Journal of Applied Analysis and Computation
Volume 10, Number 5, October 2020, 2198–2219

Website:http://jaac-online.com
DOI:10.11948/20190373

REMARKS ON THE ASYMPTOTIC BEHAVIOR OF SCALAR AUXILIARY VARIABLE (SAV) SCHEMES FOR GRADIENT-LIKE FLOWS

Anass Bouchriti[1], Morgan Pierre[2,†] and Nour Eddine Alaa[1]

Abstract We introduce a time semi-discretization of a damped wave equation by a SAV scheme with second order accuracy. The energy dissipation law is shown to hold without any restriction on the time step. We prove that any sequence generated by the scheme converges to a steady state (up to a subsequence). We notice that the steady state equation associated to the SAV scheme is a modified version of the steady state equation associated to the damped wave equation. We show that a similar result holds for a SAV fully discrete version of the Cahn-Hilliard equation and we compare numerically the two steady state equations.

Keywords Gradient flow, SAV schemes, BDF methods, sine-Gordon equation, Cahn-Hilliard equation.

MSC(2010) 65M06, 65M60, 35B40.

1. Introduction

Gradient flow equations are essential in materials science and engineering or physically motivated problems. A gradient flow is a dynamic driven by a free energy and a dissipation mechanism. It has the general form

$$\frac{\partial \phi}{\partial t} = \mathcal{G}\frac{\delta E}{\delta \phi}, \tag{1.1}$$

where ϕ is the state variable of the system and $\frac{\delta E}{\delta \phi}$ is the variational derivative of a free energy $E[\phi]$. The nonpositive symmetric operator $\mathcal{G}$ partakes in the dissipation of the energy through

$$\frac{d}{dt}E[\phi] = \frac{\delta E}{\delta \phi}\frac{\partial \phi}{\partial t} = (\frac{\delta E}{\delta \phi}, \mathcal{G}\frac{\delta E}{\delta \phi}) \leq 0.$$

In a large class of physical models, on which we focus here, $\mathcal{G}$ is a differential operator with constant coefficients (such as $-Id$, Δ, ...) and the derivative $\frac{\delta E}{\delta \phi}$ can

[†] The corresponding author.
Email address: morgan.pierre@math.univ-poitiers.fr (M. Pierre)
[1] LAMAI Laboratory, Faculty of Science and Technology, Cadi Ayyad University, Marrakesh, Morocco
[2] Laboratoire de Mathématiques et Applications, Université de Poitiers, CNRS, F-86073 Poitiers, France

be written as

$$\frac{\delta E}{\delta \phi} = \mathcal{L}\phi + F'(\phi), \tag{1.2}$$

where $\mathcal{L}$ is a linear symmetric nonnegative operator and the nonlinear part $F(\phi)$ is the free energy density. In such case, the free energy reads

$$E[\phi] = \frac{1}{2}(\mathcal{L}\phi, \phi) + (F(\phi), 1).$$

The nonlinear term $F'(\phi)$ can also include derivatives of order less than $\mathcal{L}$ thus it $(F'(\phi))$ can be perceived as the variational derivative of $F(\phi)$. This framework includes the Cahn-Hilliard equation [6], the Allen-Cahn equation [2], the phase field crystal equation [8], epitaxial growth models [19], ...

Most of the solutions of gradient flows are analytically unattainable, so that attempts to obtain concrete and reliable solutions must resort to numerical methods. Many efforts have been put on finding numerical schemes which preserve the dissipation mechanism (see [1,3,5,12,21,32] and references therein). This ensures the stability of the numerical solution on long time scales.

Regarding the time discretization of (1.1), one of the most efficient approach is to treat the linear term implicitly and the nonlinear term explicitly in (1.2). However, for Cahn-Hilliard type equations, the assumptions on the nonlinearity $F(\phi)$ and the restriction on the time step remain severe [29]. In a recent paper of Shen et al [28], these inconveniences are avoided by introducing the so-called scalar auxiliary variable (SAV). The variable expands the degree of freedom of the dynamical system.

We note that the SAV approach is an enhanced version of the invariant energy quadratization (IEQ) (see Section 2 for details). It has been applied to a great variety of situations (see, e.g., [7,28,30]) and at every time step, only two linear PDEs with constant coefficients need to be solved. This can be very efficiently done with a FFT or a spectral method. The dissipation mechanism has been proved to hold for first order and second order SAV schemes (BDF2, Crank-Nicolson) without any restriction on the time step (unconditional stability). Convergence of the SAV method as the time step goes to 0 has been proved in [24,27].

In this note, we investigate the asymptotic behavior of solutions generated by SAV schemes, as time goes to infinity. We focus on a time semi-discretization of the damped wave equation with second order accuracy, namely the backward differentiation formula of order 2 (SAV/BDF2).

Our damped wave equation includes the dissipative sine-Gordon equation and the modified Allen-Cahn equation as particular cases. It is a model problem for evolution equations of the form

$$\frac{\partial^2 \phi}{\partial t^2} + \frac{\partial \phi}{\partial t} = \mathcal{G}\frac{\delta E}{\delta \phi}, \tag{1.3}$$

with $\mathcal{G}$ and E as previously. In (1.3), the dissipation mechanism reads, assuming that $\mathcal{G}$ is invertible,

$$\begin{aligned}\frac{d}{dt}\left(\frac{1}{2}(\frac{\partial \phi}{\partial t}, (-\mathcal{G})^{-1}\frac{\partial \phi}{\partial t}) + E[\phi]\right) &= (\frac{\partial^2 \phi}{\partial t^2}, (-\mathcal{G})^{-1}\frac{\partial \phi}{\partial t}) + \frac{\delta E}{\delta \phi}\frac{\partial \phi}{\partial t} \\ &= (\frac{\partial \phi}{\partial t}, (\mathcal{G})^{-1}\frac{\partial \phi}{\partial t}) \leq 0.\end{aligned}$$

We see here that a modified energy is nonincreasing, namely the sum of a kinetic energy and of the free energy $E[\phi]$. Equations such as (1.3) are sometimes called "gradient-like flows."

Many schemes which were developed for gradient flows have also been adapted to gradient-like flows such as (1.3) (see [1, 14] and references therein). An IEQ approach has been proposed for the modified phase field crystal equation in [23]. But up to now, SAV schemes do not seem to have been considered so we explain our approach in Section 2.

Then, we prove that the SAV/BDF2 scheme for the damped wave equation satisfies an energy dissipation law where no restriction on the time step is required (unconditional stability, cf. Proposition 3.1). Moreover, we show that the scheme can be implemented into an efficient way by solving two linear second-order equations with constant coefficients at each time step (Section 3.4). In Theorem 3.1, we demonstrate that any sequence generated by the SAV scheme converges to a stationary state (up to a subsequence). To the extent of our knowledge, this is the first asymptotic result for a second order time semi-discretization of the damped wave equation. In contrast, for most of second order schemes which are known to preserve the dissipation mechanism in (1.3), obtaining the precompactness of trajectories is an issue [14]. We note that the standard BDF2 scheme applied to the gradient-like flow (1.3) does not preserve the dissipation, unlike what happens for the gradient flow [5].

Our analysis shows that, because of the additional scalar auxiliary variable, the steady state equation for the SAV scheme is a modified version of the steady state equation for (1.1). This is a drawback of the SAV method since it could drastically modify the longtime dynamics. We make this clear by examining several examples in Section 3.6. In the last section, we perform numerical simulations for the Cahn-Hilliard equation using a first order SAV scheme for the time discretization and a finite element space discretization. This allows a comparison with the theoretical results for a model gradient flow.

2. The SAV method for gradient-like flows

2.1. The IEQ and SAV methods for gradient flows

The SAV method in an enhanced version of the invariant energy quadratization (IEQ) method which was introduced in the work of [33] based on a Lagrange multiplier approach [4, 15]. In these methods, a fundamental idea is to add a variable to the gradient flow (1.1) in such a way that the dissipation mechanism is preserved.

The time continuous IEQ version of the gradient flow (1.1) associated to the energy (1.2) reads

$$\begin{cases} \dfrac{\partial \phi}{\partial t} = \mathcal{G}\left(\mathcal{L}\phi + \dfrac{q}{\sqrt{F(\phi)+C_0}} F'(\phi)\right), \\ \dfrac{\partial q}{\partial t} = \dfrac{F'(\phi)}{2\sqrt{F(\phi)+C_0}} \dfrac{\partial \phi}{\partial t}, \end{cases}$$

where $q(t,x) = \sqrt{F(\phi)+C_0}$. The free energy density $F(\phi)$ is assumed to be bounded from below as C_0 is a constant large enough such that $F(\phi)+C_0 > 0$.

For the case of SAV, the reformulated system reads

$$\begin{cases} \dfrac{\partial \phi}{\partial t} = \mathcal{G} w, \\ w = \mathcal{L}\phi + \dfrac{r}{\sqrt{\int F(\phi) dx + C_0}} F'(\phi), \\ \dfrac{dr}{dt} = \dfrac{1}{2\sqrt{\int F(\phi) dx + C_0}} \displaystyle\int F'(\phi)\phi_t dx, \end{cases} \tag{2.1}$$

where the variable $q(t, x)$ is replaced by the *scalar* variable

$$r(t) = \sqrt{\int F(\phi) dx + C_0},$$

with C_0 large enough. The intermediate variable $w(x, t)$, which does not change the dynamical system since $\frac{\partial w}{\partial t}$ is not involved, is costumary in Cahn-Hilliard type problems.

