MATHEMATICS OF COMPUTATION
Volume 91, Number 334, March 2022, Pages 785–809
https://doi.org/10.1090/mcom/3704
Article electronically published on November 9, 2021

STABILITY AND CONVERGENCE ANALYSIS FOR THE IMPLICIT-EXPLICIT METHOD TO THE CAHN-HILLIARD EQUATION

DONG LI, CHAOYU QUAN, AND TAO TANG

Abstract. Implicit-explicit methods have been successfully used for the efficient numerical simulation of phase field problems such as the Cahn-Hilliard equation or thin film type equations. Due to the lack of maximum principle and stiffness caused by the effect of small dissipation coefficient, most existing theoretical analysis relies on adding additional stabilization terms, mollifying the nonlinearity or introducing auxiliary variables which implicitly either changes the structure of the problem or trades accuracy for stability in a subtle way. In this work, we introduce a robust theoretical framework to analyze directly the stability and accuracy of the standard implicit-explicit approach without stabilization or any other modification. We take the Cahn-Hilliard equation as a model case and provide a rigorous stability *and* convergence analysis for the original semi-discrete scheme under certain time step constraints. These settle several questions which have been open since the work of Chen and Shen [Comput. Phys. Comm. 108 (1998), pp. 147–158].

1. Introduction

The Cahn-Hilliard (CH) equation was first introduced by Cahn and Hilliard in [2] to describe the complicated phase separation and coarsening phenomena in non-uniform systems such as glasses, alloys and polymer mixtures. In this work we are concerned with the numerical solutions for the Cahn-Hilliard equation in nondimensionalized form as

$$\begin{cases} \partial_t u = \Delta(-\nu\Delta u + f(u)), & (t,x) \in (0,\infty)\times\Omega, \\ u\Big|_{t=0} = u_0, \end{cases} \tag{1.1}$$

where $u = u(t,x)$ is a real-valued function which represents the concentration difference in a binary system, and $\nu > 0$ is usually called mobility coefficient. The function $f(u)$ is taken as the derivative of a standard double well potential:

$$f(u) = u^3 - u = F'(u), \quad F(u) = \frac{1}{4}(u^2-1)^2.$$

Note that with this choice equation (1.1) is invariant under the sign change $u \to -u$ which is natural since u corresponds to the difference of concentrations. The minima

Received by the editor June 16, 2020, and, in revised form, June 20, 2021, and July 6, 2021.
2020 *Mathematics Subject Classification.* Primary 35K55, 65M12.
Key words and phrases. Cahn-Hilliard equation, implicit-explicit method.
The second author's work was supported by the NSFC Grant 11901281, the Guangdong Basic and Applied Basic Research Foundation (2020A1515010336), and the Stable Support Plan Program of Shenzhen Natural Science Fund (Program Contract No. 20200925160747003). The third author's work was partially supported by the NSFC grants 11731006 and K20911001, and the science challenge project (No. TZ2018001).

©2021 American Mathematical Society

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

of the double well potential are $u = \pm 1$ which typically correspond to the formation of domains. The length scale of the transition regions between domains is typically proportional to $\sqrt{\nu}$. For simplicity we shall fix the spatial domain Ω in (1.1) as the usual 1-periodic torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d = [-\frac{1}{2}, \frac{1}{2})^d$ in physical dimensions $d \leq 3$. With this it is convenient to work with the Fourier basis $(e^{2\pi i k\cdot x})_{k\in\mathbb{Z}^d}$ and this is the convention we shall adopt for the Fourier inversion formulae to be used throughout this paper. For smooth solutions, the mass conservation law takes the form

$$\frac{d}{dt}M(t) = \frac{d}{dt}\int_\Omega u(t,x)dx \equiv 0. \tag{1.2}$$

In particular $M(t) \equiv 0$ if the initial total mass is zero. If $M(t) \equiv c$ for some c nonzero, then by a change of variable $u \to u + c$ one can work with a modified nonlinearity $f(u + c)$ in (1.1) and the corresponding analysis can be adjusted suitably. Therefore, throughout this work we will only consider mean zero initial data for simplicity. As is well known the system (1.1) is a gradient flow of a Ginzburg-Landau type energy functional $\mathcal{E}(u)$ in H^{-1}, i.e.

