Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 392 (2022) 114718
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

A new class of implicit–explicit BDFk SAV schemes for general dissipative systems and their error analysis☆

Fukeng Huang*, Jie Shen

Department of Mathematics, Purdue University, United States of America

Received 1 September 2021; received in revised form 12 December 2021; accepted 16 December 2021
Available online 17 February 2022

Abstract

We construct a new class of efficient implicit–explicit (IMEX) BDFk schemes combined with a scalar auxiliary variable (SAV) approach for general dissipative systems. We show that these schemes are unconditionally stable, and lead to a uniform bound of the numerical solution in the norm based on the principal linear operator in the free energy. Based on this uniform bound, we carry out a rigorous error analysis for the kth-order ($k = 1, 2, 3, 4, 5$) SAV schemes in a unified form for a class of general dissipative systems. We also present numerical results confirming our theoretical convergence rates.

MSC: 65M12; 35K20; 35K35; 35K55

Keywords: Dissipative system; Error analysis; SAV approach; Energy stability; High-order BDF scheme

1. Introduction

The original scalar auxiliary variable (SAV) approach proposed in [1,2] is a powerful approach to construct efficient time discretization schemes for gradient flows. Due to its simplicity, efficiency and generality, it attracted much attention and has been applied to various problems (see, for instance, [3,4] and the references therein).

Analysis of standard semi-implicit schemes for general gradient flows often requires to assume global Lipschitz condition on the nonlinear term (see, for instance, [5–7]), although for some semi-linear parabolic equations such as Allen–Cahn type equations, error estimates for some first- and/or second-order semi-implicit schemes [8,9] and first- and second-order exponential time differencing schemes [10,11] have been established recently without assuming the global Lipschitz condition. On the other hand, the convergence of SAV schemes can be established without such assumption thanks to the unconditional energy stability. For examples, rigorous error analysis of the semi-discretized first order original SAV schemes for L^2 and H^{-1} gradient flows with minimum assumptions have been presented in [12], first- and second-order error estimates have been derived for a related semi-discretized gPAV scheme for the Cahn–Hilliard equation in [13], and error analysis of fully discretized SAV schemes with finite differences and finite-elements have also been established in [14] and [15]. On the other hand, error estimates for a Fourier-spectral

☆ This research is partially supported by National Science Foundation (NSF) DMS-2012585 and AFOSR, United States of America FA9550-20-1-0309.

* Correspondence to: Department of Mathematics, National University of Singapore, 119076, Singapore.
E-mail addresses: huang972@purdue.edu, hfkeng@nus.edu.sg (F. Huang), shen7@purdue.edu (J. Shen).

https://doi.org/10.1016/j.cma.2022.114718
0045-7825/

SAV scheme for the phase-field crystal equation [16] (see also [17] for a related work). and a MAC-SAV scheme for the Navier–Stokes equation [18] are established. Note that for the original SAV approach, unconditional energy stability can only be established for first- and second-order BDF schemes, although it has been shown in [19] (see also [20]) that the SAV approach coupled with extrapolated and linearized Runge–Kutta methods can achieve arbitrarily high order unconditionally energy stable with a modified energy for the Allen–Cahn and Cahn–Hilliard equations, but require solving coupled linear systems. On the other hand, a recent work [21] shows that a third-order IMEX scheme for the MBE equation with no slope selection is stable. Note however that the nonlinear term in the MBE equation with no slope selection satisfies the Lipschitz condition.

Most recently, a new SAV approach for gradient flow is proposed in [22] which offers some essential improvements over the original SAV approach such as (i) its computational cost is about half of the original SAV approach, and (ii) its higher-order BDF versions are also unconditionally stable with a modified energy. While ample numerical results in [22] have shown that the new higher-order SAV schemes are indeed stable and can achieve higher-order accuracy, the modified energy is represented only by a SAV which does not involve any function norm, so it is difficult to carry out convergence and error analysis. See however [13] for an attempt on the error analysis for related first- and second-order gPAV-based schemes for the Cahn–Hilliard equation.

Inspired by [22], we construct in this paper a new class of explicit–implicit BDFk schemes for general dissipative systems and carry out a rigorous and unified error analysis for $1 \le k \le 5$. In particular, we choose a special control factor η_k^{n+1} (cf. (2.5d)) for the kth-order scheme which allows us to obtain a unconditional and uniform bound on the norm based on principal linear term in the energy functional of the dissipative system. This bound is essential for the error analysis in this paper.

While these IMEX schemes are applicable to a large class of dissipative systems, their error analysis is highly non-trivial, particularly at higher than second order. Since a unified analysis for general dissipative systems will involve complicated assumptions and techniques that may obscure the clarity of presentation, we shall consider the error analysis for two classes of typical dissipative systems: Allen–Cahn type and Cahn–Hilliard type equations. The key ingredients are the uniform H^1 bound derived from the general stability result (see (2.13) in Theorem 1) and a stability result in [23] (see Lemma 1) for the BDFk ($1 \le k \le 5$) schemes. With a delicate induction argument, we are able to establish optimal error estimates in $L^\infty(0, T; H^2)$ norm for our implicit–explicit BDFk ($1 \le k \le 5$) SAV schemes for both Allen-Cahn type and Cahn–Hilliard type equations.

In summary, the new class of kth-order ($1 \le k \le 5$) IMEX SAV schemes enjoy several distinct advantages, including:

- only requires solving, in most common situations, one linear system with constant coefficients at each time step, so its computational cost is essentially the same as the usual implicit–explicit (IMEX) schemes;
- applicable to general dissipative systems, and can be combined with any consistent Galerkin type spatial discretization;
- higher-order BDFk SAV schemes are unconditionally stable and amenable to adaptive time stepping without restriction on time step size;
- rigorous error estimates can be established for BDFk ($1 \le k \le 5$) SAV schemes.

The rest of the paper is organized as follows. In the next section, we describe out new SAV schemes for general dissipative systems in a unified form, prove its unconditionally stability, and provide some numerical results to demonstrate the convergence rate. In Section 3, we present the detailed proof for the kth-order schemes ($k = 1, 2, 3, 4, 5$) in a unified form for Allen–Cahn type equations. In Section 4, we present error analysis for Cahn–Hilliard type equations. Some concluding remarks are given in the last section.

We use the following notations throughout the paper. Let $\Omega \in \mathcal{R}^n$ ($n = 1, 2, 3$) be a bounded domain with sufficiently smooth boundary. We denote by $(\cdot, \cdot)$ and $\|\cdot\|$ the inner product and the norm in $L^2(\Omega)$, and by $H^s(\Omega)$ the usual Sobolev spaces with norm $\|\cdot\|_{H^s}$. Let V be a Banach space, we shall also use the standard notations $L^p(0, T; V)$ and $C([0, T]; V)$. To simplify the notation, we often omit the spatial dependence in the notation for the exact solution u, namely we denote $u(x, t)$ by $u(t)$. We shall use C to denote a constant which can change from one step to another, but is independent of δt.

2. New SAV schemes for dissipative systems

In this section, we describe the new SAV schemes for dissipative systems, show that they are unconditionally energy stable with a modified energy and derive a uniform bound for the norm based on the principal linear term in the energy functional.

Consider the following class of dissipative systems

$$\frac{\partial u}{\partial t} + \mathcal{A}u + g(u) = 0, \tag{2.1}$$

where u is a scalar or vector function, $\mathcal{A}$ is a positive differential operator and $g(u)$ is a nonlinear operator possibly with lower-order derivatives. We assume that the above equation satisfies a dissipative energy law

$$\frac{d\tilde{E}(u)}{dt} = -\mathcal{K}(u), \tag{2.2}$$

where $\tilde{E}(u) > -C_0$ for all u is an energy functional, $\mathcal{K}(u) > 0$ for all $u \neq 0$.

The above class of dissipative systems include in particular gradient flows but also other dissipative systems which do not have the gradient structure, such as viscous Burgers equation, reaction–diffusion equations etc.

2.1. The new SAV schemes

The key for the SAV approach is to introduce a scalar auxiliary variable (SAV) to rewrite (2.1) as an expanded system, and to discretize the expanded system instead of the original (2.1). In this paper, we introduce the following new SAV approach inspired by the SAV schemes introduced in [22]

Setting $r(t) = E(u)(t) := \tilde{E}(u)(t) + C_0 > 0$, we rewrite Eq. (2.1) with the energy law (2.2) as the following expanded system

$$\frac{\partial u}{\partial t} + \mathcal{A}u + g(u) = 0, \tag{2.3}$$

$$\frac{dE(u)}{dt} = -\frac{r(t)}{E(u)(t)}\mathcal{K}(u). \tag{2.4}$$

We construct the kth order new SAV schemes based on the implicit–explicit BDF-k formulae in the following unified form:

Given u^n, r^n, we compute $\bar{u}^{n+1}$, r^{n+1}, ξ^{n+1} and u^{n+1} consecutively by

$$\frac{\alpha_k\bar{u}^{n+1} - A_k(u^n)}{\delta t} + \mathcal{A}\bar{u}^{n+1} + g[B_k(\bar{u}^n)] = 0, \tag{2.5a}$$

$$\frac{1}{\delta t}\big(r^{n+1} - r^n\big) = -\frac{r^{n+1}}{E(\bar{u}^{n+1})}\mathcal{K}(\bar{u}^{n+1}), \tag{2.5b}$$

$$\xi^{n+1} = \frac{r^{n+1}}{E(\bar{u}^{n+1})}, \tag{2.5c}$$

$$u^{n+1} = \eta_k^{n+1}\bar{u}^{n+1} \ \text{ with } \eta_k^{n+1} = 1 - (1-\xi^{n+1})^{k+1}, \tag{2.5d}$$

where α_k, the operators A_k and B_k ($k = 1, 2, 3, 4, 5$) are given by:

first-order:

$$\alpha_1 = 1, \quad A_1(u^n) = u^n, \quad B_1(\bar{u}^n) = \bar{u}^n; \tag{2.6}$$

second-order:

$$\alpha_2 = \frac{3}{2}, \quad A_2(u^n) = 2u^n - \frac{1}{2}u^{n-1}, \quad B_2(\bar{u}^n) = 2\bar{u}^n - \bar{u}^{n-1}; \tag{2.7}$$

third-order:

$$\alpha_3 = \frac{11}{6}, \quad A_3(u^n) = 3u^n - \frac{3}{2}u^{n-1} + \frac{1}{3}u^{n-2}, \quad B_3(\bar{u}^n) = 3\bar{u}^n - 3\bar{u}^{n-1} + \bar{u}^{n-2}; \tag{2.8}$$

fourth-order:

$$\alpha_4 = \frac{25}{12}, \ A_4(u^n) = 4u^n - 3u^{n-1} + \frac{4}{3}u^{n-2} - \frac{1}{4}u^{n-3}, \ B_4(\bar{u}^n) = 4\bar{u}^n - 6\bar{u}^{n-1} + 4\bar{u}^{n-2} - \bar{u}^{n-3}. \tag{2.9}$$

fifth-order:

$$\alpha_5 = \frac{137}{60}, \quad A_5(u^n) = 5u^n - 5u^{n-1} + \frac{10}{3}u^{n-2} - \frac{5}{4}u^{n-3} + \frac{1}{5}u^{n-4}, \\ B_5(\bar{u}^n) = 5\bar{u}^n - 10\bar{u}^{n-1} + 10\bar{u}^{n-2} - 5\bar{u}^{n-3} + \bar{u}^{n-4}. \tag{2.10}$$

Several remarks are in order:

- Initialization: the second-order scheme can be initialized with a first-order scheme for the first step, the kth-order scheme can be initialized with a $k-1$th-order Runge–Kutta method for the first $k-1$ steps.
- We observe from (2.5b) that r^{n+1} is a first order approximation to $E(u(\cdot, t_{n+1}))$ which implies that ξ^{n+1} is a first order approximation to 1.
- we observe from (2.5a) and (2.5d) that

$$\frac{\alpha_k u^{n+1} - \eta_k^{n+1} A_k(u^n)}{\delta t} + \mathcal{A}u^{n+1} + \eta_k^{n+1} g[B_k(\bar{u}^n)] = 0, \tag{2.11}$$

  which, along with (2.5d), implies that

$$\frac{\alpha_k u^{n+1} - A_k(u^n)}{\delta t} + \mathcal{A}u^{n+1} + g[B_k(u^n)] = O(\delta t^k).$$

  Hence, both u^{n+1} and $\bar{u}^{n+1}$ are formally kth order approximations for $u(\cdot, t^{n+1})$.
- The main difference of the above scheme from the scheme in [22] is the choice of η_k^{n+1}, which can be considered as a special case in [22]. However, as we show below, this choice allows us to obtain a uniform bound on $(\mathcal{L}u^n, u^n)$, which in turn plays a crucial role in the error analysis. Another slight difference is here we use $g[B_k(\bar{u}^n)]$ in (2.5a), which makes the error analysis slightly easier, while $g[B_k(u^n)]$ is used in [22]. Thanks to (2.5d), this does not affect the kth order accuracy nor unconditional energy stability.
- Since the energy stability is achieved through only (2.5b), we can replace (2.5a) by other types of explicit-implicit multistep schemes.

The above scheme can be efficiently implemented as follows:

i. Obtain $\bar{u}^{n+1}$ from (2.5a) by solving an equation of the form

$$(\frac{\alpha_k}{\delta t} I + \mathcal{A})\bar{u}^{n+1} = f^{n+1},$$

   where f^{n+1} includes all known terms from previous time steps, and in most cases, this is a linear equation with constant coefficients;

ii. With $\bar{u}^{n+1}$ known, determine r^{n+1} explicitly from (2.5b);

iii. Compute ξ^{n+1}, η_k^{n+1} and u^{n+1} from (2.5d), goto the next step.

The main computational cost of this scheme is to solve (2.5a) once, while the main computational cost in the original SAV approach is to solve an equation similar to (2.5a) twice. So the cost of this scheme is about half of the original SAV approach while enjoying the same unconditional energy stability as we show below.

2.2. A stability result

We have the following results concerning the stability of the above schemes.