We note here that only $\int F(\phi) dx$ is required to be bounded from below. This allows the possibility of studying a relatively larger class of physical models since only a few acquire the boundedness of the free energy density [28]. We can easily obtain a modified energy dissipation law by taking the inner products of the above equations with w, $-\frac{\partial \phi}{\partial t}$ and $2r$ respectively. It reads

$$\frac{d}{dt}\left(\frac{1}{2}(\mathcal{L}\phi, \phi) + r^2\right) = (\mathcal{G}w, w) \le 0.$$

The SAV system (2.1) is then discretized in time with an explicit expression for the nonlinear terms and an implicit expression for the linear terms. For instance, the first order time discrete SAV scheme reads

$$\begin{cases} \dfrac{\phi^{n+1} - \phi^n}{\Delta t} = \mathcal{G} w^{n+1}, \\ w^{n+1} = \mathcal{L}\phi^{n+1} + \dfrac{r^{n+1}}{\sqrt{\int F(\phi^n) dx + C_0}} F'(\phi^n), \\ r^{n+1} - r^n = \dfrac{1}{2\sqrt{\int F(\phi^n) dx + C_0}} \displaystyle\int F'(\phi^n)(\phi^{n+1} - \phi^n) dx. \end{cases} \tag{2.2}$$

The discrete dissipation law is obtained by taking the inner product of the equations in (2.2) by $\Delta t w^{n+1}$, $\phi^n - \phi^{n+1}$ and $2r^{n+1}$ respectively, and using the well-known identity $2a(a-b) = a^2 - b^2 + (a-b)^2 \ge a^2 - b^2$. It reads

$$E^{n+1} - E^n \le \Delta t(\mathcal{G}w^{n+1}, w^{n+1}) \le 0,$$

where $E^n = \frac{1}{2}(\mathcal{L}\phi^n, \phi^n) + (r^n)^2$. No restriction on the time step is required.

While IEQ and SAV are very identical from a numerical analysis point of view, taking away the space dependency of the variable makes the numerical solving process **greatly simplified**. At every time step, only two linear equations with *constant coefficients* need to be solved for a SAV scheme, whereas the IEQ requires solving two equations with *variable coefficients*. We refer the reader to Sections 3.4 and 4.3 for more details.

2.2. The SAV approach for gradient-like flows

In order to derive a SAV scheme for the gradient-like flow (1.3), we write the second order equation as a first order system and we introduce the scalar variable $r(t)$ as previously. This reads

$$\begin{cases} \dfrac{\partial \phi}{\partial t} = \psi, \\ \dfrac{\partial \psi}{\partial t} = -\psi + \mathcal{G} w, \\ w = \mathcal{L}\phi + \dfrac{r}{\sqrt{\int F(\phi)dx + C_0}} F'(\phi), \\ \dfrac{dr}{dt} = \dfrac{1}{2\sqrt{\int F(\phi)dx + C_0}} \displaystyle\int F'(\phi)\phi_t dx. \end{cases} \tag{2.3}$$

In (2.3), we take the inner product of the second equation by $-\mathcal{G}^{-1}\psi$, of the third equation by $-\frac{\partial \phi}{\partial t}$ and of the last equation by $2r$. This yields the dissipation law

$$\frac{d}{dt}\left(\frac{1}{2}(\psi, (-\mathcal{G})^{-1}\psi) + \frac{1}{2}(\mathcal{L}\phi, \phi) + r^2\right) = (\psi, \mathcal{G}^{-1}\psi) \leq 0.$$

We can obtain a SAV first time semi-discrete scheme which preserves the dissipation by treating the nonlinear term explicitly and the linear term implicitly, as in (2.2). Second order schemes based on BDF2 or Crank-Nicolson can also be derived as for the gradient flows (see Section 3.2). We note that an IEQ approach has been used in a similar manner for the conservative sine-Gordon equation in [18].

3. A second order SAV scheme for the damped wave equation

3.1. Assumptions and notation

Let Ω denote a bounded domain of $\mathbb{R}^d$ $(d \geq 1)$ with sufficiently smooth boundary $\partial\Omega$. We consider the following damped wave equation,

$$u_{tt} + u_t - \Delta u + f(u) = 0 \text{ in } \Omega \times (0, +\infty), \tag{3.1}$$

subject to homogeneous Dirichlet on Ω. The well-posedness and asymptotic behavior of equation (3.1) has been thoroughly studied (see, e.g., [10, 16, 17, 31, 34]).

Let $|\cdot|_0$ and $(\cdot,\cdot)$ be the standard $L^2(\Omega)$ norm and scalar product. We denote by $|\cdot|_1 = |\nabla\cdot|_0$ the Hilbertian norm of the Sobolev space $H_0^1(\Omega)$. Recall that $-\Delta : H_0^1(\Omega) \longrightarrow (H_0^1(\Omega))'$ is an isomorphism associated to the inner product on $H_0^1(\Omega)$ through

$$\langle -\Delta u, v\rangle_{(H_0^1(\Omega))' \times H_0^1(\Omega)} = (\nabla u, \nabla v) \qquad \forall u, v \in H_0^1(\Omega).$$

We assume that the nonlinearity $f : \mathbb{R} \to \mathbb{R}$ is continuous on $\mathbb{R}$ and if $d \geq 2$, we assume that there exist a positive constant c and a nonnegative real number p such that

$$(d-2)p < 4 \quad \text{and} \quad |f(s)| \leq c(1 + |s|^{p+1}) \qquad \forall s \in \mathbb{R}. \tag{3.2}$$

If $d = 1$, no growth assumption is required on f.

We define F as the function $F(s) = \int_0^s f(x)dx$. By (3.2), we have

$$|F(s)| \leq c'(1 + |s|^{p+2}) \quad \forall s \in \mathbb{R},$$

for some positive constant c'. The assumption on p yields the Sobolev imbedding $H_0^1(\Omega) \subset L^{p+2}(\Omega)$ so that for all $u \in H_0^1(\Omega)$, $\int_\Omega F(u)dx$ is well defined [11]. We assume furthermore that there exists a constant C_0 such that

$$\int_\Omega F(u)dx + C_0 \geq \varepsilon > 0, \quad \forall u \in H_0^1(\Omega), \tag{3.3}$$

for some $\varepsilon > 0$ independent of u.

The growth assumption on f implies that the functional $u \mapsto \int_\Omega F(u)dx$ is of class C^1 on $H_0^1(\Omega)$ and that the (nonlinear) operator

$$u \mapsto f(u) \tag{3.4}$$

is continuous from $H_0^1(\Omega)$ into $L^{(p+2)/(p+1)}(\Omega)$ [20]. Moreover, by Hölder's inequality, for every $u, v \in H_0^1(\Omega)$, $f(u)v \in L^1(\Omega)$ and $L^{(p+2)/(p+1)}(\Omega)$ is continuously imbedded in $(H_0^1(\Omega))'$.

Formally, by multiplying equation (3.1) with u_t and integrating on Ω, one gets

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |u_t|^2 dx + \int_\Omega |u_t|^2 dx + \frac{d}{dt}\int_\Omega \frac{1}{2}|\nabla u|^2 + F(u)dx = 0.$$

In particular, the functional $t \mapsto \tilde{E}(u(t), u_t(t))$ is nonincreasing in time, where

$$\tilde{E}(u, v) := \int_\Omega \frac{1}{2}|\nabla u|^2 + F(u) + \frac{1}{2}|v|^2 dx. \tag{3.5}$$

Example 3.1. If $f(s) = \sin s$, then equation (3.1) is known as the sine-Gordon equation (see, e.g., [31]). We have $F(s) = 1 - \cos s$ and the nonlinearity satisfies the required assumptions (3.2)-(3.3) with $p = 0$ for any dimension $d \geq 1$.

Example 3.2. If $f(s) = s^3 - s$, equation (3.1) is sometimes called the modified Allen-Cahn equation. We have $F(s) = s^4/4 - s^2/2$ and the nonlinearity satisfies assumptions (3.2)-(3.3) with $p = 2$ for $1 \leq d \leq 3$.

3.2. The SAV/BDF2 time discretization

We formally rewrite (3.1) as a first-order system and we introduce scalar auxiliary variable (SAV) $r(t) = \sqrt{\int_\Omega F(u)dx + C_0}$ where C_0 is a constant which satisfies (3.3). This yields

$$u_t = v \tag{3.6}$$

$$v_t = -v + \Delta u - \frac{r}{\sqrt{\int_\Omega F(u)dx + C_0}} f(u) \tag{3.7}$$

$$r'(t) = \frac{1}{2\sqrt{\int_\Omega F(u)dx + C_0}} \int_\Omega f(u)u_t dx \tag{3.8}$$

Note that here, in comparison with (2.3), we do not introduce the intermediate variable w because $\mathcal{G}$ is simply $-Id$.

A second-order time semi-discretization by BDF2 reads

$$\frac{3u^{n+1}-4u^n+u^{n-1}}{2\delta t} = v^{n+1} \tag{3.9}$$

$$\frac{3v^{n+1}-4v^n+v^{n-1}}{2\delta t} = -v^{n+1}+\Delta u^{n+1} - \frac{r^{n+1}}{s^{n+1/2}}f(u^{n+1/2}) \tag{3.10}$$

$$3r^{n+1}-4r^n+r^{n-1} = \frac{1}{2s^{n+1/2}}\int_\Omega f(u^{n+1/2})(3u^{n+1}-4u^n+u^{n-1})dx, \tag{3.11}$$

where $u^{n+1/2}=2u^n-u^{n-1}$ and $s^{n+1/2}=\sqrt{\int_\Omega F(u^{n+1/2})dx+C_0}$.