$$\partial_t u = -\frac{\delta\mathcal{E}}{\delta u}\Big|_{H^{-1}} = \Delta(\frac{\delta\mathcal{E}}{\delta u}), \tag{1.3}$$

where $\frac{\delta\mathcal{E}}{\delta u}\Big|_{H^{-1}}$, $\frac{\delta\mathcal{E}}{\delta u}$ denote the standard variational derivatives in H^{-1} and L^2 respectively, and

$$\mathcal{E}(u) = \int_\Omega \left(\frac{1}{2}\nu|\nabla u|^2 + F(u)\right)dx = \int_\Omega \left(\frac{1}{2}\nu|\nabla u|^2 + \frac{1}{4}(u^2-1)^2\right)dx. \tag{1.4}$$

As such the basic energy conservation law takes the form

$$\frac{d}{dt}\mathcal{E}(u(t)) + \||\nabla|^{-1}\partial_t u\|_2^2 = \frac{d}{dt}\mathcal{E}(u(t)) + \int_\Omega |\nabla(-\nu\Delta u + f(u))|^2 dx = 0. \tag{1.5}$$

Here one should note that $\partial_t u$ has mean zero so that $|\nabla|^{-1}$ is defined as a Fourier multiplier $(2\pi|k|)^{-1}\cdot 1_{k\neq 0}$ (see (2.2) for the convention of Fourier transform). From energy conservation one immediately deduces monotonic energy decay and a priori $\dot{H}^1$ bound of the solution as

$$\mathcal{E}(u(t)) \leq \mathcal{E}(u(s)), \qquad \forall\, t \geq s; \tag{1.6a}$$

$$\|\nabla u(t)\|_2 \leq \sqrt{\frac{2}{\nu}}\mathcal{E}(u(t)) \leq \sqrt{\frac{2}{\nu}}\mathcal{E}(u_0), \qquad \forall\, t > 0. \tag{1.6b}$$

By a scaling analysis one can identify the critical spaces for CH in 2D and 3D as L^2 and $\dot{H}^{\frac{1}{2}}$ respectively. Global wellposedness in H^1 and regularity of solutions then follow easily from the a priori $\dot{H}^1$ bound and standard arguments. As the primary goal of the CH model is to understand the physics and dynamics of spinodal decomposition (especially concerning the stages after quenching), a large body of existing mathematical analysis is naturally devoted to the investigation of asymptotic behavior of solutions concerning coarsening, pattern formation, evolution of interfaces, stability and instability in various time regimes. One should keep in mind that the dynamics of CH is quite complex and takes place on a myriad of time scales ranging from short time scales $t = O(\sqrt{\nu})$ on which the solution generally develops rich structures with many internal layers and sharp gradients to metastable time scales $t = O(e^{\text{const}/\sqrt{\nu}})$ on which the unstable or super slow modes are fully developed and coarsening starts to dominate. As a result the numerical simulation

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

of CH can be quite stiff as one need to resolve layers of order $\sqrt{\nu}$ especially when $\nu \ll 1$.

There exist some natural scaling transformations which lead to slightly different forms of the CH system in the literature. A sample case considered in Liu and Shen [13] can be expressed as

$$\partial_s \tilde{u} = \alpha\Delta(-\Delta\tilde{u} + \beta f(\tilde{u})), \quad (s,y) \in (0,\infty) \times \Omega_L, \tag{1.7}$$

where $\alpha > 0$, $\beta > 0$, $\tilde{u} = \tilde{u}(s,y)$, and $\Omega_L = [-\frac{L}{2}, \frac{L}{2})^d$. Now make a change of variable $\tilde{u}(s,y) = u(s/\lambda, y/L) = u(t,x), \lambda = L^2/(\alpha\beta)$; then we can recast (1.7) into our standard form as

$$\partial_t u = \Delta(-\nu\Delta u + f(u)), \quad (t,x) \in (0,\infty) \times \mathbb{T}^d, \tag{1.8}$$

where $\nu = \frac{1}{\beta L^2}$. If we denote the typical time step for (1.7) as Δs and for (1.8) as Δt, then apparently we have $\Delta t = \Delta s/\lambda = \Delta s \cdot \alpha\beta \cdot L^{-2}$, $\nu = \beta^{-1} \cdot L^{-2}$.

We record these relations here so that one can translate many existing numerical analysis results in the literature about the system (1.7) in terms of the system (1.8) which is quite convenient for comparison purposes.

