Noname manuscript No.
(will be inserted by the editor)

Stabilized exponential-SAV schemes preserving energy dissipation law and maximum bound principle for the Allen–Cahn type equations

Lili Ju · Xiao Li · Zhonghua Qiao

Received: date / Accepted: date

Abstract It is well-known that the Allen–Cahn equation not only satisfies the energy dissipation law but also possesses the maximum bound principle (MBP) in the sense that the absolute value of its solution is pointwise bounded for all time by some specific constant under appropriate initial/boundary conditions. In recent years, the scalar auxiliary variable (SAV) method and many of its variants have attracted much attention in numerical solution for gradient flow problems due to their inherent advantage of preserving certain discrete analogues of the energy dissipation law. However, existing SAV schemes usually fail to preserve the MBP when applied to the Allen–Cahn equation. In this paper, we develop and analyze new first- and second-order stabilized exponential-SAV schemes for a class of Allen–Cahn type equations, which are shown to simultaneously preserve the energy dissipation law and MBP in discrete settings. In addition, optimal error estimates for the numerical solutions are rigorously obtained for both schemes. Extensive numerical tests and comparisons are also conducted to demonstrate the performance of the proposed schemes.

Keywords maximum bound principle · energy dissipation · stabilized method · exponential scalar auxiliary variable

Mathematics Subject Classification (2020) 65M12 · 65M15 · 35K55 · 35K57

1 Introduction

Let us consider a class of reaction-diffusion equations taking the following form

$$u_t = \varepsilon^2 \Delta u + f(u), \quad t > 0, \ \boldsymbol{x} \in \Omega, \tag{1.1}$$

where $\Omega \subset \mathbb{R}^d$ is a spatial domain, $u = u(t, \boldsymbol{x}) : [0, \infty) \times \overline{\Omega} \to \mathbb{R}$ is the unknown function, $\varepsilon > 0$ denotes an interfacial parameter, and $f(u)$ is a nonlinear reaction term with f being continuously differentiable. We also impose the initial condition

$$u(0, \cdot) = u_{\text{init}} \quad \text{on } \overline{\Omega}$$

This work is supported by the CAS AMSS-PolyU Joint Laboratory of Applied Mathematics. L. Ju's work is partially supported by US National Science Foundation grant DMS-2109633 and US Department of Energy grant DE-SC0020270. X. Li's work is partially supported by the Hong Kong Research Council GRF grant 15300821 and the Hong Kong Polytechnic University internal grants 4-ZZMK and 1-BD8N. Z. Qiao's work is partially supported by the Hong Kong Research Council RFS grant RFS2021-5S03 and GRF grants 15300417 and 15302919 and the Hong Kong Polytechnic University internal grant 4-ZZKK.

L. Ju
Department of Mathematics, University of South Carolina, Columbia, SC 29208, USA
E-mail: ju@math.sc.edu

X. Li
Corresponding author. Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
E-mail: xiao1li@polyu.edu.hk

Z. Qiao
Department of Applied Mathematics & Research Institute for Smart Energy, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
E-mail: zqiao@polyu.edu.hk

and the periodic or homogeneous Neumann boundary conditions. The equation (1.1) usually can be regarded as the L^2 gradient flow with respect to the energy functional

$$E(u)=\int_{\Omega}\Big(\frac{\varepsilon^2}{2}|\nabla u(\boldsymbol{x})|^2+F(u(\boldsymbol{x}))\Big)\,\mathrm{d}\boldsymbol{x}, \tag{1.2}$$

where F is a smooth potential function satisfying $F'=-f$, and thus, the solution to the equation (1.1) decreases the energy (1.2) along with the time, i.e., $\frac{\mathrm{d}}{\mathrm{d}t}E(u(t))\leq 0$, which is often called the *energy dissipation law*. In addition, we also assume that

$$\text{there exists a constant } \beta>0 \text{ such that } f(\beta)\leq 0\leq f(-\beta). \tag{1.3}$$

It has been proved in [10] that the equation (1.1) satisfies the *maximum bound principle* (MBP) in the sense that if the absolute value of the initial data is bounded pointwise by β, then the absolute value of the solution is also bounded by β pointwise for all time, i.e.,

$$\max_{\boldsymbol{x}\in\overline{\Omega}}|u_{\text{init}}(\boldsymbol{x})|\leq\beta \quad\Longrightarrow\quad \max_{\boldsymbol{x}\in\overline{\Omega}}|u(t,\boldsymbol{x})|\leq\beta,\quad \forall\, t>0. \tag{1.4}$$

An important and special case of (1.1) is the Allen–Cahn equation with $f(u)=u-u^3$, which was originally introduced in [1] to model the motion of anti-phase boundaries in crystalline solids. The solution represents the difference between the concentrations of two components of the alloy and thus should be evaluated between -1 and 1, which is guaranteed by the MBP. With the corresponding double-well potential $F(u)=\frac{1}{4}(u^2-1)^2$, the associated energy functional (1.2) decays in time, which reflects the energy dissipation of the phase transition process. Both the MBP and the energy dissipation are also satisfied by some variants of (1.1), such as the nonlocal Allen–Cahn equation for phase separations within long-range interactions [3, 9] and the fractional Allen–Cahn equation used to describe some anomalous diffusion processes [11,14]. To obtain stable numerical simulations and avoid nonphysical solutions for these models, it is highly desirable to design numerical schemes preserving effectively these two basic physical properties, the MBP and the energy dissipation law in time-discrete settings.

In the past decades, there has been a large amount of research denoted to energy-stable numerical schemes for time discretization of gradient flow equations, such as convex splitting schemes [5,13,30,36], stabilized semi-implicit schemes [12,34,38,40], and exponential time differencing (ETD) schemes [9,20, 21]. More recently, invariant energy quadratization (IEQ) schemes [39,41,42] and scalar auxiliary variable (SAV) schemes [31–33] were proposed to naturally provide energy-stable and linear algorithms. While the main idea for both methods is to reformulate and split the energy functional (1.2) in the quadratic form by introducing extra variables, the SAV approach is usually more efficient in terms of computations. Many variants of SAV schemes were developed later; see [2,4,6,7,15,18,26] and the references therein. In practice, a suitable stabilization term is also introduced in such splitting in order to maintain numerical stability for highly stiff problems. On the other hand, existing SAV-type schemes usually fail to preserve the MBP, and a special case is the auxiliary variable proposed in [18] which is shown to be positivity-preserving.

The MBP preservation recently has also attracted increasingly attention in the field of numerical methods for the Allen–Cahn type equations of the form (1.1). The semi-implicit schemes were extensively studied in, e.g. [16,17,25,29,35,37], for the classic, fractional, or surface Allen–Cahn equations. The first- and second-order stabilized ETD schemes were shown to preserve the MBP unconditionally for the nonlocal Allen–Cahn equation [9] and the conservative Allen–Cahn equation [22]. An abstract framework on MBP preservation of the ETD schemes for a class of semilinear parabolic equations was established in [10], where sufficient conditions for the linear and nonlinear operators are presented in order to guarantee the MBP. A family of stabilized integrating factor Runge–Kutta (IFRK) schemes, up to third order, were developed in [23], which can unconditionally preserve the MBP. In addition, a fourth-order (conditionally) MBP-preserving IFRK scheme was presented in [19]. So far, as one of the very popular methods, there is still not much systematical study on MBP-preserving SAV schemes.

The main goal of this paper is to develop first- and second-order energy dissipative and MBP-preserving SAV-type schemes for the Allen–Cahn type equation (1.1) by using an appropriate stabilization technique. Specifically, we propose new stabilized exponential-SAV (ESAV) schemes by introducing an artificial stabilization term rather than basing on the splitting of the energy functional suggested in [33]. With the effect of such stabilization, we show that the proposed first-order scheme preserves the MBP unconditionally with an appropriate stabilizing parameter and the second-order one does under a time step size constraint. A main difficulty for numerical analysis of the two schemes lies in that the coefficients of the nonlinear term and the stabilization term are varying rather than constant due to the use of the SAV approach. With the help of the energy dissipation and MBP, we are able to show that such variable coefficients are bounded from above and below by certain positive constants, and consequently, optimal error estimate are successfully obtained for the proposed schemes. To the best of our knowledge, this is the first work in the direction of designing such SAV-type methods. More importantly, the proposed stabilizing approaches can be easily generalized to deal with many other type of gradient flow problems where the SAV methods apply.

The rest of this paper is organized as follows. Section 2 is devoted to spatial discretization of the equation (1.1) and a brief summary of the classic SAV and ESAV schemes for the time integration. Then, our first- and second-order stabilized ESAV schemes are presented in Section 3, together with the energy dissipation law, MBP preservation, and convergence analysis of the resulting fully discrete systems. In Section 4, extensive numerical tests and comparisons are carried out to demonstrate the performance of the proposed schemes. Some concluding remarks are finally given in Section 5.

2 Spatial discretization and SAV schemes for time integration

For simplicity, throughout this paper, we consider the two-dimensional square domain $\Omega = (0, L) \times (0, L)$ for the equation (1.1) equipped with periodic boundary conditions. Note that the extensions to three-dimension problems and homogeneous Neumann boundary condition are straightforward. In this section, we first present some notations related to the spatial discretization by central finite difference, then briefly review the classic SAV schemes for time integration. For other feasible spatial discretization, we refer to [10] for more details.

2.1 Spatial discretization and the space-discrete problem

Given a positive integer M, we set $h = L/M$ to be the size of the uniform mesh partitioning $\overline{\Omega}$. Denote by Ω_h the set of mesh points $(x_i, y_j) = (ih, jh)$, $1 \le i, j \le M$. For a grid function v defined on Ω_h, we write $v_{ij} = v(x_i, y_j)$ for simplicity. Let $\mathcal{M}_h$ be the set of all periodic grid functions on Ω_h, i.e.,

$$\mathcal{M}_h = \{v : \Omega_h \to \mathbb{R} \,|\, v_{i+kM, j+lM} = v_{ij},\ k, l \in \mathbb{Z},\ 1 \le i, j \le M\}.$$

The discrete inner product $\langle \cdot, \cdot \rangle$, discrete L^2 norm $\|\cdot\|$, and discrete L^∞ norm $\|\cdot\|_\infty$ can be defined as usual, namely,

$$\langle v, w \rangle = h^2 \sum_{i,j=1}^{M} v_{ij} w_{ij}, \qquad \|v\| = \sqrt{\langle v, v \rangle}, \qquad \|v\|_\infty = \max_{1 \le i,j \le M} |v_{ij}|$$

for any $v, w \in \mathcal{M}_h$, and

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \langle v^1, w^1 \rangle + \langle v^2, w^2 \rangle, \qquad \|\boldsymbol{v}\| = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle}$$

for any $\boldsymbol{v} = (v^1, v^2)^T, \boldsymbol{w} = (w^1, w^2)^T \in \mathcal{M}_h \times \mathcal{M}_h$. We apply the second-order central finite difference to approximate spatial differentiation operators. For any $v \in \mathcal{M}_h$, the discrete Laplace operator Δ_h is defined by

$$\Delta_h v_{ij} = \frac{1}{h^2}(v_{i+1,j} + v_{i-1,j} + v_{i,j+1} + v_{i,j-1} - 4v_{ij}), \quad 1 \le i, j \le M,$$

and the discrete gradient operator ∇_h is defined by

$$\nabla_h v_{ij} = \Big(\frac{v_{i+1,j} - v_{ij}}{h}, \frac{v_{i,j+1} - v_{ij}}{h} \Big)^T, \quad 1 \le i, j \le M.$$

By periodic boundary conditions, the summation-by-parts formula is easy to verify:

$$\langle v, \Delta_h w \rangle = -\langle \nabla_h v, \nabla_h w \rangle = \langle \Delta_h v, w \rangle, \quad \forall\, v, w \in \mathcal{M}_h.$$

Obviously, Δ_h is self-adjoint and negative semi-definite. For any function $\varphi : \overline{\Omega} \to \mathbb{R}$, we denote by I_h the operator projecting φ on the mesh as $(I_h \varphi)_{ij} = \varphi(x_i, y_j)$ for $1 \le i, j \le M$. For example, we have

$$\max_{1 \le i,j \le N} |\Delta_h (I_h \varphi)_{ij} - \Delta \varphi(x_i, y_j)| \le C_\varphi h^2, \quad \forall\, \varphi \in C^4_{\rm per}(\overline{\Omega}).$$

For simplicity, we may directly omit the notation I_h when there is no ambiguity.

Since $\mathcal{M}_h$ is a finite-dimensional linear space, any grid function $v \in \mathcal{M}_h$ and any linear operator $Q : \mathcal{M}_h \to \mathcal{M}_h$ can be regarded as a vector in $\mathbb{R}^{M^2}$ and a matrix in $\mathbb{R}^{M^2 \times M^2}$, respectively. We still use the notations $\|\cdot\|$ and $\|\cdot\|_\infty$ to denote the matrix induced-norms consistent with $\|\cdot\|$ and $\|\cdot\|_\infty$ defined for vectors before, respectively. By regarding Δ_h as a linear operator, we know that Δ_h is the generator of a contraction semigroup on $\mathcal{M}_h$ [10]. Instead, by viewing Δ_h as a matrix, it is weakly diagonally dominant with all diagonal entries negative. Moreover, we have the following useful estimate and the proof can be found in [35].

Lemma 1 *For any $a > 0$, we have $\|(aI - \Delta_h)^{-1}\|_\infty \le a^{-1}$, where I represents the identity matrix.*

We have assumed that f is continuously differentiable, so $\|f'\|_{C[-\beta,\beta]}$ is always finite and then the following result is valid [10].

Lemma 2 *Under the assumption (1.3), if $\kappa \ge \|f'\|_{C[-\beta,\beta]}$ holds for some positive constant κ, then we have $|f(\xi) + \kappa\xi| \le \kappa\beta$ for any $\xi \in [-\beta, \beta]$.*

Next, let us introduce the space-discrete version of (1.1). The space-discrete problem is to find a function $u_h : [0, \infty) \to \mathcal{M}_h$ satisfies

$$\frac{\mathrm{d}u_h}{\mathrm{d}t} = \varepsilon^2 \Delta_h u_h + f(u_h) \tag{2.1}$$

with $u_h(0) = u_{\mathrm{init}}$. It is easy to verify the energy dissipation law for (2.1) in the sense that

$$\frac{\mathrm{d}}{\mathrm{d}t} E_h(u_h(t)) \le 0,$$

where E_h is the spatially-discretized energy functional defined as

$$E_h(v) := \frac{\varepsilon^2}{2}\|\nabla_h v\|^2 + \langle F(v), 1\rangle, \quad \forall\, v \in \mathcal{M}_h. \tag{2.2}$$

According to [10], the MBP also holds for u_h, i.e., $\|u_h(t)\|_\infty \le \beta$ for any $t > 0$ if $\|u_{\mathrm{init}}\|_\infty \le \beta$.

Let us partition the time interval into $\{t_n = n\tau\}_{n\ge 0}$ with $\tau > 0$ being a uniform time step size. In the remaining part of the paper, we will study time integration schemes for the space-discrete system (2.1). For simplicity of representation, we denote by u^n the fully discrete approximate value of $u_e(t_n)$ or $u_{h,e}(t_n)$ with u_e and $u_{h,e}$ denoting the exact solutions to the original continuous problem (1.1) and the space-discrete problem (2.1), respectively. In general, for a sequence $\{v^n\}$, we define the following notations:

$$\delta_t v^{n+1} = \frac{v^{n+1} - v^n}{\tau}, \qquad v^{n+\frac{1}{2}} = \frac{v^{n+1} + v^n}{2}.$$

2.2 Classic SAV schemes and their stabilization

Here we give a brief summary of the classic SAV schemes. The main idea of SAV is to reformulate the energy functional (1.2) in the quadratic form by introducing an appropriate SAV. The framework of the classic SAV schemes is based on a linear splitting of the energy functional and, as shown in [33], a suitable stabilization term is usually also introduced in such splitting so that the numerical simulations can provide satisfactory results for highly stiff problems in practice.

Denoting by $\kappa \ge 0$ the stabilizing constant, the energy functional (1.2) can be rewritten with a stabilization term as

$$\begin{aligned} E(u) &= \int_\Omega \Big(\frac{\varepsilon^2}{2}|\nabla u|^2 + \frac{\kappa}{2}u^2 + F(u) - \frac{\kappa}{2}u^2\Big)\,\mathrm{d}\boldsymbol{x} \\ &= \frac{\varepsilon^2}{2}\|\nabla u\|_{L^2}^2 + \frac{\kappa}{2}\|u\|_{L^2}^2 + \int_\Omega \Big(F(u) - \frac{\kappa}{2}u^2\Big)\,\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{2.3}$$

Suppose the last term in (2.3) is bounded from below, that is,

$$E_2(u) := \int_\Omega \Big(F(u) - \frac{\kappa}{2}u^2\Big)\,\mathrm{d}\boldsymbol{x} \ge -C_0 \tag{2.4}$$

for some constant $C_0 \geq 0$. Choosing $\delta > C_0$, let us define an auxiliary variable $r(t) = \sqrt{E_2(u(t)) + \delta}$, and reformulate the original problem (1.1) to the following equivalent system:

$$
\begin{aligned}
u_t &= \varepsilon^2 \Delta u - \kappa u + \frac{r}{\sqrt{E_2(u) + \delta}}(f(u) + \kappa u), \\
r_t &= -\frac{1}{2\sqrt{E_2(u) + \delta}}(f(u) + \kappa u, u_t).
\end{aligned}
$$

Then the first-order SAV scheme (SAV1) is given by [33]

$$
\delta_t u^{n+1} = \varepsilon^2 \Delta_h u^{n+1} - \kappa u^{n+1} + \frac{r^{n+1}}{\sqrt{E_{2h}(u^n) + \delta}}(f(u^n) + \kappa u^n), \tag{2.6a}
$$

$$
\delta_t r^{n+1} = -\frac{1}{2\sqrt{E_{2h}(u^n) + \delta}}\langle f(u^n) + \kappa u^n, \delta_t u^{n+1}\rangle, \tag{2.6b}
$$

and the Crank–Nicolson type second-order SAV scheme (SAV2) reads as [33]

$$
\delta_t u^{n+1} = \varepsilon^2 \Delta_h u^{n+\frac{1}{2}} - \kappa u^{n+\frac{1}{2}} + \frac{r^{n+1} + r^n}{2\sqrt{E_{2h}(\widehat{u}^{n+\frac{1}{2}}) + \delta}}(f(\widehat{u}^{n+\frac{1}{2}}) + \kappa \widehat{u}^{n+\frac{1}{2}}), \tag{2.7a}
$$

$$
\delta_t r^{n+1} = -\frac{1}{2\sqrt{E_{2h}(\widehat{u}^{n+\frac{1}{2}}) + \delta}}\langle f(\widehat{u}^{n+\frac{1}{2}}) + \kappa \widehat{u}^{n+\frac{1}{2}}, \delta_t u^{n+1}\rangle, \tag{2.7b}
$$

where $\widehat{u}^{n+\frac{1}{2}}$ is generated by solving the system

$$
\frac{\widehat{u}^{n+\frac{1}{2}} - u^n}{\tau/2} = \varepsilon^2 \Delta_h \widehat{u}^{n+\frac{1}{2}} + f(u^n) - \kappa(\widehat{u}^{n+\frac{1}{2}} - u^n).
$$

Both (2.6) and (2.7) are linear schemes and energy dissipative in the sense that $\overline{E}_h(u^{n+1}, r^{n+1}) \leq \overline{E}_h(u^n, r^n)$ with respect to the following modified energy

$$
\overline{E}_h(u^n, r^n) := \frac{\varepsilon^2}{2}\|\nabla_h u^n\|^2 + \frac{\kappa}{2}\|u^n\|^2 + (r^n)^2 - \delta.
$$

Note that in the discrete settings, $\overline{E}_h(u^n, r^n)$ is only an approximation of the original discrete energy $E_h(u^n)$ defined in (2.2) and they are not equal in general since $r^n \neq \sqrt{E_{2h}(u^n) + \delta}$ for $n \geq 1$. However, the MBP cannot be theoretically preserved by the above classic SAV schemes (2.6) and (2.7) (See the discussion in Remark 3).