Theorem 1. *Given $r^n \geq 0$, we have $r^{n+1} \geq 0$, $\xi^{n+1} \geq 0$, and the scheme (2.5) for any k is unconditionally energy stable in the sense that*

$$r^{n+1} - r^n = -\delta t \xi^{n+1} \mathcal{K}(\bar{u}^{n+1}) \leq 0. \tag{2.12}$$

Furthermore, if $E(u) = \frac{1}{2}(\mathcal{L}u, u) + E_1(u)$ with $\mathcal{L}$ positive and $E_1(u)$ bounded from below, there exists $M_k > 0$ such that

$$(\mathcal{L}u^n, u^n) \leq M_k^2, \ \forall n. \tag{2.13}$$

Proof. Given $r^n \geq 0$ and since $E[\bar{u}^{n+1}] > 0$, it follows from (2.5b) that

$$r^{n+1} = \frac{r^n}{1 + \delta t \frac{\mathcal{K}(\bar{u}^{n+1})}{E[\bar{u}^{n+1}]}} \geq 0.$$

Then we derive from (2.5c) that $\xi^{n+1} \geq 0$ and obtain (2.12).

Denote $M := r^0 = E[u(\cdot, 0)]$, then (2.12) implies $r^n \leq M$, $\forall n$.

Without loss of generality, we can assume $E_1(u) > 1$ for all u. It then follows from (2.5c) that

$$|\xi^{n+1}| = \frac{r^{n+1}}{E(\bar{u}^{n+1})} \leq \frac{2M}{(\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1}) + 2}. \tag{2.14}$$

Let $\eta_k^{n+1} = 1 - (1 - \xi^{n+1})^{k+1}$, we have $\eta_k^{n+1} = \xi^{n+1} P_k(\xi^{n+1})$ with P_k being a polynomial of degree k. Then, we derive from (2.14) that there exists $M_k > 0$ such that

$$|\eta_k^{n+1}| = |\xi^{n+1} P_k(\xi^{n+1})| \leq \frac{M_k}{(\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1}) + 2},$$

which, along with $u^{n+1} = \eta_k^{n+1} \bar{u}^{n+1}$, implies

$$\begin{aligned}(\mathcal{L}u^{n+1}, u^{n+1}) &= (\eta_k^{n+1})^2 (\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1}) \\ &\leq \Big(\frac{M_k}{(\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1}) + 2}\Big)^2 (\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1}) \leq M_k^2.\end{aligned}$$

The proof is complete. □

Remark 1. From the above proof, we observe that it is essential to introduce $\bar{u}^{n+1}$ and η_k^{n+1} in order to obtain (2.13), and that the bound constant M_k increases as k increases. So while we can replace $k+1$ in η_k^{n+1} by any larger integer without affecting the kth order accuracy, it is best to use the smallest possible integer, which is $k+1$ for kth order accuracy.

Note that (2.5a) is uniquely solvable if $\mathcal{A}$ is a linear positive operator.

Remark 2. Note that the proof of (2.13) does not depend on specific form of (2.5a), so the result is also valid if we replace (2.5a) in the scheme by other implicit–explicit multistep schemes.

2.3. Numerical examples

Before we start the error analysis, we provide numerical examples to validate the convergence rates and demonstrate the advantage of our approach with the usual IMEX scheme.

Example 1. Consider the Allen–Cahn equation

$$\frac{\partial u}{\partial t} = \alpha \Delta u - (1 - u^2)u + f, \tag{2.15}$$

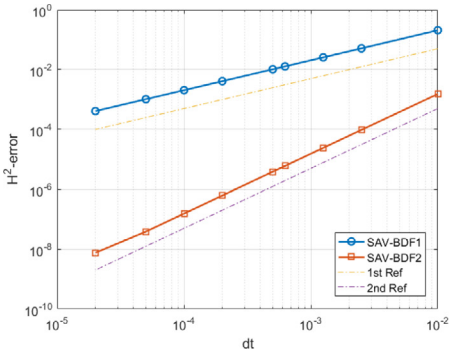
and the Cahn–Hilliard equation

$$\frac{\partial u}{\partial t} = -m_0 \Delta(\alpha \Delta u - (1 - u^2)u) + f, \tag{2.16}$$

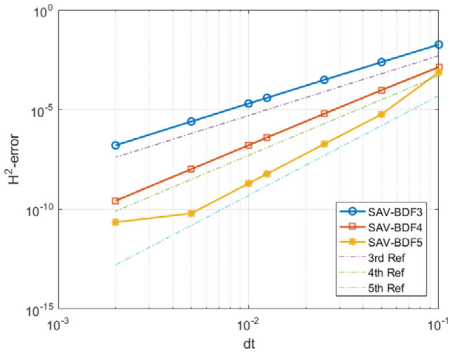
in $\Omega = (0, 2) \times (0, 2)$ with periodic boundary condition, and f is chosen such that the exact solution is

$$u(x, y, t) = \exp\big(\sin(\pi x)\sin(\pi y)\big)\sin(t). \tag{2.17}$$

We set $\alpha = 0.01^2$ in (2.15) and $\alpha = 0.04$, $m_0 = 0.005$ in (2.16), and use the Fourier spectral method with 64×64 modes for space discretization so that the spatial discretization error is negligible when compared with the time discretization error. In Figs. 1 (resp. 2), we plot the convergence rate of the H^2 error at $T = 1$ for the Allen–Cahn (resp. Cahn–Hilliard) equation. We observe the expected convergence rates for all cases.

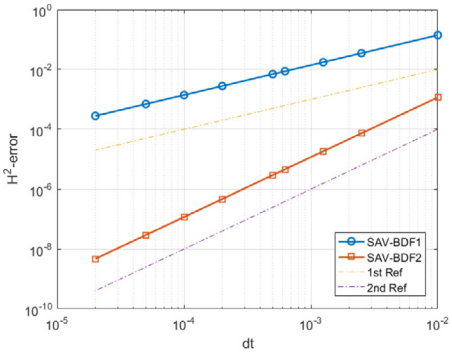


(a) BDF1 and BDF2 vs errors of u

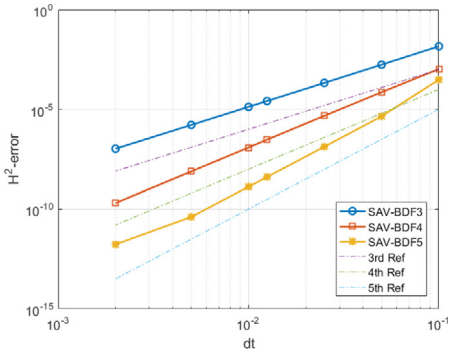


(b) BDF3, 4, 5 vs errors of u

Fig. 1. Convergence test for the Allen–Cahn equation using the new SAV/BDFk ($k = 1, 2, 3, 4, 5$). (a)–(b) H^2 errors of u as a function of Δt.



(a) BDF1 and BDF2 vs errors of u



(b) BDF3, 4, 5 vs errors of u

Fig. 2. Convergence test for the Cahn–Hilliard equation using the new SAV/BDFk ($k = 1, 2, 3, 4, 5$). (a)–(b) H^2 errors of u as a function of Δt.

Example 2. Next, we consider the 1-D Burgers equation

$$\frac{\partial u}{\partial t} - \nu u_{xx} + u u_x = 0, \tag{2.18}$$

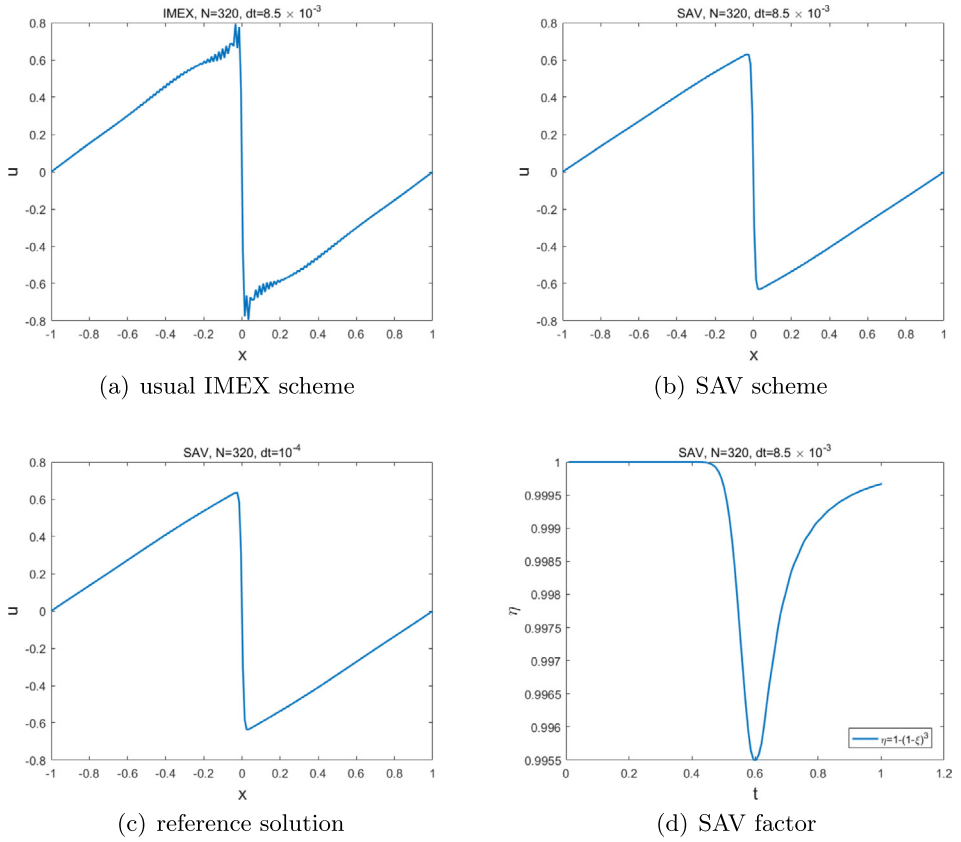
in $\Omega = (-1, 1)$ with the initial condition and Dirichlet boundary condition given as

$$u(x, 0) = -\sin(\pi x), \quad u(\pm 1, t) = 0. \tag{2.19}$$

In this test, we use the second order SAV scheme and the corresponding second-order IMEX scheme with $\nu = \frac{1}{314}$, $N = 320$, $\delta t = 8.5 \times 10^{-3}$. The numerical solutions at $T = 1$ are plotted in Fig. 3(a) solution obtained by the usual IMEX scheme and (b) solution obtained by the SAV scheme. We observe that the usual IMEX scheme produces oscillatory solutions while the SAV scheme produces the correct solution which is indistinguishable with the reference solution obtained with $\delta t = 10^{-4}$ in 3(c). We also plot in 3(d) the SAV factor $\eta^n = 1 - (1 - \xi^n)^3$. We observe that when the solution exhibits large gradients (for $t \in (0.5, 1)$), the SAV factor η^n deviates slightly from 1 so that the SAV scheme still produces correct result while the corresponding IMEX scheme produces incorrect result.

3. Error analysis for Allen-Cahn type equations

While the stability results in Theorem 1 are valid for general dissipative systems, it is cumbersome to carry out error analysis with such generality. So to simplify the presentation, we shall carry out error analysis for two class



(a) usual IMEX scheme (b) SAV scheme

(c) reference solution (d) SAV factor

Fig. 3. Burgers equation: a comparison of usual IMEX and SAV.

of typical semi-linear equations: Allen–Cahn type equation in this section and Cahn–Hilliard type equation in the next section.

We first recall the following important result. Based on Dahlquist's G-stability theory, Nevanlinna and Odeh [23] proved the following results for BDFk ($1 \le k \le 5$) schemes.

Lemma 1. *For $1 \le k \le 5$, there exist $0 \le \tau_k < 1$, a positive definite symmetric matrix $G = (g_{ij}) \in \mathcal{R}^{k,k}$ and real numbers $\delta_0, \ldots, \delta_k$ such that*

$$\Big(\alpha_k u^{n+1} - A_k(u^n),\, u^{n+1} - \tau_k u^n\Big) = \sum_{i,j=1}^{k} g_{ij}(u^{n+1+i-k},\, u^{n+1+j-k}) - \sum_{i,j=1}^{k} g_{ij}(u^{n+i-k},\, u^{n+j-k}) + \|\sum_{i=0}^{k} \delta_i u^{n+1+i-k}\|^2,$$

where the smallest possible values of τ_k are

$$\tau_1 = \tau_2 = 0, \quad \tau_3 = 0.0836, \quad \tau_4 = 0.2878, \quad \tau_5 = 0.8160,$$

and α_k, A_k are defined in (2.8)–(2.10).

The above result played a key role in proving the stability of high-order BDF schemes for nonlinear parabolic equations [24], and it plays an important role in our error analysis.

We shall also frequently use the following discrete Gronwall Lemma (see for example, [25], Lemma B.10).

Lemma 2 (*Discrete Gronwall Lemma*). *Let y^k, h^k, g^k, f^k be four nonnegative sequences satisfying*

$$y^n + \delta t \sum_{k=0}^{n} h^k \le B + \delta t \sum_{k=0}^{n} (g^k y^k + f^k) \;\; with \;\; \delta t \sum_{k=0}^{T/\delta t} g^k \le M, \; \forall\, 0 \le n \le T/\delta t.$$

We assume $\delta t\, g^k < 1$ and let $\sigma = \max_{0\le k\le T/\delta t}(1-\delta t g^k)^{-1}$. Then

$$y^n + \delta t \sum_{k=1}^{n} h^k \le \exp(\sigma M)(B + \delta t \sum_{k=0}^{n} f^k), \;\; \forall\, n \le T/\delta t.$$

Consider the Allen–Cahn type equation:

$$\frac{\partial u}{\partial t} - \Delta u + \lambda u - g(u) = 0 \quad (\boldsymbol{x}, t) \in \Omega \times (0, T], \tag{3.1}$$

where Ω is an open bounded domain in $\mathbb{R}^d$ ($d = 1, 2, 3$), with the initial condition $u(x, 0) = u^0(x)$, and boundary condition:

$$periodic, \; or \; u|_{\partial\Omega} = 0, \; or \; \frac{\partial u}{\partial \boldsymbol{n}}|_{\partial\Omega} = 0. \tag{3.2}$$

The above equation is a special case of (2.1) with $\mathcal{A} = -\Delta + \lambda I$, and satisfies the dissipation law (2.2) with $E(u) = \frac{1}{2}(\mathcal{L}u, u) + (G(u), 1)$ where $(\mathcal{L}u, u) = (\nabla u, \nabla u) + \lambda(u, u)$, $G(u) = \int^u g(v)dv$ and $\mathcal{K}(u) = (\frac{\delta E}{\delta u}, \frac{\delta E}{\delta u})$. We assume, without loss of generality,

$$\int_\Omega G(v)dx \ge \underline{C} > 0 \quad \forall v. \tag{3.3}$$

In particular, with $g(u) = (1-u^2)u$ and $\lambda = 0$, the above equation becomes the celebrated Allen–Cahn equation [26].

