed ξ values are essentially 1, with a discrepancy on the order of magnitude of 10^{-6}. This discrepancy between the computed ξ and the unit value is associated with the smaller $C_0 = 1$ and 10^3. With the larger $C_0 = 10^6$ and 10^{10}, no difference can be observed at this scale. This suggests that with a small Δt (so that the simulation result is generally accurate) a larger C_0 value tends to give rise to more accurate ξ in terms of its discrepancy from the unit value. Figs. 3.7(c) and (d) are the corresponding result obtained with a larger $\Delta t = 10^{-4}$, in which case the simulation result is no longer accurate. In this case it is observed that with the larger $C_0 = 10^6$ and 10^{10}, the energy E_{tot} history curves exhibit a bump, apparently artificial; see Fig. 3.7(c). In contrast, with the smaller $C_0 = 1$ and 10^3, such a bump is not quite obvious from the energy history curves. In addition, with the larger $C_0 = 10^6$ and 10^{10}, the computed ξ attains a very small value (close to 0), while ξ attains a value around 0.2 with the smaller $C_0 = 1$ and 10^3. This indicates that, with larger Δt (when simulation loses accuracy), the simulation results obtained with a smaller C_0 may be better than those obtained with a larger C_0, even though all the results become inaccurate. The results of this group of tests suggest the following. With small Δt values, a larger C_0 tends to give rise to more accurate results in the sense that the computed ξ tends to be closer to the unit value. However, a C_0 that is very large seems to have an adverse effect when Δt becomes large, because it can lead to computed ξ values that deviate from the unit value more severely. The majority of simulations in this section are performed using $C_0 = 1$.

The method developed in the current work can employ a general function $\mathscr{F}(R)$ (with inverse $\mathscr{G}$) to define the auxiliary variable $R(t)$, as long as $\mathscr{F}$ is a one-to-one increasing differentiable function satisfying (2.12). We observe that the choice for the specific mapping $\mathscr{F}$ seems to have very little or no influence on the simulation results using the current method. This

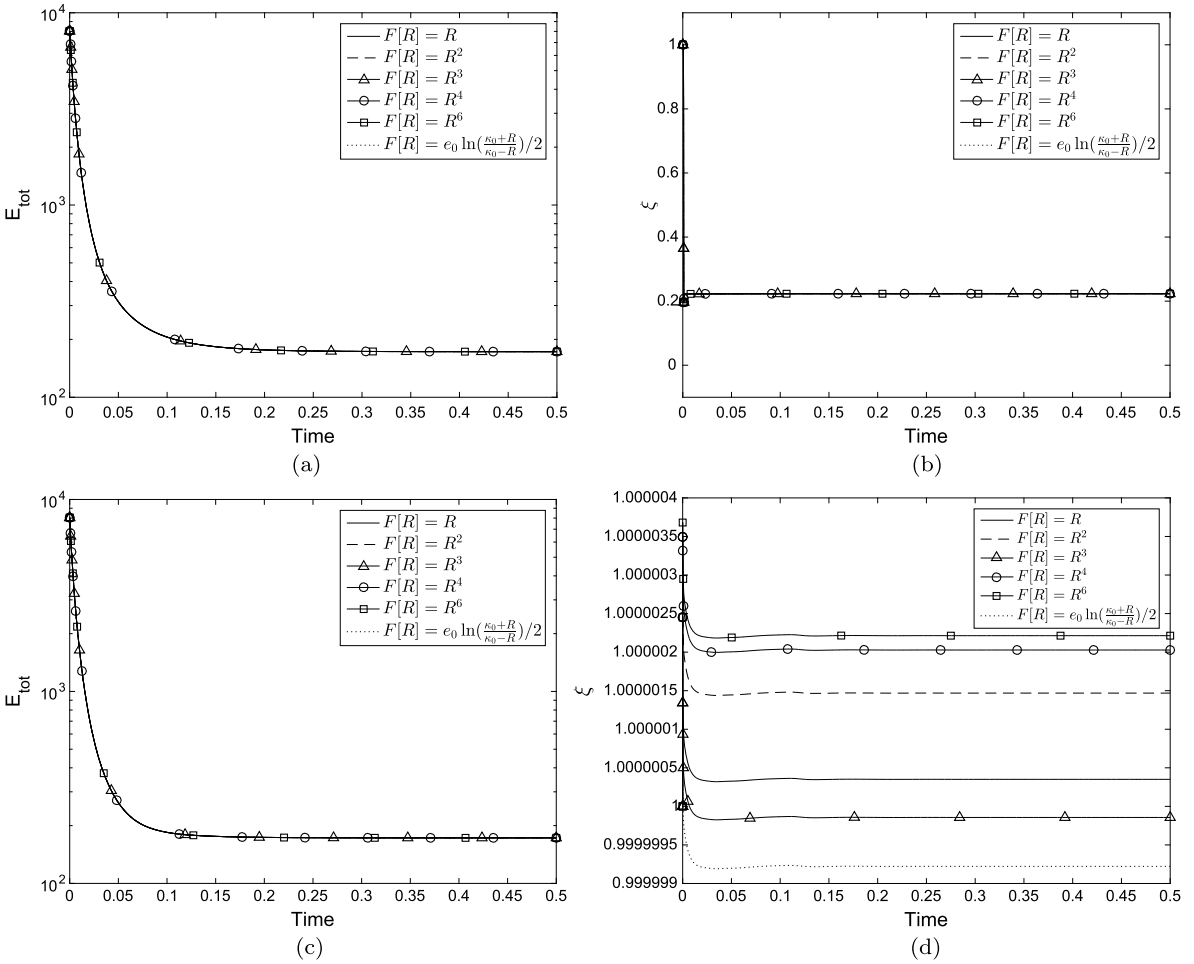


Fig. 3.8. Chemo-repulsion model: time histories of $E_{tot}(t)$ (plots (a) and (c)) and $\xi = \frac{\mathscr{F}(R)}{E}$ (plots (b) and (d)) obtained using several mapping functions $\mathscr{F}(R)$ as shown in the legend. Results in (a) and (b) correspond to $\Delta t = 10^{-4}$ and those in (c) and (d) correspond to $\Delta t = 10^{-5}$ in the simulations. Other parameters are fixed, with $C_0 = 1$, $e_0 = 8040$ and $\kappa_0 = 1000$.

point is demonstrated by Fig. 3.8. Here we have considered several functions, $\mathscr{F}(R) = R^m$ ($m = 1, 2, 3, 4, 6$) and $\mathscr{F}(R) = \frac{e_0}{2}\ln(\frac{\kappa_0+R}{\kappa_0-R})$ with $e_0 = 8040$ and $\kappa_0 = 10^3$. Fig. 3.8 shows the time histories of $E_{tot}(t)$ and ξ obtained using these mappings, together with a fixed $C_0 = 1$ and two time step sizes $\Delta t = 10^{-4}$ and 10^{-5}. It can be observed that the time history curves for both $E_{tot}(t)$ and ξ corresponding to different $\mathscr{F}$ functions overlap with one another, suggesting no or very little difference in the simulation results. In particular, Fig. 3.8(d) shows the ξ history curves corresponding to different $\mathscr{F}$ obtained with the smaller Δt, with the vertical axis ξ magnified around the unit value. It can be observed that the difference between various curves is on the order of magnitude 10^{-6}. Since little difference in the numerical results is observed with different mapping functions $\mathscr{F}(R)$ using the current method, the majority of numerical tests reported in this and subsequent sections will be carried out using the simplest mapping $\mathscr{F}(R) = R$.

In Section 2.4 we have discussed another unconditionally energy-stable scheme (referred to as "alternative method"), which is based on an alternative formulation with $\xi = \frac{R}{\mathscr{G}(E)}$. The dynamic equation for the auxiliary variable $R(t)$ is accordingly replaced by equation (2.47). Fig. 3.9 is a comparison of the time histories for $E_{tot}(t)$ and ξ obtained using these two methods. The results in Fig. 3.9(a) and (b) are obtained with a mapping function $\mathscr{F}(R) = R^2$ (or equivalently $\mathscr{G}(E) = \sqrt{E}$), and those in (c) and (d) correspond to $\mathscr{F}(R) = R^3$ (or $\mathscr{G}(E) = \sqrt[3]{E}$). We observe that there seems to be little difference in the computed total energy $E_{tot}(t)$. But some difference can be noted with the ξ histories. The computed ξ values using the current method (with $\frac{\mathscr{F}(R)}{E}$) seem to be consistently larger than those using the alternative method (with $\frac{R}{\mathscr{G}(E)}$). While all these values deviate from the unit value substantially because of the time step size $\Delta t = 10^{-4}$, the deviation with the current method appears noticeably smaller than that with the alternative method. This seems to suggest that, while the simulation results using these methods are not very much different, the formulation using $\frac{\mathscr{F}(R)}{E}$ may be somewhat better than the alternative formulation using $\frac{R}{\mathscr{G}(E)}$.

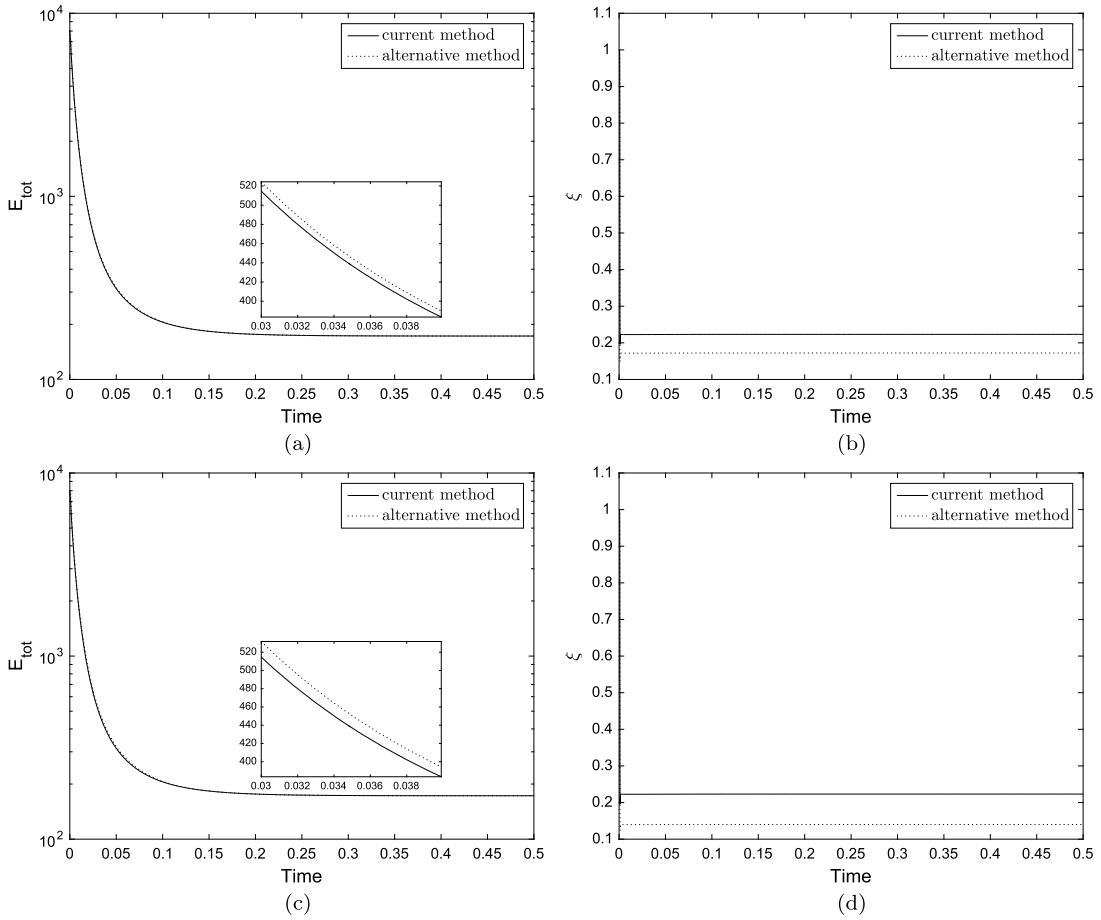


Fig. 3.9. Chemo-repulsion model: comparison of the time histories of $E_{tot}(t)$ (plots (a) and (c)) and ξ (plots (b) and (d)) computed using the current method and the alternative method from Section 2.4. In the current method $\xi = \mathscr{F}(R)/E$, and in the alternative method $\xi = R/\mathscr{G}(E)$. Plots (a) and (b) are obtained with the mapping $\mathscr{F}(R) = R^2$ (i.e. $\mathscr{G}(E) = \sqrt{E}$), and plots (c) and (d) are obtained with $\mathscr{F}(R) = R^3$ (i.e. $\mathscr{G}(E) = \sqrt[3]{E}$). Other parameters are fixed with $\Delta t = 10^{-4}$ and $C_0 = 1$.

4. Cahn-Hilliard equation with constant and variable mobility

We apply the gPAV method to simulate the Cahn-Hilliard equation [4] in this section. This equation has widespread applications in the phase-field modeling of materials science, two-phase and multiphase flows (see e.g. [32,7,30,48,26,12,16, 13,14,31,42,43], among others). Consider the Cahn-Hilliard equation on a domain Ω (with boundary Γ):

$$\frac{\partial \phi}{\partial t} = \nabla \cdot \big(m(\phi)\nabla \mu\big) + f(\boldsymbol{x}, t), \tag{4.1a}$$

$$\mu = \frac{\delta E_{tot}}{\delta \phi} = -\lambda \nabla^2 \phi + h(\phi), \tag{4.1b}$$

$$m(\phi)\boldsymbol{n} \cdot \nabla \mu = d_a(\boldsymbol{x}, t), \ \text{ on } \Gamma, \tag{4.1c}$$

$$\boldsymbol{n} \cdot \nabla \phi = d_b(\boldsymbol{x}, t) \ \text{ on } \Gamma, \tag{4.1d}$$

supplemented by the initial condition

$$\phi(\boldsymbol{x}, 0) = \phi_{\text{in}}(\boldsymbol{x}). \tag{4.2}$$

In these equations, $\phi(\boldsymbol{x}, t) \in [-1, 1]$ is the phase field function, $f(\boldsymbol{x}, t)$, $d_a(\boldsymbol{x}, t)$ and $d_b(\boldsymbol{x}, t)$ are prescribed source terms for the purpose of convergence testing only, and will be set to $f(\boldsymbol{x}, t) = d_a(\boldsymbol{x}, t) = d_b(\boldsymbol{x}, t) = 0$ in actual simulations. E_{tot} is the free energy functional,

$$E_{tot}(t) = E_{tot}[\phi, \nabla\phi] = \int_{\Omega} \left[\frac{\lambda}{2} \nabla\phi \cdot \nabla\phi + H(\phi) \right] d\Omega, \quad \text{with } H(\phi) = \frac{\lambda}{4\eta^2}(\phi^2 - 1)^2 \tag{4.3}$$

in which η is the characteristic interfacial thickness scale, and λ is referred to as the mixing energy density coefficient and is related to other physical parameters. For example, for two-phase flow problems λ is given by $\lambda = \frac{3}{2\sqrt{2}}\sigma\eta$, where σ is the surface tension. μ is referred to as the chemical potential, and the nonlinear term $h(\phi)$ is given by $h(\phi) = H'(\phi)$. $H(\phi)$ is referred to as the potential free energy density function, which can take many different forms. In this paper we only consider the double-well form as given in (4.3). $m \geqslant 0$ is the mobility, and in this work we consider two cases: (i) $m = m_0$, and (ii) $m = m(\phi) = \max(m_0(1-\phi^2), 0)$, with m_0 being a given positive constant.

We take the L^2 inner product between (4.1a) and μ, perform integration by part and impose the boundary condition (4.1d). This leads to the energy balance equation,

$$\frac{\partial}{\partial t}\int_\Omega \Big(\frac{\lambda}{2}|\nabla\phi|^2 + H(\phi)\Big)d\Omega = -\int_\Omega m(\phi)|\nabla\mu|^2 d\Omega + \int_\Omega f\mu d\Omega + \int_\Gamma m(\phi)(\boldsymbol{n}\cdot\nabla\mu)\mu d\Gamma + \lambda\int_\Gamma (\boldsymbol{n}\cdot\nabla\phi)\frac{\partial\phi}{\partial t}d\Gamma. \tag{4.4}$$

Based on equations (2.11) and (4.4), we define the shifted total energy by

$$E(t) = E[\phi] = \int_\Omega \Big(\frac{\lambda}{2}|\nabla\phi|^2 + H(\phi)\Big)d\Omega + C_0, \tag{4.5}$$

where C_0 is chosen to ensure $E(t) > 0$. Let us define $\mathscr{F}$ and $\mathscr{G}$ and $R(t)$ based on equations (2.13a)–(2.13b). Following equation (2.14) and using (4.5), we have

$$\mathscr{F}'(R)\frac{dR}{dt} = \int_\Omega \Big[-\lambda\nabla^2\phi + h(\phi)\Big]\frac{\partial\phi}{\partial t}d\Omega + \lambda\int_\Gamma d_b\frac{\partial\phi}{\partial t}d\Gamma, \tag{4.6}$$

where the boundary condition (4.1d) has been used.

4.1. Constant mobility

Assume that $m(\phi) = m_0 > 0$ is a constant. We reformulate equations (4.1a)–(4.1c) as follows,

$$\frac{\partial\phi}{\partial t} = m_0\nabla^2\left[-\lambda\nabla^2\phi + S(\phi-\phi) + \frac{\mathscr{F}(R)}{E}h(\phi)\right] + f, \tag{4.7a}$$

$$m_0\boldsymbol{n}\cdot\nabla\left[-\lambda\nabla^2\phi + S(\phi-\phi) + \frac{\mathscr{F}(R)}{E}h(\phi)\right] = d_a, \quad \text{on } \Gamma, \tag{4.7b}$$

where S is chosen constant satisfying a condition to be specified later. Note that a zero term $S(\phi-\phi)$ is added in these equations. By incorporating appropriate zero terms into the RHS, we reformulate equation (4.6) as follows,

$$\begin{aligned}\mathscr{F}'(R)\frac{dR}{dt} = &\int_\Omega \mu\frac{\partial\phi}{\partial t}d\Omega - \int_\Omega \mu\left[m_0\nabla^2\left(-\lambda\nabla^2\phi + S(\phi-\phi) + \frac{\mathscr{F}(R)}{E}h(\phi)\right) + f\right]d\Omega \\ &+ \frac{\mathscr{F}(R)}{E}\left[-\int_\Omega m_0|\nabla\mu|^2 d\Omega + \int_\Omega f\mu d\Omega + \int_\Gamma d_a\mu d\Gamma + \int_\Gamma \lambda d_b\frac{\partial\phi}{\partial t}d\Gamma\right] \\ &+ \left(1 - \frac{\mathscr{F}(R)}{E}\right)\left|\int_\Omega f\mu d\Omega + \int_\Gamma d_a\mu d\Gamma + \int_\Gamma \lambda d_b\frac{\partial\phi}{\partial t}d\Gamma\right|\end{aligned}, \tag{4.8}$$

where μ is given by (4.1b).