3.3. Energy stability

We define the discrete energy functional associated to the dynamical system

$$E(u,v,r) = \frac{1}{2}\left|u\right|_1^2 + \frac{1}{2}\left|v\right|_0^2 + r^2. \tag{3.12}$$

As a shortcut, we write

$$E^n = E(u^n,v^n,r) = \frac{1}{2}\left|u^n\right|_1^2 + \frac{1}{2}\left|v^n\right|_0^2 + (r^n)^2. \tag{3.13}$$

We introduce the modified discrete energy associated to equations (3.9)-(3.11),

$$\mathcal{E}^n = E^n + \frac{1}{2}\left|2u^n-u^{n-1}\right|_1^2 + \frac{1}{2}\left|2v^n-v^{n-1}\right|_0^2 + |2r^n-r^{n-1}|^2.$$

Proposition 3.1. *The scheme* (3.9)-(3.11) *is unconditionally stable in the sense that*

$$\begin{aligned}&\mathcal{E}^{n+1} + \delta t\left|v^{n+1}\right|_0^2 + \frac{1}{2}\left|u^{n+1}-2u^n+u^{n-1}\right|_1^2\\ &+\frac{1}{2}\left|v^{n+1}-2v^n+v^{n-1}\right|_0^2 + |r^{n+1}-2r^n+r^{n-1}|^2 \le \mathcal{E}^n,\end{aligned} \tag{3.14}$$

for all $n \ge 1$.

Proof. We multiply (3.10) by $3u^{n+1}-4u^n+u^{n-1} = 2\delta t v^{n+1}$ and equation (3.11) by $2r^{n+1}$. We integrate on Ω and sum up the two equations. The nonlinear terms cancel, yielding the following equation,

$$\begin{aligned}&(3v^{n+1}-4v^n+v^{n-1},v^{n+1}) + 2\delta\left|v^{n+1}\right|_0^2\\ &+\left(\nabla u^{n+1},\nabla(3u^{n+1}-4u^n+u^{n-1})\right) + 2(3r^{n+1}-4r^n+r^{n-1})r^{n+1} = 0.\end{aligned}$$

The inequality (3.14) is a direct result of the following expansion [28]:

$$\begin{aligned}2(x^{n+1},3x^{n+1}-4x^n+x^{n-1}) = &\left|x^{n+1}\right|^2 + \left|2x^{n+1}-x^n\right|^2 - \left|x^n\right|^2 - \left|2x^n-x^{n-1}\right|^2\\ &+\left|x^{n+1}-2x^n+x^{n-1}\right|^2.\end{aligned}$$

□

3.4. An efficient solver

It is well-known that SAV schemes can be very efficiently solved. We show that this is the case for system (3.9)-(3.11).

We assume that (u^n, v^n, r^n) and $(u^{n-1}, v^{n-1}, r^{n-1})$ are known in $H_0^1(\Omega) \times L^2(\Omega) \times \mathbb{R}$. By eliminating v^{n+1} in (3.10), thanks to (3.9), we get the following expression:

$$Au^{n+1} = 4v^n - v^{n-1} + \frac{3+2\delta t}{2\delta t}\left(4u^n - u^{n-1}\right) - 2\delta t\, r^{n+1}\, q^n, \tag{3.15}$$

where

$$A = 2\delta t\left(\frac{9+6\delta t}{4\delta t^2}\mathrm{I} - \Delta\right) \text{ and } q^n = \frac{f(u^{n+1/2})}{s^{n+1/2}}.$$

By (3.11), we know that

$$r^{n+1} = \frac{1}{3}(4r^n - r^{n-1}) + \frac{1}{6s^{n+1/2}}\left(f(u^{n+1/2}), 3u^{n+1} - 4u^n + u^{n-1}\right). \tag{3.16}$$

Thus, substituting the last equation into (3.15) yields

$$Au^{n+1} + \delta t\left(u^{n+1}, q^n\right) q^n = g^n, \tag{3.17}$$

where

$$g^n = \frac{2+3\delta t}{2\delta t}\left(4u^n - u^{n-1}\right) + \frac{\delta t}{3}\left(q^n, 4u^n - u^{n-1}\right) q^n \\ +4v^n - v^{n-1} - 2\delta t(4r^n - r^{n-1})q^n.$$

In order to solve (3.17), the idea is first to compute the term (u^{n+1}, q^n). This can be done by applying A^{-1} to (3.17) and taking the inner product with q^n (we note that A is an isomorphism from $H_0^1(\Omega)$ into $(H_0^1(\Omega))'$ and that q^n belongs to $(H_0^1(\Omega))'$ by (3.4)). We obtain

$$\left(1 + \delta t\left(A^{-1}q^n, q^n\right)\right)(u^{n+1}, q^n) = \left(A^{-1}g^n, q^n\right).$$

The numerical implementation to solve u^{n+1} at each iteration can be resumed as follows:

i) Compute $A^{-1}q^n$ and $\left(A^{-1}q^n, q^n\right)$.
"*This consists in solving a linear second order equation with constant coefficients.*"

ii) Compute $A^{-1}g^n$ and $(u^{n+1}, q^n) = \left(A^{-1}g^n, q^n\right)/\left(1 + \delta t\left(A^{-1}q^n, q^n\right)\right)$.
"*This also consists in solving another linear second order equation with constant coefficients.*"

iii) Compute $u^{n+1} = A^{-1}g^n - \delta t\left(u^{n+1}, q^n\right) A^{-1}q^n$.
"*At this stage, all terms required in* (3.17) *are known. We can then easily find* u^{n+1}*.*"

Once $u^{n+1} \in H_0^1(\Omega)$ is known, $v^{n+1} \in L^2(\Omega)$ can be explicitly computed with (3.9) and r^{n+1} with (3.11).

3.5. Asymptotic behavior

In this section, we assume that (u^0, v^0, r^0) and (u^1, v^1, r^1) are given in $H^1_0(\Omega) \times L^2(\Omega) \times \mathbb{R}$. We have seen in Section 3.4 that the SAV/BDF2 scheme (3.9)-(3.11) generates a unique sequence $\big((u^n, v^n, r^n)\big)_n$ in $H^1_0(\Omega) \times L^2(\Omega) \times \mathbb{R}$. We are interested in the asymptotic behavior of this sequence as n goes to $+\infty$.

It is easily seen that $v^n \to 0$ in $L^2(\Omega)$ (see below). Thus, we focus on the sequence $(u^n, r^n)_n$ and we introduce its ω-limit set

$$\omega\left((u^n, r^n)_n\right) = \Big\{ (u^\star, r^\star) \in H^1_0(\Omega) \times \mathbb{R} \mid \exists n_k \longrightarrow \infty, u^{n_k} \longrightarrow u^\star \text{ in } H^1_0(\Omega) \\ \text{and } r^{n_k} \longrightarrow r^\star \text{ in } \mathbb{R} \Big\}.$$

For simplicity, we denote by $\tilde{V}$ the space $H^1_0(\Omega) \times \mathbb{R}$ associated with any standard norm (in our case, we chose $\| \,.\, \|_{\tilde{V}} = |\,.\,|_1 + |\,.\,|$). We have the following result.

Theorem 3.1. *Let $\big((u^n, v^n, r^n)\big)_n$ be a sequence in $H^1_0(\Omega) \times L^2(\Omega) \times \mathbb{R}$ generated by the SAV/BDF2 scheme* (3.9)-(3.11). *Then $v^n \to 0$ in $L^2(\Omega)$ and the set $\omega\left((u^n, r^n)_n\right)$ is a compact and connected subset of $\tilde{V}$ on which $(u, r) \mapsto E(u, 0, r)$ is constant. Furthermore, for each $(u^\star, r^\star) \in \omega\left((u^n, r^n)_n\right)$ we have*

$$-\Delta u^\star + \frac{r^\star}{s^\star} f(u^\star) = 0 \quad \text{in } (H^1_0(\Omega))', \tag{3.18}$$

where $s^\star = \sqrt{\int_\Omega F(u^\star) dx + C_0}$.

Proof. We know from Proposition (3.1) that the sequence $(\mathcal{E}^n)$ is nonincreasing. Since $(\mathcal{E}^n)$ is bounded from below by 0, this sequence converges to some $\mathcal{E}^*$ in $\mathbb{R}$. As a consequence (see (3.13)), the sequences (u^n), (v^n) and (r^n) are respectively bounded in $H^1_0(\Omega)$, $L^2(\Omega)$ and $\mathbb{R}$.

We claim that the sequence (u^n) is precompact in $H^1_0(\Omega)$. The idea is to demonstrate that (u^n) is bounded in a Sobolev space $\mathbb{W}^{2,q}(\Omega)$ for an appropriate value of $q > 1$.

By (3.10), we have the following expression

$$\Delta u^{n+1} = g(v^{n+1}, v^n, v^{n-1}) + \frac{r^{n+1}}{s^{n+1/2}} f(u^{n+1/2}), \tag{3.19}$$

where g is a linear combination of v^{n+1}, v^n and v^{n-1}. In particular, the sequence $g(v^{n+1}, v^n, v^{n-1})$ is bounded in $L^2(\Omega)$ and therefore in $L^q(\Omega)$ for any $1 < q \leq 2$. For the nonlinear term, we discuss according to d and we eventually demonstrate that $(f(u^{n+1/2}))_n$ is bounded in $L^q(\Omega)$ for an appropriate choice of $q > 1$.