2.3 Exponential-SAV schemes

A variant of the classic SAV approach, called the exponential-SAV (ESAV) scheme, was studied in [26]. We below summarize the ESAV method, also with a stabilization term based on the energy splitting (2.3). Define the auxiliary variable by $r(t) = \exp\{E_2(u(t))\}$ and reformulate (1.1) as

$$
\begin{aligned}
u_t &= \varepsilon^2 \Delta u - \kappa u + \frac{r}{\exp\{E_2(u)\}}(f(u) + \kappa u), \\
(\ln r)_t &= -\frac{r}{\exp\{E_2(u)\}}(f(u) + \kappa u, u_t).
\end{aligned}
$$

Then the first-order ESAV scheme (ESAV1) reads as

$$
\delta_t u^{n+1} = \varepsilon^2 \Delta_h u^{n+1} - \kappa u^{n+1} + \frac{r^n}{\exp\{E_{2h}(u^n)\}}(f(u^n) + \kappa u^n), \tag{2.9a}
$$

$$
\frac{\ln r^{n+1} - \ln r^n}{\tau} = -\frac{r^n}{\exp\{E_{2h}(u^n)\}}\langle f(u^n) + \kappa u^n, \delta_t u^{n+1}\rangle. \tag{2.9b}
$$

Setting $\kappa = 0$, the scheme (2.9) reduces exactly to the original ESAV scheme (without stabilization) presented in [26]. The Crank–Nicolson type ESAV scheme (ESAV2) is given by

$$\delta_t u^{n+1} = \varepsilon^2 \Delta_h u^{n+\frac{1}{2}} - \kappa u^{n+\frac{1}{2}} + \frac{\widehat{r}^{n+\frac{1}{2}}}{\exp\{E_{2h}(\widehat{u}^{n+\frac{1}{2}})\}}(f(\widehat{u}^{n+\frac{1}{2}}) + \kappa \widehat{u}^{n+\frac{1}{2}}), \tag{2.10a}$$

$$\frac{\ln r^{n+1} - \ln r^n}{\tau} = -\frac{\widehat{r}^{n+\frac{1}{2}}}{\exp\{E_{2h}(\widehat{u}^{n+\frac{1}{2}})\}}\langle f(\widehat{u}^{n+\frac{1}{2}}) + \kappa \widehat{u}^{n+\frac{1}{2}}, \delta_t u^{n+1}\rangle, \tag{2.10b}$$

where the value $(\widehat{u}^{n+\frac{1}{2}}, \widehat{r}^{n+\frac{1}{2}})$ can be generated by an extrapolation as suggested in [26] or predicted by the first-order scheme (2.9) with half of the time step size:

$$\frac{\widehat{u}^{n+\frac{1}{2}} - u^n}{\tau/2} = \varepsilon^2 \Delta_h \widehat{u}^{n+\frac{1}{2}} - \kappa \widehat{u}^{n+\frac{1}{2}} + \frac{r^n}{\exp\{E_{2h}(u^n)\}}(f(u^n) + \kappa u^n), \tag{2.11a}$$

$$\ln \widehat{r}^{n+\frac{1}{2}} - \ln r^n = -\frac{r^n}{\exp\{E_{2h}(u^n)\}}\langle f(u^n) + \kappa u^n, \widehat{u}^{n+\frac{1}{2}} - u^n\rangle. \tag{2.11b}$$

We will adopt (2.11) in the numerical experiments for the comparison. Both (2.9) and (2.10) are energy dissipative in the sense that $\widetilde{E}_h(u^{n+1}, r^{n+1}) \le \widetilde{E}_h(u^n, r^n)$ with respect to the following modified energy

$$\widetilde{E}_h(u^n, r^n) := \frac{\varepsilon^2}{2}\|\nabla_h u^n\|^2 + \frac{\kappa}{2}\|u^n\|^2 + \ln r^n.$$

Similar to the classic SAV schemes, the above ESAV schemes (2.9) and (2.10) also cannot preserve the MBP (See the discussion in Remark 3).

3 New stabilized exponential-SAV schemes

From now on, we always assume the initial value u_{init} has the enough regularity as needed. By spatial discretization, there is a constant $h_0 > 0$, depending on u_{init}, F, ε, such that

$$E_h(u_{\text{init}}) \le E(u_{\text{init}}) + 1, \quad \|\Delta_h u_{\text{init}}\| \le \|\Delta u_{\text{init}}\|_{L^2} + 1, \qquad \forall\, h \in (0, h_0]. \tag{3.1}$$

The continuity of F implies that F is bounded from below on $[-\beta, \beta]$. Therefore, according to the MBP (1.4), it holds that

$$E_1(u) := \int_\Omega F(u)\,\mathrm{d}\boldsymbol{x} \ge -C_*$$

for some constant $C_* \ge 0$. Introducing $s(t) = E_1(u(t))$, we then have the following energy which is equivalent to $E(u)$:

$$\mathcal{E}(u, s) = \frac{\varepsilon^2}{2}\|\nabla u\|_{L^2}^2 + s.$$

Partially inspired by the idea of ESAV method [26], we rewrite the equation (1.1) as the following equivalent system:

$$u_t = \varepsilon^2 \Delta u + \frac{\exp\{s\}}{\exp\{E_1(u)\}} f(u),$$

$$s_t = -\frac{\exp\{s\}}{\exp\{E_1(u)\}}(f(u), u_t).$$

The corresponding space-discrete problem is to find $u_h(t) \in \mathcal{M}_h$ and $s_h(t)$ for $t > 0$ satisfies

$$\frac{\mathrm{d}u_h}{\mathrm{d}t} = \varepsilon^2 \Delta_h u_h + g(u_h, s_h) f(u_h), \tag{3.3a}$$

$$\frac{\mathrm{d}s_h}{\mathrm{d}t} = -g(u_h, s_h)\Big\langle f(u_h), \frac{\mathrm{d}u_h}{\mathrm{d}t}\Big\rangle, \tag{3.3b}$$

where

$$g(u_h, s_h) := \frac{\exp\{s_h\}}{\exp\{E_{1h}(u_h)\}} > 0, \tag{3.4}$$

and E_{1h} denotes the space-discrete version of E_1, i.e., $E_{1h}(v) := \langle F(v), 1\rangle$ for any $v \in \mathcal{M}_h$. Based on such an equivalent form, we will give the stabilized ESAV schemes in the fully discrete version. This section is devoted to the first-order scheme and the second-order one will be discussed in the next section. Recall that we use u^n to represent the fully discrete approximate value of $u_e(t_n)$, the exact solution to the problem (1.1).

3.1 First-order sESAV scheme

The first-order stabilized ESAV fully-discrete scheme (sESAV1) is given by

$$\delta_t u^{n+1} = \varepsilon^2 \Delta_h u^{n+1} + g(u^n, s^n) f(u^n) - \kappa g(u^n, s^n)(u^{n+1} - u^n), \tag{3.5a}$$

$$\delta_t s^{n+1} = -g(u^n, s^n) \langle f(u^n), \delta_t u^{n+1} \rangle, \tag{3.5b}$$

where $\kappa \geq 0$ is a stabilizing constant and $g(u^n, s^n) > 0$. The scheme (3.5) is started by $u^0 = u_{\text{init}}$ and $s^0 = E_{1h}(u^0)$. We can rewrite (3.5) equivalently as follows:

$$\Big[\Big(\frac{1}{\tau} + \kappa g(u^n, s^n)\Big) I - \varepsilon^2 \Delta_h\Big] u^{n+1} = \frac{u^n}{\tau} + g(u^n, s^n) f(u^n) + \kappa g(u^n, s^n) u^n, \tag{3.6a}$$

$$s^{n+1} = s^n - g(u^n, s^n) \langle f(u^n), u^{n+1} - u^n \rangle. \tag{3.6b}$$

Obviously, (3.6) is uniquely solvable for any $\tau > 0$ since $(\frac{1}{\tau} + \kappa g(u^n, s^n)) I - \varepsilon^2 \Delta_h$ is self-adjoint and positive definite, which makes u^{n+1} linearly determined from (3.6a) and then s^{n+1} computed explicitly by (3.6b). If we take $\kappa = 0$ and $r^n = \exp\{s^n\}$, i.e., $s^n = \ln r^n$, it is easy to verify that the scheme (3.5) gives us exactly the ESAV scheme (2.9) with $\kappa = 0$. However, they differ when $\kappa > 0$.

3.1.1 Energy dissipation and MBP

Now let us define a discrete energy as follows

$$\mathcal{E}_h(u^n, s^n) := \frac{\varepsilon^2}{2} \|\nabla_h u^n\|^2 + s^n, \tag{3.7}$$

which is clearly again an approximation of the original discrete energy $E_h(u^n)$. We first show that the sESAV1 scheme (3.5) preserves the energy dissipation law and the MBP unconditionally. Then, as an application of both properties, we also prove the uniform boundedness of the variable coefficient $g(u^n, s^n)$.

Theorem 1 (Energy dissipation of sESAV1) *For any $\kappa \geq 0$ and $\tau > 0$, the sESAV1 scheme* (3.5) *is energy dissipative in the sense that $\mathcal{E}_h(u^{n+1}, s^{n+1}) \leq \mathcal{E}_h(u^n, s^n)$.*

Proof Taking the inner product with (3.5) by $u^{n+1} - u^n$ yields

$$\Big(\frac{1}{\tau} + \kappa g(u^n, s^n)\Big) \|u^{n+1} - u^n\|^2 = \varepsilon^2 \langle \Delta_h u^{n+1}, u^{n+1} - u^n \rangle + g(u^n, s^n) \langle f(u^n), u^{n+1} - u^n \rangle. \tag{3.8}$$

Combining (3.8), (3.5b), and the identity

$$\langle \Delta_h u^{n+1}, u^{n+1} - u^n \rangle = -\frac{1}{2} \|\nabla_h u^{n+1}\|^2 + \frac{1}{2} \|\nabla_h u^n\|^2 - \frac{1}{2} \|\nabla_h u^{n+1} - \nabla_h u^n\|^2,$$

we obtain

$$\mathcal{E}_h(u^{n+1}, s^{n+1}) - \mathcal{E}_h(u^n, s^n) = -\Big(\frac{1}{\tau} + \kappa g(u^n, s^n)\Big) \|u^{n+1} - u^n\|^2 - \frac{\varepsilon^2}{2} \|\nabla_h u^{n+1} - \nabla_h u^n\|^2,$$

which completes the proof.

Theorem 1 implies that the sESAV1 scheme (3.5) is energy dissipative with respect to the modified energy $\mathcal{E}_h(u^n, s^n)$ rather than the original energy $E_h(u^n)$. Note that $s^n \neq E_{1h}(u^n)$ for $n \geq 1$ in general, thus $\mathcal{E}_h(u^n, s^n) \neq E_h(u^n)$. In fact, $\mathcal{E}_h(u^n, s^n)$ is a first-order approximation of $E_h(u^n)$, and it is easy to see that $E_h(u^n)$ is always nonnegative while $\mathcal{E}_h(u^n, s^n)$ may not be. Nevertheless, $\mathcal{E}_h(u^n, s^n)$ is always bounded from below, which is a direct result of the lower boundedness of $\{s^n\}$ to be established by (3.14) under some extra conditions of Corollary 3.

Corollary 1 *For any $\kappa \geq 0$ and $\tau > 0$, it holds $s^n \leq E_h(u_{\text{init}})$ for all n.*

Proof By Theorem 1, since $s^0 = E_{1h}(u_{\text{init}})$, we have

$$\frac{\varepsilon^2}{2} \|\nabla_h u^n\|^2 + s^n = \mathcal{E}_h(u^n, s^n) \leq \mathcal{E}_h(u^{n-1}, s^{n-1}) \leq \cdots \leq \mathcal{E}_h(u^0, s^0) = E_h(u_{\text{init}}).$$

Dropping off the nonnegative term leads to the expected result.

Theorem 2 (MBP of sESAV1) *If $\kappa \ge \|f'\|_{C[-\beta,\beta]}$, the sESAV1 scheme (3.5) preserves the MBP for $\{u^n\}$, i.e., the discrete version of (1.4) is valid as follows:*

$$\|u_{\text{init}}\|_\infty \le \beta \quad \Longrightarrow \quad \|u^n\|_\infty \le \beta, \quad \forall\, n. \tag{3.9}$$

Proof Suppose (u^n, s^n) is given and $\|u^n\|_\infty \le \beta$ for some n. From (3.6a), we have

$$u^{n+1} = \Big[\Big(\frac{1}{\tau} + \kappa g(u^n, s^n)\Big)I - \varepsilon^2 \Delta_h\Big]^{-1}\Big[\frac{1}{\tau}u^n + g(u^n, s^n)(f(u^n) + \kappa u^n)\Big].$$

Since $g(u^n, s^n) > 0$, by Lemma 1, we have

$$\Big\|\Big[\Big(\frac{1}{\tau} + \kappa g(u^n, s^n)\Big)I - \varepsilon^2 \Delta_h\Big]^{-1}\Big\|_\infty \le \Big(\frac{1}{\tau} + \kappa g(u^n, s^n)\Big)^{-1}.$$

Since $\kappa \ge \|f'\|_{C[-\beta,\beta]}$ and $\|u^n\|_\infty \le \beta$, according to Lemma 2, it holds

$$\Big\|\frac{1}{\tau}u^n + g(u^n, s^n)(f(u^n) + \kappa u^n)\Big\|_\infty \le \Big(\frac{1}{\tau} + \kappa g(u^n, s^n)\Big)\beta. \tag{3.10}$$

Therefore, we obtain

$$\|u^{n+1}\|_\infty \le \Big(\frac{1}{\tau} + \kappa g(u^n, s^n)\Big)^{-1}\Big(\frac{1}{\tau} + \kappa g(u^n, s^n)\Big)\beta = \beta.$$

By induction, we have $\|u^n\|_\infty \le \beta$ for all n.

Remark 1 The inequality (3.10) is valid if $(\tau g(u^n, s^n))^{-1} + \kappa \ge \|f'\|_{C[-\beta,\beta]}$. In other words, when $\kappa = 0$ (no stabilization), the MBP still holds for the sESAV1 scheme if the time step size satisfies $\tau \le (g(u^n, s^n)\|f'\|_{C[-\beta,\beta]})^{-1}$ for all n.

Remark 2 For the sESAV1 scheme (3.5), we know that the extra term $-\kappa g(u^n, s^n)(u^{n+1} - u^n)$ stabilizes the time stepping and $\kappa g(u^n, s^n)$ is indeed the stabilizing constant, which is an n-dependent quantity. In the proof of Theorem 2, the key ingredients to preserve the MBP for $\{u^n\}$ involve two aspects: the positivity of $\kappa g(u^n, s^n)$ and the relation of u^{n+1} and u^n. The former implies that the extra term is really a good stabilization term and the latter guarantees the balance between the linear and nonlinear parts so that the stabilized linear operator is sufficient to dominate the nonlinear term in order to preserve the MBP.

Remark 3 For the classic SAV1 scheme (2.6), the stabilization term in (2.6a) actually takes the form

$$-\kappa u^{n+1} + \frac{r^{n+1}}{\sqrt{E_{2h}(u^n) + \delta}}\kappa u^n.$$

The sign of r^{n+1}, and thus the sign of $\frac{r^{n+1}}{\sqrt{E_{2h}(u^n)+\delta}}\kappa$, is uncertain, which violates the positivity of the stabilizing constant. Even though r^{n+1} may be positive in practical computations in some specific cases, such a stabilization term leads to an imbalance between the linear and nonlinear parts since $r^{n+1} \ne \sqrt{E_{2h}(u^n) + \delta}$ for $n \ge 0$ in general, so the scheme (2.6) cannot preserve the MBP theoretically, which will be also observed later in our numerical experiments. Similarly, the stabilization term in the ESAV scheme (2.9a) reads as

$$-\kappa u^{n+1} + \frac{r^n}{\exp\{E_{2h}(u^n)\}}\kappa u^n,$$

and the imbalance also exists between the linear and nonlinear parts since $r^n \ne \exp\{E_{2h}(u^n)\}$ for $n \ge 1$ in general, and thus the ESAV scheme (2.9) also does not preserve the MBP theoretically. Nevertheless, $r^n > 0$ always holds due to the definition of the auxiliary variable, and this is the reason why we consider the ESAV approach rather than the classic one in this work.