We recall the following regularity result for (3.1) (see, for instance, [27]).

Theorem 2. *Assume $u^0 \in H^2(\Omega)$ and the following holds*

$$|g'(x)| < C(|x|^p + 1), \quad p > 0 \; arbitrary \quad if \; d = 1, 2; \quad 0 < p < 4 \quad if \; d = 3. \tag{3.4}$$

Then for any $T > 0$, the problem (3.1) has a unique solution in the space

$$C([0, T]; H^2(\Omega)) \cap L^2(0, T; H^3(\Omega)).$$

We also recall a result (see Lemma 2.3 in [12]) which is useful to deal with the nonlinear term in (3.1).

Lemma 3. *Assume that $\|u\|_{H^1} \le M$ and (3.4) holds. Then for any $u \in H^3$, there exist $0 \le \sigma < 1$ and a constant $C(M)$ such that the following inequality holds:*

$$\|\nabla g(u)\|^2 \le C(M)(1 + \|\nabla \Delta u\|^{2\sigma}).$$

We denote hereafter

$$t^n = n\,\delta t, \; \bar{e}^n = \bar{u}^n - u(\cdot, t^n), \; e^n = u^n - u(\cdot, t^n), \; s^n = r^n - r(t^n).$$

In the following, we carry out a unified error analysis for the first- to fifth-order SAV schemes described as in (2.5) with the coefficients defined in (2.8)–(2.10).

For (3.1), the kth-order version of (2.5a) and (2.11) read:

$$\frac{\alpha_k \bar{u}^{n+1} - A_k(u^n)}{\delta t} = \Delta \bar{u}^{n+1} - \lambda \bar{u}^{n+1} + g[B_k(\bar{u}^n)], \tag{3.5}$$

$$\frac{\alpha_k u^{n+1} - \eta_k^{n+1} A_k(u^n)}{\delta t} = \Delta u^{n+1} - \lambda u^{n+1} + \eta_k^{n+1} g[B_k(\bar{u}^n)], \tag{3.6}$$

where α_k, A_k, B_k defined in (2.8)–(2.10).

Theorem 3. *Given initial condition $\bar{u}^0 = u^0 = u(0)$, $r^0 = E[u^0]$. Let $\bar{u}^{n+1}$ and u^{n+1} be computed with the kth order scheme (2.5a)–(2.5d) $(1 \le k \le 5)$ for (3.1) with*

$$\eta_1^{n+1} = 1 - (1 - \xi^{n+1})^3, \quad \eta_k^{n+1} = 1 - (1 - \xi^{n+1})^{k+1} \ (k = 2, 3, 4, 5).$$

We assume (3.4) holds and

$$u^0 \in H^3, \ \frac{\partial^j u}{\partial t^j} \in L^2(0, T; H^1) \ 1 \le j \le k + 1.$$

Then for $n + 1 \le T/\delta t$ and $\delta t < \min\{\frac{1}{1+2C_0^{k+2}}, \frac{1-\tau_k}{3(k+1)}\}$, we have

$$\|\bar{e}^{n+1}\|_{H^2}, \ \|e^{n+1}\|_{H^2} \le C\delta t^k,$$

where the constants C_0, C are dependent on T, Ω, the $k \times k$ matrix $G = (g_{ij})$ in Lemma 1 and the exact solution u but are independent of δt and $0 \le \tau_k < 1$ is the constant in Lemma 1.

Proof. We assume that $\bar{u}^i$ and u^i $(i = 1, \ldots, k-1)$ are computed with a proper initialization procedure such that $\|\bar{u}^i - u(t_i)\|_{H^2} = O(\delta t^k)$ and $\|u^i - u(t_i)\|_{H^2} = O(\delta t^k)$ $(i = 1, \ldots, k-1)$. To simplify the presentation, we set $\bar{u}^i = u^i = u(t_i)$ and $r^i = E_1[u^i]$ for $i = 1, \ldots, k-1$.

The main task is to prove

$$|1 - \xi^q| \le C_0 \, \delta t, \ \forall q \le T/\delta t. \tag{3.7}$$

where the constant C_0 is dependent on T, Ω and the exact solution u but is independent of δt, and will be defined in the proof process. Below we shall prove (3.7) by induction.

Under the assumption, (3.7) certainly holds for $q = 0$. Now suppose we have

$$|1 - \xi^q| \le C_0 \, \delta t, \ \forall q \le m, \tag{3.8}$$

we shall prove below

$$|1 - \xi^{m+1}| \le C_0 \delta t. \tag{3.9}$$

We shall first consider $k = 2, 3, 4, 5$, and point out the necessary modifications for the case $k = 1$ later.

Step 1: H^2 bound for u^n and $\bar{u}^n$ for all $n \le m$. For the kth-order schemes, it follows from Theorem 1 that

$$\|u^q\|_{H^1} \le M_k, \ \forall q \le T/\delta t. \tag{3.10}$$

Under assumption (3.8), if we choose δt small enough such that

$$\delta t \le \min\{\frac{1}{2C_0^{k+1}}, 1\}, \tag{3.11}$$

we have

$$1 - \frac{\delta t^k}{2} \le |\eta_k^q| \le 1 + \frac{\delta t^k}{2}, \ |1 - \eta_k^q| \le \frac{\delta t^k}{2}, \forall q \le m, \tag{3.12}$$

and

$$\|\bar{u}^q\|_{H^1} \le 2M_k, \ \forall q \le m, \ \forall \delta t \le 1. \tag{3.13}$$

Consider (3.6) in step q:

$$\frac{\alpha_k u^q - \eta_k^q A_k(u^{q-1})}{\delta t} = \Delta u^q - \lambda u^q + \eta_k^q g[B_k(\bar{u}^{q-1})]. \tag{3.14}$$

Thanks to Lemma 3 and (3.12), we have

$$\begin{aligned}\|\nabla g[B_k(\bar{u}^{q-1})]\|^2 &\le C(M_k)(\|\nabla\Delta B_k(\bar{u}^{q-1})\|^{2\sigma}+1)\\ &\le \bar{\gamma}_k\|\nabla\Delta B_k(\bar{u}^{q-1})\|^2 + C(M_k,\bar{\gamma}_k)\\ &\le \bar{\gamma}_k\|\nabla\Delta B_k(\frac{1}{\eta_k^{q-1}}u^{q-1})\|^2 + C(M_k,\bar{\gamma}_k)\\ &\le 40\bar{\gamma}_k\sum_{i=1}^{k}\|\nabla\Delta u^{q-i}\|^2 + C(M_k,\bar{\gamma}_k)\\ &\le \gamma_k\sum_{i=1}^{k}\|\nabla\Delta u^{q-i}\|^2 + C(M_k,\gamma_k),\end{aligned}$$

where $\bar{\gamma}_k$ can be any positive constant and the constant 40 comes from the coefficients in B_k. To simplify the notation, we let $\gamma_k = 40\bar{\gamma}_k$. Taking the inner product of (3.14) with $\Delta^2u^q - \tau_k\Delta^2u^{q-1}$ and using the above inequality, it follows from Lemma 1 that there exist $0 \le \tau_k < 1$, a positive definite symmetric matrix $G = (g_{ij}) \in \mathcal{R}^{k,k}$ and $\delta_0,\dots,\delta_k$ such that

$$\begin{aligned}&\frac{1}{\delta t}\Big(\sum_{i,j=1}^{k} g_{ij}(\Delta u^{q+i-k},\Delta u^{q+j-k}) - \sum_{i,j=1}^{k} g_{ij}(\Delta u^{q-1+i-k},\Delta u^{q-1+j-k}) + \|\sum_{i=0}^{k}\delta_i\Delta u^{q+i-k}\|^2\Big)\\ &\quad+\frac{1}{2}\|\nabla\Delta u^q\|^2+\frac{\lambda}{2}\|\Delta u^q\|^2\\ &\quad\le \eta_k^q\big(g[B_k(\bar{u}^{q-1})],\Delta^2u^q-\tau_k\Delta^2u^{q-1}\big)+\frac{\tau_k}{2}\|\nabla\Delta u^{q-1}\|^2+\frac{\lambda\tau_k}{2}\|\Delta u^{q-1}\|^2\\ &\quad+\frac{(\eta_k^q-1)}{\delta t}\big(A_k(u^{q-1}),\Delta^2u^q-\tau_k\Delta^2u^{q-1}\big)\\ &\quad\le C(\varepsilon_k)|\eta_k^q|\|\nabla g[B_k(\bar{u}^{q-1})]\|^2+\varepsilon_k|\eta_k^q|\big(\|\nabla\Delta u^q\|^2+\|\nabla\Delta u^{q-1}\|^2\big)+\frac{\tau_k}{2}\|\nabla\Delta u^{q-1}\|^2\\ &\quad+\frac{\lambda\tau_k}{2}\|\Delta u^{q-1}\|^2+\frac{|1-\eta_k^q|}{\delta t}\|\nabla A_k(u^{q-1})\|^2+\frac{|1-\eta_k^q|}{\delta t}\big(\|\nabla\Delta u^q\|^2+\|\nabla\Delta u^{q-1}\|^2\big)\\ &\quad\le C(M_k,\varepsilon_k,\gamma_k)+\big(C(\varepsilon_k)|\eta_k^q|\gamma_k+\varepsilon_k|\eta_k^q|+\frac{|1-\eta_k^q|}{\delta t}\big)\sum_{i=0}^{k}\|\nabla\Delta u^{q-i}\|^2\\ &\quad+\frac{\tau_k}{2}\|\nabla\Delta u^{q-1}\|^2+\frac{\lambda\tau_k}{2}\|\Delta u^{q-1}\|^2+\frac{|1-\eta_k^q|}{\delta t}\|\nabla A_k(u^{q-1})\|^2,\end{aligned}\tag{3.15}$$

where ε_k can be any positive constant. After taking the sum on (3.15), we are supposed to choose suitable δt, ε_k and γ_k such that

$$\frac{1}{2}\sum_{q=1}^{n}\big(\|\nabla\Delta u^q\|^2-\tau_k\|\nabla\Delta u^{q-1}\|^2\big)+C_I \ge \sum_{q=1}^{n}\Big(\big(C(\varepsilon_k)|\eta_k^q|\gamma_k+\varepsilon_k|\eta_k^q|+\frac{|1-\eta_k^q|}{\delta t}\big)\sum_{i=0}^{k}\|\nabla\Delta u^{q-i}\|^2\Big)\tag{3.16}$$

with C_I is a constant only depending on the initial data. Note that $0 \le \tau_k < 1$ and we first consider $k \ge 2$, we can choose δt, ε_k and γ_k small enough such that

$$\delta t < \frac{1-\tau_k}{3(k+1)},\quad \varepsilon_k < \frac{1-\tau_k}{12(k+1)},\quad \gamma_k < \frac{1-\tau_k}{12(k+1)C(\varepsilon_k)},\tag{3.17}$$

with the estimate in (3.12), we have

$$\begin{aligned}C(\varepsilon_k)|\eta_k^q|\gamma_k+\varepsilon_k|\eta_k^q|+\frac{|1-\eta_k^q|}{\delta t} &\le 2C(\varepsilon_k)\gamma_k+2\varepsilon_k+\frac{\delta t^{k-1}}{2}\\ &\le \frac{1-\tau_k}{6(k+1)}+\frac{1-\tau_k}{6(k+1)}+\frac{1-\tau_k}{6(k+1)}\\ &\le \frac{1-\tau_k}{2(k+1)}.\end{aligned}\tag{3.18}$$

Then, taking the sum on (3.15) for q from k to n ($\leq m$), we obtain

$$\begin{aligned}\sum_{i,j=1}^{k} g_{ij}(\Delta u^{n+i-k}, \Delta u^{n+j-k}) &\leq C(M_k, \tau_k)T + C(u^0, \ldots, u^{k-1}) + C_{A_k} k\delta t^k \sum_{q=0}^{n-1} \|\nabla u^q\|^2 \\ &\leq C(M_k, \tau_k)T + C(u^0, \ldots, u^{k-1}) + C_{A_k} k\delta t^{k-1} T M_k^2,\end{aligned}$$

where $C(M_k, \tau_k)$ is a constant only depends on M_k, τ_k, $C(u^0, \ldots, u^{k-1})$ only depends on $u^0, \ldots, u^{k-1}$ and C_{A_k} only depends on the coefficients in A_k. Since $G = (g_{ij})$ is a positive definite symmetric matrix, we have

$$\begin{aligned}\lambda_G \|\Delta u^n\|^2 &\leq \sum_{i,j=1}^{k} g_{ij}(\Delta u^{n+i-k}, \Delta u^{n+j-k}) \\ &\leq C(M_k, \tau_k)T + C(u^0, \ldots, u^{k-1}) + C_{A_k} k\delta t^{k-1} T M_k^2.\end{aligned}$$

where $\lambda_G > 0$ is the minimum eigenvalue of $G = (g_{ij})$. Together with (3.10), the above implies

$$\|u^n\|_{H^2} \leq \frac{1}{\lambda_G}\sqrt{C(M_k, \tau_k)T + C(u^0, \ldots, u^{k-1}) + C_{A_k} kT M_k^2} + M_k := C_1, \ \forall \delta t < 1, \ n \leq m. \tag{3.19}$$

Noting that

$$\|u^n\|_{H^2} = |\eta_k^n| \|\bar{u}^n\|_{H_2},$$

then (3.12) implies

$$\|\bar{u}^n\|_{H^2} \leq 2C_1, \ \forall \delta t < 1, \ n \leq m. \tag{3.20}$$