If $d \geq 3$, the growth condition (3.2) on the nonlinearity implies that for any $q \in [1, 2]$,

$$||f(u^{n+1/2})||_{L^q(\Omega)} \leq C \left(||u^{n+1/2}||_{L^q(\Omega)} + ||u^{n+1/2}||_{L^{(p+1)q}(\Omega)} \right).$$

By the Sobolev imbedding $H^1_0(\Omega) \subset L^{2^*}(\Omega)$, we know that (u^{n+1}) is bounded in $L^{2^*}(\Omega)$ where $2^* = \frac{2d}{d-2}$. A sufficient condition for $(f(u^{n+1/2}))_n$ to be bounded in $L^q(\Omega)$ is that $(p+1)q \leq 2^* = \frac{2d}{d-2}$. The assumption on p reads $p(d-2) < 4$ which implies that $p+1 < \frac{d+2}{d-2}$. On choosing $q = \min\{\frac{2^*}{p+1}, 2\}$, we have $2d/(d+2) < q \leq 2$ and $(f(u^{n+1/2}))_n$ bounded in $L^q(\Omega)$, as required. Using elliptic regularity, we obtain that $(u^{n+1})_n$ is bounded in $\mathbb{W}^{2,q}(\Omega)$. For such a choice of q, $\mathbb{W}^{2,q}(\Omega)$ is compactly imbedded in $H_0^1(\Omega)$ [11]. The claim follows.

If $d = 1$ or $d = 2$, we obtain from the Sobolev imbeddings that $(f(u^{n+1/2}))_n$ is bounded in $L^q(\Omega)$, for any $1 < q < \infty$, so we may choose $q = 2$ and we conclude as previously.

We know that (r^n) is bounded in $\mathbb{R}$ and so the sequence $(u^n, r^n)_n$ is precompact in $\tilde{V}$. It is well-known that $\omega\left((u^n, r^n)_n\right)$ is closed in $\tilde{V}$ and therefore compact.

On summing (3.14) from $n = 1$ to $+\infty$, we find that

$$\sum_{n=1}^{\infty} \delta t|v^{n+1}|^2 + \frac{1}{2}|\nabla(u^{n+1} - 2u^n + u^{n-1})|^2 + |r^{n+1} - 2r^n + r^{n-1}|^2 \leq \mathcal{E}^1 < +\infty.$$

In particular,

$$(v^n) \longrightarrow 0 \qquad \text{in } L^2(\Omega), \tag{3.20}$$

$$\left(u^{n+1} - 2u^n + u^{n-1}\right) \longrightarrow 0 \qquad \text{in } H_0^1(\Omega), \tag{3.21}$$

$$(r^{n+1} - 2r^n + r^{n-1}) \longrightarrow 0 \qquad \text{in } \mathbb{R}. \tag{3.22}$$

Moreover, by (3.9),

$$3u^{n+1} - 4u^n + u^{n-1} = 2\delta t v^{n+1} \longrightarrow 0 \text{ in } L^2(\Omega).$$

Thus,

$$u^{n+1} - u^n = \frac{1}{2}\big((3u^{n+1} - 4u^n + u^{n-1}) - (u^{n+1} - 2u^n + u^{n-1})\big) \longrightarrow 0 \text{ in } L^2(\Omega).$$

In a same manner, we obtain that

$$r^{n+1} - r^n \longrightarrow 0 \text{ in } \mathbb{R}. \tag{3.23}$$

By precompactness of the sequence (u^n) in $H_0^1(\Omega)$, we deduce that

$$u^{n+1} - u^n \to 0 \text{ in } H_0^1(\Omega). \tag{3.24}$$

We conclude, using a standard result, that $\omega\left((u^n, r^n)_n\right)$ is connected in $\tilde{V}$ (see Lemma 3.1).

Next, we write

$$\begin{aligned}\left|2u^n - u^{n-1}\right|_1^2 &= \left|u^n + (u^n - u^{n-1})\right|_1^2 \\ &= |u^n|_1^2 + (\nabla u^n, \nabla(u^n - u^{n-1})) + \left|u^n - u^{n-1}\right|_1^2 \\ &= |u^n|_1^2 + \varepsilon_1^n\end{aligned}$$

where $\varepsilon_1^n \longrightarrow 0$. Similarly,

$$|2r^n - r^{n-1}|^2 = (r^n)^2 + \varepsilon_2^n$$

where $\varepsilon_2^n \longrightarrow 0$. Thus,

$$\mathcal{E}^n = |u^n|_1^2 + 2(r^n)^2 + \varepsilon^n,$$

where $\varepsilon^n \longrightarrow 0$. This implies that

$$\mathcal{E}^n - \varepsilon^n \longrightarrow \mathcal{E}^\star = \lim\left(|u^n|_1^2 + 2(r^n)^2\right),$$

and so $(u,r) \mapsto 2E(u,0,r) = |u|_1^2 + 2(r)^2$ is a constant equal to $\mathcal{E}^\star$ on the set $\omega\left((u^n,r^n)_n\right)$.

Let now $(u^\star,r^\star) \in \omega\left((u^n,r^n)_n\right)$. There exists a subsequence (u^{n_k},r^{n_k}) of $(u^n,r^n)_n$ such that $(u^{n_k},r^{n_k}) \to (u^\star,r^\star)$ in $\tilde{V}$. We have (recall (3.24))

$$u^{n_k+1/2} = 2u^{n_k} - u^{n_k-1} \to u^\star \text{ in } H_0^1(\Omega).$$

Thus, $s^{n_k+1/2} \to s^\star$ in $\mathbb{R}$. By letting n_k tend to $+\infty$ in (3.10) we obtain the steady state equation (3.18). □

For the reader's convenience, we prove the following result.

Lemma 3.1. *The set $\omega\left((u^n,r^n)_n\right)$ is connected in $\tilde{V}$.*

Proof. Assume that $\omega\left((u^n,r^n)_n\right)$ is not connected in $\tilde{V}$. Then we can find two nonempty closed sets K_1 and K_2 in $\tilde{V}$ such that

$$\omega\left((u^n,r^n)_n\right) = K_1 \cup K_2 \quad \text{and} \quad K_1 \cap K_2 = \emptyset.$$

We note that K_1 and K_2 are compact and we define

$$2\varepsilon = \inf_{x\in K_1, y\in K_2} \|x-y\|_{\tilde{V}} = \min_{(x,y)\in K_1\times K_2} \|x-y\|_{\tilde{V}} > 0.$$

For simplicity, we write $w^n = (u^n,r^n)$. By (3.23) and (3.24), $\|w^{n+1} - w^n\|_{\tilde{V}} \longrightarrow 0$, so there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $\|w^{n+1} - w^n\|_{\tilde{V}} < \varepsilon$. For $i = 1,2$, we define the infinite sets $I_i = \{n \geq N \ : \ d(w^n,K_i) < \varepsilon/2\}$ $\big(K_i$ are nonempty subset included in $\omega((u^n,r^n)_n)\big)$. We can always find a certain $n_0 \geq N$ such that $n_0 \in I_1$, $n_0+1 \in I_2$ and $n_0 + 1 \notin I_1$. The infimum is reached and so we write

$$d(w^{n_0},K_1) = \|w^{n_0} - v^1\|_{\tilde{V}} \quad \text{and} \quad d(w^{n_0+1},K_2) = \|w^{n_0+1} - v^2\|_{\tilde{V}},$$

where $v^1 \in K_1$ and $v^2 \in K_2$. We have

$$w^{n_0+1} - w^{n_0} = (v^2 - v^1) + (w^{n_0+1} - v^2) + (v^1 - w^{n_0}),$$

hence,

$$\begin{aligned}\|w^{n_0+1} - w^{n_0}\|_{\tilde{V}} &\geq \|v^2 - v^1\|_{\tilde{V}} - \left(\|w^{n_0+1} - v^2\|_{\tilde{V}} + \|v^1 - w^{n_0}\|_{\tilde{V}}\right)\\ &> 2\varepsilon - \varepsilon/2 - \varepsilon/2 = \varepsilon.\end{aligned}$$

This is absurd. The set $\omega\left((u^n,r^n)_n\right)$ is therefore connected in $\tilde{V}$, as claimed. □

Remark 3.1. Theorem 3.1 shows that the sequence generated by the SAV scheme converges to a steady state, up to a subsequence. We did not manage to prove that the whole sequence converges to a single equilibrium, even with additional assumptions on f. In contrast, in [26], we proved by means of a Lojasiewicz-Simon inequality that the sequence generated by the backward Euler scheme applied to (3.1) converges to a single equilibrium, for analytic nonlinearities satisfying a one-sided Lipschitz condition.

3.6. Analysis of the steady state equation for three examples

In this section, we compare the steady state equations for the damped wave equation (3.1) and for the SAV time semi-discrete version (3.9)-(3.11).

3.6.1. The two steady state equations

A steady state $u \in H_0^1(\Omega)$ for the damped wave equation is a solution of (3.1) which does not depend on time, that is a solution of the elliptic PDE

$$-\Delta u + f(u) = 0 \text{ in } \Omega. \tag{3.25}$$

If we consider the damped wave equation (3.1) as a first order system acting on $H_0^1(\Omega) \times L^2(\Omega)$, that is

$$\begin{cases} u_t = v, \\ v_t = -v + \Delta u - f(u), \end{cases} \tag{3.26}$$

we see that $(u^\star, v^\star) \in H_0^1(\Omega) \times L^2(\Omega)$ is a steady state for (3.26) (i.e. a solution which does not depend on time) if and only if $(u^\star, v^\star) = (u^\star, 0)$ is a critical point of the C^1 functional $\tilde{E}(u,v)$ defined by (3.5) in $H_0^1(\Omega) \times L^2(\Omega)$.