1.1. **Motivation.** On the numerical side, there is by now an enormous body of literature on the simulation and analysis of the CH equation and related phase field models (see, e.g., [3, 4, 9, 14, 19, 21, 26] and the references therein). A fundamental challenge is to design fast, efficient and accurate numerical schemes which are robust and energy stable especially in the computationally stiff small ν regime. Roughly speaking, almost all existing numerical algorithms can be classified into five categories with some possible overlaps. To elucidate the discussion below, it is useful to recast (1.1) as an abstract and more general model:

$$\partial_t u = \mathcal{L}u + \mathcal{N}(u),$$

where $\mathcal{L}$ denotes the linear operator (for example, $\mathcal{L} = -\Delta^2$ for Cahn-Hilliard and $\mathcal{L} = \Delta$ for Allen-Cahn) and $\mathcal{N}(u)$ collects the nonlinear part. In its most generality one should think of $\mathcal{N}(u)$ as a functional. For example, we write $G(u, \nabla u, \nabla^2 u, \cdots)$ (for some function G) as $\mathcal{N}(u)$. To ease the discussion we shall ignore completely the space discretization and focus momentarily on first order in time discretization. Now for $n = 0, 1, 2, \cdots$, denote by $u^n \approx u(n\tau)$ as the numerical solution at time $t = n\tau$ where τ is the time step size. Then the following are the prototypical semi-discrete schemes often considered in the literature:

$$\frac{u^{n+1} - u^n}{\tau} = \begin{cases} \mathcal{L}u^n + \mathcal{N}(u^n), & \text{(Explicit)}, \\ \mathcal{L}u^{n+1} + \mathcal{N}(u^{n+1}), & \text{(Fully implicit)}, \\ \mathcal{L}u^{n+1} + \mathcal{N}(u^n), & \text{(Semi-implicit/Implicit-Explicit)}, \\ \mathcal{L}u^{n+1} + \mathcal{N}_{\mathrm{I}}(u^n, u^{n+1}), & \text{(Partially implicit)}, \\ \mathcal{L}u^{n+1} + \widetilde{\mathcal{N}}_{\mathrm{I}}(u^n, u^{n+1}) + \mathcal{S}(u^n, u^{n+1}), & \text{(Stabilization)}. \end{cases}$$

In the above $\mathcal{N}_{\mathrm{I}}$ represents a careful splitting/interpolation of the nonlinearity term using u^n and u^{n+1}. To ensure consistency it should satisfy $\mathcal{N}_{\mathrm{I}}(u,u) = \mathcal{N}(u)$. In practical algorithms such as convex splitting one often chooses $\mathcal{N}_{\mathrm{I}}(u^n, u^{n+1}) = \mathcal{N}_+(u) - \mathcal{N}_-(u) = \mathcal{N}(u)$ where $\mathcal{N}_+$ and $\mathcal{N}_-$ are convex. The term $\mathcal{S}(u^n, u^{n+1})$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

represents certain carefully chosen additional stabilization terms which vanishes suitably fast as the time step $\tau \to 0$. In the last one we use the general notation $\widetilde{\mathcal{N}}_{\mathrm{I}}(u^n, u^{n+1})$ to include $\mathcal{N}(u^n)$ or $\mathcal{N}_{\mathrm{I}}(u^n, u^{n+1})$ as special cases.

As mentioned above, the first type in our classification is pure explicit methods such as forward Euler in time and explicit treatment of the linear dissipation and nonlinearity. For small systems and short time scales, these methods are speedy, efficient and relatively easy to implement. But due to the poor stability and low accuracy one often has to employ very small time step and spatial grid size which puts a serious limitation for large scale and long time simulations.

The second type is (fully) implicit schemes such as backward Euler in time and fully implicit of both the linear dissipation and the nonlinearity. The Crank-Nicolson (CN) and modified Crank-Nilcolson type methods also fall into this category. A representative work for the CH equation in this direction dates back to Elliot [5] (see page 55 therein) which analyzed a fully implicit CN-type scheme for the CH system (see also page 1644 of [6] by Elliot and Stuart for the idea of using secant approximation). For the phase field models such as the Cahn-Hilliard equation, it was found that the stability requirement is even worse than the explicit Euler and much worse than the semi-implicit method (the third type below), which is also the motivation of the introduction of the convex splitting method of Eyre [7]. The deficiency of fully implicit methods is also the expense of Newton's method for which efficient preconditioning is often needed in practice.

The third category is the semi-implicit methods which treats the principal linear dissipation term implicitly and the nonlinear term explicitly. In the phase-field context such methods date back to the work of Chen and Shen [3] in which a semi-implicit Fourier spectral method was implemented on an Allen-Cahn system and a CH system. These methods are quite efficient, accurate and observed to have good stability properties in practical numerical simulations. However, due to the lack of maximum principle and stiffness caused by small viscosity coefficient a rigorous stability and error analysis was a longstanding open problem.