The energy-stable scheme for the equations (4.7a)–(4.7b), (4.1d) and (4.8) is as follows:

$$\frac{\partial\phi}{\partial t}\Big|^{n+1} = m_0\nabla^2\Big[-\lambda\nabla^2\phi^{n+1} + S(\phi^{n+1} - \bar{\phi}^{n+1}) + \xi h(\bar{\phi}^{n+1})\Big] + f^{n+1}, \tag{4.9a}$$

$$m_0\boldsymbol{n}\cdot\nabla\Big[-\lambda\nabla^2\phi^{n+1} + S(\phi^{n+1} - \bar{\phi}^{n+1}) + \xi h(\bar{\phi}^{n+1})\Big] = d_a^{n+1}, \quad \text{on } \Gamma, \tag{4.9b}$$

$$\boldsymbol{n}\cdot\nabla\phi^{n+1} = d_b^{n+1}, \quad \text{on } \Gamma, \tag{4.9c}$$

$$\xi = \frac{\mathscr{F}(R^{n+\frac{3}{2}})}{E[\tilde{\phi}^{n+3/2}]}, \tag{4.9d}$$

$$E[\tilde{\phi}^{n+3/2}] = \int_\Omega \left[\frac{\lambda}{2}|\nabla\tilde{\phi}^{n+3/2}|^2 + H(\tilde{\phi}^{n+3/2})\right] d\Omega + C_0, \tag{4.9e}$$

and

$$\begin{aligned} D_{\mathscr{F}}(R)\Big|^{n+1}\frac{dR}{dt}\Big|^{n+1} &= \int_\Omega \Big[-\lambda\nabla^2\phi^{n+1} + h(\phi^{n+1})\Big]\frac{\partial\phi}{\partial t}\Big|^{n+1} d\Omega \\ &- \int_\Omega \Big[-\lambda\nabla^2\phi^{n+1} + h(\phi^{n+1})\Big]\Big\{m_0\nabla^2\Big[-\lambda\nabla^2\phi^{n+1} + S(\phi^{n+1}-\bar{\phi}^{n+1}) + \xi h(\bar{\phi}^{n+1})\Big] + f^{n+1}\Big\} d\Omega \\ &+ \xi\Big\{-\int_\Omega m_0|\nabla\tilde{\mu}^{n+1}|^2 d\Omega + \int_\Omega f^{n+1}\tilde{\mu}^{n+1} d\Omega + \int_\Gamma \left(d_a^{n+1}\tilde{\mu}^{n+1} + \lambda d_b^{n+1}\frac{\partial\phi}{\partial t}\Big|^{*,n+1}\right) d\Gamma\Big\} \\ &+ (1-\xi)\left|\int_\Omega f^{n+1}\tilde{\mu}^{n+1} d\Omega + \int_\Gamma \left(d_a^{n+1}\tilde{\mu}^{n+1} + \lambda d_b^{n+1}\frac{\partial\phi}{\partial t}\Big|^{*,n+1}\right) d\Gamma\right|. \end{aligned} \tag{4.10}$$

These are supplemented by the initial conditions

$$\phi^0(\boldsymbol{x}) = \phi_{\text{in}}(\boldsymbol{x}), \quad R^0 = \mathscr{G}(E^0), \ \text{ with } \ E^0 = \int_\Omega \Big(\frac{1}{2}|\nabla\phi_{\text{in}}|^2 + H(\phi_{\text{in}})\Big) d\Omega + C_0. \tag{4.11}$$

In the above equations, $\frac{\partial\phi}{\partial t}\Big|^{n+1}$ and $\frac{dR}{dt}\Big|^{n+1}$ are defined by (2.21b), and $\bar{\phi}^{n+1}$ is defined by (2.21c). $\tilde{\phi}^{n+1}$, $\tilde{\phi}^{n+3/2}$ and $\tilde{\mu}^{n+1}$ are second-order approximations of ϕ^{n+1}, $\phi^{n+3/2}$ and μ^{n+1}, respectively, to be specified later in (4.24)–(4.26). $\frac{\partial\phi}{\partial t}\Big|^{*,n+1}$ is an approximation of $\frac{\partial\phi}{\partial t}\Big|^{n+1}$ to be specified later in (4.25).

Theorem 4.1. *In the absence of the external force $f = 0$, and with zero boundary conditions $d_a = d_b = 0$, the scheme consisting of (4.9)-(4.10) is unconditionally energy stable in the sense that*

$$\mathscr{F}(R^{n+\frac{3}{2}}) - \mathscr{F}(R^{n+\frac{1}{2}}) = -\xi\Delta t\int_\Omega m_0|\nabla\tilde{\mu}^{n+1}|^2 \le 0, \tag{4.12}$$

if the approximation of $R(t)$ at time step $\frac{1}{2}$ is positive.

Proof. Multiplying $-\lambda\nabla^2\phi^{n+1} + h(\phi^{n+1})$ to equation (4.9a), integrating over the domain, and adding the resultant equation to equation (4.10), we obtain the energy balance relation as follows:

$$\begin{aligned} \frac{\mathscr{F}(R^{n+\frac{3}{2}}) - \mathscr{F}(R^{n+\frac{1}{2}})}{\Delta t} &= \xi\Big\{\int_\Omega f^{n+1}\tilde{\mu}^{n+1} d\Omega + \int_\Gamma \left(d_a^{n+1}\tilde{\mu}^{n+1} + \lambda d_b^{n+1}\frac{\partial\phi}{\partial t}\Big|^{*,n+1}\right) d\Gamma\Big\} \\ &- \xi\int_\Omega m_0|\nabla\tilde{\mu}^{n+1}|^2 d\Omega + (1-\xi)\left|\int_\Omega f^{n+1}\tilde{\mu}^{n+1} d\Omega + \int_\Gamma \left(d_a^{n+1}\tilde{\mu}^{n+1} + \lambda d_b^{n+1}\frac{\partial\phi}{\partial t}\Big|^{*,n+1}\right) d\Gamma\right|, \end{aligned} \tag{4.13}$$

where we have used the relation (2.28). If $f = 0$ and $d_a = d_b = 0$, then

$$\xi = \frac{\mathscr{F}(R^{n+1/2})}{E[\tilde{\phi}^{n+3/2}] + \Delta t\int_\Omega m_0|\nabla\tilde{\mu}^{n+1}|^2}. \tag{4.14}$$

If $R^{n+1/2}|_{n=0} > 0$, one can conclude by induction that $\xi > 0$ for any $n \geqslant 0$. This leads to (4.12). □

The method from the Appendix A can be employed to compute the first time step, which can ensure that the approximation of $R(t)$ at the step $\frac{1}{2}$ is positive.

To implement the scheme we note that equation (4.9a) can be transformed into

$$\nabla^2(\nabla^2\phi^{n+1}) - \frac{S}{\lambda}\nabla^2\phi^{n+1} + \frac{\gamma_0}{m_0\lambda\Delta t}\phi^{n+1} = \frac{1}{m_0\lambda}\Big[\frac{\hat{\phi}}{\Delta t} + f^{n+1}\Big] - \frac{S}{\lambda}\nabla^2\bar{\phi}^{n+1} + \xi\frac{1}{\lambda}\nabla^2 h(\bar{\phi}^{n+1}), \tag{4.15}$$

where we have used the notation in equation (2.37). This equation can be reformulated into the following two Helmholtz type equations that are de-coupled from each other (barring the unknown scalar number ξ), (see e.g. [16,47] for details)

$$\nabla^2\psi^{n+1}-\Big(\alpha+\frac{S}{\lambda}\Big)\psi^{n+1}=\frac{1}{m_0\lambda}\Big[\frac{\hat{\phi}}{\Delta t}+f^{n+1}\Big]-\frac{S}{\lambda}\nabla^2\bar{\phi}^{n+1}+\xi\frac{1}{\lambda}\nabla^2 h(\bar{\phi}^{n+1}), \tag{4.16a}$$

$$\nabla^2\phi^{n+1}+\alpha\phi^{n+1}=\psi^{n+1}, \tag{4.16b}$$

where ψ^{n+1} is an auxiliary field variable defined by (4.16b), and the constant α is given by and the chosen constant S must satisfy

$$\alpha=-\frac{S}{2\lambda}\Big(1-\sqrt{1-\frac{4\gamma_0\lambda}{m_0\Delta t S^2}}\Big);\quad S\geqslant\sqrt{\frac{4\lambda\gamma_0}{m_0\Delta t}}. \tag{4.17}$$

In light of (4.16b) and (4.9c), the boundary condition (4.9b) can be transformed into

$$\boldsymbol{n}\cdot\nabla\psi^{n+1}=\Big[\Big(\alpha+\frac{S}{\lambda}\Big)d_b^{n+1}-\frac{1}{m_0\lambda}d_a^{n+1}\Big]-\frac{S}{\lambda}\boldsymbol{n}\cdot\nabla\bar{\phi}^{n+1}+\xi\frac{1}{\lambda}\boldsymbol{n}\cdot\nabla h(\bar{\phi}^{n+1}). \tag{4.18}$$

To solve equations (4.16a)-(4.16b) together with the boundary conditions (4.18) and (4.9c), we take advantage of the fact that ξ is a scalar number and introduce two sets of field functions $(\psi_i^{n+1},\phi_i^{n+1})$ $(i=1,2)$ as solutions of the following equations:

For ψ_1^{n+1}:

$$\nabla^2\psi_1^{n+1}-\Big(\alpha+\frac{S}{\lambda}\Big)\psi_1^{n+1}=\frac{1}{m_0\lambda}\Big[\frac{\hat{\phi}}{\Delta t}+f^{n+1}\Big]-\frac{S}{\lambda}\nabla^2\bar{\phi}^{n+1}, \tag{4.19a}$$

$$\boldsymbol{n}\cdot\nabla\psi_1^{n+1}=\Big[\Big(\alpha+\frac{S}{\lambda}\Big)d_b^{n+1}-\frac{1}{m_0\lambda}d_a^{n+1}\Big]-\frac{S}{\lambda}\boldsymbol{n}\cdot\nabla\bar{\phi}^{n+1}. \tag{4.19b}$$

For ψ_2^{n+1}:

$$\nabla^2\psi_2^{n+1}-\Big(\alpha+\frac{S}{\lambda}\Big)\psi_2^{n+1}=\frac{1}{\lambda}\nabla^2 h(\bar{\phi}^{n+1}),\quad \boldsymbol{n}\cdot\nabla\psi_2^{n+1}=\frac{1}{\lambda}\boldsymbol{n}\cdot\nabla h(\bar{\phi}^{n+1}). \tag{4.20}$$

For ϕ_1^{n+1}:

$$\nabla^2\phi_1^{n+1}+\alpha\phi_1^{n+1}=\psi_1^{n+1},\quad \boldsymbol{n}\cdot\nabla\phi_1^{n+1}=d_b^{n+1}. \tag{4.21}$$

For ϕ_2^{n+1}:

$$\nabla^2\phi_2^{n+1}+\alpha\phi_2^{n+1}=\psi_2^{n+1},\quad \boldsymbol{n}\cdot\nabla\phi_2^{n+1}=0. \tag{4.22}$$

Then for given scalar number ξ, the following field functions solve the system consisting of equations (4.16), (4.18) and (4.9c):

$$\psi^{n+1}=\psi_1^{n+1}+\xi\psi_2^{n+1},\quad \phi^{n+1}=\phi_1^{n+1}+\xi\phi_2^{n+1}, \tag{4.23}$$

where $(\psi_i^{n+1},\phi_i^{n+1})$ $(i=1,2)$ are given by equations (4.19a)-(4.22).

Now we are ready to determine the unknown scalar ξ. Following equations (2.42), we define

$$\tilde{\phi}^{n+1}=\phi_1^{n+1}+\phi_2^{n+1},\quad \tilde{\psi}^{n+1}=\psi_1^{n+1}+\psi_2^{n+1},\quad \nabla^2\tilde{\phi}^{n+1}=\tilde{\psi}^{n+1}-\alpha\tilde{\phi}^{n+1} \tag{4.24}$$

where equation (4.16b) has been used. Accordingly, in light of equations (4.1b) and (2.37), we define

$$\begin{cases}\tilde{\mu}^{n+1}=-\lambda\nabla^2\tilde{\phi}^{n+1}+h(\tilde{\phi}^{n+1})=-\lambda(\tilde{\psi}^{n+1}-\alpha\tilde{\phi}^{n+1})+h(\tilde{\phi}^{n+1}),\\ \dfrac{\partial\phi}{\partial t}\Big|^{*,n+1}=\dfrac{\gamma_0\tilde{\phi}^{n+1}-\hat{\phi}}{\Delta t}.\end{cases} \tag{4.25}$$

We further define

$$\tilde{\phi}^{n+\frac{3}{2}}=\frac{3}{2}\tilde{\phi}^{n+1}-\frac{1}{2}\phi^{n}. \tag{4.26}$$

Combining equations (4.9d) and (4.13), we obtain the formula for ξ,

$$\xi=\frac{\mathscr{F}(R^{n+1/2})+\Delta t|S_0|}{E[\tilde{\phi}^{n+\frac{3}{2}}]+\Delta t m_0\int_\Omega|\nabla\tilde{\mu}^{n+1}|^2d\Omega+\Delta t(|S_0|-S_0)}, \tag{4.27}$$

where S_0 is given by

$$S_0 = \int_\Omega f^{n+1}\tilde{\mu}^{n+1}d\Omega + \int_\Gamma \left(d_a^{n+1}\tilde{\mu}^{n+1} + \lambda d_b^{n+1}\left.\frac{\partial \phi}{\partial t}\right|^{*,n+1} \right) d\Gamma. \tag{4.28}$$

Once ξ is known, ϕ^{n+1} and ψ^{n+1} can be obtained directly by equation (4.23) and R^{n+1} can be computed based on equation (2.45).

Equations (4.16a)-(4.22) are Helmholtz type equations with Neumann type boundary conditions. They can be implemented with C^0 spectral elements in a straightforward fashion.

Remark 4.1. In equation (4.9a), we have treated the nonlinear term explicitly by $h(\bar{\phi}^{n+1})$. When Δt becomes large, $\bar{\phi}^{n+1}$ can no longer approximate ϕ^{n+1} well. Thus, although the scheme (4.9)-(4.10) is unconditionally stable, the simulation will lose accuracy for large time steps. One possible approach to improve the accuracy is to replace $\xi h(\bar{\phi}^{n+1})$ in equation (4.9a) by

$$\frac{\lambda}{\eta^2}\left(\phi_0^2 - 1\right)\phi^{n+1} + \xi\left[h(\bar{\phi}^{n+1}) - \frac{\lambda}{\eta^2}\left(\phi_0^2 - 1\right)\bar{\phi}^{n+1}\right],$$

where ϕ_0 is a chosen field function close to ϕ^{n+1}, e.g. a snapshot of the ϕ field in the recent past. The first term in the above equation serves as a linearized approximation of $h(\phi^{n+1})$ and the second term serves as a correction to this approximation. By doing so, equation (4.9a) with the mentioned modification is still linear, but can no longer be decoupled straightforwardly. One needs to solve either a fourth-order linear equation or a coupled linear system. However, this treatment can result in improved accuracy besides unconditional stability. We will demonstrate this in the forthcoming case for the Cahn-Hilliard equation with variable mobility.

4.2. Variable mobility

Next, we consider the case with a variable mobility, $m(\phi) = \max(m_0(1-\phi^2), 0)$. We reformulate the equations (4.1a)–(4.1c) into

$$\frac{\partial \phi}{\partial t} = \nabla\cdot[m_c(\phi_0)\nabla C] + \frac{\mathscr{F}(R)}{E}\nabla\cdot[m(\phi)\nabla\mu - m_c(\phi_0)\nabla C] + f, \tag{4.29a}$$

$$m_c(\phi_0)\boldsymbol{n}\cdot\nabla C + \frac{\mathscr{F}(R)}{E}\boldsymbol{n}\cdot[m(\phi)\nabla\mu - m_c(\phi_0)\nabla C] = d_a. \tag{4.29b}$$

In these equations, μ is given by (4.1b), ϕ_0 is a chosen field distribution corresponding to $\phi(\boldsymbol{x}, t)$ at a certain time instant or at some time instants, and

$$\begin{cases} C = -\lambda\nabla^2\phi + S(\phi - \phi) + \kappa(\phi_0)\phi; \\ m_c(\phi_0) = m(\phi_0), \quad \text{or } m_c(\phi_0) = m_0; \\ \kappa(\phi_0) = \dfrac{\lambda}{\eta^2}(\phi_0^2 - 1), \quad \text{or } \kappa(\phi_0) = 0; \end{cases} \tag{4.30}$$

where $S \geqslant 0$ is a chosen constant. By incorporating appropriate zero terms into the RHS of (4.6), we can transform this equation into,

$$\begin{aligned} \mathscr{F}'(R)\frac{dR}{dt} = &\int_\Omega \mu\frac{\partial\phi}{\partial t}d\Omega - \int_\Omega \mu\left[\nabla\cdot(m_c(\phi_0)\nabla C) + \frac{\mathscr{F}(R)}{E}\nabla\cdot[m(\phi)\nabla\mu - m_c(\phi_0)\nabla C] + f\right]d\Omega \\ &+ \frac{\mathscr{F}(R)}{E}\left[-\int_\Omega m(\phi)\nabla\mu\cdot\nabla\mu + \int_\Omega f\mu d\Omega + \int_\Gamma d_a\mu d\Gamma + \int_\Gamma \lambda d_b\frac{\partial\phi}{\partial t}d\Gamma\right] \\ &+ \left(1 - \frac{\mathscr{F}(R)}{E}\right)\left|\int_\Omega f\mu d\Omega + \int_\Gamma d_a\mu d\Gamma + \int_\Gamma \lambda d_b\frac{\partial\phi}{\partial t}d\Gamma\right|. \end{aligned} \tag{4.31}$$