Concerning the SAV approach, we note the following.

Definition 3.1. A triple $(u^\star, v^\star, r^\star) \in H_0^1(\Omega) \times L^2(\Omega) \times \mathbb{R}$ is a stationary state for the SAV scheme (3.9)-(3.11) (that is, a constant sequence which complies with the scheme) if and only if $v^\star = 0$ and

$$-\Delta u^\star + \frac{r^\star}{s^\star} f(u^\star) = 0 \text{ in } \Omega, \tag{3.27}$$

where $s^\star = \sqrt{\int_\Omega F(u^\star) dx + C_0}$.

The functional $E(u,v,r)$ (cf. (3.12)) has a unique critical point in $H_0^1(\Omega) \times L^2(\Omega) \times \mathbb{R}$ which is $(0,0,0)$, but there are a lot more steady states than $(0,0,0)$. We have a Lyapunov functional for the dynamical system (3.9)-(3.11) which does not drive every solution to its unique critical point $(0,0,0)$ (this is also true for the continuous-in-time SAV dynamical system (3.6)-(3.8), so it is not a consequence of the time discretization: it is a consequence of the additional auxiliary variable r). However, Theorem 3.1 shows that the energy dissipation implies that, up to a subsequence, every sequence generated by the SAV scheme converges to a steady state which solves (3.27).

By setting $\alpha = r^\star / s^\star$ in (3.27), we are led to seek the functions $u \in H_0^1(\Omega)$ which solve

$$-\Delta u + \alpha f(u) = 0 \text{ in } \Omega, \tag{3.28}$$

for some constant $\alpha \in \mathbb{R}$. In general $\alpha \neq 1$ so that (3.28) is only a modified version of (3.25).

3.6.2. The linear damped wave equation

We first assume that $f(u) = u$, in which case (3.1) is the linear damped wave equation. Equation (3.25) reads

$$-\Delta u + u = 0 \text{ in } \Omega.$$

This linear PDE has a unique solution in $H_0^1(\Omega)$, namely $u = 0$. In contrast, equation (3.28) reads

$$-\Delta u + \alpha u = 0 \text{ in } \Omega. \tag{3.29}$$

This is a well-known eigenvalue problem. Let $0 < \lambda_1 \leq \lambda_2 \leq \cdots$ denote the eigenvalues of $-\Delta$ with Dirichlet boundary conditions. If $\alpha = -\lambda_i$ for some i, then (3.29) has a continuum of solutions corresponding to the eigenspace for λ_i. If $\alpha \neq -\lambda_i$ for every i, then the unique solution to (3.29) is $u = 0$. With this simple example, we see that the set of steady states is drastically different for the damped wave equation and its SAV discretization.

3.6.3. The dissipative sine-Gordon equation

Next, we assume that $f(u) = \sin u$ and $d \geq 1$ (cf. Example 3.1). If $d = 1$ we simply assume that Ω is an interval. If $d \geq 2$, we assume that Ω is a bounded domain with a nonnegative mean curvature (this holds for instance if Ω is convex or star-shaped), or that Ω is an annulus of $\mathbb{R}^d$.

Equation (3.28) reads

$$-\Delta u + \alpha \sin u = 0 \text{ in } \Omega, \tag{3.30}$$

with $\alpha \in \mathbb{R}$. If $\alpha > 0$ (in particular if $\alpha = 1$ as in (3.25)), then the unique solution to (3.30) is $u = 0$. This is a consequence of Pohozaev's identity in case Ω is star-shaped [20]. The other situations have been considered in [13]. The linearized equation of (3.30) at $u = 0$ reads

$$-\Delta w + \alpha w = 0 \text{ in } \Omega,$$

and we recover the previous eigenvalue problem. Using a bifurcation approach (see, e.g., [22]), a bifurcation branch is likely to start for (3.30) at every singular value $\alpha = -\lambda_i$ where λ_i is an eigenvalue of $-\Delta$ as previously. Moreover, for all $\alpha < -\lambda_1$, the functional

$$u \mapsto \int_\Omega \frac{1}{2}|\nabla u|^2 + \alpha(1 - \cos u) dx$$

has at least one global minimizer in $H_0^1(\Omega)$, which therefore solves (3.30), and which cannot be identically equal to 0 because 0 is unstable (cf. next example). Again, the set of steady states for (3.30) is therefore very different from the case $\alpha = 1$.

3.6.4. The modified Allen-Cahn equation

Last, we consider the case where $f(u) = u^3 - \beta u$ with $\beta > \lambda_1 (> 0)$ and $1 \leq d \leq 3$ (cf. Example 3.2). Equation (3.25) reads

$$-\Delta u + u^3 - \beta u = 0 \text{ in } \Omega. \tag{3.31}$$

Let $\hat{E}(u) = \int_\Omega \frac{1}{2}|\nabla u|^2 + F(u) dx$ with $F(s) = \int_0^s f(s) ds = s^4/4 - \beta s^2/2$. A function $u \in H_0^1(\Omega)$ solves (3.31) if and only if u is a critical point of $\hat{E}$ in $H_0^1(\Omega)$. By considering a minimizing sequence, it is easily seen that E has a global minimizer $u^\star$ in $H_0^1(\Omega)$, which is therefore a solution of (3.31). On the other hand $\hat{E}$ is of class C^2 on $H_0^1(\Omega)$ and its hessian at some point u reads

$$\langle d^2 \hat{E}(u) h, h \rangle = \int_\Omega |\nabla h|^2 + (3u^2 - \beta) h^2 dx.$$

Since $\beta > \lambda_1$, we see that the critical point 0 is unstable, so that $u^\star \not\equiv 0$ on Ω. Indeed, on choosing $h = e_1$ as an eigenvector associated to λ_1, we have

$$\langle d^2\hat{E}(0)e_1, e_1\rangle = \int_\Omega |\nabla e_1|^2 - \beta|e_1|^2 dx = (\lambda_1 - \beta)|e_1|_0^2 < 0.$$

Thus, 0 is not a global minimizer of $\hat{E}$ in $H_0^1(\Omega)$.

As a consequence, the pair $(u^\star, 0) \in H_0^1(\Omega) \times L^2(\Omega)$ is a global minimizer in $H_0^1(\Omega) \times L^2(\Omega)$ of the functional $\tilde{E}(u,v)$ defined by (3.5), whereas $(0,0)$ is a critical point of $\tilde{E}(u,v)$ in $H_0^1(\Omega) \times L^2(\Omega)$ which is *not* a global minimizer. In contrast, $(0,0,0) \in H_0^1(\Omega) \times L^2(\Omega) \times \mathbb{R}$ is the unique global minimizer of the functional $E(u,v,r)$ (cf. (3.12)) associated to the SAV time discrete scheme. This shows that the stability of a critical point can change drastically between the damped wave equation (3.1) and its SAV time discrete version.

In the following result, we still assume that $f(s) = s^3 - \beta s$ with $\beta > \lambda_1$ and $1 \le d \le 3$.

Proposition 3.2. *For every $\alpha \in (\lambda_1/\beta, +\infty) \setminus \{1\}$, there exists $u_\alpha \in H_0^1(\Omega)$ which solves* (3.28) *but not* (3.25). *Moreover, there exists a sequence $(\alpha_k)_{k\in\mathbb{N}}$ in $(\lambda_1/\beta, +\infty) \setminus \{1\}$ such that $\alpha_k \to 1$ and $u_{\alpha_k} \to u_1$ in $H_0^1(\Omega)$, where u_1 is a global minimizer of $\hat{E}$ in $H_0^1(\Omega)$.*

Proof. We have

$$F(s) + \beta^2/4 = s^4/4 - \beta s^2/2 + \beta^2/4 = (s^2 - \beta)^2/4 \ge 0,$$

so F is bounded from below. For every $\alpha > 0$ we define

$$\hat{E}_\alpha(u) = \int_\Omega |\nabla u|^2 + \alpha(F(u) + \beta^2/4)dx.$$

By considering a minimizing sequence for $\hat{E}_\alpha$ in $H_0^1(\Omega)$, we obtain a global minimizer u_α for $\hat{E}_\alpha$. As a consequence, u_α solves (3.28). If $\alpha > \lambda_1/\beta$, then by the same argument as previously, we see that 0 is unstable, so that $u_\alpha \not\equiv 0$ on Ω. Thus, u_α cannot solve both (3.28) and (3.25) unless $\alpha = 1$.