Step 2: estimate for $\|\bar{e}^{n+1}\|_{H^2}$ for all $0 \leq n \leq m$. By Theorem 2 and (3.20) we can choose C large enough such that

$$\|u(t)\|_{H^2} \leq C, \ \forall t \leq T, \ \|\bar{u}^q\|_{H^2} \leq C, \ \forall q \leq m. \tag{3.21}$$

Since $H^2 \subset L^\infty$, without loss of generality, we can adjust C such that

$$|g^{(i)}[u(t)]|_{L^\infty} \leq C, \ \forall t \leq T; \ |g^{(i)}(\bar{u}^q)|_{L^\infty} \leq C, \ \forall q \leq m, \ i = 0, 1, 2. \tag{3.22}$$

From (3.5), we can write down the error equation as

$$\alpha_k \bar{e}^{n+1} - A_k(\bar{e}^n) = A_k(u^n) - A_k(\bar{u}^n) + \delta t \Delta \bar{e}^{n+1} - \delta t \lambda \bar{e}^{n+1} + R_k^n + \delta t Q_k^n, \tag{3.23}$$

where R_k^n, Q_k^n are given by

$$\begin{aligned}R_k^n &= -\alpha_k u(t^{n+1}) + A_k(u(t^n)) + \delta t u_t(t^{n+1}) \\ &= \sum_{i=1}^{k} a_i \int_{t^{n+1-i}}^{t^{n+1}} (t^{n+1-i} - s)^k \frac{\partial^{k+1} u}{\partial t^{k+1}}(s) ds,\end{aligned} \tag{3.24}$$

with a_i being some fixed and bounded constants determined by the truncation errors, and

$$Q_k^n = g[B_k(\bar{u}^n)] - g[u(t^{n+1})]. \tag{3.25}$$

For example, in the case $k = 3$, we have

$$R_3^n = -3\int_{t^n}^{t^{n+1}} (t^n - s)^3 \frac{\partial^4 u}{\partial t^4}(s) ds + \frac{3}{2}\int_{t^{n-1}}^{t^{n+1}} (t^{n-1} - s)^3 \frac{\partial^4 u}{\partial t^4}(s) ds - \frac{1}{3}\int_{t^{n-2}}^{t^{n+1}} (t^{n-2} - s)^3 \frac{\partial^4 u}{\partial t^4}(s) ds.$$

Taking the inner product of (3.23) with $\bar{e}^{n+1} - \tau_k \bar{e}^n$, it follows from Lemma 1 that

$$\begin{aligned}
&\sum_{i,j=1}^{k} g_{ij}(\bar{e}^{n+1+i-k}, \bar{e}^{n+1+j-k}) - \sum_{i,j=1}^{k} g_{ij}(\bar{e}^{n+i-k}, \bar{e}^{n+j-k}) \\
&\quad + \|\sum_{i=0}^{k} \delta_i \bar{e}^{n+1+i-k}\|^2 + \delta t \|\nabla \bar{e}^{n+1}\|^2 + \lambda \delta t \|\bar{e}^{n+1}\|^2 \\
&\quad = (A_k(u^n) - A_k(\bar{u}^n), \bar{e}^{n+1} - \tau_k \bar{e}^n) - \delta t(\Delta \bar{e}^{n+1}, \tau_k \bar{e}^n) + \delta t \lambda(\bar{e}^{n+1}, \tau_k \bar{e}^n) \\
&\quad + (R_k^n, \bar{e}^{n+1} - \tau_k \bar{e}^n) + \delta t(Q_k^n, \bar{e}^{n+1} - \tau_k \bar{e}^n).
\end{aligned} \tag{3.26}$$

In the following, we bound the right hand side of (3.26). Note that

$$u^q = \eta_k^q \bar{u}^q, \quad |\eta_k^q - 1| \le C_0^{k+1} \delta t^{k+1}, \quad \forall q \le n.$$

Hence

$$\begin{aligned}
|(A_k(u^n) - A_k(\bar{u}^n), \bar{e}^{n+1} - \tau_k \bar{e}^n)| &\le \frac{\|A_k(u^n) - A_k(\bar{u}^n)\|^2}{2\delta t} + \frac{\delta t}{2} \|\bar{e}^{n+1} - \tau_k \bar{e}^n\|^2 \\
&\le C C_0^{2k+2} \delta t^{2k+1} + \delta t \|\bar{e}^{n+1}\|^2 + \delta t \|\bar{e}^n\|^2.
\end{aligned} \tag{3.27}$$

It follows from (3.24) that

$$\|R_k^n\|^2 \le C \delta t^{2k+1} \int_{t^{n+1-k}}^{t^{n+1}} \left\| \frac{\partial^{k+1} u}{\partial t^{k+1}}(s) \right\|^2 ds. \tag{3.28}$$

And we can bound Q_k^n based on (3.22) and (3.25) as

$$\begin{aligned}
|Q_k^n| &= \left| g[B_k(\bar{u}^n)] - g[B_k(u(t^n))] + g[B_k(u(t^n))] - g[u(t^{n+1})] \right| \\
&\le C|B_k(\bar{e}^n)| + C|B_k(u(t^n)) - u(t^{n+1})| \\
&= C|B_k(\bar{e}^n)| + C\left| \sum_{i=1}^{k} b_i \int_{t^{n+1-i}}^{t^{n+1}} (t^{n+1-i} - s)^{k-1} \frac{\partial^k u}{\partial t^k}(s) ds \right|,
\end{aligned} \tag{3.29}$$

where b_i are some fixed and bounded constants determined by the truncation error. For example, in the case $k = 3$, we have

$$\begin{aligned}
B_3(u(t^n)) - u(t^{n+1}) = &-\frac{3}{2} \int_{t^n}^{t^{n+1}} (t^n - s)^2 \frac{\partial^3 u}{\partial t^3}(s) ds + \frac{3}{2} \int_{t^{n-1}}^{t^{n+1}} (t^{n-1} - s)^2 \frac{\partial^3 u}{\partial t^3} ds \\
&- \frac{1}{2} \int_{t^{n-2}}^{t^{n+1}} (t^{n-2} - s)^2 \frac{\partial^3 u}{\partial t^3} ds.
\end{aligned}$$

Therefore,

$$\begin{aligned}
|\left(R_k^n, \bar{e}^{n+1} - \tau_k \bar{e}^n\right)| &\le \frac{1}{2\delta t} \|R_k^n\|^2 + \delta t \|\bar{e}^{n+1}\|^2 + \delta t \|\bar{e}^n\|^2, \\
&\le \delta t \|\bar{e}^{n+1}\|^2 + \delta t \|\bar{e}^n\|^2 + C \delta t^{2k} \int_{t^{n+1-k}}^{t^{n+1}} \left\| \frac{\partial^{k+1} u}{\partial t^{k+1}}(s) \right\|^2 ds.
\end{aligned} \tag{3.30}$$

$$\delta t |\left(Q_k^n, \bar{e}^{n+1} - \tau_k \bar{e}^n\right)| \le C \delta t \left(\|B_k(\bar{e}^n)\|^2 + \|\bar{e}^{n+1}\|^2 + \|\bar{e}^n\|^2 \right) + C \delta t^{2k} \int_{t^{n+1-k}}^{t^{n+1}} \left\| \frac{\partial^k u}{\partial t^k}(s) \right\|^2 ds. \tag{3.31}$$

Now, combining (3.26), (3.27), (3.30), (3.31), we arrive at

$$\begin{aligned}
&\sum_{i,j=1}^{k} g_{ij}(\bar{e}^{n+1+i-k}, \bar{e}^{n+1+j-k}) - \sum_{i,j=1}^{k} g_{ij}(\bar{e}^{n+i-k}, \bar{e}^{n+j-k}) \\
&\quad + \|\sum_{i=0}^{k} \delta_i \bar{e}^{n+1+i-k}\|^2 + \frac{1}{2}\delta t\|\nabla \bar{e}^{n+1}\|^2 + \frac{\lambda}{2}\delta t\|\bar{e}^{n+1}\|^2 \\
&\quad \le \frac{\tau_k}{2}\delta t\|\nabla \bar{e}^{n}\|^2 + \frac{\lambda\tau_k}{2}\delta t\|\bar{e}^{n}\|^2 + CC_0^{2k+2}\delta t^{2k+1} + C\delta t\sum_{i=0}^{k}\|\bar{e}^{n+1-i}\|^2 \\
&\quad + C\delta t^{2k}\int_{t^{n+1-k}}^{t^{n+1}} (\left\|\frac{\partial^k u}{\partial t^k}(s)\right\|^2 + \left\|\frac{\partial^{k+1} u}{\partial t^{k+1}}(s)\right\|^2) ds.
\end{aligned}$$

Taking the sum of the above for n from $k-1$ to m, noting that $G = (g_{ij})$ is a positive definite symmetric matrix with minimum eigenvalue λ_G, we obtain:

$$\begin{aligned}
\lambda_G\|\bar{e}^{m+1}\|^2 &\le \sum_{i,j=1}^{k} g_{ij}(\bar{e}^{m+1+i-k}, \bar{e}^{m+1+j-k}) \\
&\le C\delta t\sum_{q=0}^{m+1}\|\bar{e}^{q}\|^2 + C\delta t^{2k}\int_0^T (\left\|\frac{\partial^k u}{\partial t^k}(s)\right\|^2 + \|\frac{\partial^{k+1} u}{\partial t^{k+1}}(s)\|^2 + C_0^{2k+2}) ds
\end{aligned} \tag{3.32}$$

We can obtain similar inequalities for $\|\nabla \bar{e}^m\|$ and $\|\Delta \bar{e}^m\|$ by using essentially the same procedure. Indeed, taking the inner product of (3.23) with $-\Delta \bar{e}^{n+1} + \tau_k \Delta \bar{e}^n$, by using Lemma 1, we obtain

$$\begin{aligned}
&\sum_{i,j=1}^{k} g_{ij}(\nabla \bar{e}^{n+1+i-k}, \nabla \bar{e}^{n+1+j-k}) - \sum_{i,j=1}^{k} g_{ij}(\nabla \bar{e}^{n+i-k}, \nabla \bar{e}^{n+j-k}) + \|\sum_{i=0}^{k}\delta_i \nabla \bar{e}^{n+1+i-k}\|^2 \\
&+ \delta t\|\Delta \bar{e}^{n+1}\|^2 + \lambda\delta t\|\nabla \bar{e}^{n+1}\|^2 \\
&= (\nabla A_k(u^n) - \nabla A_k(\bar{u}^n), \nabla \bar{e}^{n+1} - \tau_k \nabla \bar{e}^n) + \delta t(\Delta \bar{e}^{n+1}, \tau_k \Delta \bar{e}^n) + \delta t\lambda(\nabla \bar{e}^{n+1}, \tau_k \nabla \bar{e}^n) \\
&+ (R_k^n, -\Delta \bar{e}^{n+1} + \tau_k \Delta \bar{e}^n) + \delta t(Q_k^n, -\Delta \bar{e}^{n+1} + \tau_k \Delta \bar{e}^n).
\end{aligned} \tag{3.33}$$

Taking the sum of the above for n from $k-1$ to m, using Lemma 1, (3.28) and (3.29), we can obtain

$$\begin{aligned}
\lambda_G\|\nabla \bar{e}^{m+1}\|^2 &\le \sum_{i,j=1}^{k} g_{ij}(\nabla \bar{e}^{m+1+i-k}, \nabla \bar{e}^{m+1+j-k}) \\
&\le C\delta t\sum_{q=0}^{m+1}\|\nabla \bar{e}^{q}\|^2 + C\delta t^{2k}\int_0^T (\left\|\frac{\partial^k u}{\partial t^k}(s)\right\|^2 + \left\|\frac{\partial^{k+1} u}{\partial t^{k+1}}(s)\right\|^2 + C_0^{2k+2}) ds.
\end{aligned} \tag{3.34}$$

On the other hand, taking the inner product of (3.23) with $\Delta^2 \bar{e}^{n+1} - \tau_k \Delta^2 \bar{e}^n$, by using Lemma 1, we obtain

$$\begin{aligned}
&\sum_{i,j=1}^{k} g_{ij}(\Delta \bar{e}^{n+1+i-k}, \Delta \bar{e}^{n+1+j-k}) - \sum_{i,j=1}^{k} g_{ij}(\Delta \bar{e}^{n+i-k}, \Delta \bar{e}^{n+j-k}) + \|\sum_{i=0}^{k}\delta_i \Delta \bar{e}^{n+1+i-k}\|^2 \\
&+ \delta t\|\nabla \Delta \bar{e}^{n+1}\|^2 + \lambda\delta t\|\Delta \bar{e}^{n+1}\|^2 \\
&= (\Delta A_k(u^n) - \Delta A_k(\bar{u}^n), \Delta \bar{e}^{n+1} - \tau_k \Delta \bar{e}^n) + \delta t(\nabla \Delta \bar{e}^{n+1}, \tau_k \nabla \Delta \bar{e}^n) + \delta t\lambda(\Delta \bar{e}^{n+1}, \tau_k \Delta \bar{e}^n) \\
&+ (\nabla R_k^n, -\nabla \Delta \bar{e}^{n+1} + \tau_k \nabla \Delta \bar{e}^n) + \delta t(\nabla Q_k^n, -\nabla \Delta \bar{e}^{n+1} + \tau_k \nabla \Delta \bar{e}^n).
\end{aligned} \tag{3.35}$$

Here, we need to pay attention to the terms with $\nabla \Delta \bar{e}^{n+1}$ or $\nabla \Delta \bar{e}^n$. Firstly, we have

$$|\delta t(\nabla \Delta \bar{e}^{n+1}, \tau_k \nabla \Delta \bar{e}^n)| \le \frac{\delta t}{2}\|\nabla \Delta \bar{e}^{n+1}\|^2 + \frac{\tau_k^2 \delta t}{2}\|\nabla \Delta \bar{e}^n\|^2.$$