The fourth group in our classification contains partially implicit methods. These are one of the most explored directions during the past decades. The most popular ones are the convex-splitting schemes which have been developed in [4, 7, 8, 10, 20] for the CH model, higher order models and related nonlocal versions. The main advantage of a typical convex-splitting scheme is its unconditional energy stability with no stringent restriction on the time step. This is in stark contrast with a standard fully implicit scheme where very small time steps need to be taken in order to ensure energy stability.

The fifth category in our classification consists of stabilized or mollified methods. The basic idea of stabilization is to introduce an additional $O(\Delta t^p)$ (for a pth-order method) term to the numerical scheme to alleviate the time step constraint. These methods were first developed in [26] for a Cahn-Hilliard-Cook equation, [12] for CH and [24] for epitaxial growth models. The work of [12, 24] relies on some conditional L^∞-bound of the numerical solution. In [21], Shen and Yang considered a modified/mollified CH system with suitable Lipschitz truncation and proved various stability results under such assumptions. Removing these conditional assumptions and proving the unconditional energy stability for such stabilized methods were known as the unconditional stability conjecture.

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Recently, Xu et al. [25] discovered an interesting reformulation of several first-order convex-splitting schemes and stabilized schemes as a version of the fully implicit scheme with a proper time re-parametrization/rescaling. As such, it was argued that these first-order methods implicitly trade numerical accuracy for stability. Note that the analysis in [25] is only true for first-order scheme. Second- and higher-order stabilized schemes do not correspond to time rescaled non-stabilized schemes. In fact, for accurate dynamic simulations, one should always use second- and higher-order schemes. In this regard, it is noticed that quite many delicate efforts have been made to develop higher-order convex splitting methods or stabilized methods; see, e.g., [1, 11]. Developing upon the second-order scheme in [16], Song and Shu [23] recently constructed a new unconditionally stable second-order stabilized semi-implicit local discontinuous Galerkin method for the CH equation. In another direction, the recent review article of Shen et al. [22] (and the references therein) fashioned another novel form of stabilization which is based on the introduction of an auxiliary variable. A nice feature of this novel framework is that it can render unconditional energy stability more easily.

Although there have been many important algorithm developments for phase field equations, there are very few rigorous stability and convergence analyses for these methods, even for the well-known implicit-explicit method which enjoys conditional stability with affordable time step constraints, and more importantly without sacrificing accuracy too much [3]. The main purpose of this work is to provide a rigorous stability *and* convergence analysis for the implicit-explicit method.

1.2. **Outline of the main results.** Consider the following semi discretization of (1.1) on $\mathbb{T}^d = [-\frac{1}{2}, \frac{1}{2})^d$ $(d \le 3)$: Given $u^n \in H^1(\mathbb{T}^d)$, find $u^{n+1} \in H^1(\mathbb{T}^d)$ such that

$$\frac{u^{n+1} - u^n}{\tau} = -\nu\Delta^2 u^{n+1} + \Delta(f(u^n)), \quad n \ge 0, \tag{1.9}$$

with u^0 as the initial data, where $\tau > 0$ is the time step. It is also possible to reformulate (1.9) in terms of the usual weak formulation, for example:

$$(\frac{u^{n+1} - u^n}{\tau}, v) + (\nabla(f(u^n)), \nabla v) + \nu(\Delta u^{n+1}, \Delta v) = 0, \quad \forall\, v \in H^1(\mathbb{T}^d), \tag{1.10}$$

where $(\cdot, \cdot)$ denotes the usual L^2 inner product (for real-valued functions) on Ω. However, in our analysis it is slightly more convenient to work with (1.9). Note that u^n has mean zero for all $n \ge 0$ since we assume u_0 has mean zero.

Proposition 1.1 albeit conditional is instrumental to understand the relationship between the time step and the L^∞-norm of the numerical solution, whose proof follows directly from the discrete energy estimate Lemma 2.6.