Now let α_k be a sequence in $(\lambda_1/\beta, +\infty)\setminus\{1\}$ such that $\alpha_k \to 1$. Since $1 \le d \le 3$, we have the Sobolev imbedding $H_0^1(\Omega) \subset L^6(\Omega)$, so that the sequence $(f(u_{\alpha_k}))_k$ is bounded in $L^2(\Omega)$. By elliptic regularity, $(u_{\alpha_k})_k$ is bounded in $H^2(\Omega) \cap H_0^1(\Omega)$, which is compactly imbedded in $H_0^1(\Omega)$ [11]. Thus, up to a subsequence, which we still denote $(\alpha_k)_k$, we have $u_{\alpha_k} \to \bar{u}$ in $H_0^1(\Omega)$

It remains to show that the limit $\bar{u}$ is a minimizer of $\hat{E}_1$. It is easily seen, by continuity, that $\hat{E}_{\alpha_k}(u_{\alpha_k}) \to \hat{E}_1(\bar{u})$. Moreover, if u_1 is a global minimizer for $\hat{E}_1$, we have $\hat{E}_{\alpha_k}(u_1) \to \hat{E}_1(u_1)$. Since $\hat{E}_{\alpha_k}(u_1) \ge \hat{E}_{\alpha_k}(u_{\alpha_k})$, by letting k tend to $+\infty$, we obtain $\hat{E}_1(u_1) \ge \hat{E}_1(\bar{u})$. Thus, $\bar{u}$ is also a minimizer of $\hat{E}_1$ in $H_0^1(\Omega)$ (and we denote $\bar{u} = u_1$). □

Remark 3.2. By setting $r_{\alpha_k} = \alpha_k s_{\alpha_k}$ where $s_{\alpha_k} = \sqrt{\int_\Omega F(u_{\alpha_k})dx + C_0}$, we have a a sequence of initial values $(u_{\alpha_k}, 0, r_{\alpha_k})$ which are steady states for the SAV scheme (3.9)-(3.11) and such that $u_{\alpha_k} \to u_1$ and $r_{\alpha_k}/s_{\alpha_k} \to 1$ but $r_{\alpha_k}/s_{\alpha_k} \ne 1$ for all k. In other words, the SAV scheme starts arbitrarily close to u_1 and does not end up on a steady state of the PDE (but only close to u_1, i.e. at u_{α_k}).

4. Asymptotic behavior of a fully discrete SAV method for the Cahn-Hilliard equation

We have seen in the previous section that the steady state equation for the SAV scheme is a modified version of the steady state equation associated to the PDE. This can be easily quantified by the ratio $r^\star/s^\star$ in (3.18), which should be equal to 1 in an ideal situation.

In this section, we want to check this numerically. In order to underline that the situation is not restricted to the damped wave equation with Dirichlet boundary conditions, we consider the Cahn-Hilliard equation with no-flux boundary conditions (a model problem for gradient flows). For the time discretization, we use a first order SAV scheme. For the space discretization, we use a finite element method which naturally inherits the properties of the continuous problem. We first prove that an asymptotic result similar to Theorem 3.1 holds for the fully discrete scheme.

4.1. First order SAV/Finite Element method

We consider the Cahn-Hilliard equation on a bounded polyhedral domain Ω of $\mathbb{R}^d$ $(1 \leq d \leq 3)$

$$u_t = \Delta w \qquad x \in \Omega,\ t \geq 0, \tag{4.1}$$

$$w = -\alpha \Delta u + f(u) \qquad x \in \Omega,\ t \geq 0, \tag{4.2}$$

endowed with Neumann boundary conditions. Here $\alpha > 0$ and the nonlinearity f is a polynomial of odd degree with positive leading coefficient; if $d = 3$, we assume moreover that the degree of f is 1 or 3. Thus, f and F satisfy assumptions (3.2)-(3.3).

The Cahn-Hilliard equation is a gradient flow for the H^{-1} scalar product. The first order SAV scheme for (4.1)-(4.2) (known as SAV/BDF1) reads (cf. (2.2))

$$\frac{u^{n+1} - u^n}{\delta t} = \Delta w^{n+1}, \tag{4.3}$$

$$w^{n+1} = -\alpha \Delta u^{n+1} + \frac{r^{n+1}}{\sqrt{\int_\Omega F(u^n)dx + C_0}} f(u^n), \tag{4.4}$$

$$r^{n+1} - r^n = \frac{1}{2\sqrt{\int_\Omega F(u^n)dx + C_0}} \int_\Omega f(u^n)(u^{n+1} - u^n)dx. \tag{4.5}$$

For the space discretization, we use piecewise linear continuous (P^1) finite elements based on a conforming triangulation of Ω [9]. The finite element space V_h is a subspace of $H^1(\Omega)$ which has finite dimension and which contains the constants. We denote $(\varphi_1, \ldots, \varphi_{N_h})$ the standard basis and we seek $u_h^n = \sum_{i=1}^{N_h} u_i^n \varphi_i$ and $w_h^n = \sum_{i=1}^{N_h} w_i^n \varphi_i$.

The space discrete variational form of (4.3)-(4.5) reads

$$(\frac{u_h^{n+1} - u_h^n}{\delta t}, \varphi_h) = -(\nabla w_h^{n+1}, \nabla \varphi_h), \tag{4.6}$$

$$(w_h^{n+1}, \chi_h) = \alpha(\nabla u_h^{n+1}, \nabla \chi_h) + \frac{r^{n+1}}{s^n}(f(u_h^n), \chi_h), \tag{4.7}$$

$$r^{n+1} - r^n = \frac{1}{2s^n}(f(u_h^n), (u_h^{n+1} - u_h^n)), \tag{4.8}$$

for all φ_h, $\chi_h \in V_h$, where $s^n = \sqrt{\int_\Omega F(u_h^n)dx + C_0}$ and $(\cdot, \cdot)$ stands for the scalar product in $L^2(\Omega)$.

4.2. Energy estimate and asymptotic behavior

It is well-known that the time semi-discrete SAV scheme (4.3)-(4.5) satisfies a stability estimate. We prove that this still holds for the fully discrete version (4.6)-(4.8).

We choose $\varphi_h = \delta t w_h$ in (4.6), $\chi_h = -(u_h^{n+1} - u_h^n)$ in (4.7), we multiply (4.8) by $2r^{n+1}$ and we add the resulting equations. The nonlinear terms cancel and we obtain

$$\alpha(\nabla u_h^{n+1}, \nabla u_h^{n+1} - \nabla u_h^n) + 2(r^{n+1} - r^n)r^{n+1} + \delta t|\nabla w_h^{n+1}|_0^2 = 0.$$

Next, we use the identity $2a(a-b) = a^2 - b^2 + (a-b)^2$. This yields

$$E(u_h^{n+1}, r^{n+1}) + \frac{\alpha}{2}|\nabla(u_h^{n+1} - u_h^n)|_0^2 + (r^{n+1} - r^n)^2 + \delta t|\nabla w_h^{n+1}|_0^2 = E(u_h^n, r^n), \tag{4.9}$$

where

$$E(u_h, r) = \frac{\alpha}{2}|\nabla u_h|_0^2 + r^2.$$

This is the *energy estimate*, which is valid for all $\delta t > 0$ (unconditional stability). Now, we choose $\varphi_h \equiv 1$ in (4.6). We obtain that $(u_h^{n+1}, 1) = (u_h^n, 1)$ for all n, so that *the mass* $(u_h^n, 1)$ *is constant*, that is

$$(u_h^n, 1) = (u_h^0, 1), \quad \forall n \geq 0. \tag{4.10}$$

A *steady state* for the discrete dynamical system (4.6)-(4.8) is a constant sequence, namely a triple $(u_h^\star, w_h^\star, r^\star) \in V_h \times V_h \times \mathbb{R}$ such that $w_h^\star \equiv C^\star$ is constant on Ω and $(u_h^\star, r^\star) \in V_h \times \mathbb{R}$ satisfies

$$\alpha(\nabla u_h^\star, \nabla \chi_h) + \frac{r^\star}{s^\star}(f(u_h^\star), \chi_h) = (C^\star, \chi_h), \quad \forall \chi_h \in V_h, \tag{4.11}$$

where $s^\star = \sqrt{\int_\Omega F(u_h^\star)dx + C_0}$.

Let $(u_h^0, w_h^0, r^0) \in V_h \times V_h \times \mathbb{R}$. Then the fully discrete scheme (4.6)-(4.8) generates a unique sequence $(u_h^n, w_h^n, r^n)_n$ in $V_h \times V_h \times \mathbb{R}$ (see Section 4.3). We have the following asymptotic result.

Theorem 4.1. *Any sequence* $(u_h^n, w_h^n, r^n)_n$ *generated by the SAV/BDF1 scheme* (4.6)-(4.8) *converges up to a subsequence in* $V_h \times V_h \times \mathbb{R}$ *to a steady state* $(u_h^\star, C^\star, r^\star)$ *which solves* (4.11) *and where* $C^\star$ *is constant in* Ω.

Proof. The energy estimate (4.9) shows that $(E(u_h^n, r^n))_n$ is nonincreasing, so $(|\nabla u_h^n|_0)_n$ and $(r^n)_n$ are bounded. By preservation of the mass (cf. (4.10)), $(u_h^n, 1)$ is constant. The Poincaré-Wirtinger inequality shows that the Hilbertian norm on $H^1(\Omega)$ defined by

$$v \mapsto (|\nabla v|_0^2 + (v, 1)^2)^{1/2}$$

is equivalent to the standard $H^1(\Omega)$ norm. Thus, $(u_h^n)_n$ is bounded in $H^1(\Omega)$ and therefore in V_h. Since V_h has finite dimension, the Bolzano-Weierstrass theorem implies that for a subsequence, $(u_h^{n_k}, r^{n_k}) \to (u_h^\star, r^\star)$ in $V_h \times \mathbb{R}$.