It follows from (3.24) and (3.25) that

$$\|\nabla R_k^n\|^2 \le C\delta t^{2k+1}\int_{t^{n+1-k}}^{t^{n+1}} \|\nabla \frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\|^2 ds, \tag{3.36}$$

and

$$\begin{aligned} |\nabla Q_k^n| &\le C(|B_k(\bar e^n)| + |\nabla B_k(\bar e^n)|) + C\Big|\sum_{i=1}^k b_i \int_{t^{n+1-i}}^{t^{n+1}} (t^{n+1-i}-s)^{k-1}\frac{\partial^k u}{\partial t^k}(s)ds\Big| \\ &+ C\Big|\sum_{i=1}^k b_i \int_{t^{n+1-i}}^{t^{n+1}} (t^{n+1-i}-s)^{k-1}\nabla\frac{\partial^k u}{\partial t^k}(s)ds\Big|. \end{aligned} \tag{3.37}$$

Therefore,

$$\begin{aligned} |(\nabla R_k^n, -\nabla\Delta\bar e^{n+1} + \tau_k\nabla\Delta\bar e^n)| &\le \frac{C}{\delta t}\|\nabla R_k^n\|^2 + \frac{\delta t(1-\tau_k^2)}{16}\|-\nabla\Delta\bar e^{n+1} + \tau_k\nabla\Delta\bar e^n\|^2 \\ &\le C\delta t^{2k}\int_{t^{n+1-k}}^{t^{n+1}} \Big\|\nabla\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\Big\|^2 ds + \frac{(1-\tau_k^2)\delta t}{8}(\|\nabla\Delta\bar e^{n+1}\|^2 + \|\nabla\Delta\bar e^n\|^2), \end{aligned}$$

and

$$\begin{aligned} \delta t|(\nabla Q_k^n, -\nabla\Delta\bar e^{n+1} + \tau_k\nabla\Delta\bar e^n)| &\le C\delta t\|\nabla Q_k^n\|^2 + \frac{(1-\tau_k^2)\delta t}{16}\|-\nabla\Delta\bar e^{n+1} + \tau_k\nabla\Delta\bar e^n\|^2 \\ &\le C\delta t\|B_k(\bar e^n)\|_{H^1}^2 + C\delta t^{2k}\int_{t^{n+1-k}}^{t^{n+1}} \Big\|\frac{\partial^k u}{\partial t^k}(s)\Big\|_{H^1}^2 ds \\ &+ \frac{(1-\tau_k^2)\delta t}{8}(\|\nabla\Delta\bar e^{n+1}\|^2 + \|\nabla\Delta\bar e^n\|^2). \end{aligned}$$

We can bound other terms on the right hand side of (3.35) as before to arrive at

$$\begin{aligned} &\sum_{i,j=1}^k g_{ij}(\Delta\bar e^{n+1+i-k}, \Delta\bar e^{n+1+j-k}) - \sum_{i,j=1}^k g_{ij}(\Delta\bar e^{n+i-k}, \Delta\bar e^{n+j-k}) \\ &+ \frac{(1+\tau_k^2)\delta t}{4}\|\nabla\Delta\bar e^{n+1}\|^2 + \frac{\lambda\delta t}{2}\|\Delta\bar e^{n+1}\|^2 \\ &\le C\delta t(\|B_k(\bar e^n)\|_{H^1}^2 + \|\Delta\bar e^{n+1}\|^2 + \|\Delta\bar e^n\|^2) + \frac{(1+\tau_k^2)\delta t}{4}\|\nabla\Delta\bar e^n\|^2 + \frac{\lambda\tau_k^2\delta t}{2}\|\Delta\bar e^n\|^2 \\ &+ C\delta t^{2k}\int_{t^{n+1-k}}^{t^{n+1}} (\Big\|\frac{\partial^k u}{\partial t^k}(s)\Big\|_{H^1}^2 + \Big\|\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\Big\|_{H^1}^2 + C_0^{2k+2})ds. \end{aligned}$$

Then, taking the sum of the above for n from $k-1$ to m, we obtain

$$\begin{aligned} \lambda_G\|\Delta\bar e^{m+1}\|^2 &\le \sum_{i,j=1}^k g_{ij}(\Delta\bar e^{m+1+i-k}, \Delta\bar e^{m+1+j-k}) \\ &\le C\delta t\sum_{q=0}^{m+1}\|\bar e^q\|_{H^2}^2 + C\delta t^{2k}\int_0^T (\Big\|\frac{\partial^k u}{\partial t^k}(s)\Big\|_{H^1}^2 + \Big\|\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\Big\|_{H^1}^2 + C_0^{2k+2})ds. \end{aligned} \tag{3.38}$$

Summing up (3.32), (3.34) and (3.38), we obtain

$$\lambda_G\|\bar e^{m+1}\|_{H^2}^2 \le C\delta t\sum_{q=0}^{m+1}\|\bar e^q\|_{H^2}^2 + C\delta t^{2k}\int_0^T (\Big\|\frac{\partial^k u}{\partial t^k}(s)\Big\|_{H^1}^2 + \Big\|\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\Big\|_{H^1}^2 + C_0^{2k+2})ds \tag{3.39}$$

Finally, we can obtain the following H^2 estimate for $\bar{e}^{m+1}$ by applying the discrete Gronwall lemma to (3.39) with $\delta t < \frac{1}{2C}$:

$$\begin{aligned}\|\bar{e}^{m+1}\|_{H^2}^2 &\le C\exp((1-\delta t C)^{-1}T)\delta t^{2k}\int_0^T (\left\|\frac{\partial^k u}{\partial t^k}(s)\right\|_{H^1}^2 + \left\|\frac{\partial^{k+1} u}{\partial t^{k+1}}(s)\right\|_{H^1}^2 + C_0^{2k+2})ds \\ &\le C_2(1+C_0^{2k+2})\delta t^{2k} \quad \forall 0 \le n \le m.\end{aligned} \tag{3.40}$$

where C_2 is independent of δt and C_0, can be defined as

$$C_2 := C\exp(2T)\max\Big(\int_0^T (\left\|\frac{\partial^k u}{\partial t^k}(s)\right\|_{H^1}^2 + \left\|\frac{\partial^{k+1} u}{\partial t^{k+1}}(s)\right\|_{H^1}^2)ds, 2, T\Big). \tag{3.41}$$

then $\delta t < \frac{1}{2C}$ can be guaranteed by

$$\delta t < \frac{1}{C_2}. \tag{3.42}$$

In particular, (3.40) implies

$$\|\bar{e}^{n+1}\|_{H^2} \le \sqrt{C_2(1+C_0^{2k+2})}\delta t^k, \quad \forall 0 \le n \le m. \tag{3.43}$$

Combining (3.21) and (3.43), under the condition (3.11) we obtain

$$\|\bar{u}^{n+1}\|_{H^2} \le \sqrt{C_2(1+C_0^{2k+2})}\delta t^2 + C \le \sqrt{C_2(1+1)} + C := \bar{C} \quad 0 \le n \le m. \tag{3.44}$$

Note that $H^2 \subset L^\infty$, without loss of generality, we can adjust $\bar{C}$ independent of C_0 and δt so that we have

$$\|g(\bar{u}^{n+1})\|,\ \|g'(\bar{u}^{n+1})\| \le \bar{C} \quad \forall 0 \le n \le m. \tag{3.45}$$

Step 3: estimate for $|1-\xi^{m+1}|$. By direct calculation,

$$r_{tt} = \int_\Omega \big(|\nabla u_t|^2 + \nabla u \cdot \nabla u_{tt} + \lambda u_t^2 + \lambda u u_{tt} + g'(u)u_t^2 + g(u)u_{tt}\big)d\boldsymbol{x}. \tag{3.46}$$

It follows from (2.5b) that the equation for the errors can be written as

$$s^{n+1} - s^n = \delta t\big(\|h[u(t^{n+1})]\|^2 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\|h(\bar{u}^{n+1})\|^2\big) + T_1^n, \tag{3.47}$$

where $h(u) = \frac{\delta E}{\delta u} = -\Delta u + \lambda u - g(u)$, and

$$T_1^n = r(t^n) - r(t^{n+1}) + \delta t r_t(t^{n+1}) = \int_{t^n}^{t^{n+1}} (s-t^n) r_{tt}(s)ds. \tag{3.48}$$

Taking the sum of (3.47) for n from 0 to m, and noting that $s^0 = 0$, we have

$$s^{m+1} = \delta t \sum_{q=0}^{m}\big(\|h[u(t^{q+1})]\|^2 - \frac{r^{q+1}}{E(\bar{u}^{q+1})}\|h(\bar{u}^{q+1})\|^2\big) + \sum_{q=0}^{m} T_1^q. \tag{3.49}$$

We can bound the terms on the right hand side of (3.49) as follows: For T_1^n, noting (3.46) we have

$$|T_1^n| \le C\delta t\int_{t^n}^{t^{n+1}} |r_{tt}(s)|ds \le C\delta t\int_{t^n}^{t^{n+1}} \big(\|u_t(s)\|_{H^1}^2 + \|u_{tt}(s)\|_{H^1}\big)ds. \tag{3.50}$$

Next,

$$\begin{aligned}&\Big|\|h[u(t^{n+1})]\|^2 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\|h(\bar{u}^{n+1})\|^2\Big| \\ &\le \|h[u(t^{n+1})]\|^2\Big|1 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big| + \frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big|\|h[u(t^{n+1})]\|^2 - \|h(\bar{u}^{n+1})\|^2\Big| \\ &:= P_1^n + P_2^n.\end{aligned} \tag{3.51}$$

For P_1^n, it follows from (3.21), $E(v) > \underline{C} > 0\,, \forall v$ and Theorem 1 that

$$\begin{aligned} P_1^n &\le C\Big|1 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big| \\ &\le C\Big|\frac{r(t^{n+1})}{E[u(t^{n+1})]} - \frac{r^{n+1}}{E[u(t^{n+1})]}\Big| + C\Big|\frac{r^{n+1}}{E[u(t^{n+1})]} - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big| \\ &\le C\Big(|E[u(t^{n+1})] - E(\bar{u}^{n+1})| + |s^{n+1}|\Big). \end{aligned} \tag{3.52}$$

For P_2^n, it follows from (3.21), (3.22), (3.44), (3.45), $E(v) > \underline{C} > 0$ and Theorem 1 that

$$\begin{aligned} P_2^n &\le C\big|\|h(\bar{u}^{n+1})\|^2 - \|h[u(t^{n+1})]\|^2\big| \\ &\le C\|h(\bar{u}^{n+1}) - h[u(t^{n+1})]\|(\|h(\bar{u}^{n+1})\| + \|h[u(t^{n+1})]\|) \\ &\le C\bar{C}\big(\|\Delta\bar{e}^{n+1}\| + \lambda\|\bar{e}^{n+1}\| + \|g(\bar{u}^{n+1}) - g[u(t^{n+1})]\|\big) \\ &\le C\bar{C}\big(\|\Delta\bar{e}^{n+1}\| + \|\bar{e}^{n+1}\|\big). \end{aligned} \tag{3.53}$$

On the other hand,

$$\begin{aligned} |E[u(t^{n+1})] - E(\bar{u}^{n+1})| \le{}& \frac{1}{2}\big(\|\nabla u(t^{n+1})\| + \|\nabla\bar{u}^{n+1}\|\big)\|\nabla u(t^{n+1}) - \nabla\bar{u}^{n+1}\| \\ &+ \frac{\lambda}{2}\big(\|u(t^{n+1})\| + \|\bar{u}^{n+1}\|\big)\|u(t^{n+1}) - \bar{u}^{n+1}\| \\ &+ \int g[u(t^{n+1})]dx - \int g(\bar{u}^{n+1})dx \\ \le{}& C\bar{C}\big(\|\nabla\bar{e}^{n+1}\| + \|\bar{e}^{n+1}\|\big). \end{aligned} \tag{3.54}$$

Now, combining (3.43), (3.49)–(3.54), we arrive at

$$\begin{aligned} |s^{m+1}| &\le \delta t\sum_{q=0}^{m}\Big|\|h[u(t^{q+1})]\|^2 - \frac{r^{q+1}}{E(\bar{u}^{q+1})}\|h(\bar{u}^{q+1})\|^2\Big| + \sum_{q=0}^{m}|T_1^q| \\ &\le C\delta t\sum_{q=0}^{m}|s^{q+1}| + C\bar{C}\delta t\sum_{q=0}^{m}\|\bar{e}^{q+1}\|_{H^2} + C\delta t\int_0^T(\|u_t(s)\|_{H^1}^2 + \|u_{tt}(s)\|_{H^1})ds \\ &\le C\delta t\sum_{q=0}^{m}|s^{q+1}| + C\bar{C}\sqrt{C_2(1 + C_0^{2k+2})}\delta t^k + C\delta t. \end{aligned}$$

Applying the discrete Gronwall lemma to the above inequality with $\delta t < \frac{1}{2C}$, we obtain

$$\begin{aligned} |s^{m+1}| &\le C\exp((1 - C\delta t)^{-1}T)\delta t(\bar{C}\sqrt{C_2(1 + C_0^{2k+2})}\delta t^{k-1} + 1) \\ &\le C_3\delta t(\bar{C}\sqrt{C_2(1 + C_0^{2k+2})}\delta t^{k-1} + 1), \end{aligned} \tag{3.55}$$

where C_3 is independent of C_0 and δt, can be defined as

$$C_3 := C\max\{\exp(2T), 2\}, \tag{3.56}$$

then $\delta t < \frac{1}{2C}$ can be guaranteed by

$$\delta t < \frac{1}{C_3}. \tag{3.57}$$

Hence, noting (3.52), (3.54), (3.55) and (3.44), we have

$$\begin{aligned} |1 - \xi^{m+1}| &\le C\big(|E[u(t^{m+1})] - E(\bar{u}^{m+1})| + |s^{m+1}|\big) \\ &\le C(\bar{C}\|\bar{e}^{m+1}\|_{H^1} + |s^{m+1}|) \\ &\le C\delta t\big(\bar{C}\sqrt{C_2(1 + C_0^{2k+2})}\delta t^{k-1} + C_3(\bar{C}\sqrt{C_2(1 + C_0^{2k+2})}\delta t^{k-1} + 1)\big) \\ &\le C_4\delta t(\sqrt{1 + C_0^{2k+2}}\delta t^{k-1} + 1), \end{aligned} \tag{3.58}$$

where the constant C_4 is independent of C_0 and δt. Without loss of generality, we assume $C_4 > \max\{C_2, C_3, 1\}$ to simplify the proof below.