Following equations (2.25a)-(2.25e), we propose the following scheme:

$$\left.\frac{\partial\phi}{\partial t}\right|^{n+1} = \nabla\cdot\left(m_c(\phi_0)\nabla C^{n+1}\right) + \xi\nabla\cdot\left[m(\bar{\phi}^{n+1})\nabla\bar{\mu}^{n+1} - m_c(\phi_0)\nabla\bar{C}^{n+1}\right] + f^{n+1}, \tag{4.32a}$$

$$C^{n+1} = -\lambda\nabla^2\phi^{n+1} + S(\phi^{n+1} - \bar{\phi}^{n+1}) + \kappa(\phi_0)\phi^{n+1}, \tag{4.32b}$$

$$\xi = \frac{\mathscr{F}(R^{n+3/2})}{E[\tilde{\phi}^{n+3/2}]}, \tag{4.32c}$$

$$E[\tilde{\phi}^{n+3/2}] = \int_\Omega \left[\frac{\lambda}{2}\left|\nabla\tilde{\phi}^{n+3/2}\right|^2 + H(\tilde{\phi}^{n+3/2})\right]d\Omega + C_0, \tag{4.32d}$$

$$m_c(\phi_0)\boldsymbol{n}\cdot\nabla C^{n+1} + \xi\boldsymbol{n}\cdot\left[m(\bar{\phi}^{n+1})\nabla\bar{\mu}^{n+1} - m_c(\phi_0)\nabla\bar{C}^{n+1}\right] = d_a^{n+1} \text{ on } \partial\Omega, \tag{4.32e}$$

and

$$\begin{aligned} D_{\mathscr{F}}(R)\Big|^{n+1}\frac{dR}{dt}\Big|^{n+1} &= \int_\Omega \left[-\lambda\nabla^2\phi^{n+1} + h(\phi^{n+1})\right]\frac{\partial\phi}{\partial t}\Big|^{n+1}d\Omega - \xi\int_\Omega m(\tilde{\phi}^{n+1})\left|\nabla\tilde{\mu}^{n+1}\right|^2 d\Omega \\ &+ \int_\Omega \left[\lambda\nabla^2\phi^{n+1} - h(\phi^{n+1})\right]\left\{\nabla\cdot\left(m_c(\phi_0)\nabla C^{n+1}\right) + \xi\nabla\cdot\left[m(\bar{\phi}^{n+1})\nabla\bar{\mu}^{n+1} - m_c(\phi_0)\nabla\bar{C}^{n+1}\right] + f^{n+1}\right\}d\Omega \\ &+ \xi\left[\int_\Omega f^{n+1}\tilde{\mu}^{n+1}d\Omega + \int_\Gamma\left(d_a^{n+1}\tilde{\mu}^{n+1} + \lambda d_b^{n+1}\frac{\partial\phi}{\partial t}\Big|^{*,n+1}\right)d\Gamma\right] \\ &+ (1-\xi)\left|\int_\Omega f^{n+1}\tilde{\mu}^{n+1}d\Omega + \int_\Gamma\left(d_a^{n+1}\tilde{\mu}^{n+1} + \lambda d_b^{n+1}\frac{\partial\phi}{\partial t}\Big|^{*,n+1}\right)d\Gamma\right|, \end{aligned} \tag{4.33}$$

together with the boundary condition (4.9c) and the initial condition (4.11). In these equations, $\frac{\partial\phi}{\partial t}\Big|^{n+1}$ and $\frac{dR}{dt}\Big|^{n+1}$ are defined in (2.21b), $\bar{\phi}^{n+1}$ is given by (2.21c), and $\bar{C}^{n+1}$ and $\bar{\mu}^{n+1}$ are computed by

$$\bar{C}^{n+1} = -\lambda\nabla^2\bar{\phi}^{n+1} + \kappa(\phi_0)\bar{\phi}^{n+1}, \quad \bar{\mu}^{n+1} = -\lambda\nabla^2\bar{\phi}^{n+1} + h(\bar{\phi}^{n+1}). \tag{4.34}$$

$\tilde{\phi}^{n+1}$, $\tilde{\phi}^{n+3/2}$, $\tilde{\mu}^{n+1}$, and $\frac{\partial\phi}{\partial t}\Big|^{*,n+1}$ are approximations to be specified later.

Theorem 4.2. *In the absence of the external source term ($f = 0$), and with zero boundary conditions ($d_a = d_b = 0$), the scheme consisting of (4.32)-(4.33) is unconditionally energy stable in the sense that*

$$\mathscr{F}(R^{n+\frac{3}{2}}) - \mathscr{F}(R^{n+\frac{1}{2}}) = -\xi\Delta t\int_\Omega m(\tilde{\phi}^{n+1})|\nabla\tilde{\mu}^{n+1}|^2 \le 0, \tag{4.35}$$

if the approximation of $R(t)$ at time step $\frac{1}{2}$ is positive.

Proof. We take the L^2 inner product between $\left(-\lambda\nabla^2\phi^{n+1} + h(\phi^{n+1})\right)$ and equation (4.32a), and add the resultant equation to equation (4.33). This leads to

$$\begin{aligned} \frac{\mathscr{F}(R^{n+\frac{3}{2}}) - \mathscr{F}(R^{n+\frac{1}{2}})}{\Delta t} &= \xi\left\{\int_\Omega f^{n+1}\tilde{\mu}^{n+1}d\Omega + \int_\Gamma\left(d_a^{n+1}\tilde{\mu}^{n+1} + \lambda d_b^{n+1}\frac{\partial\phi}{\partial t}\Big|^{*,n+1}\right)d\Gamma\right\} \\ &- \xi\int_\Omega m(\tilde{\phi}^{n+1})|\nabla\tilde{\mu}^{n+1}|^2 d\Omega + (1-\xi)\left|\int_\Omega f^{n+1}\tilde{\mu}^{n+1} + \int_\Gamma\left(d_a^{n+1}\tilde{\mu}^{n+1} + \lambda d_b^{n+1}\frac{\partial\phi}{\partial t}\Big|^{*,n+1}\right)d\Gamma\right|. \end{aligned} \tag{4.36}$$

By the same arguments as in the proof of Theorem 4.1, we arrive at the relation (4.35) based on the above equation. □

For implementation of the scheme, one notes that equation (4.32a) can be transformed into

$$\frac{\gamma_0}{\Delta t}\phi^{n+1} - \nabla\cdot\left[m_c(\phi_0)\nabla C^{n+1}\right] = \left(\frac{\hat{\phi}}{\Delta t} + f^{n+1}\right) + \xi\nabla\cdot\left[m(\bar{\phi}^{n+1})\nabla\bar{\mu}^{n+1} - m_c(\phi_0)\nabla\bar{C}^{n+1}\right]. \tag{4.37}$$

Barring the unknown scalar ξ, equations (4.37), (4.32b), (4.32e) and (4.9c) can be solved as follows. Introduce two pairs of field functions $(\phi_i^{n+1}, C_i^{n+1})$ $(i = 1, 2)$, as the solution of the following equations:

For $(\phi_1^{n+1}, C_1^{n+1})$:

$$\frac{\gamma_0}{\Delta t}\phi_1^{n+1} - \nabla\cdot\left[m_c(\phi_0)\nabla C_1^{n+1}\right] = \frac{\hat{\phi}}{\Delta t} + f^{n+1}, \tag{4.38a}$$

$$\big(\kappa(\phi_0)+S\big)\phi_1^{n+1} - \lambda\nabla^2\phi_1^{n+1} - C_1^{n+1} = S\bar{\phi}^{n+1}, \tag{4.38b}$$

$$m_c(\phi_0)\boldsymbol{n}\cdot\nabla C_1^{n+1} = d_a^{n+1}, \ \text{ on } \Gamma \tag{4.38c}$$

$$\boldsymbol{n}\cdot\nabla\phi_1^{n+1} = d_b^{n+1}, \ \text{ on } \Gamma. \tag{4.38d}$$

For $(\phi_2^{n+1}, C_2^{n+1})$:

$$\frac{\gamma_0}{\Delta t}\phi_2^{n+1} - \nabla\cdot\left[m_c(\phi_0)\nabla C_2^{n+1}\right] = \nabla\cdot\left[m(\bar{\phi}^{n+1})\nabla\bar{\mu}^{n+1} - m_c(\phi_0)\nabla\bar{C}^{n+1}\right], \tag{4.39a}$$

$$\big(\kappa(\phi_0)+S\big)\phi_2^{n+1} - \lambda\nabla^2\phi_2^{n+1} - C_2^{n+1} = 0, \tag{4.39b}$$

$$m_c(\phi_0)\boldsymbol{n}\cdot\nabla C_2^{n+1} = -\boldsymbol{n}\cdot\left[m(\bar{\phi}^{n+1})\nabla\bar{\mu}^{n+1} - m_c(\phi_0)\nabla\bar{C}^{n+1}\right], \ \text{ on } \Gamma, \tag{4.39c}$$

$$\boldsymbol{n}\cdot\nabla\phi_2^{n+1} = 0, \ \text{ on } \Gamma. \tag{4.39d}$$

Then for given scalar value ξ, the following field functions solve the system consisting of equations (4.32a)-(4.32e) and (4.9c):

$$C^{n+1} = C_1^{n+1} + \xi C_2^{n+1}, \quad \phi^{n+1} = \phi_1^{n+1} + \xi\phi_2^{n+1} \tag{4.40}$$

where $(C_i^{n+1}, \phi_i^{n+1})$ $(i=1,2)$ are given by equations (4.38)-(4.39).

The unknown scalar value ξ remains to be determined. Following equation (2.42), $\tilde{\phi}^{n+1}$, $\tilde{\mu}^{n+1}$ and $\frac{\partial\phi}{\partial t}\Big|^{*,n+1}$ are again given by equations (4.24) and (4.25), where based on equation (4.32b) we compute $\nabla^2\tilde{\phi}^{n+1}$ by

$$\nabla^2\tilde{\phi}^{n+1} = \frac{1}{\lambda}\Big[\kappa(\phi_0)\tilde{\phi}^{n+1} + S(\tilde{\phi}^{n+1} - \bar{\phi}^{n+1}) - (C_1^{n+1} + C_2^{n+1})\Big]. \tag{4.41}$$

The approximation $\tilde{\phi}^{n+\frac{3}{2}}$ is given by (4.26). As a result, ξ can be computed by,

$$\xi = \frac{\mathscr{F}(R^{n+1/2}) + \Delta t|S_0|}{E[\tilde{\phi}^{n+\frac{3}{2}}] + \Delta t\int_\Omega m(\tilde{\phi}^{n+1})|\nabla\tilde{\mu}^{n+1}|^2 d\Omega + \Delta t(|S_0| - S_0)}, \tag{4.42}$$

where S_0 is given by (4.28), and ϕ^{n+1} and R^{n+1} can be evaluated by equations (4.40) and (2.45), respectively.

Equations (4.38)-(4.39) can be discretized in space by C^0 spectral elements, and their weak forms are:

For $(\phi_1^{n+1}, C_1^{n+1})$: Find $\phi_1^{n+1}, C_1^{n+1} \in H^1(\Omega)$ such that

$$\frac{\gamma_0}{\Delta t}(\phi_1^{n+1}, \varphi)_\Omega + \Big(m_c(\phi_0)\nabla C_1^{n+1}, \nabla\varphi\Big)_\Omega = \Big(\frac{\hat{\phi}}{\Delta t} + f^{n+1}, \varphi\Big)_\Omega + \langle d_a^{n+1}\varphi\rangle_\Gamma, \tag{4.43}$$

$$\Big(\big[\kappa(\phi_0)+S\big]\phi_1^{n+1}, \varphi\Big)_\Omega + \lambda\big(\nabla\phi_1^{n+1}, \nabla\varphi\big)_\Omega - \big(C_1^{n+1}, \varphi\big)_\Omega = S\big(\bar{\phi}^{n+1}, \varphi\big)_\Omega + \lambda\langle d_b^{n+1}\varphi\rangle_\Gamma, \tag{4.44}$$

for all $\varphi \in H^1(\Omega)$.

For $(\phi_2^{n+1}, C_2^{n+1})$: Find $\phi_2^{n+1}, C_2^{n+1} \in H^1(\Omega)$ such that

$$\frac{\gamma_0}{\Delta t}(\phi_2^{n+1}, \varphi)_\Omega + \Big(m_c(\phi_0)\nabla C_2^{n+1}, \nabla\varphi\Big)_\Omega = -\Big(m(\bar{\phi}^{n+1})\nabla\bar{\mu}^{n+1} - m_c(\phi_0)\nabla\bar{C}^{n+1}, \nabla\varphi\Big)_\Omega, \tag{4.45}$$

$$\Big(\big[\kappa(\phi_0)+S\big]\phi_2^{n+1}, \varphi\Big)_\Omega + \lambda\big(\nabla\phi_2^{n+1}, \nabla\varphi\big)_\Omega - \big(C_2^{n+1}, \varphi\big)_\Omega = 0, \tag{4.46}$$

for all $\varphi \in H^1(\Omega)$.

Remark 4.2. If one chooses $\kappa(\phi_0) = 0$ and $m_c(\phi_0) = m_0 > 0$, then the scheme (4.32a)–(4.33) can also be implemented by solving four de-coupled Helmholtz type equations in a way similar to the constant mobility case in Section 4.1.

4.3. Numerical results

We next provide numerical examples to demonstrate the accuracy and unconditional stability of the proposed schemes (4.9)-(4.10) and (4.32)-(4.33) for Cahn-Hilliard equation with constant and variable mobilities. For cases with variable mobility we employ $m_c(\phi_0) = m(\phi_0) = \max(m_0(1-\phi_0^2), 0)$ in the algorithm with these tests, where m_0 and ϕ_0 will be specified below.

Table 1
Simulation parameter values for convergence tests of Cahn-Hilliard equation.

Parameter	Value	Parameter	Value
C_0	1	λ	0.01
m_0	0.01	η	0.1
t_0	0.1	t_f	0.2 (spatial tests) or 1.1 (temporal tests)
Element order	(varied)	Elements	2
Δt	(varied)	$\Delta t_{\min}$	$1e-4$
S	1 (variable mobility solver)	S	$\sqrt{\frac{4\gamma_0\lambda}{m_0\Delta t}}$ or $\sqrt{\frac{4\gamma_0\lambda}{m_0\Delta t_{\min}}}$ (constant mobility solver)
$\mathscr{F}(R)$	R	ϕ_0	ϕ_{in}
$m_c(\phi_0)$	$m(\phi_0)$		

4.3.1. Convergence rates

Consider domain $\Omega = [0,2]\times[-1,1]$ and a contrived solution in this domain:

$$\phi(\boldsymbol{x},t) = \cos(\pi x)\cos(\pi y)\sin(t). \tag{4.47}$$

The external force and boundary source terms $f(\boldsymbol{x},t)$, $d_a(\boldsymbol{x},t)$ and $d_b(\boldsymbol{x},t)$ in (4.1a), (4.1c) and (4.1d) are chosen such that the analytic expression (4.47) satisfies (4.1).

The computational domain Ω is discretized with two equal-sized quadrilateral elements. The algorithms (4.9)-(4.10) for the constant-mobility case and (4.32)-(4.33) for the variable-mobility case are employed to numerically integrate the Cahn-Hilliard equation from $t=t_0$ to $t=t_f$. The initial field function ϕ_{in} is obtained by setting $t=t_0$ in the contrived solution (4.47). The numerical errors are computed by comparing the numerical solution against the analytic solution (4.47) at $t=t_f$. In the following convergence tests, we fix $\mathscr{F}(R)=R$, $C_0=1$, and $\phi_0=\phi_{in}(\boldsymbol{x})$ in (4.32). The values for the simulation parameters are summarized in Table 1.

In the spatial convergence test, we fix $\Delta t=0.001$, $t_0=0.1$ and $t_f=0.2$, and vary the element order systematically from 2 to 20. The numerical errors in L^∞ and L^2 norms at $t=t_f$ are then recorded. For the algorithm with constant mobility, S in equation (4.9) is chosen as $S=\sqrt{\frac{4\gamma_0\lambda}{m_0\Delta t}}$, while for the algorithm with variable mobility we use $S=1$. Figs. 4.1(a) and (b) show the numerical errors as a function of the element order from these tests. It can be observed that the errors decrease exponentially with increasing element order and that the error curves level off at around 10^{-5} and 10^{-6} beyond element order 8 and 10, respectively for these two solvers, due to the saturation of temporal errors.

In the temporal convergence test, we fix the element order at a large value 18, $t_0=0.1$, and $t_f=1.1$, and vary Δt systematically from 0.2 to 1.953125×10^{-4} to study the behavior of numerical errors. For the constant-mobility case, $S=\sqrt{\frac{4\gamma_0\lambda}{m_0\Delta t_{\min}}}$ (where $\Delta t_{\min}=10^{-4}$), while for the variable-mobility case $S=1$. Figs. 4.1(c) and (d) show the numerical errors as a function of Δt for these cases. We observe a second-order convergence rate in time for both cases.

Fig. 4.2 shows the spatial and temporal convergence behaviors of the method in terms of the L^2 errors of ϕ corresponding to several different mapping functions $\mathscr{F}(R)$. Note that $e_0=\kappa_0=10$ with the mapping $\mathscr{F}(R)=\frac{e_0}{2}\ln\left(\frac{\kappa_0+R}{\kappa_0-R}\right)$. These results demonstrate the spatial exponential convergence rate and the second-order temporal convergence rate of the method with various mapping functions, and also the insensitivity of the simulation results with respect to $\mathscr{F}(R)$.