Next, we sum the energy estimate (4.9) from $n = 0$ to $n = +\infty$. We obtain

$$\sum_{n=0}^{+\infty} \frac{\alpha}{2}|\nabla(u_h^{n+1} - u_h^n)|_0^2 + (r^{n+1} - r^n)^2 + \delta t|\nabla w_h^{n+1}|_0^2 \leq E(u_h^0, r^0) < +\infty.$$

In particular, $|\nabla w_h^{n+1}|_0 \to 0$ and $u_h^{n+1} - u_h^n \to 0$ in V_h. On choosing $\chi_h \equiv 1$ in (4.7) and using the continuity of $u \mapsto f(u)$ in $H^1(\Omega)$ (cf. (3.4)), we see that

$$(w_h^{n_k}, 1) = \frac{r_k^n}{s^{n_k-1}}(f(u_h^{n_k-1}), 1) \to \frac{r^\star}{s^\star}(f(u_h^\star), 1),$$

where $s^\star = \sqrt{\int_\Omega F(u_h^\star)dx + C_0}$. This shows that $(w_h^{n_k})$ converges to some constant $C^\star \in \mathbb{R}$. By choosing $n+1 = n_k$ in (4.7) and letting k tend to $+\infty$, for an arbitrary $\chi_h \in V_h$, we obtain the steady state equation (4.11). The proof is complete. □

Remark 4.1. As in Theorem 3.1, the ω-limit set of $(u_h^n, r^n)_n$ is a compact and connected subset of $V_h \times \mathbb{R}$ on which E is constant and equal to the limit $E^\star = \lim_{n\to+\infty} E(u_h^n, r^n)$.

Remark 4.2. If we replace Δw by $-w$ in (4.1), then (4.1)-(4.2) becomes the Allen-Cahn equation. A similar asymptotic convergence result can be obtained in that case. However, since the mass is not conserved for the Allen-Cahn equation, some coercivity must be added in the linear term in order to deal with the Neumann boundary conditions [27]. This can be done for instance by writing equation (4.2) as

$$w = -\alpha\Delta u + \alpha u + \tilde{f}(u),$$

with $\tilde{f}(u) = f(u) - \alpha u$.

4.3. The linear solver

It is well-known that the SAV scheme for the Cahn-Hilliard equation requires only the resolution of two linear systems with constant coefficients [28]. We show this for the fully discrete finite element version (4.6)-(4.8).

For this purpose, we introduce the matrix version of the scheme, which reads

$$\frac{BU^{n+1} - BU^n}{\delta t} = -AW^{n+1}, \tag{4.12}$$

$$BW^{n+1} = \alpha AU^{n+1} + \frac{r^{n+1}}{s^n}F^n, \tag{4.13}$$

$$r^{n+1} - r^n = \frac{1}{2s^n}\langle F^n, U^{n+1} - U^n\rangle. \tag{4.14}$$

Here, $u_h^n = \sum_{i=1}^{N_h} u_i^n \varphi_i$ is identified with the column vector $(u_1^n, \ldots, u_{N_h}^n)^t$ in $\mathbb{R}^{N_h}$, and $w_h^n = \sum_{i=1}^{N_h} w_i^n \varphi_i$ with $(w_1^n, \ldots, w_{N_h}^n)^t$ as well, where $(\varphi_1, \ldots, \varphi_{N_h})$ is the usual finite element basis of V_h. We have set

$$s^n = \sqrt{\int_\Omega F(u_h^n)dx + C_0},$$

the matrix $A = (A_{ij})_{1\le i,j\le N_h}$ with $A_{ij} = (\nabla\varphi_i, \nabla\varphi_j)$ is the discrete Laplacian and $B = (B_{ij})_{1\le i,j\le N_h}$ with $B_{ij} = (\varphi_i, \varphi_j)$ is the mass matrix. The nonlinear term is the column vector $F^n = (F^n_1, \dots, F^n_{N_h})^t$ with

$$F^n_i = \int_\Omega f(u^n_h(x))\varphi_i(x)dx. \tag{4.15}$$

In (4.14) we denote $\langle\cdot,\cdot\rangle$ the usual scalar product in $\mathbb{R}^{N_h}$.

We first eliminate W^{n+1}. We get

$$\frac{BU^{n+1} - BU^n}{\delta t} = -\alpha AB^{-1}AU^{n+1} - \frac{r^{n+1}}{s^n}AB^{-1}F^n, \tag{4.16}$$

$$r^{n+1} - r^n = \frac{1}{2s^n}\langle F^n, U^{n+1} - U^n\rangle. \tag{4.17}$$

Next, by eliminating r^{n+1} in (4.16) thanks to (4.17), we get

$$A_{sav}U^{n+1} + \frac{\delta t}{2}\langle Q^n, U^{n+1}\rangle AB^{-1}Q^n = G^n, \tag{4.18}$$

where $A_{sav} = B + \alpha\delta tAB^{-1}A$,

$$Q^n = \frac{F^n}{s^n} \quad \text{and} \quad G^n = BU^n - \delta tr^nAB^{-1}Q^n + \frac{\delta t}{2}\langle Q^n, U^n\rangle AB^{-1}Q^n. \tag{4.19}$$

The idea consists in computing $\langle U^{n+1}, Q^n\rangle$. Thus, we apply the operator A^{-1}_{sav} to equation (4.18) and we take the scalar product with Q^n. This yields

$$\left(1 + \frac{\delta t}{2}\langle A^{-1}_{sav}AB^{-1}Q^n, Q^n\rangle\right)\langle U^{n+1}, Q^n\rangle = \langle A^{-1}_{sav}G^n, Q^n\rangle. \tag{4.20}$$

In order to compute U^{n+1}, we proceed as follows:

(i) We first compute $A^{-1}_{sav}G^n$ and $A^{-1}_{sav}AB^{-1}Q^n$. These are *two linear systems.*

(ii) We obtain $\langle U^{n+1}, Q^n\rangle$ from (4.20) by computing two scalar products and a division (cf. Lemma 4.1).

(iii) From (4.18), we can compute U^{n+1} through

$$U^{n+1} = A^{-1}_{sav}G^n - \frac{\delta t}{2}\langle Q^n, U^{n+1}\rangle A^{-1}_{sav}AB^{-1}Q^n.$$

At this stage, all the required terms are known.

Lemma 4.1. *The matrix $A^{-1}_{sav}AB^{-1}$ is symmetric and positive semi-definite. In particular,*

$$1 + \frac{\delta t}{2}\langle A^{-1}_{sav}AB^{-1}Q^n, Q^n\rangle \ge 1 > 0.$$

Proof. The matrix A is symmetric and positive semi-definite, but it is not invertible, so we set $A_\varepsilon = A + \varepsilon I$ which is positive definite for $\varepsilon > 0$. The symmetric matrix B is also positive definite. Thus, by letting $A_{sav,\varepsilon} = B + \alpha\delta tA_\varepsilon B^{-1}A_\varepsilon$, we have

$$(A^{-1}_{sav,\varepsilon}A_\varepsilon B^{-1})^{-1} = BA^{-1}_\varepsilon A_{sav,\varepsilon} = BA^{-1}_\varepsilon B + \alpha\delta tA_\varepsilon.$$

This is a symmetric and positive definite matrix, so its inverse $A^{-1}_{sav,\varepsilon}A_\varepsilon B^{-1}$ is also symmetric and positive definite. By letting $\varepsilon \to 0^+$, the claim follows. □

Remark 4.3. The steady state equation for the dynamical system (4.16)-(4.17) reads

$$\alpha AB^{-1}AU^\star + \frac{r^\star}{s^\star}AB^{-1}F^\star = 0, \tag{4.21}$$

for some $r^\star \in \mathbb{R}$, where $s^\star = \sqrt{\int_\Omega F(u_h^\star)dx + C_0}$, $F_i^\star = \int_\Omega f(u_h^\star)\varphi_i dx$ $(1 \le i \le N_h)$ and $u_h^\star = \sum_{i=1}^{N_h} u_i^\star \varphi_i$. This is the matrix version of (4.11), since the matrix A has a kernel of dimension 1 corresponding to the constant functions in V_h.

4.4. Numerical simulations

We perform numerical simulations in one space dimension on $\Omega = (0,1)$ with `Matlab`®. The nonlinearity is $f(s) = s^3 - s$ and $\alpha = 0.01$. We consider an initial datum $u_0(x) = 0.9\cos(\pi x)$.

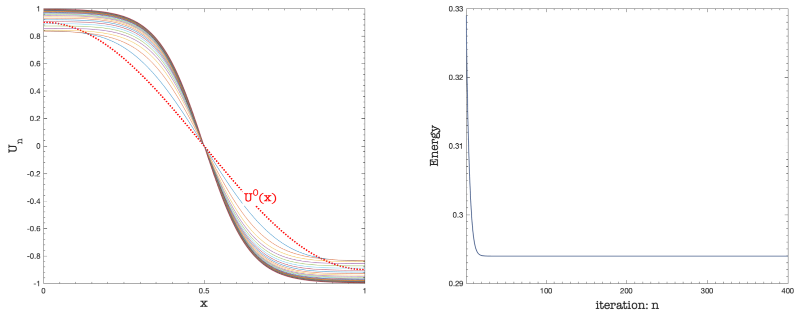

Figure 1. Convergence of $(u_h^n)_n$ (left) and of the energy $(E^n)_n$ right).

For the space discretization with P^1 finite elements, a uniform subdivision with a space step equal to $h = 1/(M-1)$ with $M = 400$ is applied. The nonlinear term F_i^n defined by (4.15) is computed with Gauss formula of order 5, so that the calculation is exactly up to the double precision. We note that the matrices A and B are tridiagonal, but B^{-1} is a full matrix. And so, the SAV operator $A_{sav} = B - \alpha\delta t AB^{-1}A$ is also a full matrix. It would be more efficient to use a lumped mass (diagonal) matrix as an approximation of B, but this would lead to an additional consistency error that we prefer to avoid.

Table 1. Values of $r^\star/s^\star - 1$ for the SAV/Cahn-Hilliard scheme.