As a result of (3.58), $|1-\xi^{m+1}| \le C_0\delta t$ if we define C_0 such that

$$C_4(\sqrt{1+C_0^{2k+2}}\delta t^{k-1}+1) \le C_0. \tag{3.59}$$

For the cases $k \ge 2$, the above can be satisfied if we choose $C_0 = 3C_4$ and $\delta t \le \frac{1}{1+C_0^{k+1}}$:

$$C_4(\sqrt{1+C_0^{2k+2}}\delta t^{k-1}+1) \le C_4[(1+C_0^{k+1})\delta t+1] \le 3C_4 = C_0. \tag{3.60}$$

For the case $k=1$, we cannot define C_0 satisfying (3.59) if $\eta_1^{n+1} = 1-(1-\xi^{n+1})^2$. However, if we choose $\eta_1^{n+1} = 1-(1-\xi^{n+1})^3$, we can repeat the same process above and arrive at a similar version of (3.59) for the first order case:

$$C_4(\sqrt{1+C_0^{6}}\delta t+1) \le C_0. \tag{3.61}$$

The above can be satisfied if we choose $C_0 = 3C_4$ and $\delta t < \frac{1}{C_0^3}$ so that

$$C_4(\sqrt{1+C_0^{6}\delta t^2}+1) \le C_4[1+C_0^3\delta t+1] \le 3C_4 = C_0.$$

To summarize, under the condition

$$\delta t \le \frac{1}{1+C_0^{k+2}}, \quad 1 \le k \le 5, \tag{3.62}$$

we have $|1-\xi^{m+1}| \le C_0\delta t$. Note that with $C_4 > \max\{C_2, C_3, 1\}$, (3.62) also implies (3.42) and (3.57). The induction process for (3.7) is complete.

Finally, thanks to (3.43), it remains to show $\|e^{m+1}\|_{H^2} \le C\delta t^k$.

We derive from (2.5d) and (3.44) that

$$\|u^{m+1}-\bar{u}^{m+1}\|_{H^2} \le |\eta_k^{m+1}-1|\|\bar{u}^{m+1}\|_{H^2} \le |\eta_k^{m+1}-1|\bar{C}. \tag{3.63}$$

On the other hand, we derive from (3.7) that

$$|\eta_k^{m+1}-1| \le C_0^{k+1}\delta t^{k+1}. \tag{3.64}$$

Then it follows from (3.43), (3.63) and (3.64) and combine the condition (3.11), (3.17) and (3.62) on δt that

$$\begin{aligned}\|e^{m+1}\|_{H^2}^2 &\le 2\|\bar{e}^{m+1}\|_{H^2}^2 + 2\|u^{m+1}-\bar{u}^{m+1}\|_{H^2}^2\\ &\le 2C_2(1+C_0^{2(k+1)})\delta t^{2k} + 2\bar{C}^2C_0^{2(k+1)}\delta t^{2(k+1)}\end{aligned}$$

holds under the condition $\delta t < \min\{\frac{1}{1+2C_0^{k+2}}, \frac{1-\tau_k}{3(k+1)}\}$. The proof is complete. □

Remark 3. Note that we set $\eta_1^{n+1} = 1-(1-\xi^{n+1})^3$ purely for technical reasons in the proof. It is clear that $\eta_1^{n+1} = 1-(1-\xi^{n+1})^2$ leads to first-order accuracy which is confirmed by our numerical tests.

4. Error analysis for Cahn–Hilliard type equations

In this section, we consider the Cahn–Hilliard type equation

$$\frac{\partial u}{\partial t} = -\Delta^2 u + \lambda\Delta u - \Delta g(u) \quad (\boldsymbol{x}, t) \in \Omega \times (0, T], \tag{4.1}$$

where Ω is an open bounded domain in $\mathbb{R}^d$ $(d = 1, 2, 3)$, with the initial condition $u(\boldsymbol{x}, 0) = u^0(\boldsymbol{x})$ and boundary conditions

$$\text{periodic, or, } \frac{\partial u}{\partial \boldsymbol{n}}|_{\partial\Omega} = \frac{\partial \Delta u}{\partial \boldsymbol{n}}|_{\partial\Omega} = 0. \tag{4.2}$$

The above equation is a special case of (2.1) with $\mathcal{A} = \Delta^2 - \lambda\Delta$ and $g(u)$ replaced by $-\Delta g(u)$. It satisfies the dissipation law (2.2) with $E(u) = \frac{1}{2}(\mathcal{L}u, u) + (G(u), 1)$ where $(\mathcal{L}u, u) = (\nabla u, \nabla u) + \lambda(u, u)$, $G(u) = \int^u g(v)dv$ and $\mathcal{K}(u) = (\nabla\frac{\delta E}{\delta u}, \nabla\frac{\delta E}{\delta u})$.

In particular, with $g(u) = (1-u^2)u$ and $\lambda = 0$, the above equation becomes the celebrated Cahn–Hilliard equation [28].

We first recall the following result (cf. for instance [27]).

Theorem 4. *Let $u^0 \in H^2$, and (3.4) holds. We assume additionally*

$$|g''(x)| < C(|x|^{p'}+1), \quad p' > 0 \ arbitrary \quad if\ n = 1,2; \quad 0 < p' < 3 \quad if\ n = 3. \tag{4.3}$$

Then for any $T > 0$, there exists a unique solution u for (4.1) such that

$$u \in C([0,T]; H^2) \cap L^2(0,T; H^4).$$

We also recall the following result (see Lemma 2.3 in [12]) which we shall use to deal with the nonlinear term.

Lemma 4. *Assume that $\|u\|_{H^1} \le M$, and that (3.4) and (4.3) hold. Then for any $u \in H^4$, there exist $0 \le \sigma < 1$ and a constant $C(M)$ such that the following inequality holds:*

$$\|\Delta g(u)\|^2 \le C(M)(1 + \|\Delta^2 u\|^{2\sigma}).$$

For (4.1), the kth-order version of (2.5a) and (2.11) read:

$$\frac{\alpha_k \bar{u}^{n+1} - A_k(u^n)}{\delta t} = -\Delta\big(\Delta \bar{u}^{n+1} - \lambda \bar{u}^{n+1} + g[B_k(\bar{u}^n)]\big), \tag{4.4}$$

and

$$\frac{\alpha_k u^{n+1} - \eta_k^{n+1} A_k(u^n)}{\delta t} = -\Delta\big(\Delta u^{n+1} - \lambda u^{n+1} + \eta_k^{n+1} g[B_k(\bar{u}^n)]\big), \tag{4.5}$$

where α_k, A_k, B_k defined in (2.8)–(2.10).

Theorem 5. *Given initial condition $\bar{u}^0 = u^0 = u(0)$, $r^0 = E[u^0]$. Let $\bar{u}^{n+1}$ and u^{n+1} be computed with the kth order scheme (2.5a)–(2.5d) ($1 \le k \le 5$) for (4.1) with*

$$\eta_1^{n+1} = 1 - (1-\xi^{n+1})^3, \quad \eta_k^{n+1} = 1 - (1-\xi^{n+1})^{k+1} \ (k = 2,3,4,5).$$

We assume (3.4) and (4.3) hold, and

$$u \in C([0,T]; H^3), \ \frac{\partial^j u}{\partial t^j} \in L^2(0,T; H^2)\ 1 \le j \le k, \ \frac{\partial^{k+1} u}{\partial t^{k+1}} \in L^2(0,T; H^1).$$

Then for $n+1 \le T/\delta t$ and $\delta t \le min\{\frac{1}{1+4C_0^{k+2}}, \frac{1-\tau_k}{3(k+1)}\}$, we have

$$\|\bar{e}^{n+1}\|_{H^2}, \ \|e^{n+1}\|_{H^2} \le C\delta t^k,$$

where the constants C_0, C are dependent on T, Ω, the $k \times k$ matrix $G = (g_{ij})$ in Lemma 1 and the exact solution u but are independent of δt.

Since the proof of this theorem shares some similar procedures with the proof of Theorem 3, we shall defer its proof to the appendix.

5. Concluding remarks

We constructed a new class of implicit–explicit BDFk SAV schemes for general linear systems. This class of schemes enjoys the following advantages: (i) it only requires solving, in most common situations, one linear system with constant coefficients at each time step, which is the same as the usual IMEX schemes; (ii) it is not restricted to gradient flows and is applicable to general dissipative systems; and (iii) it can be high-order with unconditional stability and suitable for adaptive time stepping without restriction on time step size; and most importantly, (iv) it leads to a unconditional uniform bound for the numerical solution, for any order k on the norm based on the principal linear term in the energy functional, which is of critical importance for the convergence and error analysis. We presented numerical results which validated the stability and convergence rates of our schemes, and showed that

the SAV scheme is at least as accurate as the usual IMEX scheme, and may lead to more accurate solutions in some critical situations (solutions with large gradients or near singularities).

Using the uniform bound on the norm based on the principal linear operator that we derived for the BDFk SAV schemes and to a stability result in [23] for the BDFk ($k = 1, 2, 3, 4, 5$) schemes, we were able to establish, with a delicate inductive argument, rigorous error estimates for the BDFk ($k = 1, 2, 3, 4, 5$) SAV schemes in a unified form for the typical Allen–Cahn and Cahn–Hilliard type equations. We note that recently the result in Lemma 1 was extended to the six-order BDF scheme in [29], and it is expected that the results in Theorems 3 and 5 can be extended to the six-order case using the result in [29].

As mentioned in Remark 2, we can replace the BDKk scheme in (2.5a) by other IMEX multistep schemes, and the stability result in Theorem 1 will still hold. However, error analysis for other implicit–explicit multistep SAV schemes needs to be investigated separately.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jie Shen reports financial support was provided by National Science Foundation. Jie Shen reports financial support was provided by Air Force Office of Scientific Research. Fukeng Huang reports a relationship with National University of Singapore that includes: employment.

Appendix A. Proof of Theorem 5

For the sake of brevity, we shall only carry out in detail the error analysis for the first-order case. The analysis for the higher-order cases can be carried out by combining the procedures for the first-order case below and for the high-order cases in the proof Theorem 3. The detail will be left for the interested readers.

As in the proof of Theorem 3, we will first prove the following by induction:

$$|1 - \xi^q| \le C_0\,\delta t,\ \ \forall q \le T/\delta t, \tag{A.1}$$

where the constant C_0 is dependent on T, Ω and the exact solution u but is independent of δt, and will be defined in the proof process.

Under the assumptions, (A.1) certainly holds for $q = 0$. Now suppose we have

$$|1 - \xi^q| \le C_0\,\delta t,\ \ \forall q \le m, \tag{A.2}$$

we shall prove below that (A.1) holds for $q = m + 1$, namely,

$$|1 - \xi^{m+1}| \le C_0 \delta t. \tag{A.3}$$

We will carry out this proof in three steps.

Step 1: H^2 bound for u^n and $\bar{u}^n$ for all $n \le m$. It follows from Theorem 1 and under condition

$$\delta t \le \min\{\frac{1}{4C_0^3}, 1\}, \tag{A.4}$$

we have

$$\frac{3}{4} \le |\eta_1^q| \le 2,\ \ |1 - \eta_1^q| \le \frac{\delta t^2}{4}, \forall q \le m, \tag{A.5}$$

and

$$\|u^q\|_{H^1} \le M_2,\ \ \forall q \le T/\delta t,\ \ \|\bar{u}^q\|_{H^1} \le \frac{4}{3}M_2,\ \forall q \le m. \tag{A.6}$$

Now, consider (4.5) at step q:

$$\frac{u^q - \eta_1^q u^{n-1}}{\delta t} = -\Delta^2 u^q + \lambda \Delta u^q - \eta_1^q \Delta g[\bar{u}^{q-1}] \tag{A.7}$$

Multiply (A.7) with $\Delta^2 u^q$, and by the similar process as step 1 in Theorem 3, we can obtain

$$\|\Delta u^q\|^2 - \|\Delta u^{q-1}\|^2 + \delta t\|\Delta^2 \bar{u}^q\|^2 - \frac{\delta t}{2}\|\Delta^2 \bar{u}^{q-1}\|^2 \le C(M_2)\delta t + |1 - \eta_1^q|\|u^{q-1}\|^2 \tag{A.8}$$

Taking the sum from 1 to $n\,(\leq m)$ of (A.8), we obtain

$$\begin{aligned}\|\Delta u^n\|^2+\frac{\delta t}{2}\sum_{q=0}^{n}\|\Delta^2\bar{u}^q\|^2 &\leq C(M_2)T+C(u^0)+\delta t^2\sum_{q=1}^{n-1}\|u^q\|^2\\ &\leq C(M_2)T+C(u^0)+\delta t T M_2^2.\end{aligned}$$

with $C(M_2)$ is a constant only depends on M_2 and $C(u^0)$ only depends on u^0. Then together with (A.6) implies

$$\|u^n\|_{H^2}\leq\sqrt{C(M)T+C(u^0)+TM_2^2}+M_2:=C_1,\quad\forall n\leq m. \tag{A.9}$$

As $\|u^n\|_{H^2}=\eta_1^n\|\bar{u}^n\|_{H^2}$, (A.5) implies

$$\|\bar{u}^n\|_{H^2}\leq\frac{4}{3}C_1,\quad\forall n\leq m. \tag{A.10}$$

Step 2: estimates for $\|\bar{e}^{n+1}\|_{H^2}$ and $\|\bar{e}^{n+1}\|_{H^3}$ for all $0\leq n\leq m$. By given assumption on the exact solution u and (A.10), we can choose C large enough such that

$$\|u(t)\|_{H^3}\leq C,\ \forall t\leq T,\|\bar{u}^q\|_{H^2}\leq C,\ \forall q\leq m, \tag{A.11}$$

and since $H^2\subset L^\infty$, without loss of generality, we can adjust C such that

$$|g^{(i)}[u(t)]|_{L^\infty}\leq C,\ \forall t\leq T;\ |g^{(i)}(\bar{u}^q)|_{L^\infty}\leq C,\ \forall q\leq m;\ i=0,1,2,3. \tag{A.12}$$