4.3.2. Constant mobility: coalescence of two drops

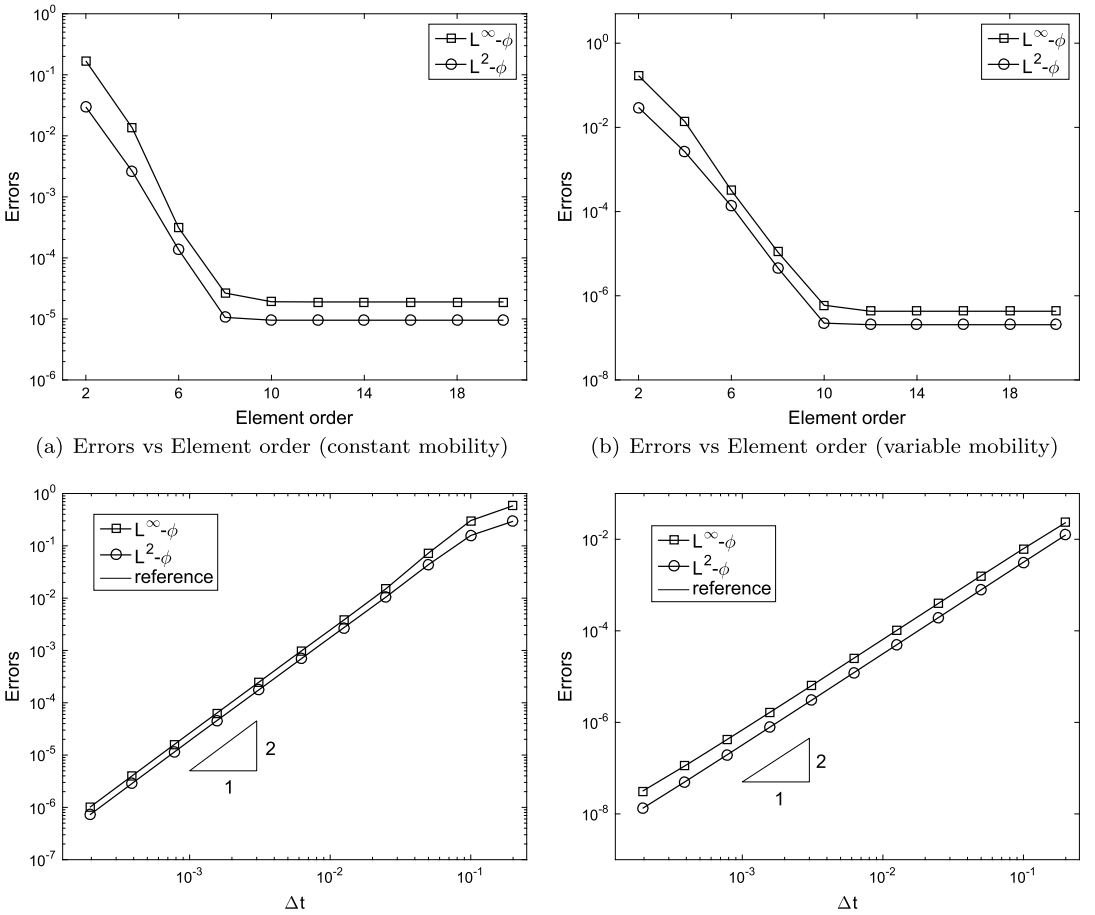
We next consider the coalescence of two drops to demonstrate the numerical properties of the proposed scheme (4.9)-(4.10) for problems with constant mobility. Consider a square domain $\Omega=[0,1]^2$ and two materials contained in this domain. It is assumed that the dynamics of the material regions is governed by the Cahn-Hilliard equation with a constant mobility, $m(\phi)=m_0>0$, and that $\phi=1$ and $\phi=-1$ correspond to the bulk of the first and second materials, respectively. We assume that at $t=0$ the first material occupies two circular regions that are right next to each other and the rest of the domain is filled by the second material.

To be more specific, the initial distribution of the material takes the form

$$\phi_{in}(\boldsymbol{x}) = 1-\tanh\frac{|\boldsymbol{x}-\boldsymbol{x}_0|-R_0}{\sqrt{2}\eta}-\tanh\frac{|\boldsymbol{x}-\boldsymbol{x}_1|-R_0}{\sqrt{2}\eta}, \tag{4.48}$$

where $\boldsymbol{x}_0=(x_0,y_0)=(0.3,0.5)$ and $\boldsymbol{x}_1=(0.7,0.5)$ are the centers of the circular regions for the first material, and $R_0=0.19$ is the radius of these circles. The external force and the boundary source terms in (4.1) are set to $f(\boldsymbol{x},t)=d_a(\boldsymbol{x},t)=d_b(\boldsymbol{x},t)=0$. We discretize the domain using 400 equal-sized quadrilateral elements with element order 10. We employ a mapping function $\mathscr{F}(R)=R^2$ for this problem. The simulation parameters are listed as follows:

$$\eta=0.01,\quad \sigma=151.15,\quad \lambda=\frac{3}{2\sqrt{2}}\sigma\eta,\quad m_0=\frac{10^{-6}}{\lambda},\quad S=\sqrt{\frac{4\gamma_0\lambda}{m_0\Delta t}},\quad C_0=10^6. \tag{4.49}$$



(a) Errors vs Element order (constant mobility)

(b) Errors vs Element order (variable mobility)

(c) Errors vs Δt (constant mobility)

(d) Errors vs Δt (variable mobility)

Fig. 4.1. Spatial/temporal convergence tests for Cahn-Hilliard equation. L^2 and L^∞ errors of ϕ versus element order for (a) constant mobility, (b) variable mobility (fixed $\Delta t = 0.001$, $t_0 = 0.1$, $t_f = 0.2$). L^2 and L^∞ errors of ϕ versus Δt for (c) constant mobility, (d) variable mobility (fixed element order 18 and $t_0 = 0.1$, $t_f = 1.1$). Numerical results correspond to $\phi_0 = \phi_{in}(\boldsymbol{x})$ in (4.32) for cases with variable mobility.

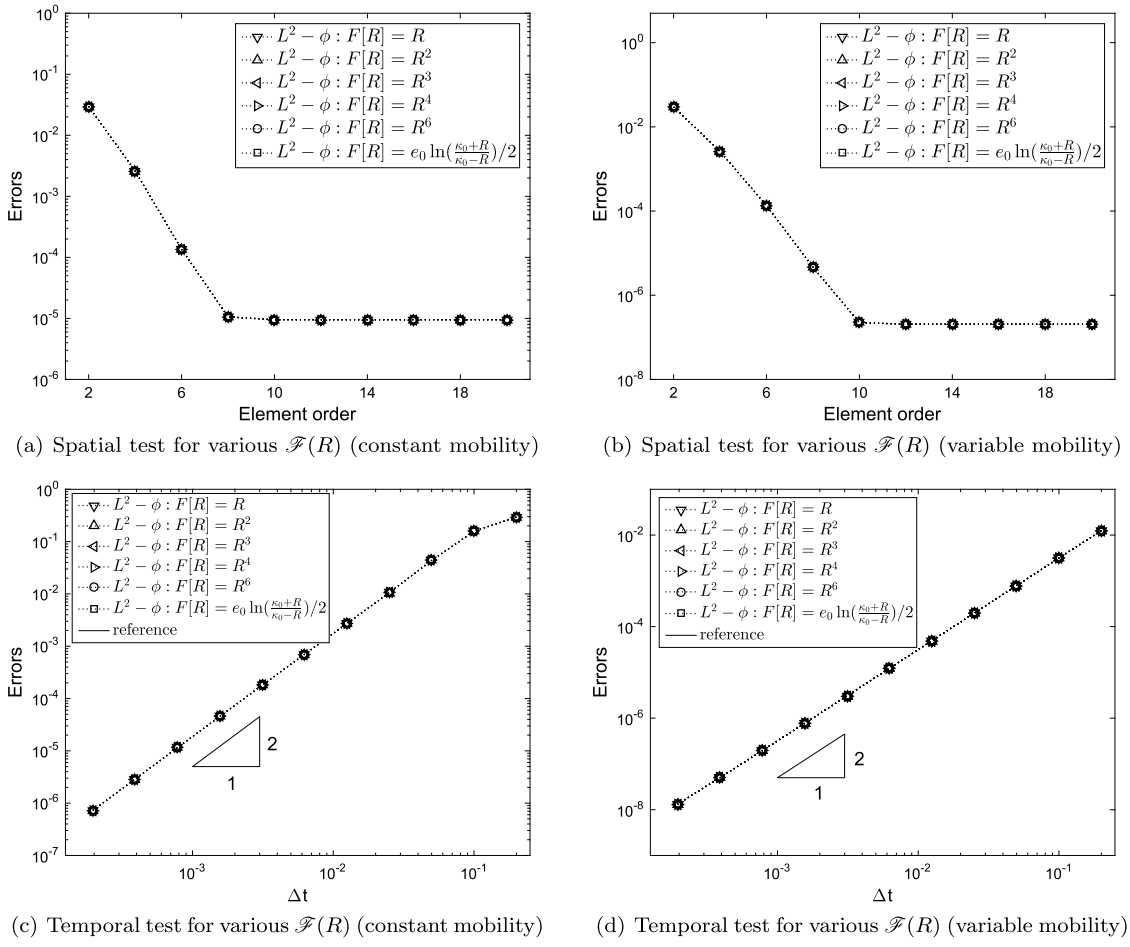
Fig. 4.3 shows the evolution of the two material regions with a temporal sequence of snapshots of the interfaces between these two materials visualized by the contour level $\phi = 0$. It can be observed that the two separate regions of the first material gradually coalescence with each other to form a single drop under the Cahn-Hilliard dynamics.

To investigate the effect of time step size on the accuracy of the simulation results, in Fig. 4.4 we compare the distributions of the material interfaces at $t = 50$ obtained with several time step sizes, ranging from $\Delta t = 10^{-1}$ to $\Delta t = 10^{-4}$. The distribution computed with $\Delta t = 10^{-2}$, 10^{-3} and 10^{-4} are essentially the same. With the larger time step size $\Delta t = 10^{-1}$, some difference can be noticed in the material distribution compared with those obtained using smaller Δt values. This suggests the simulation is starting to lose accuracy with time step sizes $\Delta t = 10^{-1}$ and larger.

Fig. 4.5 shows the time histories of the total energy $E_{tot}(t)$ (see equation (4.3)) and the ratio $\xi = \frac{\mathscr{F}(R)}{E}$ obtained using time step sizes $\Delta t = 10^{-2}$ to $\Delta t = 10^{-4}$. It can be observed that the history curves essentially overlap with one another for different time step sizes. The computed values for $\xi = \frac{\mathscr{F}(R)}{E}$ are very close to 1 for each Δt, suggesting that $\mathscr{F}(R)$ is a good approximation for $E(t)$ and the numerical approximation is accurate with these time steps.

Thanks to the energy stability property of the current method, we can use fairly large time step sizes for the simulations. In Fig. 4.6, we depict some longer time histories (up to $t = 10000$) of the total energy $E_{tot}(t)$ and the ratio $\xi = \frac{\mathscr{F}(R)}{E}$ obtained using several large time step sizes $\Delta t = 0.1, 1, 10$. At these large Δt values we can no longer expect the results to be accurate. Indeed, in Fig. 4.6(a), E_{tot} increases initially, and levels off over time at around $E_{tot} \approx 2000$. Meanwhile, ξ decreases rapidly to a smaller number close to 0, suggesting that there is a large discrepancy between $\mathscr{F}(R)$ and $E(t)$. While these computation results are not accurate, they nonetheless demonstrate the proposed method is stable and robust with large time steps.

As discussed in previous sections, the current scheme guarantees the positivity of the computed ξ and $R(t)$ values, regardless of the time step size or the external forces. In Fig. 4.7, we compare the time histories of the computed auxiliary variable $R(t)$ obtained using the current method and the auxiliary variable obtained by the scalar auxiliary variable



(a) Spatial test for various $\mathscr{F}(R)$ (constant mobility)

(b) Spatial test for various $\mathscr{F}(R)$ (variable mobility)

(c) Temporal test for various $\mathscr{F}(R)$ (constant mobility)

(d) Temporal test for various $\mathscr{F}(R)$ (variable mobility)

Fig. 4.2. Spatial/temporal convergence tests for the Cahn-Hilliard equation obtained using several mapping functions $\mathscr{F}(R)$ as given in the legend. L^2 errors of ϕ versus the element order for (a) constant mobility, (b) variable mobility (fixed $\Delta t = 0.001$, $t_0 = 0.1$, $t_f = 0.2$). L^2 errors of ϕ versus Δt for (c) constant mobility, (d) variable mobility (fixed element order 18 and $t_0 = 0.1$, $t_f = 1.1$).

(SAV) method from [38]. Note that in the current method $\mathscr{F}(R) = R^2$ (hence $R(t) = \sqrt{E(t)}$) has been used. In the SAV method the auxiliary variable, denoted by $R_1(t)$ in the figure for clarity, is computed by a dynamic equation stemming from the relation $R_1(t) = \sqrt{E_1(t)}$, where $E_1(t) = \int_\Omega H(\phi) d_\Omega + C_0 > 0$. Therefore, $R_1(t)$ should physically be positive. In reality, however, the discrete solutions for $R_1(t)$ computed by the SAV method can become negative. This is evident from Fig. 4.7(b), where the result obtained using the SAV method with a large $\Delta t = 1$ is shown. On the other hand, the discrete solutions for $R(t)$ from the current method are guaranteed to be positive, which is evident from Fig. 4.7(a). It should be pointed out that since the definitions of the auxiliary variables in the current method and in the SAV method are different, the specific values of $R(t)$ and $R_1(t)$ are not comparable. Nevertheless, the positivity of these variables are both expected physically.

4.3.3. Variable mobility: evolution of a drop

We next consider the evolution of a square drop governed by the Cahn-Hilliard equation with a variable mobility. The computational domain and the settings follow those for the coalescence of two drops discussed above. The difference lies in the initial distribution of the materials. To be precise, the initial distribution of field function is set as follows:

$$\phi_{in}(\boldsymbol{x}) = \frac{1}{2}\Big[\tanh\frac{x-x_0+h_0}{\sqrt{2}\eta} - \tanh\frac{x-x_0-h_0}{\sqrt{2}\eta}\Big]\cdot\Big[\tanh\frac{y-y_0+h_0}{\sqrt{2}\eta} - \tanh\frac{y-y_0-h_0}{\sqrt{2}\eta}\Big] - 1, \tag{4.50}$$

where $(x_0, y_0) = (0.5, 0.5)$ is the center of the domain and $h_0 = 0.2$.

Fig. 4.8 shows the evolution of the system with a temporal sequence of snapshots of the interfaces between the two materials. These results are computed with a time step size $\Delta t = 0.01$, $S = 1$, $C_0 = 10^6$, and the mapping function $\mathscr{F}(R) = R^2$. The ϕ_0 in the algorithm is taken as the field $\phi(\boldsymbol{x}, t)$ at every fifth time step, i.e. $\phi_0(\boldsymbol{x}) = \phi^{5k}(\boldsymbol{x})$ $(k = 0, 1, 2\ldots)$. In other words, the ϕ_0 field and also the coefficient matrices of the system are updated every 5 time steps in this set of tests. These

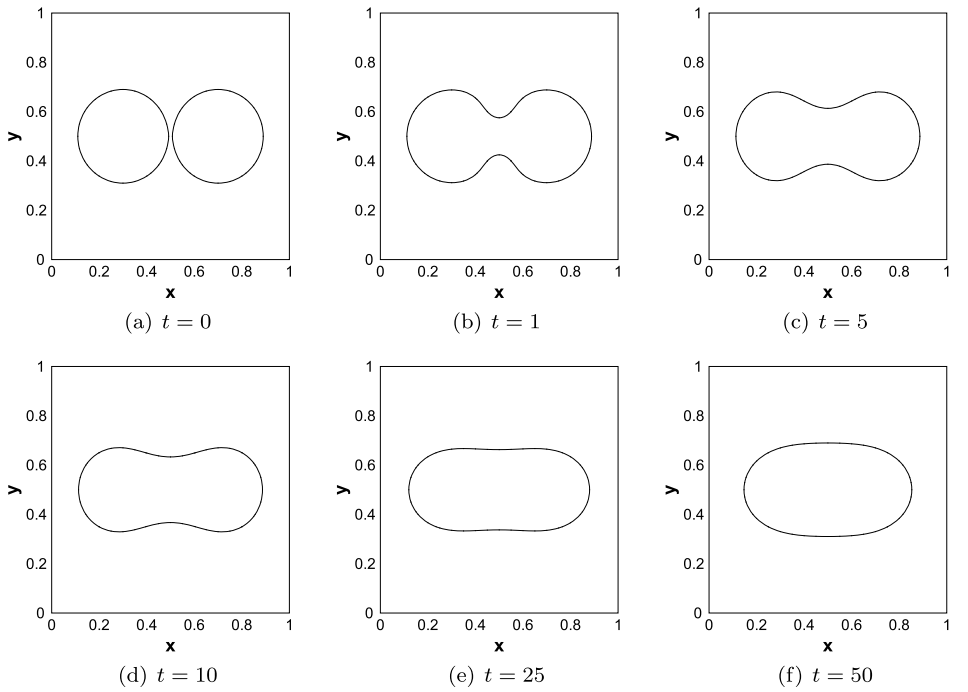


Fig. 4.3. Temporal sequence of snapshots showing the coalescence of two circular drops visualized by the contour level $\phi = 0$ governed by Cahn-Hilliard equation with constant mobility.

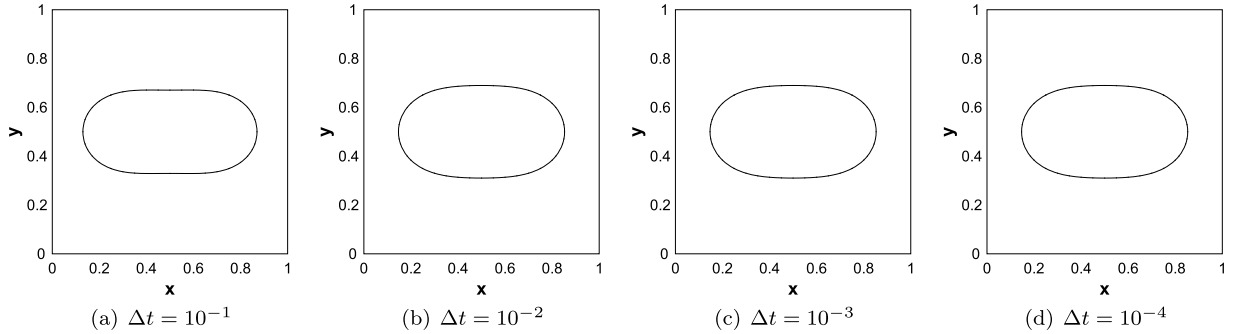


Fig. 4.4. Coalescence of two drops: snapshots of material interfaces at $t = 50$ computed using (a) $\Delta t = 10^{-1}$, (b) $\Delta t = 10^{-2}$, (c) $\Delta t = 10^{-3}$, (d) $\Delta t = 10^{-4}$.