C_0	$\delta t = \frac{1}{100}$	$\delta t = \frac{1}{200}$	$\delta t = \frac{1}{400}$	$\delta t = \frac{1}{1000}$
0.1	-0.00053	-0.000162	-0.000032	+0.000004
0.2	-0.00065	-0.000267	-0.000106	-0.000033
0.4	-0.00051	-0.000227	-0.000099	-0.000034
0.8	-0.00033	-0.000149	-0.000067	-0.000024
100	-0.00000321	-0.00000157	-0.00000072	-0.00000026
1000	-0.00000033	-0.00000015	-0.00000007	-0.00000002

Figure 1 illustrates that the generated sequence $(u_h^n)_n$ converges to a state of equilibrium and that the modified energy $E^n = E(u_h^n, r^n)$ associated to (4.6)-(4.8) is nonincreasing in time (cf. (4.9)).

In order to show numerically that we have no guaranty that $r^\star/s^\star = 1$ in (4.21), we study the relative difference $(r^\star - s^\star)/s^\star = r^\star/s^\star - 1$. The results are shown in Table 1 for various choices of the time step δt and of the constant C_0.

The iterations are stopped at the final time $T = 4$, at which time the steady state is considered to be numerically reached. This is confirmed in Figure 2 where the difference $|U^{n+1} - U^n|_\infty = \|u_h^{n+1} - u_h^n\|_\infty$ is plotted as a function of the iteration n. We see that the difference $|U^{n+1} - U^n|_\infty$ rapidly decreases from 10^{-1} to 10^{-10}, then stabilizes around 10^{-10}. The time step used for this particular simulation is $\delta t = \frac{1}{100}$ and the SAV constant is $C_0 = 0.2$. The corresponding value of $\frac{r^\star}{s^\star} - 1$ equals 0.00065 as shown in Table 1.

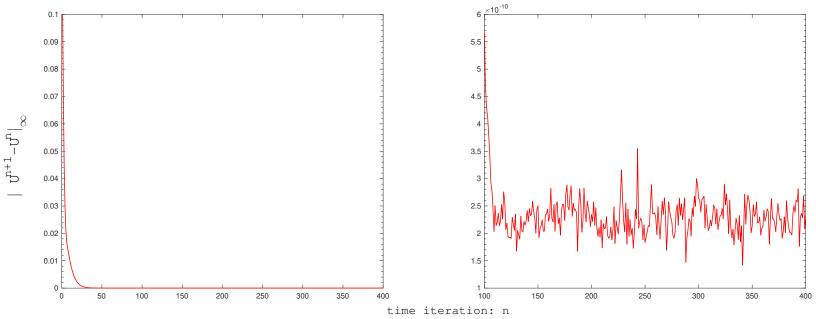

Figure 2. Values of $|U^{n+1} - U^n|_\infty$.

Table 1 confirms that the ratio $r^\star/s^\star$ is never exactly equal to 1. In all cases except one (in red), we have found $r^\star < s^\star$. However, $r^\star/s^\star$ is very close to 1, all the more since our initial value is close to a steady state. For a given constant C_0, we observe that the relative error for $r^\star$ has an order close to $O(\delta t)$, especially for large values of C_0. This is consistent with the first order approximation of the SAV scheme. Note that the space discretization does not appear here, because we work in a fixed space V_h.

For a fixed time step δt, the absolute value of the ratio seems to grow from $C_0 = 0.1$ to $C_0 = 0.2$, and then it monotonically decreases as C_0 increases. This does not mean C_0 should be chosen very large, because the numerical errors due to a very large constant C_0 could significantly change the numerical solution. We should rather seek an optimal value of C_0, possibly small. An approach has been proposed in [25].

References

[1] N. E. Alaa and M. Pierre, *Convergence to equilibrium for discretized gradient-like systems with analytic features*, IMA J. Numer. Anal., 2013, 33(4), 1291–1321.

[2] S. Allen and J. Cahn, *A microscopic theory for antiphase boundary motion and*

its application to antiphase domain coarsing, Acta. Metall., 1979, 27, 1084–1095.

[3] P. F. Antonietti, B. Merlet, M. Pierre and M. Verani, *Convergence to equilibrium for a second-order time semi-discretization of the Cahn-Hilliard equation*, AIMS Mathematics, 2016, 1(3), 178–194.

[4] S. Badia, F. Guillén-González and J. V. Gutiérrez-Santacreu, *Finite element approximation of nematic liquid crystal flows using a saddle-point structure*, J. Comput. Physics, 2011, 230, 1686–1706.

[5] A. Bouchriti, M. Pierre and N. E. Alaa, *Gradient stability of high-order BDF methods and some applications*, J. Difference Equ. Appl., 2020, 0(0), 1–30.

[6] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system. I. Interfacial free energy*, J. Chem. Phys., 1958, 28, 258–267.

[7] Q. Cheng, J. Shen and X. Yang, *Highly efficient and accurate numerical schemes for the epitaxial thin film growth models by using the SAV approach*, J. Sci. Comput., 2019, 78(3), 1467–1487.

[8] K. R. Elder, M. Katakowski, M. Haataja and M. Grant, *Modeling elasticity in crystal growth*, Phys. Rev. Lett., 2002, 88, 245701.

[9] A. Ern and J. L. Guermond, *Theory and practice of finite elements*, 159 of *Applied Mathematical Sciences*, Springer-Verlag, New York, 2004.

[10] S. Gatti, M. Grasselli, A. Miranville and V. Pata, *A construction of a robust family of exponential attractors*, Proc. Amer. Math. Soc., 2006, 134(1), 117–127.

[11] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, Classics in Mathematics, Springer-Verlag, Berlin, 2001.

[12] H. Gomez and T. J. R. Hughes, *Provably unconditionally stable, second order time-accurate, mixed variational methods for phase-field models*, J. Comput. Phys., 2011, 230(13), 5310–5327.

[13] O. Goubet, *Remarks on some dissipative sine-Gordon equations*, Complex Var. Elliptic Equ., 2019, 0(0), 1–7.

[14] M. Grasselli and M. Pierre, *Energy stable and convergent finite element schemes for the modified phase field crystal equation*, ESAIM Math. Model. Numer. Anal., 2016, 50(5), 1523–1560.

[15] F. Guillén-González and G. Tierra, *On linear schemes for a Cahn-Hilliard diffuse interface model*, J. Comput. Physics, 2013, 234, 140–171.

[16] A. Haraux, *Systèmes dynamiques dissipatifs et applications*, 17 of *Recherches en Mathématiques Appliquées*, Masson, Paris, 1991.

[17] A. Haraux and M. A. Jendoubi, *The convergence problem for dissipative autonomous systems*, SpringerBriefs in Mathematics, Springer, Cham; BCAM Basque Center for Applied Mathematics, Bilbao, 2015.

[18] C. Jiang, W. Cai and Y. Wang, *A linearly implicit and local energy-preserving scheme for the sine-Gordon equation based on the invariant energy quadratization approach*, J. Sci. Comput., 2019, 80(3), 1629–1655.

[19] M. D. Johnson, C. Orme, A. W. Hunt et al., *Stable and unstable growth in molecular beam epitaxy*, Phys. Rev. Lett., 1994, 72, 116–119.

[20] O. Kavian, *Introduction à la théorie des points critiques et applications aux problèmes elliptiques*, 13 of *Mathématiques & Applications (Berlin)*, Springer-Verlag, Paris, 1993.

[21] H. Khalfi, M. Pierre, N. E. Alaa and M. Guedda, *Convergence to equilibrium of a DC algorithm for an epitaxial growth model*, Int. J. Numer. Anal. Model., 2019, 16(3), 398–411.

[22] H. Kielhöfer, *Bifurcation theory*, 156 of *Applied Mathematical Sciences*, 2nd Edn, Springer, New York, 2012.

[23] Q. Li, L. Mei, X. Yang and Y. Li, *Efficient numerical schemes with unconditional energy stabilities for the modified phase field crystal equation*, Adv. Comput. Math., 2019, 45(3), 1551–1580.

[24] X. Li, J. Shen and H. Rui, *Energy stability and convergence of SAV block-centered finite difference method for gradient flows*, Math. Comp., 2019, 88(319), 2047–2068.

[25] Z. Liu and X. Li, *Efficient modified techniques of invariant energy quadratization approach for gradient flows*, Appl. Math. Lett., 2019, 98, 206–214.

[26] M. Pierre and P. Rogeon, *Convergence to equilibrium for a time semi-discrete damped wave equation*, J. Appl. Anal. Comput., 2016, 6(4), 1041–1048.

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

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

[29] J. Shen and X. Yang, *Numerical approximations of Allen-Cahn and Cahn-Hilliard equations*, Discrete Contin. Dyn. Syst., 2010, 28(4), 1669–1691.

[30] S. Sun, X. Jing and Q. Wang, *Error estimates of energy stable numerical schemes for Allen-Cahn equations with nonlocal constraints*, J. Sci. Comput., 2018, 79, 593–623.

[31] R. Temam, *Infinite-dimensional dynamical systems in mechanics and physics*, 68 of *Applied Mathematical Sciences*, 2nd Edn, Springer-Verlag, New York, 1997.

[32] G. Tierra and F. Guillén-González, *Numerical methods for solving the Cahn-Hilliard equation and its applicability to related energy-based models*, Arch. Comput. Methods Eng., 2015, 22(2), 269–289.

[33] X. Yang, *Linear, first and second-order, unconditionally energy stable numerical schemes for the phase field model of homopolymer blends*, J. Comput. Physics, 2016, 327, 294–316.

[34] S. Zelik, *Asymptotic regularity of solutions of a nonautonomous damped wave equation with a critical growth exponent*, Commun. Pure Appl. Anal., 2004, 3(4), 921–934.