From (4.4), we can write down the equation for error as

$$\bar{e}^{n+1}-\bar{e}^n=(\eta_1^n-1)\bar{u}^n-\delta t\Delta^2\bar{e}^{n+1}+\lambda\delta t\Delta\bar{e}^{n+1}+R_1^n+\delta t\Delta R_2^n, \tag{A.13}$$

where $R_1^n,\ R_2^n$ are given by

$$R_1^n=u(t^n)-u(t^{n+1})+\delta t u_t(t^{n+1})=\int_{t^n}^{t^{n+1}}(s-t^n)u_{tt}ds, \tag{A.14}$$

and

$$R_2^n=-g(\bar{u}^n)+g[u(t^{n+1})]. \tag{A.15}$$

Taking inner product with $\bar{e}^{n+1}-\Delta\bar{e}^{n+1}+\Delta^2\bar{e}^{n+1}$ on both sides of (A.13), we obtain

$$\begin{aligned}&\frac{1}{2}\big(\|\bar{e}^{n+1}\|^2-\|\bar{e}^n\|^2\big)+\frac{1}{2}\|\bar{e}^{n+1}-\bar{e}^n\|^2+\delta t\|\Delta\bar{e}^{n+1}\|^2+\lambda\delta t\|\nabla\bar{e}^{n+1}\|^2\\ &\quad+\frac{1}{2}\big(\|\nabla\bar{e}^{n+1}\|^2-\|\nabla\bar{e}^n\|^2\big)+\frac{1}{2}\|\nabla(\bar{e}^{n+1}-\bar{e}^n)\|^2+\delta t\|\nabla\Delta\bar{e}^{n+1}\|^2+\lambda\delta t\|\Delta\bar{e}^{n+1}\|^2\\ &\quad+\frac{1}{2}\big(\|\Delta\bar{e}^{n+1}\|^2-\|\Delta\bar{e}^n\|^2\big)+\frac{1}{2}\|\Delta(\bar{e}^{n+1}-\bar{e}^n)\|^2+\delta t\|\Delta^2\bar{e}^{n+1}\|^2+\lambda\delta t\|\nabla\Delta\bar{e}^{n+1}\|^2\\ &=(\eta_1^n-1)\big(\bar{u}^n,\bar{e}^{n+1}\big)+\big(R_1^n,\bar{e}^{n+1}\big)-\delta t\big(\nabla R_2^n,\nabla\bar{e}^{n+1}\big)\\ &\quad+(\eta_1^n-1)\big(\nabla\bar{u}^n,\nabla\bar{e}^{n+1}\big)+\big(R_1^n,-\Delta\bar{e}^{n+1}\big)+\delta t\big(\nabla R_2^n,\nabla\Delta\bar{e}^{n+1}\big)\\ &\quad+(\eta_1^n-1)\big(\Delta\bar{u}^n,\Delta\bar{e}^{n+1}\big)+\big(R_1^n,\Delta^2\bar{e}^{n+1}\big)+\delta t\big(\Delta R_2^n,\Delta^2\bar{e}^{n+1}\big).\end{aligned} \tag{A.16}$$

In the following, we bound the right hand side of (A.16). Noting that $|\eta_1^n-1|\leq C_0^3\,\delta t^3$, hence

$$|(\eta_1^n-1)\big(\bar{u}^n,\bar{e}^{n+1}\big)|\leq\frac{\|(\eta_1^n-1)\bar{u}^n\|^2}{\delta t}+\frac{\delta t}{4}\|\bar{e}^{n+1}\|^2\leq CC_0^6\delta t^5+\frac{\delta t}{4}\|\bar{e}^{n+1}\|^2, \tag{A.17}$$

$$|(\eta_1^n-1)\big(\nabla\bar{u}^n,\nabla\bar{e}^{n+1}\big)|\leq CC_0^6\delta t^5+\frac{\delta t}{4}\|\nabla\bar{e}^{n+1}\|^2, \tag{A.18}$$

and

$$|(\eta_1^n-1)\big(\Delta\bar{u}^n,\Delta\bar{e}^{n+1}\big)|\leq CC_0^6\delta t^5+\frac{\delta t}{4}\|\Delta\bar{e}^{n+1}\|^2. \tag{A.19}$$

It follows from (A.14) that

$$\|R_1^n\|^2 \le C\delta t^3 \int_{t^n}^{t^{n+1}} \|u_{tt}(s)\|^2 ds. \tag{A.20}$$

Therefore,

$$|\big(R_1^n, \bar{e}^{n+1}\big)| \le \frac{1}{2\delta t}\|R_1^n\|^2 + \frac{\delta t}{2}\|\bar{e}^{n+1}\|^2 \le \frac{\delta t}{2}\|\bar{e}^{n+1}\|^2 + C\delta t^2 \int_{t^n}^{t^{n+1}} \|u_{tt}(s)\|^2 ds, \tag{A.21}$$

$$|\big(R_1^n, -\Delta\bar{e}^{n+1}\big)| \le \frac{\delta t}{2}\|\Delta\bar{e}^{n+1}\|^2 + C\delta t^2 \int_{t^n}^{t^{n+1}} \|u_{tt}(s)\|^2 ds, \tag{A.22}$$

and

$$|\big(R_1^n, \Delta^2\bar{e}^{n+1}\big)| \le \frac{\delta t}{2}\|\Delta^2\bar{e}^{n+1}\|^2 + C\delta t^2 \int_{t^n}^{t^{n+1}} \|u_{tt}(s)\|^2 ds. \tag{A.23}$$

Noting that

$$\begin{aligned}
|\nabla R_2^n| &= |\nabla g(\bar{u}^n) - \nabla g[u(t^n)] + \nabla g[u(t^n)] - \nabla g[u(t^{n+1})]| \\
&\le |g'(\bar{u}^n)\nabla\bar{u}^n - g'[u(t^n)]\nabla u(t^n)| + |g'[u(t^n)]\nabla u(t^n) - g'[u(t^{n+1})]\nabla u(t^{n+1})| \\
&\le |g'(\bar{u}^n)|\big|\nabla\bar{u}^n - \nabla u(t^n)\big| + \big|g'(\bar{u}^n) - g'[u(t^n)]\big|\big|\nabla u(t^n)\big| \\
&+ \big|g'[u(t^n)] - g'[u(t^{n+1})]\big|\big|\nabla u(t^n)\big| + \big|g'[u(t^{n+1})]\big|\big|\nabla u(t^n) - \nabla u(t^{n+1})\big| \\
&\le C\big(|\nabla\bar{e}^n| + |\bar{e}^n| + \int_{t^n}^{t^{n+1}} \big(|u_t(s)| + |\nabla u_t(s)|\big)ds\big),
\end{aligned} \tag{A.24}$$

then for the terms with ∇R_2^n, it follows from (A.24) that

$$\begin{aligned}
\delta t|(\nabla R_2^n, \nabla\bar{e}^{n+1})| &\le \frac{\delta t}{2}\|\nabla R_2^n\|^2 + \frac{\delta t}{2}\|\nabla\bar{e}^{n+1}\|^2 \\
&\le C\delta t(\|\nabla\bar{e}^{n+1}\|^2 + \|\bar{e}^n\|_{H^1}^2) + C\delta t^2 \int_{t^n}^{t^{n+1}} \|u_t(s)\|_{H^1}^2 ds,
\end{aligned} \tag{A.25}$$

and

$$\begin{aligned}
\delta t|(\nabla R_2^n, \nabla\Delta\bar{e}^{n+1})| &\le \frac{\delta t}{2}\|\nabla R_2^n\|^2 + \frac{\delta t}{2}\|\nabla\Delta\bar{e}^{n+1}\|^2 \\
&\le C\delta t\|\bar{e}^n\|_{H^1}^2 + C\delta t^2 \int_{t^n}^{t^{n+1}} \|u_t(s)\|_{H^1}^2 ds + \frac{\delta t}{2}\|\nabla\Delta\bar{e}^{n+1}\|^2.
\end{aligned} \tag{A.26}$$

For the term with ΔR_2^n, since

$$|\Delta R_2^n| \le |-\Delta g(\bar{u}^n) + \Delta g[u(t^n)]| + |-\Delta g[u(t^n)] + \Delta g[u(t^{n+1})]| := Q_1^n + Q_2^n,$$

and note that

$$\Delta g(u) = g''(u)|\nabla u|^2 + g'(u)\Delta u,$$

by using (A.11) and (A.12), we have

$$\begin{aligned}
Q_1^n &\le \big|g''(\bar{u}^n)(|\nabla\bar{u}^n|^2 - |\nabla u(t^n)|^2)\big| + \big||\nabla u(t^n)|^2(g''(\bar{u}^n) - g''[u(t^n)])\big| \\
&+ |g'(\bar{u}^n)(\Delta\bar{u}^n - \Delta u(t^n))| + |\Delta u(t^n)(g'(\bar{u}^n) - g'[u(t^n)])| \\
&\le C\big(|\nabla\bar{e}^n| + |\bar{e}^n| + |\Delta\bar{e}^n|\big),
\end{aligned}$$

and

$$Q_2^n \le C\big(\int_{t^n}^{t^{n+1}} |\nabla u(s)||\nabla u_t(s)|ds + \int_{t^n}^{t^{n+1}} |\Delta u_t(s)|ds\big).$$

Therefore,

$$\begin{aligned}\delta t|\big(\Delta R_2^n, \Delta^2\bar{e}^{n+1}\big)| &\le \delta t|\big(Q_1^n, \Delta^2\bar{e}^{n+1}\big)| + \delta t|\big(Q_2^n, \Delta^2\bar{e}^{n+1}\big)|\\ &\le \delta t\|Q_1^n\|^2 + \frac{\delta t}{4}\|\Delta^2\bar{e}^{n+1}\|^2 + \delta t\|Q_2^n\|^2 + \frac{\delta t}{4}\|\Delta^2\bar{e}^{n+1}\|^2\\ &\le C\delta t(\|\bar{e}^n\|^2 + \|\nabla\bar{e}^n\|^2 + \|\Delta\bar{e}^n\|^2) + \frac{\delta t}{2}\|\Delta^2\bar{e}^{n+1}\|^2\\ &+ C\delta t^2\int_{t^n}^{t^{n+1}}\|u_t(s)\|_{H^2}^2 ds,\end{aligned} \tag{A.27}$$

Now, combining (A.16)–(A.26) and (A.27) and dropping some unnecessary terms, we arrive at

$$\begin{aligned}&\|\bar{e}^{n+1}\|^2 - \|\bar{e}^n\|^2 + \|\nabla\bar{e}^{n+1}\|^2 - \|\nabla\bar{e}^n\|^2 + \|\Delta\bar{e}^{n+1}\|^2 - \|\Delta\bar{e}^n\|^2 + \delta t\|\nabla\Delta\bar{e}^{n+1}\|^2\\ &\quad\le CC_0^6\delta t^5 + C\delta t(\|\nabla\bar{e}^{n+1}\|^2 + \|\bar{e}^{n+1}\|^2 + \|\Delta\bar{e}^n\|^2 + \|\nabla\bar{e}^n\|^2 + \|\bar{e}^n\|^2)\\ &\quad+ C\delta t^2\int_{t^n}^{t^{n+1}}(\|u_t(s)\|_{H^2}^2 + \|u_{tt}(s)\|^2)ds.\end{aligned} \tag{A.28}$$

Taking the sum of the above for n from 0 to m, we obtain

$$\|\bar{e}^{m+1}\|_{H^2}^2 + \delta t\sum_{q=0}^{m}\|\nabla\Delta\bar{e}^{q+1}\|^2 \le C\delta t\sum_{q=0}^{m+1}\|\bar{e}^q\|_{H^2}^2 + C\delta t^2\int_0^T(\|u_t(s)\|_{H^2}^2 + \|u_{tt}(s)\|^2 + C_0^6\delta t^2)ds. \tag{A.29}$$

Finally, we can obtain the following estimate for $\bar{e}^{m+1}$ by applying the discrete Gronwall's inequality to (A.29) with $\delta t < \frac{1}{2C}$:

$$\begin{aligned}\|\bar{e}^{n+1}\|_{H^2}^2 + \delta t\sum_{q=0}^{n}\|\nabla\Delta\bar{e}^{q+1}\|^2 &\le C\exp((1-\delta tC)^{-1}T)\delta t^2\int_0^T(\|u_t(s)\|_{H^2}^2 + \|u_{tt}(s)\|^2 + C_0^6\delta t^2)ds\\ &\le C_2(1 + C_0^6\delta t^2)\delta t^2, \quad \forall\, 0 \le n \le m.\end{aligned} \tag{A.30}$$

where C_2 is independent of δt and C_0, can be defined as

$$C_2 := C\exp(2T)\max\Big(\int_0^T(\|u_t(s)\|_{H^2}^2 + \|u_{tt}(s)\|^2)ds, 2, T\Big), \tag{A.31}$$

and hence $\delta t < \frac{1}{2C}$ can be guaranteed by $\delta t < \frac{1}{C_2}$. In particular, (A.30) implies

$$\|\bar{e}^{n+1}\|_{H^2},\ \big(\delta t\sum_{q=0}^{n}\|\nabla\Delta\bar{e}^{q+1}\|^2\big)^{1/2} \le \sqrt{C_2(1 + C_0^6\delta t^2)}\delta t, \quad \forall\, 0 \le n \le m. \tag{A.32}$$

Combining (A.11) and (A.32), we obtain that for all $\forall\, 0 \le n \le m$ and under the condition on δt in (A.4), we have

$$\|\bar{u}^{n+1}\|_{H^2},\ \big(\delta t\sum_{q=0}^{n}\|\nabla\Delta\bar{u}^{q+1}\|^2\big)^{1/2} \le \sqrt{C_2(1 + C_0^6\delta t^2)}\delta t + C \le \sqrt{C_2(1+1)} + C := \bar{C}. \tag{A.33}$$

Note that $H^2 \subset L^\infty$, without loss of generality, we can adjust $\bar{C}$ so that we have

$$\|g(\bar{u}^{n+1})\|,\ \|g'(\bar{u}^{n+1})\| \le \bar{C}, \quad \forall 0 \le n \le m. \tag{A.34}$$

Step 3: estimate for $|1 - \xi^{n+1}|$. It follows from (2.5b) that the equation for the error $\{s^j\}$ can be written as

$$s^{n+1} - s^n = \delta t\big(\|\nabla h[u(t^{n+1})]\|^2 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\|\nabla h(\bar{u}^{n+1})\|^2\big) + T_1^n, \tag{A.35}$$

where $h(u) = \frac{\delta E}{\delta u} = -\Delta u + \lambda u - g(u)$ and truncation errors T_1^n are given in (3.48) with a bound given in (3.50).