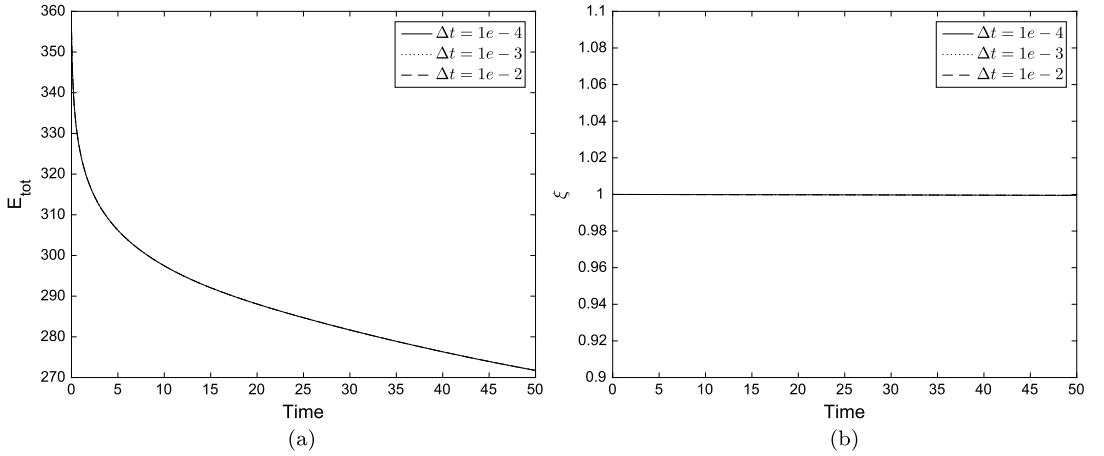
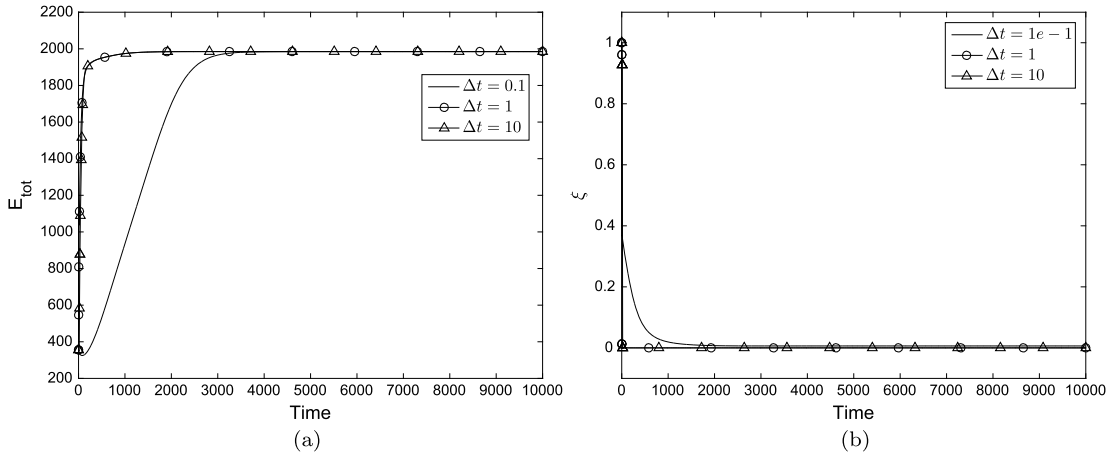
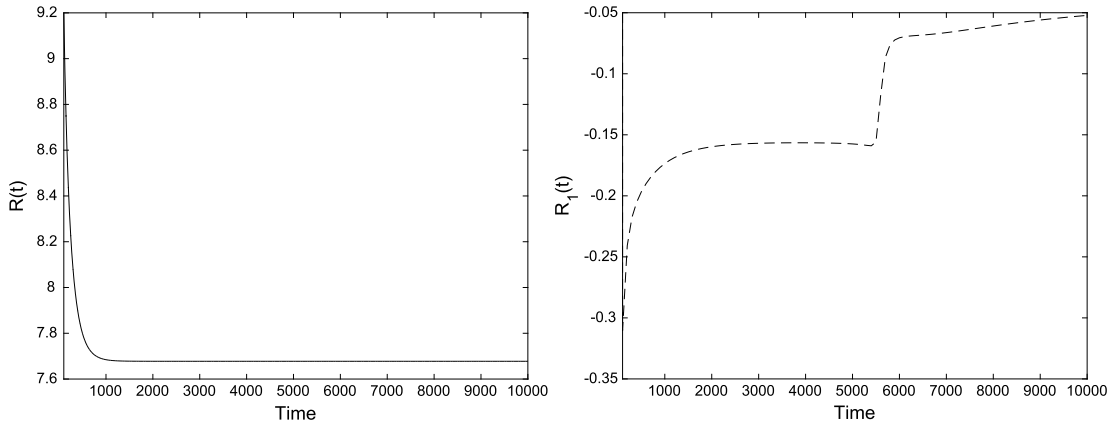


Fig. 4.5. Coalescence of two drops: time histories of (a) $E_{tot}(t)$ and (b) ξ corresponding to a range of smaller time step sizes $\Delta t = 10^{-2}, 10^{-3}, 10^{-4}$.



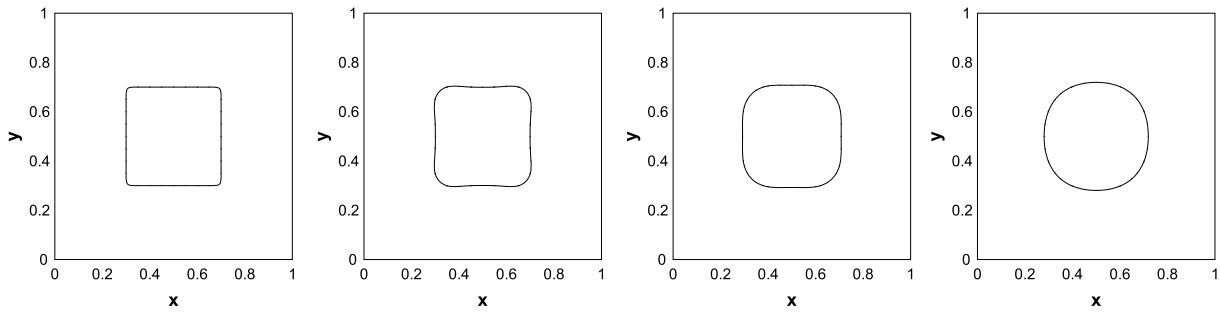
(a) (b)

Fig. 4.6. Coalescence of two drops: time history of (a) $E_{tot}(t)$ and (b) ξ for several large time step sizes $\Delta t = 0.1, 1, 10$.



(a) Current method (b) SAV method

Fig. 4.7. Positivity of computed auxiliary variables: (a) time history of the auxiliary variable ($R(t)$) obtained by the current method, and (b) time history of the auxiliary variable (denoted by $R_1(t)$) obtained by the SAV method of [38]. $\Delta t = 1$ in the simulations of coalescence of two drops (Cahn-Hilliard equation with constant mobility). Note that the auxiliary variable in the current work and that in [38] are not the same in definition, and so their specific values are not comparable. But both should physically be positive.



(a) $t = 0$ (b) $t = 1$ (c) $t = 10$ (d) $t = 30$

Fig. 4.8. Evolution of a square drop (Cahn-Hilliard equation with variable mobility): Temporal snapshots of the material interface visualized by $\phi = 0$. $\Delta t = 10^{-2}$ in the simulations.

results illustrate the process for the evolution of the initial square region into a circular region under the Cahn-Hilliard dynamics.

In Fig. 4.9, we show the time histories of the total energy $E_{tot}(t)$ and $\xi = \frac{\mathscr{F}(R)}{E(t)}$ obtained with several time step sizes ranging from $\Delta t = 10^{-2}$ to $\Delta t = 10^{-4}$. Note that the variable mobility is $m(\phi) = \max(m_0(1-\phi^2), 0)$. Here we have considered two ways to simulate the problem:

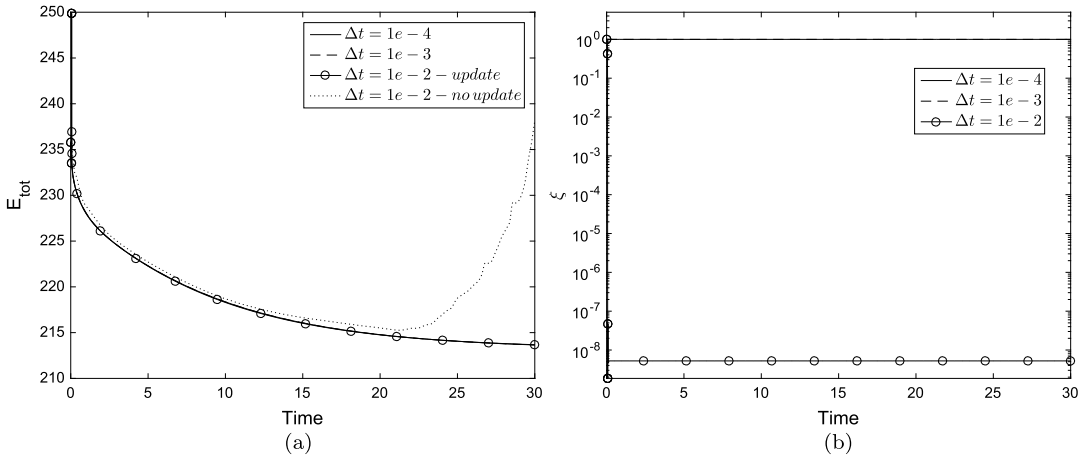


Fig. 4.9. Evolution of a square drop (Cahn-Hilliard equation with variable mobility): time histories of (a) $E_{tot}(t)$ and (b) ξ for various time step sizes $\Delta t = 10^{-2}, 10^{-3}, 10^{-4}$. In these tests, $\phi_0 = 0$ for $\Delta t = 10^{-3}$ and 10^{-4}, while for $\Delta t = 10^{-2}$, we set $\phi_0 = 0$, referred to as "no update" in the legend of (a), and also update ϕ_0 to ϕ^n every 5 time steps, referred to as "update". The case $\Delta t = 10^{-2}$ in (b) corresponds to the "update" case.

- by setting $\phi_0 = 0$ in the algorithm. This leads to $m_c(\phi_0) = m(\phi_0) = m_0$ and $\kappa(\phi_0) = -\frac{\lambda}{\eta^2}$, and a time-independent coefficient matrix for the system, which can be pre-computed. We refer to this setting as the standard way.
- by setting $\phi_0 = \phi^{5k}$ $(k = 0, 1, 2, \dots)$ in the algorithm. The ϕ_0 field and the coefficient matrix are quasi time-independent, and they are updated every 5 time steps.

With the smaller time step sizes $\Delta t = 10^{-3}$ and 10^{-4}, we set $\phi_0 = 0$ in the algorithm (the standard way) when performing simulations. With the larger $\Delta t = 10^{-2}$, we have conducted simulations in both ways with the algorithm. In Fig. 4.9(a) the results from these two settings are marked by "no update" (standard way) and "update" (second way) in the legend corresponding to $\Delta t = 10^{-2}$. It is observed that the energy histories corresponding to $\Delta t = 10^{-4}$ and 10^{-3}, and $\Delta t = 10^{-2}$ with ϕ_0 updated periodically, essentially overlap with each other. However, the energy history corresponding to $\Delta t = 10^{-2}$ with $\phi_0 = 0$ exhibits a pronounced discrepancy compared with the other cases. These results indicate that with the standard way (by setting $\phi_0 = 0$) in the algorithm the simulation result would cease to be accurate when the time step size increases to $\Delta t = 10^{-2}$. However, if one uses the second way (by updating ϕ_0 periodically), accurate simulation result can be obtained even with $\Delta t = 10^{-2}$. In other words, by updating ϕ_0 in the algorithm from time to time, one can improve the accuracy of the simulations even at larger time step sizes. We depict in Fig. 4.9(b) the time histories of $\xi = \frac{\mathscr{F}(R)}{E}$ corresponding to these time step sizes. Shown for $\Delta t = 10^{-2}$ in this plot is the result with ϕ_0 updated periodically. It is observed that the computed ξ is essentially 1 with $\Delta t = 10^{-3}$ and 10^{-4}. With $\Delta t = 10^{-2}$ (and ϕ_0 updated periodically), the computed ξ is substantially smaller than 1. But interestingly, the simulation results for the field function ϕ are still quite accurate with this larger Δt. This group of tests suggests that one possible way to improve the accuracy of the proposed energy-stable scheme is to update the ϕ_0 in the algorithm periodically, e.g. every N time steps. By choosing an appropriate N for a given problem, one can enhance the simulation accuracy even at large or fairly large time step sizes. Because ϕ_0 and the coefficient matrix for the system only need to be updated infrequently, the cost associated with updating the coefficient matrix can be manageable. There is a drawback with this, however. The computations using the second way (updating ϕ_0 periodically) seems not as robust as the standard way (by setting $\phi_0 = 0$) for large Δt. Because of the non-zero ϕ_0 field in the algorithm, the conditioning of the system coefficient matrix using the second way seems to become worse for large Δt. We observe that for larger $\Delta t \geqslant 0.1$ the system coefficient matrix using the second way can become singular and the computation may break down.

5. Nonlinear Klein-Gordon equation

We consider an energy-conserving system, the nonlinear Klein-Gordon equation, in this section and apply the gPAV method to this system. Consider the nonlinear Klein-Gordon equation [39] on a domain Ω (with boundary Γ)

$$\frac{\partial u}{\partial t} = v, \tag{5.1}$$

$$\varepsilon^2 \frac{\partial v}{\partial t} - \alpha^2 \nabla^2 u + \varepsilon_1^2 u + g(u) = f(\boldsymbol{x}, t), \tag{5.2}$$

$$u = d_a(\boldsymbol{x}, t), \ \text{ on } \Gamma, \tag{5.3}$$

where ε, α and ε_1 are positive constants. These equations are supplemented by the initial conditions

$$u(\boldsymbol{x},0)=u_{in}(\boldsymbol{x}),\quad v(\boldsymbol{x},0)=v_{in}(\boldsymbol{x})\ \text{ in }\Omega. \tag{5.4}$$

In these equations $g(u)=G'(u)$ and $G(u)$ is a potential energy function with $G(u)\geqslant 0$. The above system satisfies the following energy balance law:

$$\frac{\partial}{\partial t}\int_\Omega\Big(\frac{\alpha^2}{2}|\nabla u|^2+\frac{\varepsilon_1^2}{2}|u|^2+\frac{\varepsilon^2}{2}|v|^2+G(u)\Big)d\Omega=\int_\Omega f v d\Omega+\alpha^2\int_{\partial\Omega}(\boldsymbol{n}\cdot\nabla u)v d\Gamma. \tag{5.5}$$

We define a shifted total energy according to equation (2.11),

$$E(t)=E[u,v]=\int_\Omega\Big(\frac{\alpha^2}{2}|\nabla u|^2+\frac{\varepsilon_1^2}{2}|u|^2+\frac{\varepsilon^2}{2}|v|^2+G(u)\Big)d\Omega+C_0, \tag{5.6}$$

where C_0 is chosen such that $E(t)>0$. Choose $\mathscr{F}$ and $\mathscr{G}$, and define the auxiliary variable $R(t)$ based on equation (2.13a). Following equation (2.14), we have

$$\mathscr{F}'(R)\frac{dR}{dt}=\int_\Omega\Big(-\alpha^2\nabla^2u+\varepsilon_1^2u+g(u)\Big)\frac{\partial u}{\partial t}d\Omega+\int_\Omega\varepsilon^2v\frac{\partial v}{\partial t}d\Omega+\alpha^2\int_\Gamma(\boldsymbol{n}\cdot\nabla u)\frac{\partial u}{\partial t}d\Gamma, \tag{5.7}$$

where integration by part has been used.

Following equations (2.17)-(2.19), we reformulate equations (5.2) and (5.7) into

$$\frac{\partial v}{\partial t}=\Big(\frac{\alpha}{\varepsilon}\Big)^2\nabla^2u-\Big(\frac{\varepsilon_1}{\varepsilon}\Big)^2u-\frac{\mathscr{F}(R)}{E}\frac{1}{\varepsilon^2}g(u)+\frac{1}{\varepsilon^2}f, \tag{5.8a}$$

$$\begin{aligned}\mathscr{F}'(R)\frac{dR}{dt}=&\int_\Omega\Big(-\alpha^2\nabla^2u+\varepsilon_1^2u+g(u)\Big)\frac{\partial u}{\partial t}+\int_\Omega\varepsilon^2v\frac{\partial v}{\partial t}d\Omega\\&+\frac{\mathscr{F}(R)}{E}\Big(\int_\Omega f v d\Omega+\alpha^2\int_\Gamma(\boldsymbol{n}\cdot\nabla u)v d\Gamma\Big)+\Big[1-\frac{\mathscr{F}(R)}{E}\Big]\Big|\int_\Omega f v d\Omega+\alpha^2\int_\Gamma d_a v d\Gamma\Big|\\&-\int_\Omega\Big(-\alpha^2\nabla^2u+\varepsilon_1^2u+g(u)\Big)v d\Omega-\int_\Omega\varepsilon^2v\Big[\Big(\frac{\alpha}{\varepsilon}\Big)^2\nabla^2u-\Big(\frac{\varepsilon_1}{\varepsilon}\Big)^2u-\frac{\mathscr{F}(R)}{E}\frac{1}{\varepsilon^2}g(u)+\frac{1}{\varepsilon^2}f\Big]d\Omega.\end{aligned} \tag{5.8b}$$

The reformulated system consists of equations (5.1), (5.8a)-(5.8b) and (5.3)-(5.4), which is equivalent to the original system (5.1)-(5.4).