Note that the coefficient $g(u^n, s^n)$ may vary step-by-step, which is different from the continuous case that $g(u, s) \equiv 1$ exactly. Fortunately, the change of $g(u^n, s^n)$ is controllable in the sense that it can be bounded by some constants, which is illustrated in the following.

Corollary 2 *If $h \le h_0$, $\kappa \ge \|f'\|_{C[-\beta,\beta]}$, and $\|u_{\text{init}}\|_\infty \le \beta$, then there exists a constant $G^* = G^*(u_{\text{init}}, C_*)$ such that $0 < g(u^n, s^n) \le G^*$ for all n.*

Proof The positivity of $g(u^n, s^n)$ comes from its definition. We know from Corollary 1 and Theorem 2 that $g(u^n, s^n) \le \exp\{E_h(u_{\text{init}}) + C_*\}$ for all n. The h-dependence of upper bound can be removed by (3.1), which completes the proof.

Actually, it also holds that $g(u^n, s^n)$ has a positive lower bound uniformly in n for any fixed terminal time $T > 0$. To show it, we first prove an estimate on the discrete H^2 semi-norm of the numerical solution.

Lemma 3 *Given a fixed time $T > 0$. If $h \le h_0$, $\kappa \ge \|f'\|_{C[-\beta,\beta]}$, and $\|u_{\text{init}}\|_\infty \le \beta$, there exists a constant $M > 0$ depending on C_*, $|\Omega|$, T, u_{init}, κ, ε, and $\|f\|_{C^1[-\beta,\beta]}$, such that*

$$\|\delta_t u^{n+1}\| + \|\Delta_h u^{n+1}\| \le M, \quad 0 \le n \le \lfloor T/\tau \rfloor - 1.$$

Proof Taking the discrete inner product of (3.5a) with $2\tau\Delta_h^2 u^{n+1}$, we obtain

$$\begin{aligned}&(1 + \kappa g(u^n, s^n)\tau)\langle \Delta_h u^{n+1} - \Delta_h u^n, 2\Delta_h u^{n+1}\rangle + 2\varepsilon^2\tau\|\nabla_h \Delta_h u^{n+1}\|^2 \\ &\quad = -2g(u^n, s^n)\tau\langle \nabla_h f(u^n), \nabla_h \Delta_h u^{n+1}\rangle.\end{aligned}$$

Using the facts that

$$\begin{aligned}\langle \Delta_h u^{n+1} - \Delta_h u^n, 2\Delta_h u^{n+1}\rangle &= \|\Delta_h u^{n+1}\|^2 - \|\Delta_h u^n\|^2 + \|\Delta_h u^{n+1} - \Delta_h u^n\|^2, \\ -2g(u^n, s^n)\tau\langle \nabla_h f(u^n), \nabla_h \Delta_h u^{n+1}\rangle &\le \frac{(g(u^n, s^n))^2}{2\varepsilon^2}\tau\|\nabla_h f(u^n)\|^2 + 2\varepsilon^2\tau\|\nabla_h \Delta_h u^{n+1}\|^2,\end{aligned}$$

we obtain

$$(1 + \kappa g(u^n, s^n)\tau)(\|\Delta_h u^{n+1}\|^2 - \|\Delta_h u^n\|^2) \le \frac{(g(u^n, s^n))^2}{2\varepsilon^2}\tau\|\nabla_h f(u^n)\|^2. \tag{3.11}$$

By Theorem 2, we have $\|u^n\|_\infty \le \beta$, and thus,

$$\|\nabla_h f(u^n)\| \le \|f'\|_{C[-\beta,\beta]}\|\nabla_h u^n\| \le \|f'\|_{C[-\beta,\beta]} C_\Omega \|\Delta_h u^n\|, \tag{3.12}$$

where the second step comes from the discrete Poincaré's inequality with C_Ω being a constant depending only on $|\Omega|$ (since $\nabla_h u^n$ has a zero mean due to the periodic boundary condition). Then, by Corollary 2, (3.11) and (3.12), we obtain

$$\begin{aligned}\|\Delta_h u^{n+1}\|^2 &\le (1 + \kappa g(u^n, s^n)\tau)\|\Delta_h u^{n+1}\|^2 \\ &\le \Big[1 + \Big(\kappa G^* + \frac{(G^*\|f'\|_{C[-\beta,\beta]} C_\Omega)^2}{2\varepsilon^2}\Big)\tau\Big]\|\Delta_h u^n\|^2.\end{aligned} \tag{3.13}$$

By recursion, we obtain

$$\begin{aligned}\|\Delta_h u^{n+1}\|^2 &\le \Big[1 + \Big(\kappa G^* + \frac{(G^*\|f'\|_{C[-\beta,\beta]} C_\Omega)^2}{2\varepsilon^2}\Big)\tau\Big]^{n+1}\|\Delta_h u^0\|^2 \\ &\le \mathrm{e}^{\big(\kappa G^* + \frac{(G^*\|f'\|_{C[-\beta,\beta]} C_\Omega)^2}{2\varepsilon^2}\big)T}\|\Delta_h u_{\text{init}}\|^2.\end{aligned}$$

Then, using Corollary 2 again, we derive from (3.5a) directly to get

$$\begin{aligned}\|\delta_t u^{n+1}\| &\le (1 + \kappa g(u^n, s^n)\tau)\|\delta_t u^{n+1}\| \\ &\le \varepsilon^2\|\Delta_h u^{n+1}\| + g(u^n, s^n)\|f(u^n)\| \le \varepsilon^2\|\Delta_h u^{n+1}\| + G^* F_0 |\Omega|^{\frac{1}{2}},\end{aligned}$$

where $F_0 := \|f\|_{C[-\beta,\beta]}$. This completes the proof.

Corollary 3 *Given a fixed time $T > 0$. If $h \le h_0$, $\kappa \ge \|f'\|_{C[-\beta,\beta]}$, and $\|u_{\text{init}}\|_\infty \le \beta$, there exists a constant $G_* > 0$ such that $g(u^n, s^n) \ge G_*$ for $0 \le n \le \lfloor T/\tau \rfloor$, where G_* depends on C_*, $|\Omega|$, T, u_{init}, κ, ε, and $\|f\|_{C^1[-\beta,\beta]}$.*

Proof According to the definition of $g(u^n, s^n)$ in (3.4) and the MBP for $\{u^n\}$, it suffices to show the existence of the lower bound of $\{s^n\}$. Using Lemma 3, we have

$$\langle f(u^n), u^{n+1} - u^n \rangle \leq \tau \|f(u^n)\| \|\delta_t u^{n+1}\| \leq F_0 |\Omega|^{\frac{1}{2}} M\tau,$$

where M is the constant defined in Lemma 3. Then, from (3.6b), we have

$$s^{n+1} \geq s^n - G^* F_0 |\Omega|^{\frac{1}{2}} M\tau.$$

By recursion, noting that $s^0 = E_{1h}(u_{\text{init}}) \geq -C_*$, we obtain

$$s^n \geq s^0 - G^* F_0 |\Omega|^{\frac{1}{2}} Mn\tau \geq -C_* - G^* F_0 |\Omega|^{\frac{1}{2}} MT, \tag{3.14}$$

which completes the proof.

The combination of Corollaries 2 and 3 implies that $0 < G_* \leq g(u^n, s^n) \leq G^*$ for any fixed terminal time $T > 0$, which will play an important role in error estimates of the sESAV1 scheme (3.5) in the next subsection.

3.1.2 Error estimates

In the following error analysis, as well as that for the second-order scheme presented later, we will use many generic constants, and for simplicity of notations, we may denote the constants with the same dependence but different values by the same notation.

If the exact solution u_e to (1.1) is smooth sufficiently, letting $s_e(t) = E_1(u_e(t))$, we have

$$\begin{aligned} \frac{u_e(t_{n+1}) - u_e(t_n)}{\tau} &= \varepsilon^2 \Delta_h u_e(t_{n+1}) + g(u_e(t_n), s_e(t_n)) f(u_e(t_n)) \\ &\quad - \kappa g(u_e(t_n), s_e(t_n))(u_e(t_{n+1}) - u_e(t_n)) + R_{1u}^n, \end{aligned} \tag{3.15a}$$

$$\frac{s_e(t_{n+1}) - s_e(t_n)}{\tau} = -g(u_e(t_n), s_e(t_n)) \Big\langle f(u_e(t_n)), \frac{u_e(t_{n+1}) - u_e(t_n)}{\tau} \Big\rangle + R_{1s}^n, \tag{3.15b}$$

where the truncation errors R_{1u}^n and R_{1s}^n satisfy

$$\|R_{1u}^n\| \leq C_e(\tau + h^2), \qquad |R_{1s}^n| \leq C_e(\tau + h^2) \tag{3.16}$$

with $C_e > 0$ depending only on u_e, κ, ε, and $\|f\|_{C^1[-\beta,\beta]}$. Define the error functions as

$$e_u^n = u^n - u_e(t_n), \qquad e_s^n = s^n - s_e(t_n). \tag{3.17}$$

We first show a lemma on the error estimate for the nonlinear term.

Lemma 4 *If $h \leq h_0$ and $\|u^n\|_\infty \leq \beta$, we have*

$$|g(u^n, s^n) - g(u_e(t_n), s_e(t_n))| \leq C_g(\|e_u^n\| + |e_s^n|), \tag{3.18}$$

and

$$\|g(u^n, s^n) f(u^n) - g(u_e(t_n), s_e(t_n)) f(u_e(t_n))\| \leq C_g(\|e_u^n\| + |e_s^n|), \tag{3.19}$$

where the constant $C_g > 0$ depends on C_*, $|\Omega|$, u_{init}, and $\|f\|_{C^1[-\beta,\beta]}$.

Proof For the exact solutions $u_e(t_n)$ and $s_e(t_n)$, we have $\|u_e(t_n)\|_\infty \leq \beta$ by the MBP and $s_e(t_n) \leq E(u_{\text{init}})$ by the energy dissipation law. Some careful calculations yield

$$\begin{aligned} |g(u^n, s^n) - g(u^n, s_e(t_n))| &= \frac{1}{\exp\{E_{1h}(u^n)\}} |\exp\{s^n\} - \exp\{s_e(t_n)\}| \\ &\leq \frac{\exp\{\xi^n\}}{\exp\{E_{1h}(u^n)\}} |s^n - s_e(t_n)| \\ &\leq G^* |s^n - s_e(t_n)| \end{aligned}$$

with ξ^n being a number between s^n and $s_e(t_n)$, and

$$
\begin{aligned}
&|g(u^n, s_e(t_n)) - g(u_e(t_n), s_e(t_n))| \\
&\quad = \exp\{s_e(t_n)\}\Big|\frac{1}{\exp\{E_{1h}(u^n)\}} - \frac{1}{\exp\{E_{1h}(u_e(t_n))\}}\Big| \\
&\quad \le (\exp\{E(u_{\text{init}}) + C_*\})|E_{1h}(u^n) - E_{1h}(u_e(t_n))| \\
&\quad \le |\Omega|^{\frac{1}{2}}(\exp\{E(u_{\text{init}}) + C_*\})\|F(u^n) - F(u_e(t_n))\| \\
&\quad \le F_0|\Omega|^{\frac{1}{2}}(\exp\{E(u_{\text{init}}) + C_*\})\|u^n - u_e(t_n)\|.
\end{aligned}
$$

By combining both of the above inqualities, we obtain (3.18). In addition, we have

$$
\begin{aligned}
&\|g(u^n, s^n)f(u_e(t_n)) - g(u_e(t_n), s_e(t_n))f(u_e(t_n))\| \\
&\quad \le \|f(u_e(t_n))\||g(u^n, s^n) - g(u_e(t_n), s_e(t_n))| \\
&\quad \le F_0|\Omega|^{\frac{1}{2}}C(\|e_u^n\| + |e_s^n|).
\end{aligned}
$$

According to Corollary 2, it holds

$$
\begin{aligned}
\|g(u^n, s^n)f(u^n) - g(u^n, s^n)f(u_e(t_n))\| &\le G^*\|f(u^n) - f(u_e(t_n))\| \\
&\le G^*\|f'\|_{C[-\beta,\beta]}\|u^n - u_e(t_n)\|.
\end{aligned}
$$

Then, we obtain (3.19) with the application of the triangular inequality to the above two inequalities.

Theorem 3 (Error estimate of sESAV1) *Given a fixed time $T > 0$ and suppose the exact solution u_e is smooth enough on $[0, T] \times \overline{\Omega}$. Assume that $\kappa \ge \|f'\|_{C[-\beta,\beta]}$ and $\|u_{\text{init}}\|_\infty \le \beta$. If τ and h are small sufficiently, then we have the error estimate for the sESAV1 scheme* (3.5) *as follows:*

$$
\|e_u^n\| + \|\nabla_h e_u^n\| + |e_s^n| \le C(\tau + h^2), \qquad 0 \le n \le \lfloor T/\tau \rfloor,
$$

where the constant $C > 0$ depends on C_*, $|\Omega|$, T, u_e, κ, ε, and $\|f\|_{C^1[-\beta,\beta]}$ but is independent of τ and h.

Proof The difference between (3.5) and (3.15) leads to

$$
\begin{aligned}
\delta_t e_u^{n+1} &= \varepsilon^2\Delta_h e_u^{n+1} + g(u^n, s^n)f(u^n) - g(u_e(t_n), s_e(t_n))f(u_e(t_n)) - \kappa g(u^n, s^n)(e_u^{n+1} - e_u^n) \\
&\quad + \kappa(g(u_e(t_n), s_e(t_n)) - g(u^n, s^n))(u_e(t_{n+1}) - u_e(t_n)) - R_{1u}^n, && \text{(3.20a)} \\
\delta_t e_s^{n+1} &= \Big\langle g(u_e(t_n), s_e(t_n))f(u_e(t_n)) - g(u^n, s^n)f(u^n), \frac{u_e(t_{n+1}) - u_e(t_n)}{\tau}\Big\rangle \\
&\quad - g(u^n, s^n)\langle f(u^n), \delta_t e_u^{n+1}\rangle - R_{1s}^n. && \text{(3.20b)}
\end{aligned}
$$

Taking the discrete inner product of (3.20a) with $2\tau\delta_t e_u^{n+1}$ and rearranging the terms give us

$$
\begin{aligned}
&2\tau\|\delta_t e_u^{n+1}\|^2 + 2\kappa g(u^n, s^n)\|e_u^{n+1} - e_u^n\|^2 - 2\varepsilon^2\langle\Delta_h e_u^{n+1}, e_u^{n+1} - e_u^n\rangle \\
&\quad = 2\tau\langle g(u^n, s^n)f(u^n) - g(u_e(t_n), s_e(t_n))f(u_e(t_n)), \delta_t e_u^{n+1}\rangle \\
&\qquad + 2\kappa\tau(g(u_e(t_n), s_e(t_n)) - g(u^n, s^n))\langle u_e(t_{n+1}) - u_e(t_n), \delta_t e_u^{n+1}\rangle - 2\tau\langle R_{1u}^n, \delta_t e_u^{n+1}\rangle.
\end{aligned}
$$

Since $g(u^n, s^n) \ge G_* > 0$ by Corollary 3, using the identities

$$
\begin{aligned}
\langle\Delta_h e_u^{n+1}, e_u^{n+1} - e_u^n\rangle &= -\frac{1}{2}\|\nabla_h e_u^{n+1}\|^2 + \frac{1}{2}\|\nabla_h e_u^n\|^2 - \frac{1}{2}\tau^2\|\nabla_h\delta_t e_u^{n+1}\|^2, \\
\|e_u^{n+1} - e_u^n\|^2 &= \|e_u^{n+1}\|^2 - \|e_u^n\|^2 - 2\tau\langle e_u^n, \delta_t e_u^{n+1}\rangle, && \text{(3.21)}
\end{aligned}
$$

we obtain

$$
\begin{aligned}
&2G_*\kappa\|e_u^{n+1}\|^2 - 2G_*\kappa\|e_u^n\|^2 + \varepsilon^2\|\nabla_h e_u^{n+1}\|^2 - \varepsilon^2\|\nabla_h e_u^n\|^2 + 2\tau\|\delta_t e_u^{n+1}\|^2 \\
&\quad \le 2\tau\langle g(u^n, s^n)f(u^n) - g(u_e(t_n), s_e(t_n))f(u_e(t_n)), \delta_t e_u^{n+1}\rangle \\
&\qquad + 2\kappa\tau(g(u_e(t_n), s_e(t_n)) - g(u^n, s^n))\langle u_e(t_{n+1}) - u_e(t_n), \delta_t e_u^{n+1}\rangle \\
&\qquad + 4G_*\kappa\tau\langle e_u^n, \delta_t e_u^{n+1}\rangle - 2\tau\langle R_{1u}^n, \delta_t e_u^{n+1}\rangle. && \text{(3.22)}
\end{aligned}
$$

For the first term in the right-hand side of (3.22), by the Schwartz's inequality and Lemma 4, we have

$$\begin{aligned}
&2\tau\langle g(u^n,s^n)f(u^n)-g(u_e(t_n),s_e(t_n))f(u_e(t_n)),\delta_t e_u^{n+1}\rangle \\
&\qquad\le 2C_g\tau(\|e_u^n\|+|e_s^n|)\|\delta_t e_u^{n+1}\| \le 4C_g^2\tau(\|e_u^n\|^2+|e_s^n|^2)+\frac{\tau}{2}\|\delta_t e_u^{n+1}\|^2,
\end{aligned}\tag{3.23}$$

where $C_g>0$ is the constant in Lemma 4. For the second term in the right-hand side of (3.22), we have

$$\begin{aligned}
&2\kappa\tau(g(u_e(t_n),s_e(t_n))-g(u^n,s^n))\langle u_e(t_{n+1})-u_e(t_n),\delta_t e_u^{n+1}\rangle \\
&\qquad\le 2\kappa\tau|g(u_e(t_n),s_e(t_n))-g(u^n,s^n)|\|u_e(t_{n+1})-u_e(t_n)\|\|\delta_t e_u^{n+1}\| \\
&\qquad\le 2C_g\kappa\tau(\|u_e(t_{n+1})\|+\|u_e(t_n)\|)(\|e_u^n\|+|e_s^n|)\|\delta_t e_u^{n+1}\| \\
&\qquad\le C_1\kappa^2\tau(\|e_u^n\|^2+|e_s^n|^2)+\frac{\tau}{2}\|\delta_t e_u^{n+1}\|^2,
\end{aligned}\tag{3.24}$$

where $C_1>0$ depends on C_*, $|\Omega|$, u_e, and $\|f\|_{C^1[-\beta,\beta]}$. Using the Young's inequality, the third and fourth terms in the right-hand side of (3.22) can be bounded respectively as

$$4G_*\kappa\tau\langle e_u^n,\delta_t e_u^{n+1}\rangle \le 4G_*\kappa\tau\|e_u^n\|\|\delta_t e_u^{n+1}\| \le 8G_*^2\kappa^2\tau\|e_u^n\|^2+\frac{\tau}{2}\|\delta_t e_u^{n+1}\|^2,\tag{3.25}$$

$$-2\tau\langle R_{1u}^n,\delta_t e_u^{n+1}\rangle \le 2\tau\|R_{1u}^n\|\|\delta_t e_u^{n+1}\| \le 4\tau\|R_{1u}^n\|^2+\frac{\tau}{4}\|\delta_t e_u^{n+1}\|^2.\tag{3.26}$$