Taking the sum of (A.35) for n from 0 to m, since $s^0 = 0$, we have

$$s^{m+1} = \delta t\sum_{q=0}^{m}\big(\|\nabla h[u(t^{q+1})]\|^2 - \frac{r^{q+1}}{E(\bar{u}^{q+1})}\|\nabla h(\bar{u}^{q+1})\|^2\big) + \sum_{q=0}^{m}T_1^q. \tag{A.36}$$

For $\|\nabla h[u(t^{n+1})]\|^2 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\|\nabla h(\bar{u}^{n+1})\|^2$, we have

$$\begin{aligned}
&\Big|\|\nabla h[u(t^{n+1})]\|^2 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\|\nabla h(\bar{u}^{n+1})\|^2\Big| \\
&\le \|\nabla h[u(t^{n+1})]\|^2\Big|1 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big| + \frac{r^{n+1}}{E(\bar{u}^{n+1})}\big|\|\nabla h[u(t^{n+1})]\|^2 - \|\nabla h(\bar{u}^{n+1})\|^2\big| \\
&:= K_1^n + K_2^n.
\end{aligned}$$

For K_1^n, it follows from (A.11), $E(\bar{u}^{n+1}) > \underline{C} > 0$ and Theorem 1 that

$$\begin{aligned}
K_1^n &\le C\Big|1 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big| \\
&= C\Big|\frac{r(t^{n+1})}{E[u(t^{n+1})]} - \frac{r^{n+1}}{E[u(t^{n+1})]}\Big| + C\Big|\frac{r^{n+1}}{E[u(t^{n+1})]} - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big| \\
&\le C\big(|E[u(t^{n+1})] - E(\bar{u}^{n+1})| + |s^{n+1}|\big).
\end{aligned}$$

For K_2^n, it follows from (A.11), (A.12), (A.33), (A.34), $E(\bar{u}^{n+1}) > \underline{C} > 0$ and Theorem 1 that

$$\begin{aligned}
K_2^n &\le C\big|\|\nabla h(\bar{u}^{n+1})\|^2 - \|\nabla h[u(t^{n+1})]\|^2\big| \\
&\le C\|\nabla h(\bar{u}^{n+1}) - \nabla h[u(t^{n+1})]\|(\|\nabla h(\bar{u}^{n+1})\| + \|\nabla h[u(t^{n+1})]\|) \\
&\le C\bar{C}(1 + \|\nabla\Delta\bar{u}^{n+1}\|)\Big(\|\nabla\Delta\bar{e}^{n+1}\| + \lambda\|\nabla\bar{e}^{n+1}\| + \|\nabla\big(g(\bar{u}^{n+1}) - g[u(t^{n+1})]\big)\|\Big) \\
&\le C\bar{C}\big(\|\nabla\Delta\bar{e}^{n+1}\| + \|\nabla\bar{e}^{n+1}\|\big) + C\bar{C}\|\nabla\Delta\bar{u}^{n+1}\|\|\nabla\Delta\bar{e}^{n+1}\| + C\bar{C}\|\nabla\Delta\bar{u}^{n+1}\|\|\nabla\bar{e}^{n+1}\|.
\end{aligned}$$

It then follows from (A.32), (A.33) and the Cauchy–Schwarz inequality that

$$\delta t\sum_{q=1}^{n+1}\|\nabla\Delta\bar{u}^q\|\|\nabla\bar{e}^q\| \le \big(\delta t\sum_{q=1}^{n+1}\|\nabla\Delta\bar{u}^q\|^2\delta t\sum_{q=1}^{n+1}\|\nabla\bar{e}^q\|^2\big)^{1/2} \le C\bar{C}\sqrt{C_2(1 + C_0^6\delta t^2)}\delta t,$$

and

$$\delta t\sum_{q=1}^{n+1}\|\nabla\Delta\bar{u}^q\|\|\nabla\Delta\bar{e}^q\| \le \big(\delta t\sum_{q=1}^{n+1}\|\nabla\Delta\bar{u}^q\|^2\delta t\sum_{q=1}^{n+1}\|\nabla\Delta\bar{e}^q\|^2\big)^{1/2} \le C\bar{C}\sqrt{C_2(1 + C_0^6\delta t^2)}\delta t.$$

For $E[u(t^{n+1})] - E(\bar{u}^{n+1})$, we have estimate (3.54).

Now, we are ready to estimate s^{m+1}. Combine the estimate obtained above, (A.36) leads to

$$\begin{aligned}
|s^{m+1}| &\le \delta t\sum_{q=0}^{m}\Big|\|\nabla h[u(t^{q+1})]\|^2 - \frac{E_0(\bar{u}^{q+1}) + r^{q+1}}{E(\bar{u}^{q+1})}\|\nabla h(\bar{u}^{q+1})\|^2\Big| + \sum_{q=0}^{m}|T_1^q| \\
&\le C\delta t\sum_{q=0}^{m}|s^{q+1}| + C\bar{C}\delta t\sum_{q=0}^{m}\|\bar{e}^{q+1}\|_{H^1} + C\bar{C}\delta t\sum_{q=0}^{m}\|\nabla\Delta e^{q+1}\| \\
&+ C\bar{C}\delta t\sum_{q=1}^{m+1}\|\nabla\Delta\bar{u}^q\|\|\nabla\bar{e}^q\| + C\bar{C}\delta t\sum_{q=1}^{m+1}\|\nabla\Delta\bar{u}^q\|\|\nabla\Delta\bar{e}^q\| \\
&+ C\delta t\int_0^{t^{m+1}}(\|u_t(s)\|_{H^1}^2 + \|u_{tt}(s)\|_{H^1})ds \\
&\le C\delta t\sum_{q=0}^{m}|s^{q+1}| + C\bar{C}^2\delta t\big(\sqrt{C_2(1 + C_0^6\delta t^2)} + 1\big)
\end{aligned} \tag{A.37}$$

Finally, applying the discrete Gronwall's inequality on (A.37) with $\delta t < \frac{1}{2C}$, we obtain the following estimate for s^{n+1}:

$$\begin{aligned} |s^{n+1}| &\le C\exp((1-\delta t C)^{-1}T)\bar{C}^2\delta t\Big(\sqrt{C_2(1+C_0^6\delta t^2)}+1\Big) \\ &\le C_3\delta t\Big(\sqrt{C_2(1+C_0^6\delta t^2)}+1\Big), \forall\, 0\le n\le m, \end{aligned} \tag{A.38}$$

where C_3 is independent of δt and C_0, can be defined as

$$C_3 := \max\{C\bar{C}^2\exp(2T), 2\}. \tag{A.39}$$

Thanks to (A.38), we can define C_0 and then prove (A.3) by following exactly the same procedure as **Step 3 in** Theorem 3 with the condition

$$\delta t \le \frac{1}{1+C_0^3} \tag{A.40}$$

The induction process for (A.1) is completed.

Finally, thanks to (A.32), it remains to show $\|e^{m+1}\|_{H^2} \le C\delta t$.

We derive from (A.33) that

$$\|u^{m+1}-\bar{u}^{m+1}\|_{H^2} \le |\eta_1^{m+1}-1|\|\bar{u}^{m+1}\|_{H^2} \le |\eta_1^{m+1}-1|\bar{C}. \tag{A.41}$$

On the other hand, (A.1) implies

$$|\eta_1^{m+1}-1| \le C_0^3\delta t^3. \tag{A.42}$$

Then it follows from (A.32), (A.41) and (A.42) that

$$\begin{aligned} \|e^{m+1}\|_{H^2}^2 &\le 2\|\bar{e}^{m+1}\|_{H^2}^2 + 2\|u^{m+1}-\bar{u}^{m+1}\|_{H^2}^2 \\ &\le 2C_2(1+C_0^6\delta t^2)\delta t^2 + 2\bar{C}^2C_0^6\delta t^6. \end{aligned}$$

To summarize, combine the condition (A.4) and (A.40) on δt, we obtain $\|e^{m+1}\|_{H^2} \le C\delta t$ with $\delta t < \frac{1}{1+4C_0^3}$. The proof for the case $k=1$ is complete.

References

[1] Jie Shen, Jie Xu, Jiang Yang, The scalar auxiliary variable (SAV) approach for gradient flows, J. Comput. Phys. 353 (2018) 407–416.

[2] Jie Shen, Jie Xu, Jiang Yang, A new class of efficient and robust energy stable schemes for gradient flows, SIAM Rev. 61 (3) (2019) 474–506.

[3] Jie Shen, Efficient and accurate structure preserving schemes for complex nonlinear systems, in: Processing, Analyzing and Learning of Images, Shapes, and Forms. Part 2, in: Handb. Numer. Anal., vol. 20, Elsevier/North-Holland, Amsterdam, 2019, pp. 647–669.

[4] Jie Shen, Xiaofeng Yang, The IEQ and SAV approaches and their extensions for a class of highly nonlinear gradient flow systems, in: 75 Years of Mathematics of Computation, in: Contemp. Math., vol. 754, Amer. Math. Soc., Providence, RI, 2020, pp. 217–245.

[5] Daniel Kessler, Ricardo H. Nochetto, Alfred Schmidt, A posteriori error control for the Allen–Cahn problem: circumventing Grönwall's inequality, ESAIM Math. Model. Numer. Anal. 38 (1) (2004) 129–142.

[6] Jie Shen, Xiaofeng Yang, Numerical approximations of allen-cahn and cahn-hilliard equations, Discrete Contin. Dyn. Syst.-A 28 (4) (2010) 1669.

[7] Nicolas Condette, Christof Melcher, Endre Süli, Spectral approximation of pattern-forming nonlinear evolution equations with double-well potentials of quadratic growth, Math. Comp. 80 (273) (2011) 205–223.

[8] Dong Li, Zhonghua Qiao, Tao Tang, Characterizing the stabilization size for semi-implicit Fourier-spectral method to phase field equations, SIAM J. Numer. Anal. 54 (3) (2016) 1653–1681.

[9] Tao Tang, On effective numerical methods for phase-field models, in: Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018, World Scientific, 2018, pp. 3669–3690.

[10] Qiang Du, Lili Ju, Xiao Li, Zhonghua Qiao, Maximum principle preserving exponential time differencing schemes for the nonlocal allen-cahn equation, SIAM J. Numer. Anal. 57 (2) (2019) 875–898.

[11] Qiang Du, Lili Ju, Xiao Li, Zhonghua Qiao, Maximum bound principles for a class of semilinear parabolic equations and exponential time-differencing schemes, SIAM Rev. 63 (2) (2021) 317–359.

[12] Jie Shen, Jie Xu, Convergence and error analysis for the scalar auxiliary variable (SAV) schemes to gradient flows, SIAM J. Numer. Anal. 56 (5) (2018) 2895–2912.

[13] Yanxia Qian, Zhiguo Yang, Fei Wang, Suchuan Dong, GPAV-based unconditionally energy-stable schemes for the cahn–hilliard equation: Stability and error analysis, Comput. Methods Appl. Mech. Engrg. 372 (2020) 113444.

[14] Xiaoli Li, Jie Shen, Hongxing Rui, Energy stability and convergence of SAV block-centered finite difference method for gradient flows, Math. Comp. 88 (319) (2019) 2047–2068.
[15] Hongtao Chen, Jingjing Mao, Jie Shen, Optimal error estimates for the scalar auxiliary variable finite-element schemes for gradient flows, Numer. Math. 145 (1) (2020) 167–196.
[16] Xiaoli Li, Jie Shen, Stability and error estimates of the SAV Fourier-spectral method for the phase field crystal equation, Adv. Comput. Math. 46 (3) (2020) Paper No. 48, 20.
[17] Min Wang, Qiumei Huang, Cheng Wang, A second order accurate scalar auxiliary variable (SAV) numerical method for the square phase field crystal equation, J. Sci. Comput. 88 (2) (2021) Paper No. 33, 36.
[18] Xiaoli Li, Jie Shen, Error analysis of the SAV-MAC scheme for the Navier–stokes equations, SIAM J. Numer. Anal. 58 (5) (2020) 2465–2491.
[19] Georgios Akrivis, Buyang Li, Dongfang Li, Energy-decaying extrapolated RK-SAV methods for the Allen-Cahn and Cahn-Hilliard equations, SIAM J. Sci. Comput. 41 (6) (2019) A3703–A3727.
[20] Dongfang Li, Weiwei Sun, Linearly implicit and high-order energy-conserving schemes for nonlinear wave equations, J. Sci. Comput. 83 (3) (2020) Paper No. 65, 17.
[21] Dong Li, Chaoyu Quan, Wen Yang, The BDF3/EP3 scheme for MBE with no slope selection is stable, J. Sci. Comput. 89 (2) (2021) Paper No. 33, 24.
[22] Fukeng Huang, Jie Shen, Zhiguo Yang, A highly efficient and accurate new scalar auxiliary variable approach for gradient flows, SIAM J. Sci. Comput. 42 (4) (2020) A2514–A2536.
[23] Olavi Nevanlinna, F. Odeh, Multiplier techniques for linear multistep methods, Numer. Funct. Anal. Optim. 3 (4) (1981) 377–423.
[24] Georgios Akrivis, Stability of implicit-explicit backward difference formulas for nonlinear parabolic equations, SIAM J. Numer. Anal. 53 (1) (2015) 464–484.
[25] Jie Shen, Tao Tang, Li-Lian Wang, Spectral Methods: Algorithms, Analysis and Applications, Vol. 41, Springer Science & Business Media, 2011.
[26] S.M. Allen, J.W. Cahn, A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening, Acta Metall. Mater. 27 (1979) 1085–1095.
[27] Roger Temam, Infinite-Dimensional Dynamical Systems in Mechanics and Physics, Vol. 68, Springer Science & Business Media, 2012.
[28] John W. Cahn, John E. Hilliard, Free energy of a nonuniform system. I. Interfacial free energy, J. Chem. Phys. 28 (2) (1958) 258–267.
[29] Georgios Akrivis, Minghua Chen, Fan Yu, Zhi Zhou, The energy technique for the six-step BDF method, SIAM J. Numer. Anal. 59 (5) (2021) 2449–2472.