Since the Klein-Gordon equation is conservative (in the absence of external source term and with appropriate boundary condition), we will employ the Crank-Nicolson method for time discretization of the field variables, by enforcing the discretized equations at step $(n+1/2)$. This corresponds to the approximations (2.33a)–(2.34) with $\theta=\frac{1}{2}$ and $\beta=0$. So the method here is slightly different than the one presented in Section 2.2, which corresponds to $\theta=1$ and $\beta=\frac{1}{4}$ in the approximations (2.33a)–(2.34). The energy-stable scheme for the nonlinear Klein-Gordon equation is then as follows:

$$\frac{u^{n+1}-u^n}{\Delta t}=v^{n+\frac{1}{2}}, \tag{5.9a}$$

$$\frac{v^{n+1}-v^n}{\Delta t}=\Big(\frac{\alpha}{\varepsilon}\Big)^2\nabla^2u^{n+\frac{1}{2}}-\Big(\frac{\varepsilon_1}{\varepsilon}\Big)^2u^{n+\frac{1}{2}}-\xi\frac{1}{\varepsilon^2}g(\bar{u}^{n+\frac{1}{2}})+\frac{1}{\varepsilon^2}f^{n+\frac{1}{2}}, \tag{5.9b}$$

$$\xi=\frac{\mathscr{F}(R^{n+1})}{\tilde{E}[\tilde{u}^{n+1},\tilde{v}^{n+1}]}, \tag{5.9c}$$

$$E[\tilde{u}^{n+1},\tilde{v}^{n+1}]=\int_\Omega\Big(\frac{\alpha^2}{2}|\nabla\tilde{u}^{n+1}|^2+\frac{\varepsilon_1^2}{2}|\tilde{u}^{n+1}|^2+\frac{\varepsilon^2}{2}|\tilde{v}^{n+1}|^2+G(\tilde{u}^{n+1})\Big)d\Omega+C_0, \tag{5.9d}$$

$$u^{n+1}=d_a^{n+1},\ \text{ on }\Gamma, \tag{5.9e}$$

together with

$$
\begin{aligned}
D_{\mathscr{F}}(R)|^{n+\frac{1}{2}}\frac{R^{n+1}-R^n}{\Delta t} &= \int_\Omega \Big(-\alpha^2\nabla^2 u^{n+\frac{1}{2}}+\varepsilon_1^2 u^{n+\frac{1}{2}}+g(u^{n+\frac{1}{2}})\Big)\frac{u^{n+1}-u^n}{\Delta t}d\Omega \\
&+\int_\Omega \varepsilon^2 v^{n+\frac{1}{2}}\frac{v^{n+1}-v^n}{\Delta t}d\Omega-\int_\Omega \Big(-\alpha^2\nabla^2 u^{n+\frac{1}{2}}+\varepsilon_1^2 u^{n+\frac{1}{2}}+g(u^{n+\frac{1}{2}})\Big)v^{n+\frac{1}{2}}d\Omega \\
&-\int_\Omega \varepsilon^2 v^{n+\frac{1}{2}}\Big\{\Big(\frac{\alpha}{\varepsilon}\Big)^2\nabla^2 u^{n+\frac{1}{2}}-\Big(\frac{\varepsilon_1}{\varepsilon}\Big)^2 u^{n+\frac{1}{2}}-\xi\frac{1}{\varepsilon^2}g(\bar{u}^{n+\frac{1}{2}})+\frac{1}{\varepsilon^2}f^{n+\frac{1}{2}}\Big\}d\Omega \\
&+\xi\Big(\int_\Omega f^{n+\frac{1}{2}}\tilde{v}^{n+\frac{1}{2}}d\Omega+\alpha^2\int_\Gamma(\boldsymbol{n}\cdot\nabla\tilde{u}^{n+\frac{1}{2}})\tilde{v}^{n+\frac{1}{2}}d\Gamma\Big) \\
&+(1-\xi)\Big|\int_\Omega f^{n+\frac{1}{2}}\tilde{v}^{n+\frac{1}{2}}d\Omega+\alpha^2\int_\Gamma(\boldsymbol{n}\cdot\nabla\tilde{u}^{n+\frac{1}{2}})\tilde{v}^{n+\frac{1}{2}}d\Gamma\Big|.
\end{aligned}
\tag{5.10}
$$

These equations are supplemented by the initial conditions

$$
u^0=u_{\rm in}(\boldsymbol{x}),\quad v^0=v_{\rm in}(\boldsymbol{x}),\quad R^0=\mathscr{G}(E^0),
\tag{5.11}
$$

where E^0 is evaluated by

$$
E^0=\int_\Omega\Big(\frac{\alpha^2}{2}|\nabla u_{\rm in}|^2+\frac{\varepsilon_1^2}{2}|u_{\rm in}|^2+\frac{\varepsilon^2}{2}|v_{\rm in}|^2+G(u_{\rm in})\Big)d\Omega+C_0.
\tag{5.12}
$$

In the above equations, $D_{\mathscr{F}}(R)|^{n+\frac{1}{2}}$ is defined by (2.34) with $\theta=1/2$, and

$$
\begin{cases}
u^{n+1/2}=\dfrac{1}{2}(u^{n+1}+u^n),\quad v^{n+1/2}=\dfrac{1}{2}(v^{n+1}+v^n),\\
\bar{u}^{n+\frac{1}{2}}=\dfrac{3}{2}u^n-\dfrac{1}{2}u^{n-1},\quad \bar{v}^{n+\frac{1}{2}}=\dfrac{3}{2}v^n-\dfrac{1}{2}v^{n-1}.
\end{cases}
\tag{5.13}
$$

$\tilde{u}^{n+1}$, $\tilde{v}^{n+1}$, $\tilde{u}^{n+\frac{1}{2}}$ and $\tilde{v}^{n+\frac{1}{2}}$ are second-order approximations of u^{n+1}, v^{n+1}, $u^{n+\frac{1}{2}}$ and $v^{n+\frac{1}{2}}$, respectively, defined later in (5.22)-(5.23).

Theorem 5.1. *In the absence of the external force $f=0$, and with homogeneous boundary condition ($d_a=0$) and suppose that the initial condition $v_{\rm in}$ satisfies the compatibility condition $v_{in}|_\Gamma=0$, the scheme consisting of (5.9)-(5.11) conserves the modified energy $\mathscr{F}(R)$ in the sense that:*

$$
\mathscr{F}(R^{n+1})-\mathscr{F}(R^n)=0.
\tag{5.14}
$$

Proof. Multiplying $\big(-\alpha^2\nabla^2 u^{n+\frac{1}{2}}+\varepsilon_1^2 u^{n+\frac{1}{2}}+g(u^{n+\frac{1}{2}})\big)$ to equation (5.9a), $\varepsilon^2 v^{n+\frac{1}{2}}$ to equation (5.9b), taking the L^2 integrals, and summing up the resultant equations with equation (5.10), we arrive at the relation,

$$
\begin{aligned}
\frac{\mathscr{F}(R^{n+1})-\mathscr{F}(R^n)}{\Delta t} &= \xi\Big(\int_\Omega f^{n+\frac{1}{2}}\tilde{v}^{n+\frac{1}{2}}d\Omega+\alpha^2\int_\Gamma(\boldsymbol{n}\cdot\nabla\tilde{u}^{n+\frac{1}{2}})\tilde{v}^{n+\frac{1}{2}}d\Gamma\Big)\\
&+(1-\xi)\Big|\int_\Omega f^{n+\frac{1}{2}}\tilde{v}^{n+\frac{1}{2}}d\Omega+\alpha^2\int_\Gamma(\boldsymbol{n}\cdot\nabla\tilde{u}^{n+\frac{1}{2}})\tilde{v}^{n+\frac{1}{2}}d\Gamma\Big|,
\end{aligned}
\tag{5.15}
$$

where we have used equations (2.21b)-(2.22). If $d_a=0$, then $u^n|_\Gamma=0$ and $v^n|_\Gamma=0$ for all $n>0$. Based on the definition of $\tilde{v}^{n+\frac{1}{2}}$ in the equation (5.23) below, it is straightforward to verify that $\tilde{v}^{n+\frac{1}{2}}|_\Gamma=0$ as long as $v^0|_\Gamma=0$. Furthermore, if $f=0$, the volume integrals in equation (5.15) vanish. This leads to equation (5.14). □

Remark 5.1. Since $\mathscr{F}(R)$ is an approximation of $E(t)$, the discrete conservation for $\mathscr{F}(R)$ in equation (5.14) does not imply the conservation for $E(t)$ on the discrete level. However, it does lead to an unconditionally energy stable scheme for long time simulations.

Despite the complication caused by the unknown scalar variable ξ, the proposed scheme can be solved in a decoupled fashion. Combining equations (5.9a) and (5.13), we get

$$v^{n+1} = \frac{2}{\Delta t}u^{n+1} - \frac{2}{\Delta t}u^n - v^n. \tag{5.16}$$

Inserting equation (5.16) into (5.9b) leads to

$$\begin{aligned}\Big[\Big(\frac{2\varepsilon}{\alpha\Delta t}\Big)^2 + \Big(\frac{\varepsilon_1}{\alpha}\Big)^2\Big]u^{n+1} - \nabla^2 u^{n+1} =& \Big(\frac{\varepsilon}{\alpha}\Big)^2\Big\{\Big[\big(\frac{2}{\Delta t}\big)^2 - \big(\frac{\varepsilon_1}{\varepsilon}\big)^2\Big]u^n + \frac{4}{\Delta t}v^n + \frac{2}{\varepsilon^2}f^{n+\frac{1}{2}}\Big\}\\ &- \xi\frac{2}{\alpha^2}g(\bar{u}^{n+\frac{1}{2}}) + \nabla^2 u^n.\end{aligned} \tag{5.17}$$

To solve this equation, we introduce u_1^{n+1} and u_2^{n+1} as solutions of the following two equations:

$$\Big[\Big(\frac{2\varepsilon}{\alpha\Delta t}\Big)^2 + \Big(\frac{\varepsilon_1}{\alpha}\Big)^2\Big]u_1^{n+1} - \nabla^2 u_1^{n+1} = \Big(\frac{\varepsilon}{\alpha}\Big)^2\Big\{\Big[\big(\frac{2}{\Delta t}\big)^2 - \big(\frac{\varepsilon_1}{\varepsilon}\big)^2\Big]u^n + \frac{4}{\Delta t}v^n + \frac{2}{\varepsilon^2}f^{n+\frac{1}{2}}\Big\} + \nabla^2 u^n, \tag{5.18}$$

$$u_1^{n+1} = d_a^{n+1} \ \text{ on } \Gamma, \tag{5.19}$$

and

$$\Big[\Big(\frac{2\varepsilon}{\alpha\Delta t}\Big)^2 + \Big(\frac{\varepsilon_1}{\alpha}\Big)^2\Big]u_2^{n+1} - \nabla^2 u_2^{n+1} = -\frac{2}{\alpha^2}g(\bar{u}^{n+\frac{1}{2}}), \quad u_2^{n+1} = 0 \ \text{ on } \Gamma. \tag{5.20}$$

Then the solution to equation (5.17), together with the boundary condition (5.9e), is given by

$$u^{n+1} = u_1^{n+1} + \xi u_2^{n+1}. \tag{5.21}$$

where ξ is to be determined.

We define

$$\tilde{u}^{n+1} = u_1^{n+1} + u_2^{n+1}, \quad \tilde{u}^{n+1/2} = \frac{1}{2}(\tilde{u}^{n+1} + u^n), \tag{5.22}$$

$$\tilde{v}^{n+1} = \frac{2}{\Delta t}(\tilde{u}^{n+1} - u^n) - v^n, \quad \tilde{v}^{n+1/2} = \frac{1}{2}(\tilde{v}^{n+1} + v^n). \tag{5.23}$$

By combining equations (5.9c) and (5.15), we can determine ξ,

$$\xi = \frac{\mathscr{F}(R^n) + \Delta t|S_0|}{E[\tilde{u}^{n+1}, \tilde{v}^{n+1}] + \Delta t(|S_0| - S_0)}, \quad \text{with } S_0 = \Big(\int_\Omega f^{n+1/2}\tilde{v}^{n+1/2}d\Omega + \alpha^2\int_\Gamma (\boldsymbol{n}\cdot\nabla\tilde{u}^{n+1/2})\tilde{v}^{n+1/2}d\Gamma\Big). \tag{5.24}$$

With ξ known, u^{n+1} and v^{n+1} can be computed by equations (5.21) and (5.16), respectively. R^{n+1} can be computed by,

$$R^{n+1} = \mathscr{G}(\xi E[\tilde{u}^{n+1}, \tilde{v}^{n+1}]). \tag{5.25}$$

The weak formulations for equations (5.18) and (5.20) are: Find $(u_1^{n+1}, u_2^{n+1}) \in H^1(\Omega)$ such that

$$\begin{aligned}&\big(\nabla u_1^{n+1}, \nabla\varphi\big)_\Omega + \Big[\Big(\frac{2\varepsilon}{\alpha\Delta t}\Big)^2 + \Big(\frac{\varepsilon_1}{\alpha}\Big)^2\Big]\big(u_1^{n+1}, \varphi\big)_\Omega = -(\nabla u^n, \nabla\varphi)_\Omega\\ &+ \Big(\frac{\varepsilon}{\alpha}\Big)^2\Big(\Big[\big(\frac{2}{\Delta t}\big)^2 - \big(\frac{\varepsilon_1}{\varepsilon}\big)^2\Big]u^n + \frac{4}{\Delta t}v^n + \frac{2}{\varepsilon^2}f^{n+\frac{1}{2}}, \varphi\Big)_\Omega, \quad \forall\varphi \in H_0^1(\Omega) := \Big\{w \in H^1(\Omega) :\ w|_\Gamma = 0\Big\};\end{aligned} \tag{5.26}$$

$$\big(\nabla u_2^{n+1}, \nabla\varphi\big)_\Omega + \Big[\Big(\frac{2\varepsilon}{\alpha\Delta t}\Big)^2 + \Big(\frac{\varepsilon_1}{\alpha}\Big)^2\Big]\big(u_2^{n+1}, \varphi\big)_{\Omega_h} = -\frac{2}{\alpha^2}\big(g(\bar{u}^{n+\frac{1}{2}}), \varphi\big)_\Omega, \quad \forall\varphi \in H_0^1(\Omega). \tag{5.27}$$

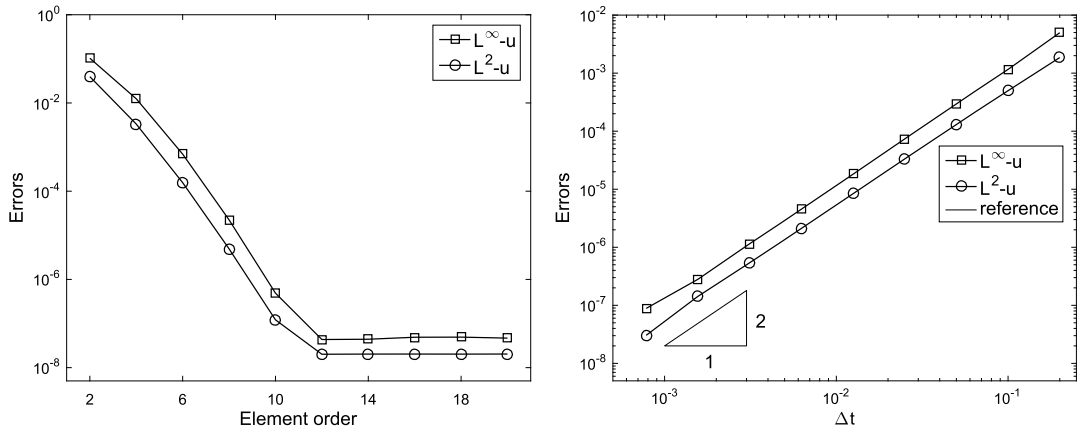
These can be implemented with C^0 spectral elements in a straightforward fashion.

5.1. Numerical results

We next provide numerical examples to demonstrate the accuracy and unconditional stability of the proposed scheme to the Klein-Gordon equation (5.1)-(5.3). Specifically, we fix the parameters therein and the potential energy function as

$$\varepsilon = \varepsilon_1 = \alpha = 1, \quad G(u) = 1 - \cos(u), \quad g(u) = G'(u) = \sin(u). \tag{5.28}$$

This corresponds to the dimensionless relativistic Sine-Gordon equation (DRSG) (see e.g. [3]).



(a) Errors vs Element order (b) Errors vs Δt

Fig. 5.1. Spatial/temporal convergence tests for DRSG equation: L^2 and L^∞ errors of u versus (a) element order (fixed $\Delta t = 0.001$ and $t_f = 0.1$), and (b) Δt (fixed element order 18 and $t_f = 1$).

5.1.1. Convergence rates

To study the convergence rates in space and time of the proposed method, we employ the following manufactured analytic solution

$$u = \cos(\pi x)\cos(\pi y)\sin(t). \tag{5.29}$$

The external force $f(\boldsymbol{x}, t)$ in (5.2) and the external boundary source term $d_a(\boldsymbol{x}, t)$ are chosen such that the above expression (5.29) satisfies equations (5.1)-(5.3).

The computational domain $\Omega = [0, 2] \times [-1, 1]$ is discretized using two equal-sized quadrilateral elements, with the element order and the time step size Δt varied systematically in the spatial and temporal tests. The algorithm presented in this section is employed to numerically integrate the DRSG equation from $t = 0$ to $t = t_f$. The mapping $\mathscr{F}(R) = R$ and $C_0 = 1$ are used in these computations. The initial condition u_{in} and v_{in} are obtained by setting $t = 0$ in the analytic expression (5.29) and using (5.1). We then record the numerical errors in different norms by comparing the numerical solution with the analytic solution at $t = t_f$.

To conduct the spatial convergence test, we vary systematically the element order from 2 to 20 and depict in Fig. 5.1(a) the L^∞ and L^2 errors of u as a function of the element order with a fixed $\Delta t = 0.001$ and $t_f = 0.1$. It is observed that the numerical errors decay exponentially with increasing element order, and levels off beyond element order 12, caused by the saturation of temporal errors.

To study the temporal convergence rate, we fix the element order at a large value 18 and $t_f = 1.0$. The time step size Δt is varied systematically from 0.2 to 7.8125×10^{-4} and the numerical errors in L^∞ and L^2 norms are depicted in Fig. 5.1(b). A second-order convergence rate in time is clearly observed.

Fig. 5.2 shows the spatial and temporal convergence behaviors of the current method, in terms of the L^2 errors, corresponding to several different mapping functions $\mathscr{F}(R)$. Note that $e_0 = \kappa_0 = 10$ with the mapping $\mathscr{F}(R) = \frac{e_0}{2}\ln\left(\frac{\kappa_0 + R}{\kappa_0 - R}\right)$. It is evident that with different mapping functions the current method exhibits a spatial exponential convergence rate and temporal second-order convergence rate. We again observe the insensitivity of the simulation results with respect to $\mathscr{F}(R)$, similar to those with the other tests in previous sections.