Then, substituting (3.23)–(3.26) into (3.22) leads to

$$\begin{aligned}
&2G_*\kappa\|e_u^{n+1}\|^2-2G_*\kappa\|e_u^n\|^2+\varepsilon^2\|\nabla_h e_u^{n+1}\|^2-\varepsilon^2\|\nabla_h e_u^n\|^2+\frac{\tau}{4}\|\delta_t e_u^{n+1}\|^2 \\
&\qquad\le (4C_g^2+C_1\kappa^2+8G_*^2\kappa^2)\tau\|e_u^n\|^2+(4C_g^2+C_1\kappa^2)\tau|e_s^n|^2+4\tau\|R_{1u}^n\|^2.
\end{aligned}\tag{3.27}$$

Multiplying (3.20b) by $2\tau e_s^{n+1}$ yields

$$\begin{aligned}
&|e_s^{n+1}|^2-|e_s^n|^2+|e_s^{n+1}-e_s^n|^2 \\
&\qquad= 2e_s^{n+1}\langle g(u_e(t_n),s_e(t_n))f(u_e(t_n))-g(u^n,s^n)f(u^n),u_e(t_{n+1})-u_e(t_n)\rangle \\
&\qquad\quad -2\tau e_s^{n+1}g(u^n,s^n)\langle f(u^n),\delta_t e_u^{n+1}\rangle-2\tau R_{1s}^n e_s^{n+1}.
\end{aligned}\tag{3.28}$$

For the first term in the right-hand side of (3.28), by the Schwartz's inequality and Lemma 4, we have

$$\begin{aligned}
&2e_s^{n+1}\langle g(u_e(t_n),s_e(t_n))f(u_e(t_n))-g(u^n,s^n)f(u^n),u_e(t_{n+1})-u_e(t_n)\rangle \\
&\qquad\le 2|e_s^{n+1}|\|g(u_e(t_n),s_e(t_n))f(u_e(t_n))-g(u^n,s^n)f(u^n)\|\|u_e(t_{n+1})-u_e(t_n)\| \\
&\qquad\le 2C_g\tau|e_s^{n+1}|(\|e_u^n\|+|e_s^n|)\|(u_e)_t(\theta_n)\| \qquad \text{(for some } t_n<\theta_n<t_{n+1}) \\
&\qquad= 2C_g\tau\|(u_e)_t(\theta_n)\|(\|e_u^n\||e_s^{n+1}|+|e_s^n||e_s^{n+1}|) \\
&\qquad\le C_2\tau(\|e_u^n\|^2+|e_s^n|^2+|e_s^{n+1}|^2),
\end{aligned}\tag{3.29}$$

where $C_2>0$ depends on C_*, $|\Omega|$, u_e, and $\|f\|_{C^1[-\beta,\beta]}$. For the second term in the right-hand side of (3.28), using Corollary 2, we obtain

$$\begin{aligned}
-2\tau e_s^{n+1}g(u^n,s^n)\langle f(u^n),\delta_t e_u^{n+1}\rangle &\le 2G^*\tau\|f(u^n)\||e_s^{n+1}|\|\delta_t e_u^{n+1}\| \\
&\le C_3\tau|e_s^{n+1}|^2+\frac{\tau}{4}\|\delta_t e_u^{n+1}\|^2,
\end{aligned}\tag{3.30}$$

where $C_3>0$ depends on C_*, $|\Omega|$, u_{init}, and $\|f\|_{C[-\beta,\beta]}$. For the third term in the right-hand side of (3.28), we have

$$-2\tau R_{1s}^n e_s^{n+1} \le \tau|R_{1s}^n|^2+\tau|e_s^{n+1}|^2.\tag{3.31}$$

Then, substituting (3.29)–(3.31) into (3.28) leads to

$$|e_s^{n+1}|^2-|e_s^n|^2 \le C_2\tau\|e_u^n\|^2+C_2\tau|e_s^n|^2+(1+C_2+C_3)\tau|e_s^{n+1}|^2+\frac{\tau}{4}\|\delta_t e_u^{n+1}\|^2+\tau|R_{1s}^n|^2.\tag{3.32}$$

Adding (3.27) and (3.32), we obtain

$$\begin{aligned}
&2G_*\kappa(\|e_u^{n+1}\|^2-\|e_u^n\|^2)+\varepsilon^2(\|\nabla_h e_u^{n+1}\|^2-\|\nabla_h e_u^n\|^2)+(|e_s^{n+1}|^2-|e_s^n|^2) \\
&\qquad\le (4C_g^2+C_1\kappa^2+8G_*^2\kappa^2+C_2)\tau\|e_u^n\|^2+(4C_g^2+C_1\kappa^2+C_2)\tau|e_s^n|^2 \\
&\qquad\quad +(1+C_2+C_3)\tau|e_s^{n+1}|^2+4\tau\|R_{1u}^n\|^2+\tau|R_{1s}^n|^2.
\end{aligned}$$

Then, using (3.16), we reach

$$
\begin{aligned}
2G_*\kappa(\|e_u^{n+1}\|^2 - \|e_u^n\|^2) + \varepsilon^2(\|\nabla_h e_u^{n+1}\|^2 - \|\nabla_h e_u^n\|^2) + (|e_s^{n+1}|^2 - |e_s^n|^2)& \\
\le C_4\tau(\|e_u^n\|^2 + |e_s^n|^2 + |e_s^{n+1}|^2) + 5C_e^2\tau(\tau+h^2)^2,&
\end{aligned}
$$

where the constant C_4 depends on C_*, $|\Omega|$, T, u_e, κ, ε, and $\|f\|_{C^1[-\beta,\beta]}$.

Letting $W^n := 2G_*\kappa\|e_u^n\|^2 + \varepsilon^2\|\nabla_h e_u^n\|^2 + |e_s^n|^2$, we have

$$
W^{n+1} - W^n \le \widetilde{C}_4\tau(W^n + W^{n+1}) + 5C_e^2\tau(\tau+h^2)^2,
$$

where $\widetilde{C}_4$ depends on C_4 and κ. When $\tau \le \frac{1}{2\widetilde{C}_4}$, noting that $\frac{1+\widetilde{C}_4\tau}{1-\widetilde{C}_4\tau} \le 1 + 4\widetilde{C}_4\tau$, we obtain

$$
W^{n+1} \le (1+4\widetilde{C}_4\tau)W^n + 10C_e^2\tau(\tau+h^2)^2.
$$

Using the discrete Gronwall's inequality, we obtain

$$
2G_*\kappa\|e_u^n\|^2 + \varepsilon^2\|\nabla_h e_u^n\|^2 + |e_s^n|^2 = W^n \le 10C_e^2 \mathrm{e}^{4\widetilde{C}_4 T}(\tau+h^2)^2,
$$

which completes the proof.

Remark 4 For any fixed $h>0$, let us recall the space-discrete problem (3.3) and denote by $u_{h,e}(t)$ the exact solution. By similar analysis as Theorem 3, one can obtain the error estimates for sufficiently small τ as follows:

$$
\|u^n - u_{h,e}(t_n)\| + \|\nabla_h u^n - \nabla_h u_{h,e}(t_n)\| + |s^n - E_{1h}(u_{h,e}(t))| \le C_h\tau,
$$

where the constant $C_h > 0$ depends on C_*, $|\Omega|$, T, $u_{h,e}$, κ, ε, and $\|f\|_{C^1[-\beta,\beta]}$ but is independent of τ.

3.2 Second-order sESAV scheme

For the space-discrete system (3.3), the second-order stabilized ESAV scheme (sESAV2) is given by

$$
\delta_t u^{n+1} = \varepsilon^2\Delta_h u^{n+\frac12} + g(\widehat{u}^{n+\frac12},\widehat{s}^{n+\frac12})f(\widehat{u}^{n+\frac12}) - \kappa g(\widehat{u}^{n+\frac12},\widehat{s}^{n+\frac12})(u^{n+\frac12} - \widehat{u}^{n+\frac12}), \tag{3.33a}
$$

$$
\delta_t s^{n+1} = -g(\widehat{u}^{n+\frac12},\widehat{s}^{n+\frac12})\langle f(\widehat{u}^{n+\frac12}), \delta_t u^{n+1}\rangle + \kappa g(\widehat{u}^{n+\frac12},\widehat{s}^{n+\frac12})\langle u^{n+\frac12} - \widehat{u}^{n+\frac12}, \delta_t u^{n+1}\rangle, \tag{3.33b}
$$

where $g(\widehat{u}^{n+\frac12},\widehat{s}^{n+\frac12}) > 0$ with $(\widehat{u}^{n+\frac12},\widehat{s}^{n+\frac12})$ being generated by the first-order scheme (3.5) with the time step size $\tau/2$, i.e.,

$$
\frac{\widehat{u}^{n+\frac12} - u^n}{\tau/2} = \varepsilon^2\Delta_h\widehat{u}^{n+\frac12} + g(u^n,s^n)f(u^n) - \kappa g(u^n,s^n)(\widehat{u}^{n+\frac12} - u^n), \tag{3.34a}
$$

$$
\widehat{s}^{n+\frac12} - s^n = -g(u^n,s^n)\langle f(u^n), \widehat{u}^{n+\frac12} - u^n\rangle. \tag{3.34b}
$$

The scheme (3.33) is started by $u^0 = u_{\text{init}}$ and $s^0 = E_{1h}(u^0)$. By the definition of $u^{n+\frac12}$, the last term in (3.33a) is actually $-\frac12\kappa g(\widehat{u}^{n+\frac12},\widehat{s}^{n+\frac12})(u^{n+1} - 2\widehat{u}^{n+\frac12} + u^n)$, which provides a second-order truncation error in time. We can rewrite (3.33) in the following form:

$$
\begin{aligned}
&\Big[\Big(\frac{2}{\tau} + \kappa g(\widehat{u}^{n+\frac12},\widehat{s}^{n+\frac12})\Big)I - \varepsilon^2\Delta_h\Big]u^{n+1} \\
&\quad = \Big[\Big(\frac{2}{\tau} - \kappa g(\widehat{u}^{n+\frac12},\widehat{s}^{n+\frac12})\Big)I + \varepsilon^2\Delta_h\Big]u^n + 2g(\widehat{u}^{n+\frac12},\widehat{s}^{n+\frac12})[f(\widehat{u}^{n+\frac12}) + \kappa\widehat{u}^{n+\frac12}],
\end{aligned} \tag{3.35a}
$$

$$
s^{n+1} = s^n - g(\widehat{u}^{n+\frac12},\widehat{s}^{n+\frac12})\langle f(\widehat{u}^{n+\frac12}) - \kappa(u^{n+\frac12} - \widehat{u}^{n+\frac12}), u^{n+1} - u^n\rangle. \tag{3.35b}
$$

It is then easy to see that the system (3.35) is linear and uniquely solvable for any $\tau > 0$ since $(\frac{2}{\tau} + \kappa g(\widehat{u}^{n+\frac12},\widehat{s}^{n+\frac12}))I - \varepsilon^2\Delta_h$ is self-adjoint and positive definite.

3.2.1 Energy dissipation and MBP

The energy dissipation law and the MBP preservation of the sESAV2 scheme (3.33) are stated below.

Theorem 4 (Energy dissipation of sESAV2) *For any $\kappa \ge 0$ and $\tau > 0$, the sESAV2 scheme* (3.33) *is energy dissipative in the sense that $\mathcal{E}_h(u^{n+1}, s^{n+1}) \le \mathcal{E}_h(u^n, s^n)$, where $\mathcal{E}_h(u^n, s^n)$ is given by* (3.7). *Moreover, it holds that $s^n \le E_h(u_{\text{init}})$ and $\widehat{s}^{n+\frac{1}{2}} \le E_h(u_{\text{init}})$ for all n.*

Proof Taking the inner product of (3.33) with $u^{n+1} - u^n$ yields

$$\frac{1}{\tau}\|u^{n+1} - u^n\|^2 = -\frac{\varepsilon^2}{2}\|\nabla_h u^{n+1}\|^2 + \frac{\varepsilon^2}{2}\|\nabla_h u^n\|^2 + g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})\langle f(\widehat{u}^{n+\frac{1}{2}}), u^{n+1} - u^n\rangle \\ - \kappa g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})\langle u^{n+\frac{1}{2}} - \widehat{u}^{n+\frac{1}{2}}, u^{n+1} - u^n\rangle. \tag{3.36}$$

Combining (3.36) and (3.33b), we obtain

$$\mathcal{E}_h(u^{n+1}, s^{n+1}) - \mathcal{E}_h(u^n, s^n) = -\frac{1}{\tau}\|u^{n+1} - u^n\|^2 \le 0.$$

Similar to the proof of Corollary 1, the uniform upper boundedness of $\{s^n\}$ is a direct result of the energy stability. Since $\widehat{s}^{n+\frac{1}{2}}$ is generated by the sESAV1 scheme (3.34), according to Theorem 1, we have $\widehat{s}^{n+\frac{1}{2}} \le \mathcal{E}_h(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \le \mathcal{E}_h(u^n, s^n)$, and thus, we also have $\widehat{s}^{n+\frac{1}{2}} \le E_h(u_{\text{init}})$.

Theorem 4 means that the sESAV2 scheme (3.33) is energy dissipative with respect to the modified energy $\mathcal{E}_h(u^n, s^n)$. Furthermore, $\mathcal{E}_h(u^n, s^n)$ is always bounded from below as a direct result of the lower boundedness of s^n and $\widehat{s}^{n+\frac{1}{2}}$ shown later in the proof of Corollary 4.

Theorem 5 (MBP of sESAV2) *If $h \le h_0$, $\kappa \ge \|f'\|_{C[-\beta,\beta]}$, and*

$$\tau \le \Big(\frac{\kappa G^*}{2} + \frac{\varepsilon^2}{h^2}\Big)^{-1}, \tag{3.37}$$

where G^* is the positive constant defined in Corollary 2, then the sESAV2 scheme (3.33) *preserves the MBP for $\{u^n\}$, i.e.,* (3.9) *is valid.*

Proof Suppose (u^n, s^n) is given and $\|u^n\|_\infty \le \beta$ for some n. By Theorems 2 and 4, we have $\|\widehat{u}^{n+\frac{1}{2}}\|_\infty \le \beta$ and $\widehat{s}^{n+\frac{1}{2}} \le E_h(u_{\text{init}})$. Then, we know that $0 < g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \le G^*$ by the similar analysis as Corollary 2. The condition (3.37) implies

$$\frac{2}{\tau} - \kappa g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \ge \frac{2\varepsilon^2}{h^2}.$$

According to the definition of the matrix ∞-norm, we have

$$\Big\|\Big(\frac{2}{\tau} - \kappa g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})\Big)I + \varepsilon^2 \Delta_h\Big\|_\infty = \frac{2}{\tau} - \kappa g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}).$$

Since $\kappa \ge \|f'\|_{C[-\beta,\beta]}$ and $\|\widehat{u}^{n+\frac{1}{2}}\|_\infty \le \beta$, according to Lemma 2, we have

$$\|f(\widehat{u}^{n+\frac{1}{2}}) + \kappa \widehat{u}^{n+\frac{1}{2}}\|_\infty \le \kappa\beta.$$

Therefore, using Lemma 1, we obtain from (3.35a) that

$$\|u^{n+1}\|_\infty \le \Big(\frac{2}{\tau} + \kappa g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})\Big)^{-1}\Big[\Big(\frac{2}{\tau} - \kappa g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})\Big)\beta + 2\kappa g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})\beta\Big] = \beta.$$

By induction, we have $\|u^n\|_\infty \le \beta$ for all n.

Remark 5 Theorem 5 implies that $0 < g(u^n, s^n) \le G^*$ and $0 < g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \le G^*$ hold for all n.

Remark 6 The condition (3.37) on the time step size implies $\tau = \mathcal{O}(h^2/\varepsilon^2)$, which is the same as those enforced in [16,17]. This restriction comes essentially from the explicit term $\Delta_h u^n$ due to the use of the Crank–Nicolson approximation, which also means that the second-order scheme (3.33) cannot preserve the MBP unconditionally even though we introduce the stabilization term. In practical computations, $g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \approx 1$ so that the requirement for the time step size can be set to be $\tau \le (\frac{\kappa}{2} + \frac{\varepsilon^2}{h^2})^{-1}$ in order to preserve the MBP, which is later used in our numerical experiments.

Similar to the analysis for the sESAV1 scheme, we can show that both $\{g(u^n, s^n)\}$ and $\{g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})\}$ have uniform positive lower bounds.

Lemma 5 *Given a fixed time $T > 0$. If $h \le h_0$, $\kappa \ge \|f'\|_{C[-\beta,\beta]}$, $\|u_{\text{init}}\|_\infty \le \beta$, and $\tau \le 1$ satisfying* (3.37), *there exists a constant $M > 0$ depending on C_*, $|\Omega|$, T, u_{init}, κ, ε, and $\|f\|_{C^1[-\beta,\beta]}$ such that*

$$\tau^{-1}\|\widehat{u}^{n+\frac{1}{2}} - u^n\| + \|\Delta_h \widehat{u}^{n+\frac{1}{2}}\| \le M,$$
$$\tau^{-1}\|u^{n+1} - u^n\| + \|\Delta_h u^{n+1}\| \le M,$$

for $0 \le n \le \lfloor T/\tau \rfloor - 1$.