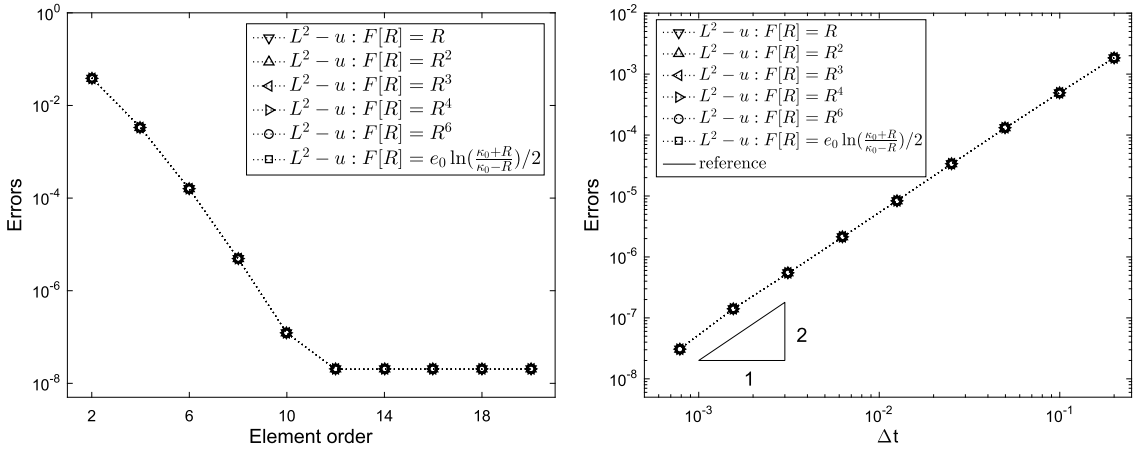
5.1.2. Study of method properties

We next study the remarkable stability of the proposed method with the DRSG equation. Consider the DRSG equation on the domain $\Omega = [0, 14]^2$, with zero external force $f(\boldsymbol{x}, t) = 0$ and zero boundary source term $d_a(\boldsymbol{x}, t) = 0$ in (5.3). The initial conditions are set to

$$u_{in}(\boldsymbol{x}) = \frac{2}{\exp\left((x-7)^2 + (y-7)^2\right) + \exp\left(-(x-7)^2 - (y-7)^2\right)}, \quad v_{in}(\boldsymbol{x}) = 0. \tag{5.30}$$

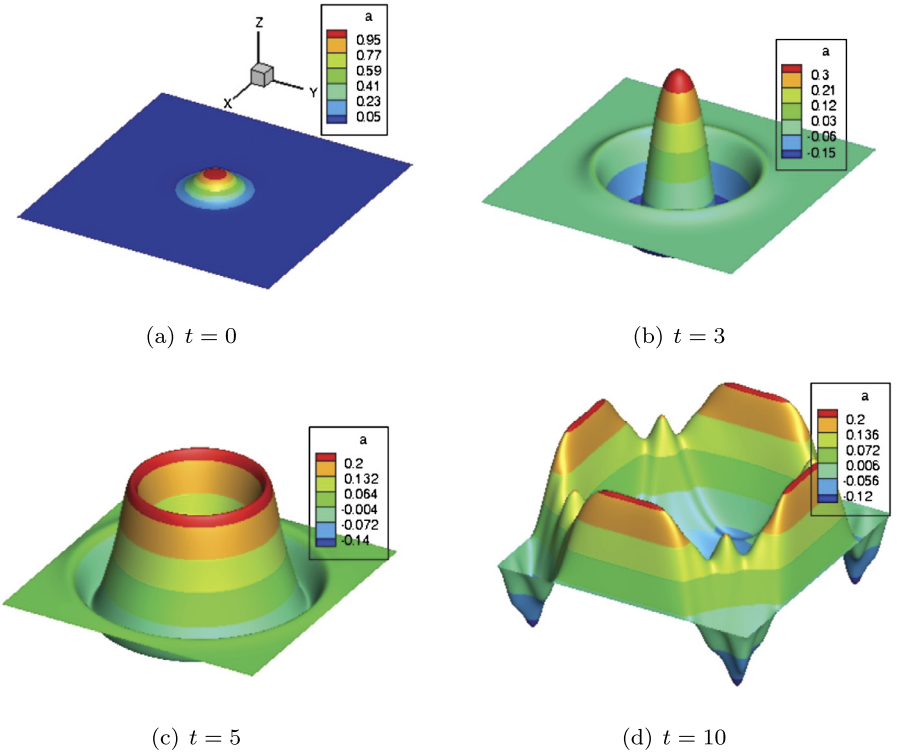
With these initial and boundary conditions, the DRSG equation is energy conserving.

The domain Ω is discretized with 400 equal-sized quadrilateral elements with a fixed element order 10. We employ a mapping function $\mathscr{F}(R) = \frac{e_0}{2}\ln(\frac{\kappa_0 + R}{\kappa_0 - R})$ ($e_0 = 10$, $\kappa_0 = 100$) and the energy constant $C_0 = 1$ in the algorithm. Fig. 5.3 illustrates the evolution of u by a sequence of snapshots of its contour levels. One can observe a circular wave pattern starting from the center of the domain and propagating outward toward the boundaries. As the wave reaches the boundaries, the interaction with the Dirichlet boundary ($u = 0$) gives rise to an extremely complicated wave pattern; see Fig. 5.3(d).



(a) Spatial convergence for various $\mathscr{F}(R)$ (b) Temporal test for various $\mathscr{F}(R)$

Fig. 5.2. Spatial/temporal convergence for the DRSG equation obtained with various mapping functions $\mathscr{F}(R)$: L^2 errors of u versus (a) element order (fixed $\Delta t = 0.001$ and $t_f = 0.1$), and (b) Δt (fixed element order 18 and $t_f = 1$).



(a) $t = 0$ (b) $t = 3$

(c) $t = 5$ (d) $t = 10$

Fig. 5.3. DRSG equation: Temporal sequence of snapshots for u distribution. Simulation results are obtained with $\Delta t = 10^{-4}$, and the mapping $\mathscr{F}(R) = \frac{e_0}{2} \ln\left(\frac{\kappa_0+R}{\kappa_0-R}\right)$ (with $e_0 = 10$, $\kappa_0 = 100$).

Fig. 5.4(a) shows the time histories of the energy errors, $|E(t) - E(0)|$, obtained using several time step sizes ($\Delta t = 10^{-4}$, 10^{-3} and 10^{-2}). One can observe oscillations in the history curves about their respective mean values that are consistent with a second order accuracy in time. It should again be noted that the current algorithm conserves the modified energy $\mathscr{F}(R)$ discretely, not the original energy $E(t)$. Fig. 5.4(b) shows time histories of the ratio $\xi = \frac{\mathscr{F}(R)}{E}$ corresponding to these Δt values. The computed ξ values are essentially 1, indicative of the accuracy of these simulations.

We then increase the time step size to $\Delta t = 0.1, 1$ and 10, and depict in Fig. 5.5(a) the time histories of $E(t)$ and $\mathscr{F}(R)$ for a long time simulation to $t = 1000$. Large discrepancies between the energy $E(t)$ and $\mathscr{F}(R)$ can be observed, especially for $\Delta t = 1$ and 10, suggesting that $\mathscr{F}(R)$ no longer approximates well the energy $E(t)$ with these time step sizes. Note that the $\mathscr{F}(R)$ histories obtained by different large Δt values overlap with one another. This is consistent with Theorem 5.1 that

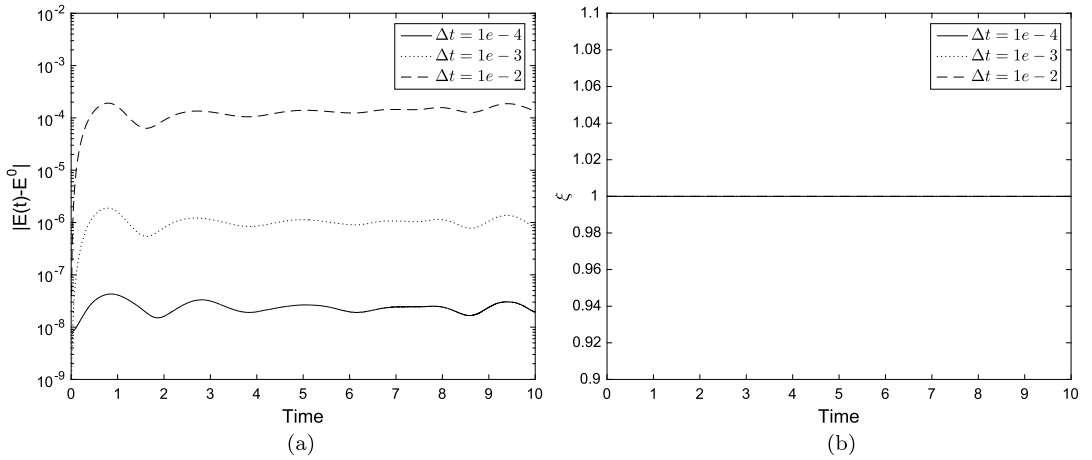


Fig. 5.4. DRSG equation: Time histories of (a) $|E(t) - E(0)|$ and (b) $\xi = \mathscr{F}(R)/E$ obtained with several time step sizes, $\Delta t = 10^{-4}, 10^{-3}, 10^{-2}$. Numerical results correspond to $\mathscr{F}(R) = \frac{e_0}{2}\ln(\frac{\kappa_0+x}{\kappa_0-x})$ ($e_0 = 10$, $\kappa_0 = 100$).

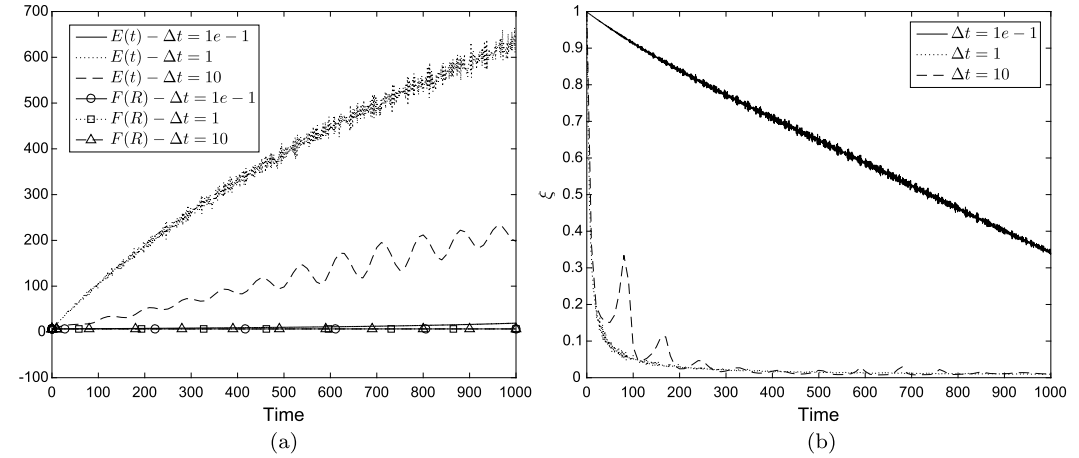


Fig. 5.5. Time histories of (a) $E(t)$ and $\mathscr{F}(R)$ and (b) $\xi = \mathscr{F}(R)/E$ versus large $\Delta t = 0.1, 1, 10$ for DRSG equation. The numerical results are obtained with $\mathscr{F}(R) = \frac{e_0}{2}\ln(\frac{\kappa_0+x}{\kappa_0-x})$ ($e_0 = 10$, $\kappa_0 = 100$).

the current scheme conserves the modified energy $\mathscr{F}(R)$. It can be observed from Fig. 5.5(b) that the computed $\xi = \frac{\mathscr{F}(R)}{E}$ becomes significantly smaller than 1, indicative of large errors in the simulations with these large time step sizes. However, the computations are evidently stable, even with these large Δt values.

6. Concluding remarks

In this paper we have presented a framework (gPAV) for developing unconditionally energy-stable schemes for general dissipative systems. The scheme is based on a generalized auxiliary variable (which is a scalar number) associated with the energy functional of the system. We find that the square root function, which is critical to previous auxiliary-variable approaches, is not essential to devising energy-stable schemes. In the current method, the auxiliary variable can be defined by a rather general class of functions, not limited to the square-root function. The gPAV method is applicable to general dissipative systems, and a unified procedure for discretely treating the dissipative governing equations and the generalized auxiliary variable has been presented. The discrete energy stability of the proposed scheme has been proven for general dissipative systems. The presented method has two attractive properties:

- The scheme requires only the solution of linear algebraic equations within a time step, and no nonlinear solver is needed. Furthermore, with appropriate choice of the $\boldsymbol{F}_L$ operator in the algorithm, the resultant linear algebraic systems upon discretization involve only constant and time-independent coefficient matrices, which only need to be computed once and can be pre-computed. In terms of the computational cost, the scheme is computationally very competitive and attractive.

- The generalized auxiliary variable can be computed directly by a well-defined explicit formula. The computed values for the auxiliary variable are guaranteed to be positive, irrespective of the time step size or the external forces or source terms.

Three specific dissipative systems (a chemo-repulsion model, Cahn-Hilliard equation with constant and variable mobility, and the nonlinear Klein-Gordon equation) have been studied in relative detail to demonstrate the gPAV framework developed herein. Ample numerical experiments have been presented for each system to demonstrate the performance of the method, the effects of algorithmic parameters, and the stability of the scheme with large time step sizes.

Time step size adaptivity can be a useful property for numerical schemes. When the approximations (2.33a)–(2.34) with $\theta = \frac{1}{2}$ and $\beta = 0$ are employed in the current scheme, just like those presented in Section 5 for the Klein-Gordon equation, the current method becomes a single-step scheme. Only the field data in the previous time step ($\boldsymbol{u}^n$) is needed when marching in time. In this case, it is quite straightforward to adaptively change the time step size. In the more general case ($\theta \neq \frac{1}{2}$), the approximations presented in the current work become a two-step method, requiring the field data in the two previous time steps ($\boldsymbol{u}^n$ and $\boldsymbol{u}^{n-1}$) for time-marching. In this case, it will be much more involved to adapt the time step size and simultaneously ensure the other attractive properties (such as positivity, unconditional energy stability, and second-order accuracy). This is an interesting problem that will be explored in a future work.

Numerical results in the current work suggest that the simulation results are not sensitive to the choice of the mapping function $\mathscr{F}(R)$ using the gPAV method. The difference between the simulation results corresponding to different $\mathscr{F}(R)$ seems very small and basically negligible. In terms of which mapping function to use in actual applications, we would like to recommend the function $\mathscr{F}(R) = R$, since it is perhaps the simplest one available. A sizable portion of the numerical experiments in the current work has been performed using this mapping function.

All physically meaningful systems in the real world are energy dissipative (or conserving) due to the second law of thermodynamics, and these systems are typically nonlinear. The design of energy-stable and computationally-efficient schemes for such systems is critical to their numerical simulations, and this is in general a very challenging task. The gPAV framework presented here lays out a roadmap for devising discretely energy-stable schemes for general dissipative systems. The computational efficiency (e.g. involving linear equations with pre-computable coefficient matrices) and the guaranteed positivity of the computed auxiliary variable of the method are particularly attractive, in the sense that the gPAV method is not only unconditionally energy-stable but also can be computationally efficient and competitive. We anticipate that the gPAV method will be useful and instrumental in numerical simulations of a number of computational science and engineering disciplines.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially supported by NSF (DMS-1522537).

Appendix A. Approximation for the first time step

We present a method on how to deal with the first time step such that the approximation for the auxiliary variable $R(t)$ at time step $\frac{1}{2}$ shall be positive. We consider below only the formulation based on $\frac{\mathscr{F}(R)}{E}$. It is noted that for the alternative formulation based on $\frac{R}{\mathscr{G}(E)}$ (see Section 2.4) one can modify the following scheme in a straightforward fashion to achieve the same property. The notations here follow those employed in the main text.

Consider the system consisting of equations (2.17), (2.19), the boundary condition (2.2), and the initial conditions (2.3) and (2.20). Define

$$\begin{cases} \boldsymbol{u}^0 = \boldsymbol{u}_{in}(\boldsymbol{x}), \\ R^0 = \mathscr{G}(E^0), \quad \text{with } E^0 = \displaystyle\int_\Omega e(\boldsymbol{u}_{in}) d\Omega + C_0. \end{cases} \tag{A.1}$$

One notes that $E^0 > 0$ and $R^0 > 0$.

We compute the first time step in two substeps. In substep one we compute an approximation of $(\boldsymbol{u}^1, R^1)$, denoted by $(\boldsymbol{u}_a^1, R_a^1)$, and in substep two we compute the final $(\boldsymbol{u}^1, R^1)$. More specifically, the scheme is as follows:

Substep One:

$$\frac{\boldsymbol{u}_a^1 - \boldsymbol{u}^0}{\Delta t} = \boldsymbol{F}_L(\boldsymbol{u}_a^1) + \xi_a \Big[\boldsymbol{F}(\boldsymbol{u}^0) - \boldsymbol{F}_L(\boldsymbol{u}^0) \Big] + \boldsymbol{f}^1, \tag{A.2a}$$

$$\xi_a = \frac{\mathscr{F}(R_a^1)}{E[\tilde{\boldsymbol{u}}_a^1]}, \tag{A.2b}$$

$$E[\tilde{\boldsymbol{u}}_a^1] = \int_\Omega e(\tilde{\boldsymbol{u}}_a^1)d\Omega + C_0, \tag{A.2c}$$

$$\boldsymbol{B}(\boldsymbol{u}_a^1) = \boldsymbol{f}_b^1, \quad \text{on } \Gamma, \tag{A.2d}$$

$$\begin{aligned}\frac{\mathscr{F}(R_a^1)-\mathscr{F}(R^0)}{R_a^1-R^0}\frac{R_a^1-R^0}{\Delta t} &= \int_\Omega e'(\boldsymbol{u}_a^1)\cdot\frac{\boldsymbol{u}_a^1-\boldsymbol{u}^0}{\Delta t}d\Omega \\ &- \int_\Omega e'(\boldsymbol{u}_a^1)\cdot\left(\boldsymbol{F}_L(\boldsymbol{u}_a^1)+\xi_a\left[\boldsymbol{F}(\boldsymbol{u}^0)-\boldsymbol{F}_L(\boldsymbol{u}^0)\right]+\boldsymbol{f}^1\right)d\Omega \\ &+ \xi_a\left[-\int_\Omega V(\tilde{\boldsymbol{u}}_a^1)d\Omega+\int_\Omega V_s(\boldsymbol{f}^1,\tilde{\boldsymbol{u}}_a^1)d\Omega+\int_\Gamma B_s(\boldsymbol{f}_b^1,\tilde{\boldsymbol{u}}_a^1)d\Gamma\right] \\ &+ (1-\xi_a)\left|\int_\Omega V_s(\boldsymbol{f}^1,\tilde{\boldsymbol{u}}_a^1)d\Omega+\int_\Gamma B_s(\boldsymbol{f}_b^1,\tilde{\boldsymbol{u}}_a^1)d\Gamma\right|.\end{aligned} \tag{A.2e}$$