Proof Since $\widehat{u}^{n+\frac{1}{2}}$ is the solution to the sESAV1 substep (3.34), according to (3.13), we have

$$\|\Delta_h \widehat{u}^{n+\frac{1}{2}}\|^2 \le \Big(1 + \frac{G^*\kappa}{2} + \frac{(G^*\|f'\|_{C[-\beta,\beta]}C_\Omega)^2}{4\varepsilon^2}\Big)\|\Delta_h u^n\|^2, \tag{3.38}$$

where we used $\tau \le 1$.

Taking the discrete inner product of (3.33a) with $2\tau\Delta_h^2 u^{n+\frac{1}{2}}$, using the fact $0 < g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \le G^*$, and conducting the similar analysis as the proof of Lemma 3, we can obtain

$$\|\Delta_h u^{n+1}\|^2 \le \|\Delta_h u^n\|^2 + \Big(G^*\kappa + \frac{(G^*\|f'\|_{C[-\beta,\beta]}C_\Omega)^2}{2\varepsilon^2}\Big)\tau\|\Delta_h \widehat{u}^{n+\frac{1}{2}}\|^2.$$

Substituting (3.38) into the above inequality, we have

$$\begin{aligned}\|\Delta_h u^{n+1}\|^2 \le \Big[1 + \Big(G^*\kappa + \frac{(G^*\|f'\|_{C[-\beta,\beta]}C_\Omega)^2}{2\varepsilon^2}\Big) \\ \cdot \Big(1 + \frac{G^*\kappa}{2} + \frac{(C_\Omega\|f'\|_{C[-\beta,\beta]}G^*)^2}{4\varepsilon^2}\Big)\tau\Big]\|\Delta_h u^n\|^2.\end{aligned}$$

By recursion, we can obtain a uniform upper bound for $\|\Delta_h u^{n+1}\|$. Then, by (3.38) we also can get the upper bound for $\|\Delta_h \widehat{u}^{n+\frac{1}{2}}\|$.

Finally, as a consequence of the above analysis, using (3.34a) and (3.33a), we also get the boundedness of $\tau^{-1}\|\widehat{u}^{n+\frac{1}{2}} - u^n\|$ and $\tau^{-1}\|u^{n+1} - u^n\|$, and the proof is completed.

Corollary 4 *Given a fixed time $T > 0$. If $h \le h_0$, $\kappa \ge \|f'\|_{C[-\beta,\beta]}$, $\|u_{\text{init}}\|_\infty \le \beta$, and $\tau \le 1$ satisfying* (3.37), *there exists a constant $\widetilde{G}_* > 0$ such that $g(u^n, s^n) \ge \widetilde{G}_*$ and $g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \ge \widetilde{G}_*$, where $\widetilde{G}_*$ depends on C_*, $|\Omega|$, T, u_{init}, κ, ε, and $\|f\|_{C^1[-\beta,\beta]}$.*

Proof It suffices to show the existence of the lower bounds of $\{s^n\}$ and $\{\widehat{s}^{n+\frac{1}{2}}\}$. By Lemma 5, we have $\|u^{n+1} - u^n\| \le M\tau$. From (3.33b), the similar analysis as Corollary 3 leads to the lower boundedness of $\{s^n\}$. Then, since $\|\widehat{u}^{n+\frac{1}{2}} - u^n\| \le M\tau \le M$, we can derive from (3.34b) to give

$$\widehat{s}^{n+\frac{1}{2}} \ge s^n - G^* F_0 |\Omega|^{\frac{1}{2}} M,$$

which completes the proof.

3.2.2 Error estimates

It is easy to check that the exact solution u_e to (1.1) with $s_e(t) = E_1(u_e(t))$ satisfies

$$\begin{aligned}\frac{u_e(t_{n+1}) - u_e(t_n)}{\tau} &= \frac{\varepsilon^2}{2}\Delta_h(u_e(t_{n+1}) + u_e(t_n)) + g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}}))f(u_e(t_{n+\frac{1}{2}})) \\ &\quad - \kappa g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}}))\Big(\frac{u_e(t_{n+1}) + u_e(t_n)}{2} - u_e(t_{n+\frac{1}{2}})\Big) + R_{2u}^n,\end{aligned} \tag{3.39a}$$

$$\begin{aligned}\frac{s_e(t_{n+1}) - s_e(t_n)}{\tau} &= -g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}}))\Big\langle f(u_e(t_{n+\frac{1}{2}})), \frac{u_e(t_{n+1}) - u_e(t_n)}{\tau}\Big\rangle \\ &\quad + \kappa g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}}))\Big\langle \frac{u_e(t_{n+1}) + u_e(t_n)}{2} - u_e(t_{n+\frac{1}{2}}), \frac{u_e(t_{n+1}) - u_e(t_n)}{\tau}\Big\rangle + R_{2s}^n,\end{aligned} \tag{3.39b}$$

where the truncation errors R_{2u}^n and R_{2s}^n satisfy

$$\|R_{2u}^n\| \le C_e(\tau^2+h^2), \qquad |R_{2s}^n| \le C_e(\tau^2+h^2). \tag{3.40}$$

Apart from the numerical error functions e_u^n and e_s^n defined by (3.17), let us also define

$$\widehat{e}_u^{n+\frac{1}{2}} = \widehat{u}^{n+\frac{1}{2}} - u_e(t_{n+\frac{1}{2}}), \qquad \widehat{e}_s^{n+\frac{1}{2}} = \widehat{s}^{n+\frac{1}{2}} - s_e(t_{n+\frac{1}{2}}).$$

We first present an estimate for $\widehat{e}_u^{n+\frac{1}{2}}$ and $\widehat{e}_s^{n+\frac{1}{2}}$, which will be used in the proof of the error estimate for the sESAV2 scheme (3.33). Recalling the proof of Theorem 3 for the sESAV1 scheme, the error equations with respect to $\widehat{e}_u^{n+\frac{1}{2}}$ and $\widehat{e}_s^{n+\frac{1}{2}}$ read as

$$\begin{aligned}\frac{\widehat{e}_u^{n+\frac{1}{2}} - e_u^n}{\tau/2} &= \varepsilon^2 \Delta_h \widehat{e}_u^{n+\frac{1}{2}} + g(u^n,s^n)f(u^n) - g(u_e(t_n),s_e(t_n))f(u_e(t_n)) - \kappa g(u^n,s^n)(\widehat{e}_u^{n+\frac{1}{2}} - e_u^n)\\ &\quad + \kappa(g(u_e(t_n),s_e(t_n)) - g(u^n,s^n))(u_e(t_{n+\frac{1}{2}}) - u_e(t_n)) - \widehat{R}_{1u}^n, \end{aligned}\tag{3.41a}$$

$$\begin{aligned}\widehat{e}_s^{n+\frac{1}{2}} - e_s^n &= \langle g(u_e(t_n),s_e(t_n))f(u_e(t_n)) - g(u^n,s^n)f(u^n), u_e(t_{n+\frac{1}{2}}) - u_e(t_n)\rangle\\ &\quad - g(u^n,s^n)\langle f(u^n), \widehat{e}_u^{n+\frac{1}{2}} - e_u^n\rangle - \frac{\tau}{2}\widehat{R}_{1s}^n,\end{aligned}\tag{3.41b}$$

where

$$\|\widehat{R}_{1u}^n\| \le C_e(\tau+h^2), \qquad |\widehat{R}_{1s}^n| \le C_e(\tau+h^2). \tag{3.42}$$

Lemma 6 *Suppose that $h \le h_0$ and $\|u^n\|_\infty \le \beta$. If τ is small sufficiently, we have*

$$\|\widehat{e}_u^{n+\frac{1}{2}}\|^2 + |\widehat{e}_s^{n+\frac{1}{2}}|^2 \le \widehat{C}(\|e_u^n\|^2 + |e_s^n|^2) + \widehat{C}C_e^2(\tau^2+\tau h^2)^2, \tag{3.43}$$

where the constant $\widehat{C} > 0$ depends on C_*, $|\Omega|$, u_e, κ, and $\|f\|_{C^1[-\beta,\beta]}$.

Proof Taking the discrete inner product of (3.41a) with $\tau\widehat{e}_u^{n+\frac{1}{2}}$ and rearranging the terms, we have

$$\begin{aligned}&\Big(1 + \frac{\kappa g(u^n,s^n)}{2}\tau\Big)\big(\|\widehat{e}_u^{n+\frac{1}{2}}\|^2 - \|e_u^n\|^2 + \|\widehat{e}_u^{n+\frac{1}{2}} - e_u^n\|^2\big) + \varepsilon^2\tau\|\nabla_h \widehat{e}_u^{n+\frac{1}{2}}\|^2\\ &\quad = \tau\langle g(u^n,s^n)f(u^n) - g(u_e(t_n),s_e(t_n))f(u_e(t_n)), \widehat{e}_u^{n+\frac{1}{2}}\rangle\\ &\qquad + \kappa\tau\langle (g(u_e(t_n),s_e(t_n)) - g(u^n,s^n))(u_e(t_{n+\frac{1}{2}}) - u_e(t_n)), \widehat{e}_u^{n+\frac{1}{2}}\rangle - \tau\langle \widehat{R}_{1u}^n, \widehat{e}_u^{n+\frac{1}{2}}\rangle.\end{aligned}$$

Using Young's inequality, we then get

$$-\tau\langle \widehat{R}_{1u}^n, \widehat{e}_u^{n+\frac{1}{2}}\rangle \le \tau\|\widehat{R}_{1u}^n\|\|\widehat{e}_u^{n+\frac{1}{2}}\| \le \tau^2\|\widehat{R}_{1u}^n\|^2 + \frac{1}{4}\|\widehat{e}_u^{n+\frac{1}{2}}\|^2.$$

Similar to the deductions of (3.23) and (3.24), applying Lemma 4 leads to

$$\tau\langle g(u^n,s^n)f(u^n) - g(u_e(t_n),s_e(t_n))f(u_e(t_n)), \widehat{e}_u^{n+\frac{1}{2}}\rangle \le 2C_g^2\tau^2(\|e_u^n\|^2 + |e_s^n|^2) + \frac{1}{4}\|\widehat{e}_u^{n+\frac{1}{2}}\|^2,$$

$$\kappa\tau\langle (g(u_e(t_n),s_e(t_n)) - g(u^n,s^n))(u_e(t_{n+\frac{1}{2}}) - u_e(t_n)), \widehat{e}_u^{n+\frac{1}{2}}\rangle \le C_1\kappa^2\tau^2(\|e_u^n\|^2 + |e_s^n|^2) + \frac{1}{4}\|\widehat{e}_u^{n+\frac{1}{2}}\|^2.$$

Then, we have

$$\begin{aligned}&\Big(\frac{1}{4} + \frac{\kappa g(u^n,s^n)}{2}\tau\Big)\|\widehat{e}_u^{n+\frac{1}{2}}\|^2 + \Big(1 + \frac{\kappa g(u^n,s^n)}{2}\tau\Big)\|\widehat{e}_u^{n+\frac{1}{2}} - e_u^n\|^2\\ &\quad \le \Big(1 + \frac{\kappa g(u^n,s^n)}{2}\tau\Big)\|e_u^n\|^2 + (2C_g^2 + C_1\kappa^2)\tau^2(\|e_u^n\|^2 + |e_s^n|^2) + \tau^2\|\widehat{R}_{1u}^n\|^2.\end{aligned}$$

By Remark 5, we then can simplify the above equation to get

$$\|\widehat{e}_u^{n+\frac{1}{2}}\|^2 + 4\|\widehat{e}_u^{n+\frac{1}{2}} - e_u^n\|^2 \le (4 + 2G^*\kappa\tau)\|e_u^n\|^2 + (8C_g^2 + 4C_1\kappa^2)\tau^2(\|e_u^n\|^2 + |e_s^n|^2) + 4\tau^2\|\widehat{R}_{1u}^n\|^2.$$

When $\tau \le 1$, using (3.42), we obtain

$$\begin{aligned}\|\widehat{e}_u^{n+\frac{1}{2}}\|^2 + 4\|\widehat{e}_u^{n+\frac{1}{2}} - e_u^n\|^2 &\le (4 + 8C_g^2 + 2G^*\kappa + 4C_1\kappa^2)\|e_u^n\|^2\\ &\quad + (8C_g^2 + 4C_1\kappa^2)|e_s^n|^2 + 4C_e^2\tau^2(\tau+h^2)^2.\end{aligned}\tag{3.44}$$

Multiplying (3.41b) by $2\widehat{e}_s^{n+\frac{1}{2}}$ yields

$$
\begin{aligned}
&|\widehat{e}_s^{n+\frac{1}{2}}|^2 - |e_s^n|^2 + |\widehat{e}_s^{n+\frac{1}{2}} - e_s^n|^2 \\
&\quad = 2\widehat{e}_s^{n+\frac{1}{2}} \langle g(u_e(t_n), s_e(t_n)) f(u_e(t_n)) - g(u^n, s^n) f(u^n), u_e(t_{n+\frac{1}{2}}) - u_e(t_n)\rangle \\
&\quad\quad - 2\widehat{e}_s^{n+\frac{1}{2}} g(u^n, s^n) \langle f(u^n), \widehat{e}_u^{n+\frac{1}{2}} - e_u^n\rangle - \tau \widehat{R}_{1s}^n \widehat{e}_s^{n+\frac{1}{2}}.
\end{aligned}
$$

We can bound the third and second terms in the right-hand side of the above equation repectively as

$$
\begin{aligned}
-\tau \widehat{R}_{1s}^n \widehat{e}_s^{n+\frac{1}{2}} &\le \tau^2 |\widehat{R}_{1s}^n|^2 + \frac{1}{4} |\widehat{e}_s^{n+\frac{1}{2}}|^2, \\
-2\widehat{e}_s^{n+\frac{1}{2}} g(u^n, s^n) \langle f(u^n), \widehat{e}_u^{n+\frac{1}{2}} - e_u^n\rangle &\le \frac{1}{4} |\widehat{e}_s^{n+\frac{1}{2}}|^2 + C_5 \|\widehat{e}_u^{n+\frac{1}{2}} - e_u^n\|^2,
\end{aligned}
$$

where $C_5 > 0$ depends on C_*, $|\Omega|$, u_{init}, and $\|f\|_{C[-\beta,\beta]}$. By estimating the first term in the similar way as (3.29), we obtain

$$
\begin{aligned}
&|\widehat{e}_s^{n+\frac{1}{2}}|^2 - |e_s^n|^2 + |\widehat{e}_s^{n+\frac{1}{2}} - e_s^n|^2 \\
&\quad \le C_2 \tau (\|e_u^n\|^2 + |e_s^n|^2 + |\widehat{e}_s^{n+\frac{1}{2}}|^2) + C_5 \|\widehat{e}_u^{n+\frac{1}{2}} - e_u^n\|^2 + \frac{1}{2} |\widehat{e}_s^{n+\frac{1}{2}}|^2 + \tau^2 |\widehat{R}_{1s}^n|^2,
\end{aligned}
$$

and thus,

$$
(1 - 2C_2\tau) |\widehat{e}_s^{n+\frac{1}{2}}|^2 \le 2|e_s^n|^2 + 2C_2\tau (\|e_u^n\|^2 + |e_s^n|^2) + 2C_5 \|\widehat{e}_u^{n+\frac{1}{2}} - e_u^n\|^2 + 2\tau^2 |\widehat{R}_{1s}^n|^2.
$$

When $\tau \le \frac{1}{4C_2}$, using (3.42), we can get

$$
|\widehat{e}_s^{n+\frac{1}{2}}|^2 \le \|e_u^n\|^2 + 5|e_s^n|^2 + 4C_5 \|\widehat{e}_u^{n+\frac{1}{2}} - e_u^n\|^2 + 4C_e^2 \tau^2 (\tau + h^2)^2. \tag{3.45}
$$

The sum of (3.44) multiplied by C_5 and (3.45) leads to (3.43).

Theorem 6 (Error estimate of sESAV2) *Given a fixed time $T > 0$ and suppose the exact solution u_e is smooth enough on $[0, T] \times \overline{\Omega}$. Assume that $\kappa \ge \|f'\|_{C[-\beta,\beta]}$ and $\|u_{\text{init}}\|_\infty \le \beta$. If τ and h are small sufficiently and satisfy (3.37), then we have the error estimate for the sESAV2 scheme (3.33) as follows:*

$$
\|e_u^n\| + \|\nabla_h e_u^n\| + |e_s^n| \le C(\tau^2 + h^2), \qquad 0 \le n \le \lfloor T/\tau \rfloor,
$$

where the constant $C > 0$ depends on C_*, $|\Omega|$, T, u_e, κ, ε, and $\|f\|_{C^1[-\beta,\beta]}$ but is independent of τ and h.

Proof The difference between (3.33) and (3.39) leads to

$$
\begin{aligned}
\delta_t e_u^{n+1} &= \varepsilon^2 \Delta_h e_u^{n+\frac{1}{2}} + g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) f(\widehat{u}^{n+\frac{1}{2}}) - g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}})) f(u_e(t_{n+\frac{1}{2}})) \\
&\quad + \kappa \big(g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}})) - g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \big) \Big(\frac{u_e(t_{n+1}) + u_e(t_n)}{2} - u_e(t_{n+\frac{1}{2}}) \Big) \\
&\quad - \kappa g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) (e_u^{n+\frac{1}{2}} - \widehat{e}_u^{n+\frac{1}{2}}) - R_{2u}^n,
\end{aligned} \tag{3.46a}
$$

$$
\begin{aligned}
\delta_t e_s^{n+1} &= \Big\langle g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}})) f(u_e(t_{n+\frac{1}{2}})) - g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) f(\widehat{u}^{n+\frac{1}{2}}), \frac{u_e(t_{n+1}) - u_e(t_n)}{\tau} \Big\rangle \\
&\quad + \kappa \big(g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) - g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}})) \big) \Big\langle \frac{u_e(t_{n+1}) + u_e(t_n)}{2} - u_e(t_{n+\frac{1}{2}}), \frac{u_e(t_{n+1}) - u_e(t_n)}{\tau} \Big\rangle \\
&\quad - g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \langle f(\widehat{u}^{n+\frac{1}{2}}), \delta_t e_u^{n+1}\rangle + \kappa g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \langle u^{n+\frac{1}{2}} - \widehat{u}^{n+\frac{1}{2}}, \delta_t e_u^{n+1}\rangle \\
&\quad + \kappa g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \Big\langle e_u^{n+\frac{1}{2}} - \widehat{e}_u^{n+\frac{1}{2}}, \frac{u_e(t_{n+1}) - u_e(t_n)}{\tau} \Big\rangle - R_{2s}^n.
\end{aligned} \tag{3.46b}
$$

Taking the discrete inner product of (3.46a) with $2\tau\delta_t e_u^{n+1}$ and rearranging the term yield

$$
\begin{aligned}
&\varepsilon^2 \|\nabla_h e_u^{n+1}\|^2 - \varepsilon^2 \|\nabla_h e_u^n\|^2 + 2\tau \|\delta_t e_u^{n+1}\|^2 \\
&\quad = 2\tau \langle g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) f(\widehat{u}^{n+\frac{1}{2}}) - g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}})) f(u_e(t_{n+\frac{1}{2}})), \delta_t e_u^{n+1}\rangle \\
&\quad\quad + 2\kappa\tau \big(g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}})) - g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \big) \Big\langle \frac{u_e(t_{n+1}) + u_e(t_n)}{2} - u_e(t_{n+\frac{1}{2}}), \delta_t e_u^{n+1} \Big\rangle \\
&\quad\quad - 2\kappa\tau g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) \langle e_u^{n+\frac{1}{2}} - \widehat{e}_u^{n+\frac{1}{2}}, \delta_t e_u^{n+1}\rangle - 2\tau \langle R_{2u}^n, \delta_t e_u^{n+1}\rangle.
\end{aligned} \tag{3.47}
$$

Since $g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}}) \geq \widetilde{G}_* > 0$, we get

$$
\begin{aligned}
&2\kappa\tau g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})\langle e_u^{n+\frac{1}{2}} - \widehat{e}_u^{n+\frac{1}{2}}, \delta_t e_u^{n+1}\rangle \\
&\quad = 2\kappa\tau g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})\Big\langle \frac{e_u^{n+1} - e_u^n}{2} + e_u^n - \widehat{e}_u^{n+\frac{1}{2}}, \delta_t e_u^{n+1}\Big\rangle \\
&\quad = \kappa g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})\|e_u^{n+1} - e_u^n\|^2 + 2\kappa\tau g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})\langle e_u^n - \widehat{e}_u^{n+\frac{1}{2}}, \delta_t e_u^{n+1}\rangle \\
&\quad \geq \kappa\widetilde{G}_*\|e_u^{n+1} - e_u^n\|^2 + 2\kappa\tau g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})\langle e_u^n - \widehat{e}_u^{n+\frac{1}{2}}, \delta_t e_u^{n+1}\rangle \\
&\quad = \kappa\widetilde{G}_*\|e_u^{n+1}\|^2 - \kappa\widetilde{G}_*\|e_u^n\|^2 - 2\kappa\widetilde{G}_*\tau\langle e_u^n, \delta_t e_u^{n+1}\rangle + 2\kappa\tau g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})\langle e_u^n - \widehat{e}_u^{n+\frac{1}{2}}, \delta_t e_u^{n+1}\rangle,
\end{aligned}
$$

where we have used (3.21) in the last step. Then, we obtain from (3.47) that

$$
\begin{aligned}
&\widetilde{G}_*\kappa\|e_u^{n+1}\|^2 - \widetilde{G}_*\kappa\|e_u^n\|^2 + \varepsilon^2\|\nabla_h e_u^{n+1}\|^2 - \varepsilon^2\|\nabla_h e_u^n\|^2 + 2\tau\|\delta_t e_u^{n+1}\|^2 \\
&\quad = 2\tau\langle g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})f(\widehat{u}^{n+\frac{1}{2}}) - g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}}))f(u_e(t_{n+\frac{1}{2}})), \delta_t e_u^{n+1}\rangle \\
&\qquad + 2\kappa\tau\big(g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}})) - g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})\big)\Big\langle \frac{u_e(t_{n+1}) + u_e(t_n)}{2} - u_e(t_{n+\frac{1}{2}}), \delta_t e_u^{n+1}\Big\rangle \\
&\qquad + 2\widetilde{G}_*\kappa\tau\langle e_u^n, \delta_t e_u^{n+1}\rangle + 2\kappa\tau g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})\langle \widehat{e}_u^{n+\frac{1}{2}} - e_u^n, \delta_t e_u^{n+1}\rangle - 2\tau\langle R_{2u}^n, \delta_t e_u^{n+1}\rangle. \qquad (3.48)
\end{aligned}
$$

For the last three terms in the right-hand side of (3.48), we have respectively

$$
\begin{aligned}
2\kappa\tau g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})\langle \widehat{e}_u^{n+\frac{1}{2}} - e_u^n, \delta_t e_u^{n+1}\rangle &\leq 2G^*\kappa\tau(\|\widehat{e}_u^{n+\frac{1}{2}}\| + \|e_u^n\|)\|\delta_t e_u^{n+1}\| \\
&\leq 6{G^*}^2\kappa^2\tau(\|\widehat{e}_u^{n+\frac{1}{2}}\|^2 + \|e_u^n\|^2) + \frac{\tau}{3}\|\delta_t e_u^{n+1}\|^2, \qquad (3.49)\\
2\widetilde{G}_*\kappa\tau\langle e_u^n, \delta_t e_u^{n+1}\rangle &\leq 3\widetilde{G}_*^2\kappa^2\tau\|e_u^n\|^2 + \frac{\tau}{3}\|\delta_t e_u^{n+1}\|^2, \qquad (3.50)\\
-2\tau\langle R_{2u}^n, \delta_t e_u^{n+1}\rangle &\leq 3\tau\|R_{2u}^n\|^2 + \frac{\tau}{3}\|\delta_t e_u^{n+1}\|^2. \qquad (3.51)
\end{aligned}
$$

By the energy dissipation and MBP of the sESAV1 substep (3.34), we know that $\widehat{u}^{n+\frac{1}{2}}$ and $\widehat{s}^{n+\frac{1}{2}}$ are bounded uniformly. By conducting the similar deductions to the proof of Lemma 4, we can obtain

$$
|g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}}) - g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}}))| \leq C_g(\|\widehat{e}_u^{n+\frac{1}{2}}\| + |\widehat{e}_s^{n+\frac{1}{2}}|), \qquad (3.52a)
$$

$$
\|g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})f(\widehat{u}^{n+\frac{1}{2}}) - g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}}))f(u_e(t_{n+\frac{1}{2}}))\| \leq C_g(\|\widehat{e}_u^{n+\frac{1}{2}}\| + |\widehat{e}_s^{n+\frac{1}{2}}|), \qquad (3.52b)
$$

where $C_g > 0$ is the same constant defined in Lemma 4. Then, the first and second terms in the right-hand side of (3.48) can be bounded respectively as

$$
\begin{aligned}
&2\tau\langle g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})f(\widehat{u}^{n+\frac{1}{2}}) - g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}}))f(u_e(t_{n+\frac{1}{2}})), \delta_t e_u^{n+1}\rangle \\
&\quad \leq 2C_g\tau(\|\widehat{e}_u^{n+\frac{1}{2}}\| + |\widehat{e}_s^{n+\frac{1}{2}}|)\|\delta_t e_u^{n+1}\| \\
&\quad \leq 6C_g^2\tau(\|\widehat{e}_u^{n+\frac{1}{2}}\|^2 + |\widehat{e}_s^{n+\frac{1}{2}}|^2) + \frac{\tau}{3}\|\delta_t e_u^{n+1}\|^2, \qquad (3.53)
\end{aligned}
$$

and

$$
\begin{aligned}
&2\kappa\tau\big(g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}})) - g(\widehat{u}^{n+\frac{1}{2}},\widehat{s}^{n+\frac{1}{2}})\big)\Big\langle \frac{u_e(t_{n+1}) + u_e(t_n)}{2} - u_e(t_{n+\frac{1}{2}}), \delta_t e_u^{n+1}\Big\rangle \\
&\quad \leq C_g\kappa\tau(\|\widehat{e}_u^{n+\frac{1}{2}}\| + |\widehat{e}_s^{n+\frac{1}{2}}|)(\|u_e(t_{n+1})\| + \|u_e(t_n)\| + 2\|u_e(t_{n+\frac{1}{2}})\|)\|\delta_t e_u^{n+1}\| \\
&\quad \leq C_1\kappa^2\tau(\|\widehat{e}_u^{n+\frac{1}{2}}\|^2 + |\widehat{e}_s^{n+\frac{1}{2}}|^2) + \frac{\tau}{3}\|\delta_t e_u^{n+1}\|^2, \qquad (3.54)
\end{aligned}
$$

where $C_1 > 0$ has the same dependence as the constant C_1 used in (3.24) but may have a different value. Substituting (3.49)–(3.54) into (3.48) leads to

$$
\begin{aligned}
&\widetilde{G}_*\kappa\|e_u^{n+1}\|^2 - \widetilde{G}_*\kappa\|e_u^n\|^2 + \varepsilon^2\|\nabla_h e_u^{n+1}\|^2 - \varepsilon^2\|\nabla_h e_u^n\|^2 + \frac{\tau}{3}\|\delta_t e_u^{n+1}\|^2 \\
&\quad \leq (6{G^*}^2\kappa^2 + 6C_g^2 + C_1\kappa^2)\tau\|\widehat{e}_u^{n+\frac{1}{2}}\|^2 + (6C_g^2 + C_1\kappa^2)\tau|\widehat{e}_s^{n+\frac{1}{2}}|^2 \\
&\qquad + (3\widetilde{G}_*^2 + 6{G^*}^2)\kappa^2\tau\|e_u^n\|^2 + 3\tau\|R_{2u}^n\|^2. \qquad (3.55)
\end{aligned}
$$

Multiplying (3.46b) by $2\tau e_s^{n+1}$ yields

$$\begin{aligned}
&|e_s^{n+1}|^2 - |e_s^n|^2 + |e_s^{n+1} - e_s^n|^2 \\
&\quad = 2e_s^{n+1}\langle g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}}))f(u_e(t_{n+\frac{1}{2}})) - g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})f(\widehat{u}^{n+\frac{1}{2}}), u_e(t_{n+1}) - u_e(t_n)\rangle \\
&\qquad + \kappa e_s^{n+1}\big(g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}) - g(u_e(t_{n+\frac{1}{2}}), s_e(t_{n+\frac{1}{2}}))\big)\langle u_e(t_{n+1}) + u_e(t_n) - 2u_e(t_{n+\frac{1}{2}}), u_e(t_{n+1}) - u_e(t_n)\rangle \\
&\qquad - 2\tau e_s^{n+1} g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})\langle f(\widehat{u}^{n+\frac{1}{2}}), \delta_t e_u^{n+1}\rangle + 2\kappa\tau e_s^{n+1} g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})\langle u^{n+\frac{1}{2}} - \widehat{u}^{n+\frac{1}{2}}, \delta_t e_u^{n+1}\rangle \\
&\qquad + 2\kappa e_s^{n+1} g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})\langle e_u^{n+\frac{1}{2}} - \widehat{e}_u^{n+\frac{1}{2}}, u_e(t_{n+1}) - u_e(t_n)\rangle - 2\tau R_{2s}^n e_s^{n+1}.
\end{aligned} \tag{3.56}$$

The last term in the right-hand side of (3.56) can be estimated by

$$-2\tau R_{2s}^n e_s^{n+1} \le \tau|e_s^{n+1}|^2 + \tau|R_{2s}^n|^2. \tag{3.57}$$

By the boundedness of $u^{n+\frac{1}{2}}$, $\widehat{u}^{n+\frac{1}{2}}$, and $\widehat{s}^{n+\frac{1}{2}}$, the sum of the third and fourth terms in the right-hand side of (3.56) can be estimated similarly to (3.30) as follows:

$$\begin{aligned}
&- 2\tau e_s^{n+1} g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})\langle f(\widehat{u}^{n+\frac{1}{2}}), \delta_t e_u^{n+1}\rangle + 2\kappa\tau e_s^{n+1} g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}})\langle u^{n+\frac{1}{2}} - \widehat{u}^{n+\frac{1}{2}}, \delta_t e_u^{n+1}\rangle \\
&\quad \le 2G^*\kappa\tau(\|u^{n+\frac{1}{2}}\| + \|\widehat{u}^{n+\frac{1}{2}}\|)|e_s^{n+1}|\|\delta_t e_u^{n+1}\| + 2G^*\tau\|f(\widehat{u}^{n+\frac{1}{2}})\||e_s^{n+1}|\|\delta_t e_u^{n+1}\| \\
&\quad \le C_6\tau|e_s^{n+1}|^2 + \frac{\tau}{3}\|\delta_t e_u^{n+1}\|^2,
\end{aligned} \tag{3.58}$$

where $C_6 > 0$ depends on C_*, $|\Omega|$, u_{init}, κ, and $\|f\|_{C[-\beta,\beta]}$. Then, using the facts that $\|u_e(t_{n+1}) - u_e(t_n)\| \le C\tau$, $\|u_e(t_{n+1}) + u_e(t_n) - 2u_e(t_{n+\frac{1}{2}})\| \le C\tau^2$ (where $C > 0$ is a constant due to smoothness of u_e), and the inequalities (3.52), in the similar spirit of deriving (3.29), the sum of the first, second and fifth terms in the right-hand side of (3.56) can be bounded above by

$$\tau\big(\|\widehat{e}_u^{n+\frac{1}{2}}\|^2 + |\widehat{e}_s^{n+\frac{1}{2}}|^2 + \|e_u^n\|^2 + \|e_u^{n+1}\|^2 + |e_s^{n+1}|^2\big) \tag{3.59}$$

multiplied with a positive constant depending on C_*, $|\Omega|$, u_e, κ, and $\|f\|_{C^1[-\beta,\beta]}$. Combining (3.56) with (3.57)–(3.59), we obtain

$$\begin{aligned}
|e_s^{n+1}|^2 - |e_s^n|^2 &\le C_7\tau\big(\|\widehat{e}_u^{n+\frac{1}{2}}\|^2 + |\widehat{e}_s^{n+\frac{1}{2}}|^2 + \|e_u^n\|^2 + \|e_u^{n+1}\|^2 + |e_s^{n+1}|^2\big) \\
&\quad + \frac{\tau}{3}\|\delta_t e_u^{n+1}\|^2 + \tau|R_{2s}^n|^2
\end{aligned} \tag{3.60}$$

with C_7 depending on C_*, $|\Omega|$, u_e, κ, and $\|f\|_{C^1[-\beta,\beta]}$.

Adding (3.55) and (3.60), we obtain

$$\begin{aligned}
&\widetilde{G}_*\kappa(\|e_u^{n+1}\|^2 - \|e_u^n\|^2) + \varepsilon^2(\|\nabla_h e_u^{n+1}\|^2 - \|\nabla_h e_u^n\|^2) + (|e_s^{n+1}|^2 - |e_s^n|^2) \\
&\quad \le C_8\tau\big(\|\widehat{e}_u^{n+\frac{1}{2}}\|^2 + |\widehat{e}_s^{n+\frac{1}{2}}|^2 + \|e_u^n\|^2 + \|e_u^{n+1}\|^2 + |e_s^{n+1}|^2\big) + 3\tau\|R_{2u}^n\|^2 + \tau|R_{2s}^n|^2,
\end{aligned} \tag{3.61}$$

where $C_8 > 0$ depends on C_*, $|\Omega|$, u_e, κ, and $\|f\|_{C^1[-\beta,\beta]}$. Substituting (3.43) into (3.61) and using the estimate (3.40), when $\tau \le 1$, we have

$$\begin{aligned}
&\widetilde{G}_*\kappa(\|e_u^{n+1}\|^2 - \|e_u^n\|^2) + \varepsilon^2(\|\nabla_h e_u^{n+1}\|^2 - \|\nabla_h e_u^n\|^2) + (|e_s^{n+1}|^2 - |e_s^n|^2) \\
&\quad \le C_8(\widehat{C} + 1)\tau(\|e_u^n\|^2 + \|e_u^{n+1}\|^2 + |e_s^{n+1}|^2) + (C_8\widehat{C} + 4)C_e^2\tau(\tau^2 + h^2)^2.
\end{aligned}$$

When τ is small sufficiently, similar to the last paragraph in the proof of Theorem 3, applying the discrete Gronwall's inequality yields

$$\widetilde{G}_*\kappa\|e_u^n\|^2 + \varepsilon^2\|\nabla_h e_u^n\|^2 + |e_s^n|^2 \le C(\tau^2 + h^2)^2,$$

which completes the proof.

4 Numerical experiments

This section is devoted to numerical tests and comparisons between the proposed sESAV schemes and existing SAV schemes listed in Sections 2.2 and 2.3. We consider the Allen–Cahn equation (1.1) in two-dimensional spatial domain $\Omega = (0,1) \times (0,1)$ equipped with periodic boundary conditions, so that the schemes can be solved efficiently by the fast Fourier transform. We take two types of commonly-used nonlinear functions $f(u)$. One is given by

$$f(u) = -F'(u) = u - u^3 \tag{4.1}$$

with F being the *double-well potential*

$$F(u) = \frac{1}{4}(u^2 - 1)^2.$$

In this case, one has $\beta = 1$ and $\|f'\|_{C[-1,1]} = 2$. The constant C_0 in (2.4) is $C_0 = \frac{1}{4}(\kappa^2 + 2\kappa)$. The other one is determined by the *Flory–Huggins potential*

$$F(u) = \frac{\theta}{2}[(1+u)\ln(1+u) + (1-u)\ln(1-u)] - \frac{\theta_c}{2}u^2,$$

which gives

$$f(u) = -F'(u) = \frac{\theta}{2}\ln\frac{1-u}{1+u} + \theta_c u, \tag{4.2}$$

where $\theta_c > \theta > 0$. In the following experiments, we set $\theta = 0.8$ and $\theta_c = 1.6$, then the positive root of $f(\rho) = 0$ gives us $\beta \approx 0.9575$, and $\|f'\|_{C[-\beta,\beta]} \approx 8.02$. The constant C_0 in (2.4) is then determined by $C_0 = -F(\alpha) + \frac{\kappa}{2}\alpha^2$, where $\alpha > 0$ solves $f(\alpha) + \kappa\alpha = 0$.