Substep Two:

$$\frac{\boldsymbol{u}^1-\boldsymbol{u}^0}{\Delta t} = \boldsymbol{F}_L(\boldsymbol{u}^1)+\xi\left[\boldsymbol{F}(\boldsymbol{u}^0)-\boldsymbol{F}_L(\boldsymbol{u}^0)\right]+\boldsymbol{f}^1, \tag{A.3a}$$

$$\xi = \frac{\mathscr{F}(R^{3/2})}{E[\tilde{\boldsymbol{u}}^{3/2}]}, \tag{A.3b}$$

$$E[\tilde{\boldsymbol{u}}^{3/2}] = \int_\Omega e(\tilde{\boldsymbol{u}}^{3/2})d\Omega + C_0, \tag{A.3c}$$

$$\boldsymbol{B}(\boldsymbol{u}^1) = \boldsymbol{f}_b^1, \quad \text{on } \Gamma, \tag{A.3d}$$

$$\begin{aligned}\frac{\mathscr{F}(R^{3/2})-\mathscr{F}(R^{1/2})}{R^{3/2}-R^{1/2}}\frac{R^{3/2}-R^{1/2}}{\Delta t} &= \int_\Omega e'(\boldsymbol{u}^1)\cdot\frac{\boldsymbol{u}^1-\boldsymbol{u}^0}{\Delta t}d\Omega \\ &- \int_\Omega e'(\boldsymbol{u}^1)\cdot\left(\boldsymbol{F}_L(\boldsymbol{u}^1)+\xi\left[\boldsymbol{F}(\boldsymbol{u}^0)-\boldsymbol{F}_L(\boldsymbol{u}^0)\right]+\boldsymbol{f}^1\right)d\Omega \\ &+ \xi\left[-\int_\Omega V(\tilde{\boldsymbol{u}}^1)d\Omega+\int_\Omega V_s(\boldsymbol{f}^1,\tilde{\boldsymbol{u}}^1)d\Omega+\int_\Gamma B_s(\boldsymbol{f}_b^1,\tilde{\boldsymbol{u}}^1)d\Gamma\right] \\ &+ (1-\xi)\left|\int_\Omega V_s(\boldsymbol{f}^1,\tilde{\boldsymbol{u}}^1)d\Omega+\int_\Gamma B_s(\boldsymbol{f}_b^1,\tilde{\boldsymbol{u}}^1)d\Gamma\right|.\end{aligned} \tag{A.3e}$$

Note that in the above equations the superscript of a variable such as $(\cdot)^{1/2}$ and $(\cdot)^{3/2}$ denotes the time step index. In (A.2b) and (A.2e) $\tilde{\boldsymbol{u}}_a^1$ is an approximation of $\boldsymbol{u}_a^1$ and will be specified later in (A.12). In (A.3e) $\tilde{\boldsymbol{u}}^1$ is an approximation of $\boldsymbol{u}^1$ and will be specified later also in (A.12). In (A.3b), (A.3c) and (A.3e), $\tilde{\boldsymbol{u}}^{3/2}$, $R^{1/2}$ and $R^{3/2}$ are defined by

$$\begin{cases}\tilde{\boldsymbol{u}}^{3/2} = \dfrac{3}{2}\boldsymbol{u}_a^1-\dfrac{1}{2}\boldsymbol{u}^0, \\ R^{3/2} = \dfrac{3}{2}R^1-\dfrac{1}{2}R^0, \\ R^{1/2} = \dfrac{1}{2}\left(R_a^1+R^0\right).\end{cases} \tag{A.4}$$

It can be noted that the above scheme represents a first-order approximation of $(\boldsymbol{u}^1, R^1)$ for the first time step.

Combine equations (A.2a) and (A.2e) and we have

$$\mathscr{F}(R_a^1) - \mathscr{F}(R^0) = -\xi_a \Delta t \int_\Omega V(\tilde{\boldsymbol{u}}_a^1) d\Omega - \xi_a \Delta t(|S_a| - S_a) + |S_a|\Delta t \tag{A.5}$$

where $S_a = \int_\Omega V_s(\boldsymbol{f}^1, \tilde{\boldsymbol{u}}_a^1) d\Omega + \int_\Gamma B_s(\boldsymbol{f}_b^1, \tilde{\boldsymbol{u}}_a^1) d\Gamma$. In light of (A.2b), this leads to

$$\begin{cases} \xi_a = \dfrac{\mathscr{F}(R^0) + |S_a|\Delta t}{E[\tilde{\boldsymbol{u}}_a^1] + \Delta t \int_\Omega V(\tilde{\boldsymbol{u}}_a^1) d\Omega + (|S_a| - S_a)\Delta t}, \\ R_a^1 = \mathscr{G}(\xi_a E[\tilde{\boldsymbol{u}}_a^1]). \end{cases} \tag{A.6}$$

Since $R^0 > 0$, we conclude that $\xi_a > 0$ and $R_a^1 > 0$ based on these equations. It follows that $R^{1/2} = \frac{1}{2}(R_a^1 + R^0) > 0$ in light of equation (A.4).

Similarly, combining equations (A.3a) and (A.3e) gives rise to

$$\mathscr{F}(R^{3/2}) - \mathscr{F}(R^{1/2}) = -\xi \Delta t \int_\Omega V(\tilde{\boldsymbol{u}}^1) d\Omega - \xi \Delta t(|S_0| - S_0) + |S_0|\Delta t \tag{A.7}$$

where $S_0 = \int_\Omega V_s(\boldsymbol{f}^1, \tilde{\boldsymbol{u}}^1) d\Omega + \int_\Gamma B_s(\boldsymbol{f}_b^1, \tilde{\boldsymbol{u}}^1) d\Gamma$. In light of (A.3b) and (A.4), we have

$$\begin{cases} \xi = \dfrac{\mathscr{F}(R^{1/2}) + |S_0|\Delta t}{E[\tilde{\boldsymbol{u}}^{3/2}] + \Delta t \int_\Omega V(\tilde{\boldsymbol{u}}^1) d\Omega + (|S_0| - S_0)\Delta t}, \\ R^{3/2} = \mathscr{G}(\xi E[\tilde{\boldsymbol{u}}^{3/2}]), \\ R^1 = \dfrac{2}{3} R^{3/2} + \dfrac{1}{3} R^0. \end{cases} \tag{A.8}$$

We therefore conclude that $\xi > 0$, $R^{3/2} > 0$ and $R^1 > 0$.

We still need to determine $\boldsymbol{u}_a^1$ and $\boldsymbol{u}^1$, and specify $\tilde{\boldsymbol{u}}_a^1$ and $\tilde{\boldsymbol{u}}^1$. Note that $\boldsymbol{F}_L(\boldsymbol{u})$ and $\boldsymbol{B}(\boldsymbol{u})$ are linear operators. Equations (A.2a) and (A.2d), and also equations (A.3a) and (A.3d), can be solved as follows. Define two variables $\boldsymbol{u}_1^1$ and $\boldsymbol{u}_2^1$ as solutions to the following systems, respectively:

For $\boldsymbol{u}_1^1$:

$$\frac{1}{\Delta t}\boldsymbol{u}_1^1 - \boldsymbol{F}_L(\boldsymbol{u}_1^1) = \frac{\boldsymbol{u}^0}{\Delta t} + \boldsymbol{f}^1, \tag{A.9a}$$

$$\boldsymbol{B}(\boldsymbol{u}_1^1) = \boldsymbol{f}_b^1, \quad \text{on } \Gamma. \tag{A.9b}$$

For $\boldsymbol{u}_2^1$:

$$\frac{1}{\Delta t}\boldsymbol{u}_2^1 - \boldsymbol{F}_L(\boldsymbol{u}_2^1) = \boldsymbol{F}(\boldsymbol{u}^0) - \boldsymbol{F}_L(\boldsymbol{u}^0), \tag{A.10a}$$

$$\boldsymbol{B}(\boldsymbol{u}_2^1) = 0, \quad \text{on } \Gamma. \tag{A.10b}$$

Then it is straightforward to verify that, for given ξ_a and ξ, the following functions respectively solve the equations (A.2a) and (A.2d), and equations (A.3a) and (A.3d),

$$\boldsymbol{u}_a^1 = \boldsymbol{u}_1^1 + \xi_a \boldsymbol{u}_2^1, \tag{A.11a}$$

$$\boldsymbol{u}^1 = \boldsymbol{u}_1^1 + \xi \boldsymbol{u}_2^1. \tag{A.11b}$$

We then specify $\tilde{\boldsymbol{u}}_a^1$ and $\tilde{\boldsymbol{u}}^1$ as follows,

$$\tilde{\boldsymbol{u}}_a^1 = \tilde{\boldsymbol{u}}^1 = \boldsymbol{u}_1^1 + \boldsymbol{u}_2^1. \tag{A.12}$$

The solution for $(\boldsymbol{u}^1, R^1)$ at the first time step consists of the following procedure:

- Solve equations (A.9a)–(A.9b) for $\boldsymbol{u}_1^1$;
  Solve equations (A.10a)–(A.10b) for $\boldsymbol{u}_2^1$.

- Compute $\tilde{\boldsymbol{u}}_a^1$ and $\tilde{\boldsymbol{u}}^1$ by equation (A.12);
  Compute ξ_a and R_a^1 by equation (A.6);
  Compute $\boldsymbol{u}_a^1$ by equation (A.11a).
- Compute $\tilde{\boldsymbol{u}}^{3/2}$ and $R^{1/2}$ based on equation (A.4);
  Compute ξ and R^1 based on equation (A.8);
  Compute $\boldsymbol{u}^1$ by equation (A.11b).

We can make the following conclusion based on the above discussions.

Theorem A.1. *The scheme represented by* (A.2a)–(A.3e) *for computing the first time step has the property that*

$$R^1 > 0, \quad R^{1/2} > 0, \quad \text{and } R^{3/2} > 0, \tag{A.13}$$

where $R^{1/2}$ *and* $R^{3/2}$ *are given by* (A.4)*, regardless of the time step size* Δt *and the external forces* $\boldsymbol{f}$ *and* $\boldsymbol{f}_b$.

References

[1] H. Abels, H. Garcke, G. Grün, Thermodynamically consistent, frame indifferent diffuse interface models for incompressible two-phase flows with different densities, Math. Models Methods Appl. Sci. 22 (2012) 1150013.
[2] D.M. Anderson, G.B. McFadden, A.A. Wheeler, Diffuse-interface methods in fluid mechanics, Annu. Rev. Fluid Mech. 30 (1998) 139–165.
[3] W. Bao, X. Dong, Analysis and comparison of numerical methods for the Klein-Gordon equation in the nonrelativistic limit regime, Numer. Math. 120 (2012) 189–229.
[4] J.W. Cahn, J.E. Hilliard, Free energy of a nonuniform system. I interfacial free energy, J. Chem. Phys. 28 (1958) 258–267.
[5] W. Cai, H. Li, Y. Wang, Partitioned averaged vector field methods, J. Comput. Phys. 370 (2018) 25–42.
[6] E. Celledoni, V. Grimm, R.I. McLachlan, D.I. McLaren, D. O'Neale, B. Brown, G.R.W. Quispel, Preserving energy resp. dissipation in numerical PDEs using the "average vector field" method, J. Comput. Phys. 231 (2012) 6770–6789.
[7] L.Q. Chen, Phase-field models for microstructure evolution, Annu. Rev. Mater. Res. 32 (2002) 113–140.
[8] Q. Cheng, J. Shen, Multiple scalar auxiliary variable (sav) approach and its application to the phase-field vesicle membrane model, SIAM J. Sci. Comput. 40 (2018) A3982–A4006.
[9] R. Courant, K. Friedrichs, H. Lewy, Über die partiellen differenzengleichungen der mathematischen physik, Math. Ann. 100 (1928) 32–74.
[10] M. Dahlby, B. Owren, A general framework for deriving integral preserving numerical methods for PDEs, SIAM J. Sci. Comput. 33 (2011) 2318–2340.
[11] S.R. de Groot, P. Mazur, Non-Equilibrium Thermodynamics, Dover, New York, 1984.
[12] H. Ding, P.D.M. Spelt, C. Shu, Diffuse interface model for incompressible two-phase flows with large density ratios, J. Comput. Phys. 226 (2007) 2078–2095.
[13] S. Dong, On imposing dynamic contact-angle boundary conditions for wall-bounded liquid-gas flows, Comput. Methods Appl. Mech. Eng. 247–248 (2012) 179–200.
[14] S. Dong, An efficient algorithm for incompressible N-phase flows, J. Comput. Phys. 276 (2014) 691–728.
[15] S. Dong, Multiphase flows of N immiscible incompressible fluids: a reduction-consistent and thermodynamically-consistent formulation and associated algorithm, J. Comput. Phys. 361 (2018) 1–49.
[16] S. Dong, J. Shen, A time-stepping scheme involving constant coefficient matrices for phase field simulations of two-phase incompressible flows with large density ratios, J. Comput. Phys. 231 (2012) 5788–5804.
[17] S. Eidnes, B. Owren, T. Ringholm, Adaptive energy preserving methods for partial differential equations, Adv. Comput. Math. 44 (2018) 815–839.
[18] D. Furihata, Finite difference schemes for $\frac{\partial u}{\partial t} = \left(\frac{\partial}{\partial x}\right)^{\alpha} \frac{\delta g}{\delta u}$ that inherit energy conservation or dissipation property, J. Comput. Phys. 156 (1999) 181–205.
[19] D. Furihata, T. Matsuo, Discrete Variational Derivative Method, Chapman & Hall/CRC Numerical Analysis and Scientific Computing, CRC Press, Boca Raton, 2011.
[20] Y. Gong, J. Zhao, X. Yang, Q. Wang, Fully discrete second-order linear schemes for hydrodynamic phase field models of binary viscous fluid flows with variable densities, SIAM J. Sci. Comput. 40 (2018) B138–B167.
[21] F. Gonzalez, M. Bellido, D. Gomez, Study of a chemo-repulsion model with quadratic production. Part I: analysis of the continuous problem and time-discrete numerical schemes, arXiv:1803.02386, 2018.
[22] O. Gonzalez, Time integration and discrete Hamiltonian systems, J. Nonlinear Sci. 6 (1996) 449–467.
[23] E. Hairer, C. Lubich, G. Wanner, Geometric Numerical Integration, Springer, 2006.
[24] A. Iserles, A First Course in the Numerical Analysis of Differential Equations, 2nd edn., Cambridge University Press, 2009.
[25] G.E. Karniadakis, S.J. Sherwin, Spectral/hp Element Methods for Computational Fluid Dynamics, 2nd edn., Oxford University Press, 2005.
[26] J. Kim, J. Lowengrub, Phase field modeling and simulation of three-phase flows, Interfaces Free Bound. 7 (2005) 435–466.
[27] J. Kou, S. Sun, Y. Wang, Linearly decoupled energy-stable numerical methods for multicomponent two-phase compressible flow, SIAM J. Numer. Anal. 56 (2018) 3219–3248.
[28] J. Li, J. Zhao, Q. Wang, Energy and entropy preserving numerical approximations of thermodynamically consistent crystal growth models, J. Comput. Phys. 382 (2019) 202–220.
[29] L. Lin, Z. Yang, S. Dong, Numerical approximation of incompressible Navier-Stokes equations based on an auxiliary energy variable, J. Comput. Phys. 388 (2019) 1–22.
[30] C. Liu, J. Shen, A phase field model for the mixture of two incompressible fluids and its approximation by a Fourier-spectral method, Physica D 179 (2003) 211–228.
[31] C. Liu, J. Shen, X. Yang, Decoupled energy stable schemes for a phase-field model of two-phase incompressible flows with variable density, J. Sci. Comput. 62 (2015) 601–622.
[32] J. Lowengrub, L. Truskinovsky, Quasi-incompressible Cahn-Hilliard fluids and topological transitions, Proc. R. Soc. Lond. A 454 (1998) 2617–2654.
[33] R.I. McLachlan, G.R.W. Quispel, N. Robidoux, Geometric integration using discrete gradients, Philos. Trans. R. Soc. Lond. A 357 (1999) 1021–1045.
[34] Y. Miyatake, T. Matsuo, A general framework for finding energy dissipative/conservative H^1-Galerkin schemes and their underlying h^1-weak forms for nonlinear evolution, BIT Numer. Math. 54 (2014) 1119–1154.
[35] H.C. Ottinger, Beyond Equilibrium Thermodynamics, Wiley, 2005.
[36] G.R.W. Quispel, D.I. McLaren, A new class of energy-preserving numerical integration methods, J. Phys. A, Math. Theor. 41 (2008).

[37] J. Shen, J. Xu, Convergence and error analysis for the scalar auxiliary variable (sav) schemes to gradient flows, SIAM J. Numer. Anal. 56 (2018) 2895–2912.
[38] J. Shen, J. Xu, J. Yang, The scalar auxiliary variable (sav) approach for gradient flows, J. Comput. Phys. 353 (2018) 407–416.
[39] W. Strauss, Numerical solution of nonlinear Klein-Gordon equation, J. Comput. Phys. 28 (1978) 271–278.
[40] J.C. Willems, Dissipative dynamical systems Part I: general theory, Arch. Ration. Mech. Anal. 45 (1972) 321–351.
[41] J.C. Willems, Dissipative dynamical systems, Eur. J. Control 13 (2007) 134–151.
[42] S. Wu, J. Xu, Multiphase Allen-Cahn and Cahn-Hilliard models and their discretizations with the effect of pairwise surface tensions, J. Comput. Phys. 343 (2017) 10–32.
[43] J. Xu, Y. Li, S. Wu, A. Bousquet, On the stability and accuracy of partially and fully implicit schemes for phase field modeling, Comput. Methods Appl. Mech. Eng. 345 (2019) 826–853.
[44] X. Yang, Linear first and second-order, unconditionally energy stable numerical schemes for the phase field model of homopolymer blends, J. Comput. Phys. 327 (2016) 294–316.
[45] X. Yang, Efficient linear, stabilized, second-order time marching schemes for an anisotropic phase field dendritic crystal growth model, Comput. Methods Appl. Mech. Eng. 347 (2019) 316–339.
[46] Z. Yang, S. Dong, An unconditionally energy-stable scheme based on an implicit auxiliary energy variable for incompressible two-phase flows with different densities involving only precomputable coefficient matrices, arXiv:1811.07888, 2018.
[47] Z. Yang, L. Lin, S. Dong, A family of second-order energy-stable schemes for Cahn-Hilliard type equations, J. Comput. Phys. 383 (2019) 24–54.
[48] P. Yue, J.J. Feng, C. Liu, J. Shen, A diffuse-interface method for simulating two-phase flows of complex fluids, J. Fluid Mech. 515 (2004) 293–317.
[49] J. Zhao, X. Yang, Y. Gong, X. Zhao, X. Yang, J. Li, Q. Wang, A general strategy for numerical approximations of non-equilibrium models – Part I: thermodynamical systems, Int. J. Numer. Anal. Model. 15 (2018) 884–918.