4.1 Convergence in time

We first verify the convergence order in time for the proposed sESAV schemes. Let us set $\varepsilon = 0.01$ in (1.1) and take a smooth initial value

$$u_{\text{init}}(x,y) = 0.1\sin(2\pi x)\sin(2\pi y).$$

The temporal convergence tests are conducted by fixing the spatial mesh size $h = 1/512$. As requested by the stabilizing condition $\kappa \geq \|f'\|_{C[-\beta,\beta]}$, we set $\kappa = 2$ for the double-well potential case (i.e., $f(u)$ given by (4.1)) and $\kappa = 8.02$ for the Flory–Huggins potential case (i.e., $f(u)$ given by (4.2)). We compute the numerical solutions at $t = 2$ using the sESAV1 and sESAV2 schemes with various time step sizes $\tau = 2^{-k}$, $k = 4, 5, \ldots, 12$. To compute the numerical errors, we treat the sESAV2 solution obtained by $\tau = 0.1 \times 2^{-12}$ as the benchmark solution. Figure 1 shows the relation between the L^2-norm error and the time step size, where the left picture corresponds to the double-well potential case and the right one for the Flory–Huggins potential case. The first-order temporal accuracy for sESAV1 and the second-order for sESAV2 are observed for both cases as expected.

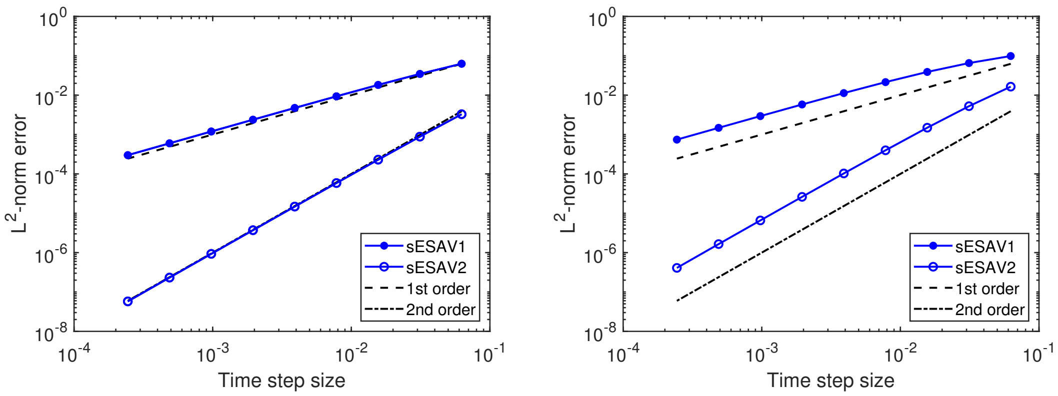

Fig. 1 The L^2-norm errors vs. the time step size produced by the proposed sESAV1 and sESAV2 schemes for the double-well potential case (4.1) (left) and the Flory–Huggins potential case (4.2) (right).

4.2 Comparisons with existing SAV schemes

In the following numerical experiments, we compare the proposed sESAV schemes with classic SAV and ESAV schemes by focusing on the MBP and energy dissipation law. While various modified energies are introduced as approximations of the original energy in discrete settings in order to facilitate the proof of energy dissipation law, the original one possesses the most accurate physical meaning for the model problem. Therefore, we are concerned about the behavior of the original (discrete) energy $E_h(u)$ defined in (2.2) for reflecting the phase transition process. The dynamic process considered usually needs a long-time evolution to reach the steady state; here we conduct simulations in a short time interval for the comparison among these schemes.

Let us still consider the problem (1.1) with $\varepsilon = 0.01$. We adopt the uniform spatial mesh with $h = 1/512$ and give the initial value by random numbers between -0.8 and 0.8 on each mesh point. We then set the time step size $\tau = 0.01$, and compute the numerical solutions by using the sESAV schemes, the classic SAV schemes (SAV1 and SAV2), and the ESAV schemes (ESAV1 and ESAV2). Note that we set $\delta = C_0 + 0.01$ for the classic SAV schemes (2.6) and (2.7). For all comparison experiments, we will consider two settings for the stabilizing parameter: $\kappa = \|f'\|_{C[-\beta,\beta]}$ and $\kappa = \frac{1}{2}\|f'\|_{C[-\beta,\beta]}$, where the former one satisfies the requirement for the MBP preservation for the sESAV schemes and the latter one was adopted in [33] for the classic SAV schemes. In addition, we take the numerical results obtained by the IFRK4 scheme [19] with the small time step size 10^{-4} as the benchmark solution.

First, we test the double-well potential case (4.1), and correspondingly, set the stabilizing parameter $\kappa = 1$ and $\kappa = 2$ respectively to carry out the experiments. Figure 2 shows the evolutions of the supremum norms and the energies of simulated solutions computed by the sESAV1, SAV1, and ESAV1 schemes. For either $\kappa = 1$ or $\kappa = 2$, the sESAV1 scheme preserves the MBP, while the supremum norms of the SAV1 and ESAV1 solutions obviously evolve beyond 1, which means that the MBP is violated. The energy dissipation are observed for these three schemes, where the sESAV1 scheme provides the most accurate result. In addition, the larger κ leads to larger errors in the results, especially for the ESAV1 scheme. Figure 3 plots corresponding results computed by the second-order schemes. Again, only the sESAV2 scheme preserves the MBP and the energy dissipation perfectly. The SAV2 and ESAV2 solutions evolve beyond 1 but closer to 1 than their first-order results due to the higher-order temporal accuracy.

Next, we test the Flory–Huggins potential case (4.2) and correspondingly set $\kappa = 4.01$ and $\kappa = 8.02$ respectively. Figures 4 and 5 present the evolutions of the supremum norms and the energies of simulated solutions obtained by the first- and second-order schemes, respectively. Similar to the double-well potential case, only the sESAV schemes preserve the MBP and the energy dissipation law as expected. The SAV1, ESAV1, and ESAV2 schemes, having the supremum norms beyond the theoretical bound 0.9575, lead to inaccurate dynamic processes. Especially, the ESAV1 solution with $\kappa = 8.02$ evolves beyond 1, which yields complex numbers due to the existence of the logarithmic term and gives the completely wrong dynamics. For the SAV2 solutions, the dynamic processes look moderately correct according to the energy evolutions. Moreover, it is interesting that the supremum norm goes larger than the desired bound for $\kappa = 8.02$ while it does not exceed for $\kappa = 4.01$, but both results are still a bit away from the expected value 0.9575.

4.3 Long-time coarsening dynamics simulations

Now we study the coarsening dynamics driven by the Allen–Cahn equation (1.1) with $\varepsilon = 0.01$. The spatial mesh size is $h = 1/512$ and the initial state is given by random numbers between -0.8 and 0.8. We adopt the sESAV2 scheme with $\tau = 0.01$ to simulate the long-time coarsening process. By the comparisons shown above, we know that $\tau = 0.01$ is sufficient to provide accurate numerical results. The steady state of the coarsening dynamics is a constant state $u \equiv \beta$ or $u \equiv -\beta$. When the absolute difference between the energies at the two consecutive moments is smaller than the tolerance value 10^{-8}, we regard the dynamics as reaching its steady state.

For the double-well potential case $f(u)$, we set $\kappa = 2$ and the phase structures captured at some moments are presented in Figure 6, and the constant steady state $u \equiv -1$ is reached at around $t = 604$. The left picture given in Figure 7 implies the preservation of the MBP during the whole phase transition process. The energy evolution is plotted in the right graph of Figure 7, which states the energy dissipation of the process. For the Flory–Huggins potential case, we set $\kappa = 8.02$ and the simulated results are shown in Figures 8 and 9. We observe that the steady state is reached at around $t = 602$ and the whole process of phase separation is similar to that of the double-well potential case. Those results are almost identical to those produced using the IFRK4 scheme in [19].

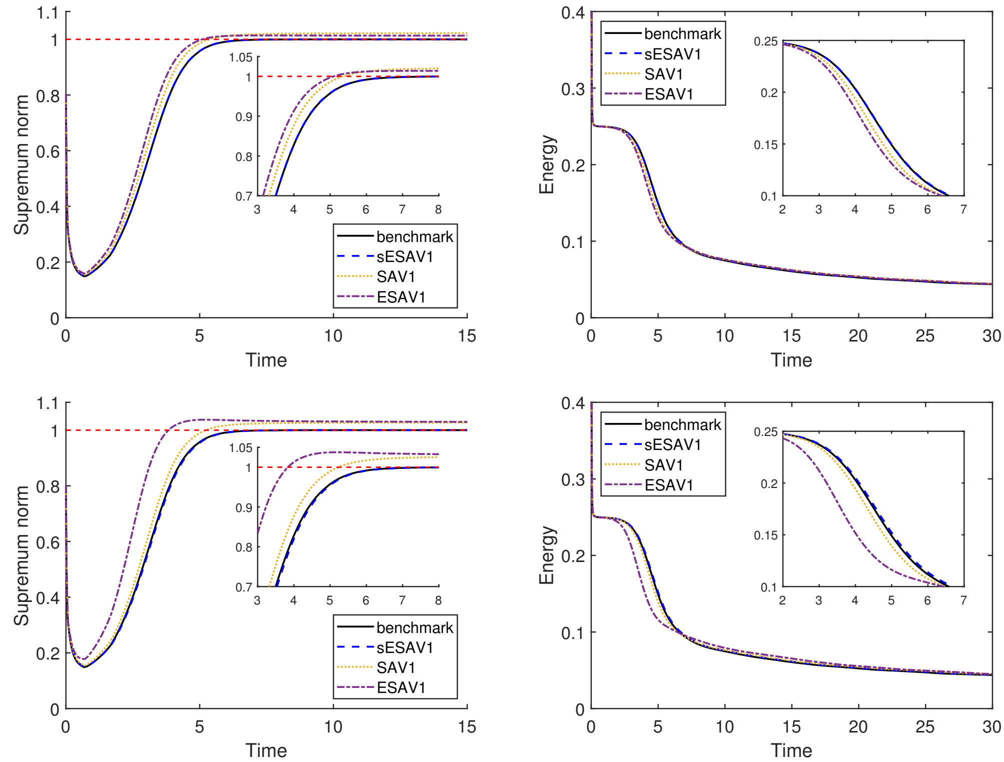

Fig. 2 Evolutions of the supremum norms and the energies of simulated solutions computed by the sESAV1, SAV1, and ESAV1 schemes with $\tau = 0.01$ and $\kappa = 1$ (top row) or $\kappa = 2$ (bottom row) for the double-well potential case.

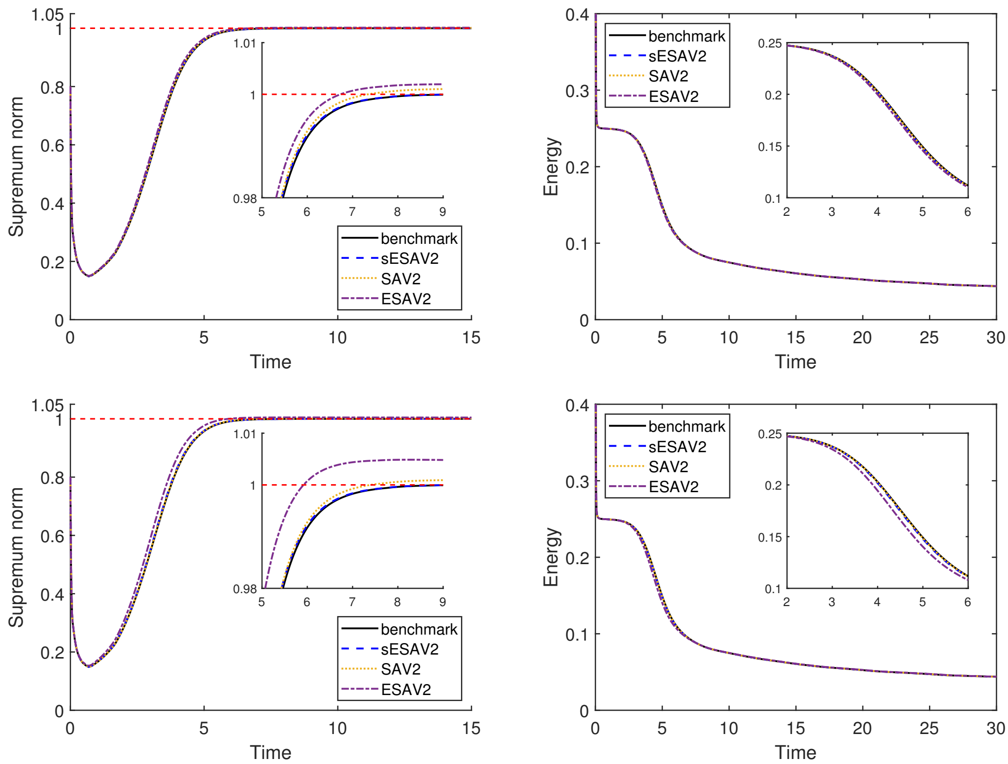

Fig. 3 Evolutions of the supremum norms and the energies of simulated solutions computed by the sESAV2, SAV2, and ESAV2 schemes with $\tau = 0.01$ and $\kappa = 1$ (top row) or $\kappa = 2$ (bottom row) for the double-well potential case.

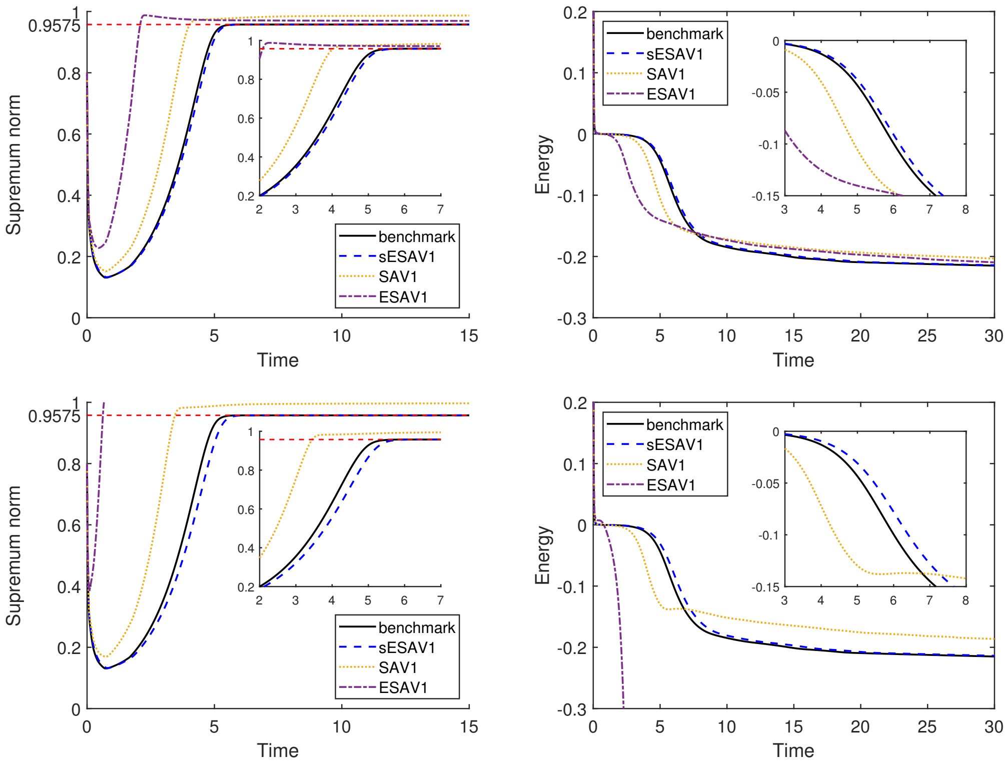

Fig. 4 Evolutions of the supremum norms and the energies of simulated solutions computed by the sESAV1, SAV1, and ESAV1 schemes with $\tau = 0.01$ and $\kappa = 4.01$ (top row) or $\kappa = 8.02$ (bottom row) for the Flory–Huggins potential case.

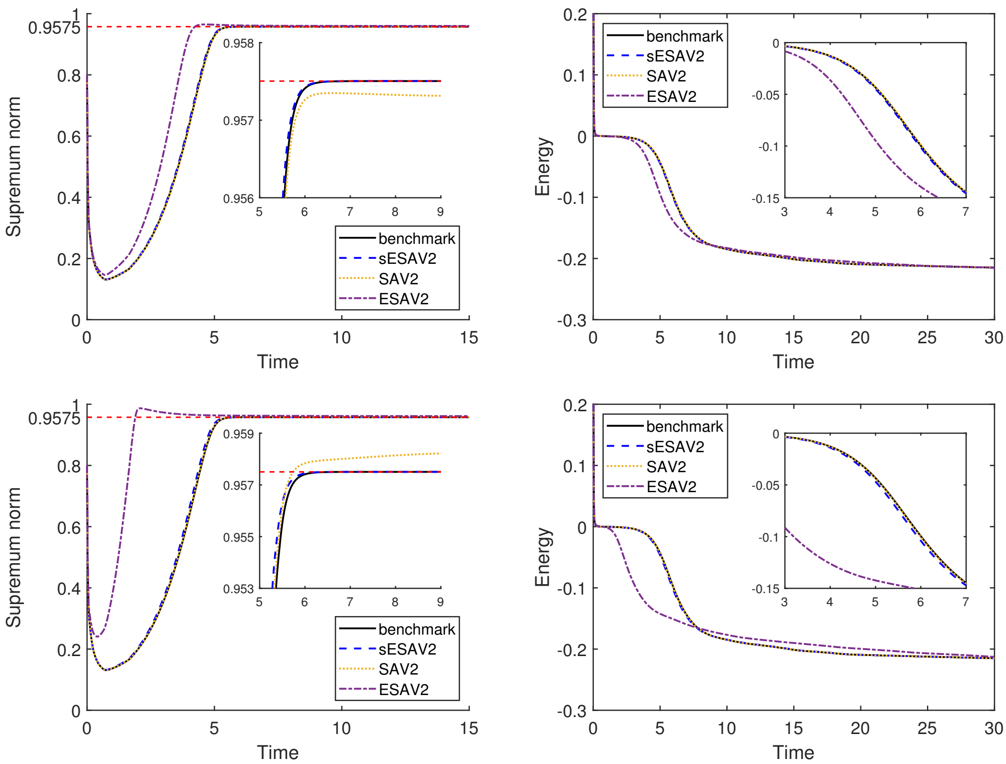

Fig. 5 Evolutions of the supremum norms and the energies of simulated solutions computed by the sESAV2, SAV2, and ESAV2 schemes with $\tau = 0.01$ and $\kappa = 4.01$ (top row) or $\kappa = 8.02$ (bottom row) for the Flory–Huggins potential case.

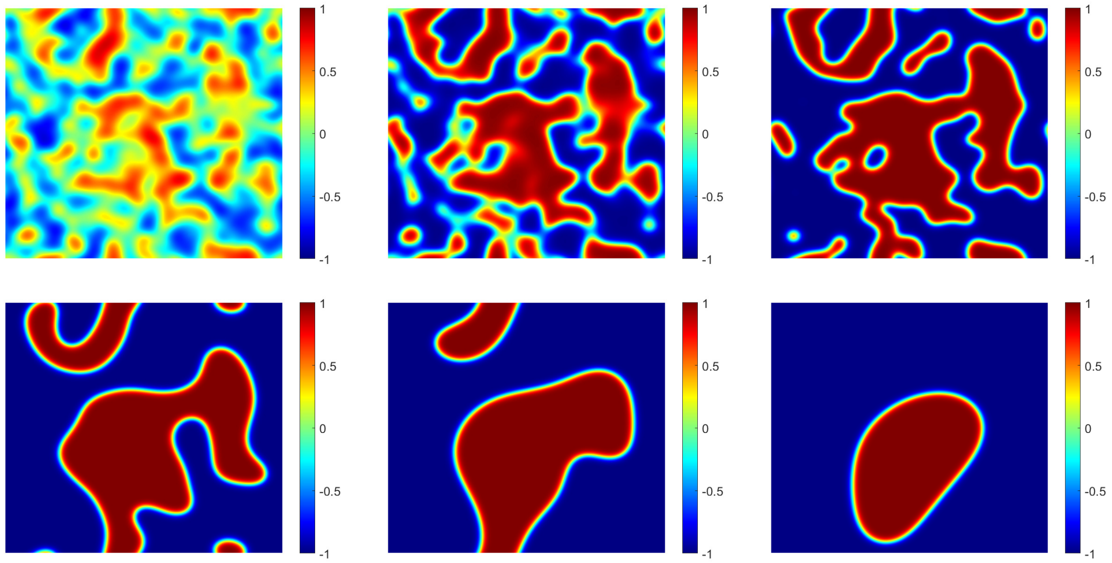

Fig. 6 Simulated phase structures at $t = 4$, 6, 10, 30, 100, and 300, respectively (left to right and top to bottom) by the sESAV2 scheme with $\tau = 0.01$ and $\kappa = 2$ for the coarsening dynamics of the double-well potential case.

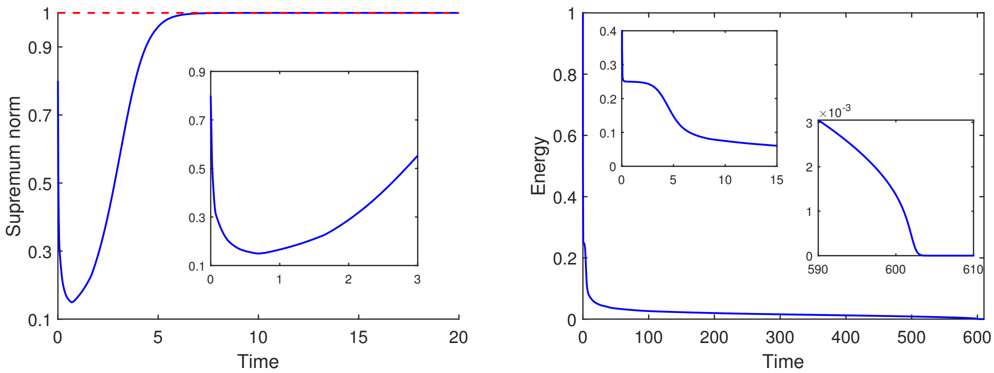

Fig. 7 Evolutions of the supremum norm (left) and the energy (right) for the coarsening dynamics of the double-well potential case.

5 Conclusion

In this paper, we study MBP-preserving and energy dissipative schemes for the Allen–Cahn type equations by combining the ESAV approach with stabilizing technique. We present first- and second-order sESAV schemes and prove their MBP preservation, energy dissipation, and error estimates. The main results and observations include two aspects. First, we choose the ESAV approach rather than the classic SAV approach, since the coefficient $(g(u^n, s^n)$ or $g(\widehat{u}^{n+\frac{1}{2}}, \widehat{s}^{n+\frac{1}{2}}))$ of the nonlinear term is positive automatically in the former one while the sign of the corresponding coefficient is uncertain for the later one. Second, to guarantee the MBP-preserving property, we add the stabilization term as an extra artificial term, that is, add and subtract a linear term in the scheme instead of a quadratic term in the energy functional; they are equivalent mutually for the classic stabilization or convex splitting method, but not for the SAV approach. Moreover, we find that the MBP preservation and the energy dissipation of the sESAV schemes can be established in parallel and independently, unlike the purely stabilized semi-implicit scheme discussed in [35] where the MBP is needed first to bound the nonlinear term in the proof of the stability with respect to the original energy. Since the schemes we studied are all one-step methods, adaptive time-stepping strategies (such as [27]) can be inherently adopted to accelerate the computation.

Some generalizations can be carried out by replacing the Laplace operator in (1.1) by some analogues, for instance, the nonlocal diffusion [8] and the fractional Laplace operators [28] which also satisfy the semi-group property with their discretizations satisfying the analogues of Lemma 1. Furthermore, the proposed

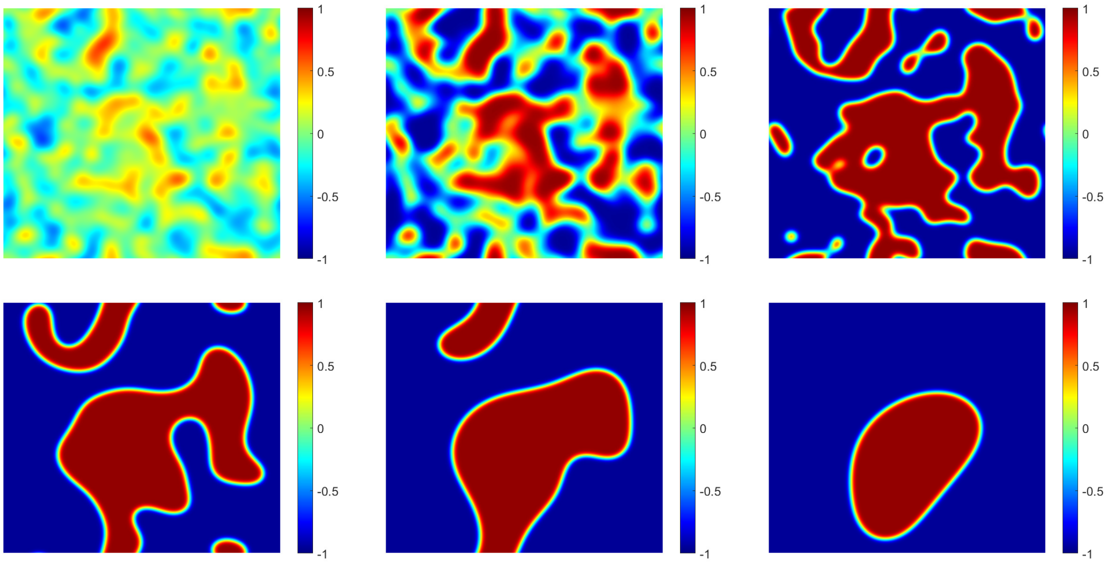

Fig. 8 Simulated phase structures at $t = 4$, 6, 10, 30, 100, and 300, respectively (left to right and top to bottom) by the sESAV2 scheme with $\tau = 0.01$ and $\kappa = 8.02$ for the coarsening dynamics of the Flory–Huggins potential case.

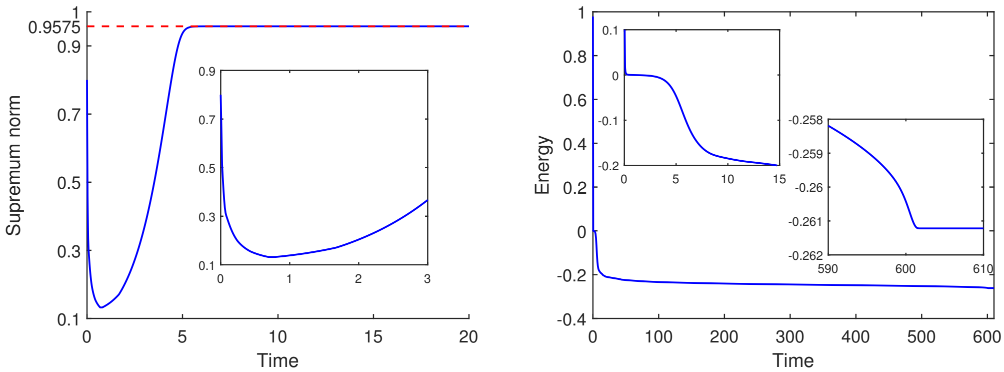

Fig. 9 Evolutions of the supremum norm (left) and the energy (right) for the coarsening dynamics of the Flory–Huggins potential case.

stabilizing approaches in this paper can be naturally extended to many other type of gradient flow problems, which can be handled by the existing SAV schemes. For example, the fourth-order Cahn–Hilliard equation is the H^{-1} gradient flow of the energy functional (1.2), and satisfies the same energy dissipation law as the Allen–Cahn equation. The MBP is not valid anymore, but the solution is still L^∞ stable. In the similar spirit of this paper, it is interesting to develop the sESAV schemes for the Cahn–Hilliard equation, and the discrete L^∞ stability of the sESAV solution can be established by combining the high-order consistency analysis and stability estimate, as done in [13,24], which will be one of our future works.

References

1. S. M. Allen and J. W. Cahn, *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metall., 27 (1979), 1085–1095.
2. G. Akrivis, B. Y. Li, and D. F. Li, *Energy-decaying extrapolated RK-SAV methods for the Allen–Cahn and Cahn–Hilliard equations*, SIAM J. Sci. Comput., 41 (2019), A3703–A3727.
3. P. W. Bates, *On some nonlocal evolution equations arising in materials science*, Fields Inst. Commun., 48 (2006), 13–52.
4. C. J. Chen and X. F. Yang, *Fast, provably unconditionally energy stable, and second-order accurate algorithms for the anisotropic Cahn–Hilliard model*, Comput. Methods Appl. Mech. Engrg., 351 (2019), 35–59.
5. W. B. Chen, X. M. Wang, Y. Yan, and Z. Y. Zhang, *A second order BDF numerical scheme with variable steps for the Cahn–Hilliard equation*, SIAM J. Numer. Anal., 57 (2019), 495–525.
6. Q. Cheng, C. Liu, and J. Shen, *A new Lagrange multiplier approach for gradient flows*, Comput. Methods Appl. Mech. Engrg., 367 (2020), 113070.

7. Q. Cheng, C. Liu, and J. Shen, *Generalized SAV approaches for gradient systems*, J. Comput. Appl. Math., 394 (2021), 113532.
8. Q. Du, M. Gunzburger, R. B. Lehoucq, and K. Zhou, *Analysis and approximation of nonlocal diffusion problems with volume constraints*, SIAM Rev., 54 (2012), 667–696.
9. Q. Du, L. Ju, X. Li, and Z. H. Qiao, *Maximum principle preserving exponential time differencing schemes for the nonlocal Allen–Cahn equation*, SIAM J. Numer. Anal., 57 (2019), 875–898.
10. Q. Du, L. Ju, X. Li, and Z. H. Qiao, *Maximum bound principles for a class of semilinear parabolic equations and exponential time-differencing schemes*, SIAM Rev., 63 (2021), 317–359.
11. Q. Du, J. Yang, and Z. Zhou, *Time-fractional Allen–Cahn equations: analysis and numerical methods*, J. Sci. Comput., 85 (2020), 42.
12. X. L. Feng, T. Tang, and J. Yang, *Stabilized Crank–Nicolson/Adams–Bashforth schemes for phase field models*, East Asian J. Appl. Math., 3 (2013), 59–80.
13. Z. Guan, C. Wang, and S. M. Wise, *A convergent convex splitting scheme for the periodic nonlocal Cahn–Hilliard equation*, Numer. Math., 128 (2014), 377–406.
14. C. F. Gui and M. F. Zhao, *Traveling wave solutions of Allen–Cahn equation with a fractional Laplacian*, Ann. Inst. H. Poincaré-An., 32 (2015), 785–812.
15. D. M. Hou, M. Azaiez, and C. J. Xu, *A variant of scalar auxiliary variable approaches for gradient flows*, J. Comput. Phys., 395 (2019), 307–332.
16. T. L. Hou and H. T. Leng, *Numerical analysis of a stabilized Crank–Nicolson/Adams–Bashforth finite difference scheme for Allen–Cahn equations*, Appl. Math. Lett., 102 (2020), 106150.
17. T. L. Hou, T. Tang, and J. Yang, *Numerical analysis of fully discretized Crank–Nicolson scheme for fractional-in-space Allen–Cahn equations*, J. Sci. Comput., 72 (2017), 1214–1231.
18. F. K. Huang, J. Shen, and Z. G. Yang, *A highly efficient and accurate new scalar auxiliary variable approach for gradient flows*, SIAM J. Sci. Comput., 42 (2020), A2514–A2536.
19. L. Ju, X. Li, Z. H. Qiao, and J. Yang, *Maximum bound principle preserving integrating factor Runge–Kutta methods for semilinear parabolic equations*, J. Comput. Phys., 439 (2021), 110405.
20. L. Ju, X. Li, Z. H. Qiao, and H. Zhang, *Energy stability and error estimates of exponential time differencing schemes for the epitaxial growth model without slope selection*, Math. Comp., 87 (2018), 1859–1885.
21. L. Ju, J. Zhang, and Q. Du, *Fast and accurate algorithms for simulating coarsening dynamics of Cahn–Hilliard equations*, Comput. Mater. Sci., 108 (2015), 272–282.
22. J. W. Li, L. Ju, Y. Y. Cai, and X. L. Feng, *Unconditionally maximum bound principle preserving linear schemes for the conservative Allen–Cahn equation with nonlocal constraint*, J. Sci. Comput., 87 (2021), 98.
23. J. W. Li, X. Li, L. Ju, and X. L. Feng, *Stabilized integrating factor Runge–Kutta method and unconditional preservation of maximum bound principle*, SIAM J. Sci. Comput., 43 (2021), A1780–A1802.
24. X. Li, Z. H. Qiao, and C. Wang, *Convergence analysis for a stabilized linear semi-implicit numerical scheme for the nonlocal Cahn–Hilliard equation*, Math. Comp., 90 (2021), 171–188.
25. H. L. Liao, T. Tang, and T. Zhou, *On energy stable, maximum-principle preserving, second-order BDF scheme with variable steps for the Allen–Cahn equation*, SIAM J. Numer. Anal., 58 (2020), 2294–2314.
26. Z. G. Liu and X. L. Li, *The exponential scalar auxiliary variable (E-SAV) approach for phase field models and its explicit computing*, SIAM J. Sci. Comput., 42 (2020), B630–B655.
27. Z. H. Qiao, Z. R. Zhang, and T. Tang, *An adaptive time-stepping strategy for the molecular beam epitaxy models*, SIAM J. Sci. Comput., 33 (2011), 1395–1414.
28. S. G. Samko, A. A. Kilbas, and O. I. Marichev, *Fractional Integrals and Derivatives*, Gordon and Breach, Yverdon, 1993.
29. J. Shen, T. Tang, and J. Yang, *On the maximum principle preserving schemes for the generalized Allen–Cahn equation*, Commun. Math. Sci., 14 (2016), 1517–1534.
30. J. Shen, C. Wang, X. M. Wang, and S. M. Wise, *Second-order convex splitting schemes for gradient flows with Ehrlich–Schwoebel type energy: application to thin film epitaxy*, SIAM J. Numer. Anal., 50 (2012), 105–125.
31. J. Shen and J. Xu, *Convergence and error analysis for the scalar auxiliary variable (SAV) schemes to gradient flows*, SIAM J. Numer. Anal., 56 (2018), 2895–2912.
32. J. Shen, J. Xu, and J. Yang, *The scalar auxiliary variable (SAV) approach for gradient flows*, J. Comput. Phys., 353 (2018), 407–416.
33. J. Shen, J. Xu, and J. Yang, *A new class of efficient and robust energy stable schemes for gradient flows*, SIAM Rev., 61 (2019), 474–506.
34. J. Shen and X. F. Yang, *Numerical approximations of Allen–Cahn and Cahn–Hilliard equations*, Discrete Contin. Dyn. Syst., 28 (2010), 1669–1691.
35. T. Tang and J. Yang, *Implicit-explicit scheme for the Allen–Cahn equation preserves the maximum principle*, J. Comput. Math., 34 (2016), 471–481.
36. S. M. Wise, C. Wang, and J. S. Lowengrub, *An energy stable and convergent finite difference scheme for the phase field crystal equation*, SIAM J. Numer. Anal., 47 (2009), 2269–2288.
37. X. F. Xiao, X. L. Feng, and J. Y. Yuan, *The stabilized semi-implicit finite element method for the surface Allen–Cahn equation*, Discrete Contin. Dyn. Syst. Ser. B, 22 (2017), 2857–2877.
38. C. J. Xu and T. Tang, *Stability analysis of large time-stepping methods for epitaxial growth models*, SIAM J. Numer. Anal., 44 (2006), 1759–1779.
39. Z. Xu, X. F. Yang, H. Zhang, and Z. Q. Xie, *Efficient and linear schemes for anisotropic Cahn–Hilliard model using the stabilized-invariant energy quadratization (S-IEQ) approach*, Comput. Phys. Commun., 238 (2019), 36–49.
40. Y. Yan, W. B. Chen, C. Wang, S. M. Wise, *A second-order energy stable BDF numerical scheme for the Cahn–Hilliard equation*, Commun. Comput. Phys., 23 (2018), 572–602.
41. X. F. 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.
42. X. F. Yang and G. D. Zhang, *Convergence analysis for the invariant energy quadratization (IEQ) schemes for solving the Cahn–Hilliard and Allen–Cahn equations with general nonlinear potential*, J. Sci. Comput., 82 (2020), 55.