Proposition 1.1 (Conditional stability for semi-implicit discretization, practical version)**.** *Let $d \le 3$, $\nu > 0$ and $\tau > 0$. Assume $u_0 \in H^1(\mathbb{T}^d)$ and has mean zero. Suppose up to $n = N_1$ the time step $\tau > 0$ satisfies*

$$\sqrt{\frac{2\nu}{\tau}} \ge \frac{3}{2} \max_{0 \le n \le N_1+1} \|u^n\|^2_{L^\infty(\mathbb{T}^d)} - \frac{1}{2}. \tag{1.11}$$

Then the semi-implicit scheme (1.9) *is conditionally energy stable up to $n = N_1$, i.e.*

$$\mathcal{E}(u^{n+1}) \le \mathcal{E}(u^n), \qquad \forall\, 0 \le n \le N_1. \tag{1.12}$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

In practical numerical simulations, it is observed that typical numerical solutions satisfy $\|u^n\|_\infty = O(1)$ even for small $\nu \ll 1$. As such if we assume the boundedness of the numerical solution then by Proposition 1.1, the time step constraint for τ is roughly $\tau \lesssim \nu$. In this sense, Proposition 1.1 is already useful for guiding practical numerical simulations at the price of monitoring closely the $\|u^n\|_\infty$ at each iteration. On the other hand, even for the PDE exact solution there is no ν-independent L^∞ bound for the nonlinearity without any mollification. Therefore, some trade-off must be made in order to obtain a stability result without assuming the boundedness of the numerical solution. Our next theorem is a first result in this direction.

Theorem 1.1 (Conditional stability for semi-implicit discretization)**.** *Let $d = 2$, $\nu > 0$, $\tau > 0$. Assume $u_0 \in H^1(\mathbb{T}^2)$ and has mean zero. Assume $\|u^0\|_{L^\infty(\mathbb{T}^2)} \leq L_0 < \infty$ (we may assume $L_0 \neq 0$). Then the semi-implicit scheme* (1.9) *is conditionally energy stable, i.e.*

$$\mathcal{E}(u^{n+1}) \leq \mathcal{E}(u^n), \qquad \forall\, n \geq 0, \tag{1.13}$$

provided the following time step constraint is satisfied:

$$0 < \tau \leq \tau_{\max} = \min\Big\{\frac{8\nu}{9L_0^4},\, C_1\nu^3\Big\}, \tag{1.14}$$

where $C_1 > 0$ is a constant depending only on the initial energy $E_0 = \mathcal{E}(u^0)$.

Our next result establishes the basic L^2 error estimate for the semi-implicit scheme applied to the Cahn-Hilliard equation.

Theorem 1.2 (L^2 error estimate for CH)**.** *Assume $d = 2$, $\nu > 0$. Let $\tau > 0$ satisfy the condition in Theorem* 1.1*. Let $u_0 \in H^s(\mathbb{T}^2)$, $s \geq 4$ with mean zero. Let $u(t)$ be the solution to* (1.1) *with initial data u_0. Let u^n be defined according to* (1.9) *with initial data $u^0 = u_0$. Define $t_m = m\tau$, $m \geq 1$. Then*

$$\|u(t_m) - u^m\|_2 \leq e^{C_1 t_m} \cdot C_2 \cdot \tau. \tag{1.15}$$

Here $C_1 > 0$, $C_2 > 0$ depend on (u_0, ν, s).

1.3. **Main idea of the proof.** Our proof of Theorem 1.1 proceeds by a simple yet powerful Trade-Energy-For-L^∞ (TEFL) scheme which is a refinement of our earlier works [15–17]. In several cases, we even manage to calculate explicit constants and identified nearly optimal parametric dependences which seem to be the first done in the literature. These will be instrumental for future refined analysis on these algorithms. It is expected that this new streamlined proof can be adapted to higher order cases and generalized to many other models and settings.

For the first-order IMEX scheme (1.9), our TEFL recipe consists of three steps.

Step 1 (Discrete energy estimate). We show that

$$\begin{aligned} &\mathcal{E}(u^{n+1}) - \mathcal{E}(u^n) + \left(\frac{1}{2} + \sqrt{\frac{2\nu}{\tau}}\right)\|u^{n+1} - u^n\|_2^2 \\ \leq\ & \frac{3}{2}\max\{\|u^n\|_\infty^2,\ \|u^{n+1}\|_\infty^2\} \cdot \|u^{n+1} - u^n\|_2^2. \end{aligned} \tag{1.16}$$

Thus to show energy monotonicity it suffices to show

$$\frac{1}{2} + \sqrt{\frac{2\nu}{\tau}} \geq \frac{3}{2}\max\{\|u^n\|_\infty^2,\ \|u^{n+1}\|_\infty^2\}. \tag{1.17}$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Step 2 (Trade energy for L^∞). This is the key step. We split u^{n+1} as

$$u^{n+1} = \mathcal{L}_1 u^n + \mathcal{L}_2\Big(\mathcal{N}(u^n)\Big), \tag{1.18}$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ are both linear operators mimicking the resolvent of elliptic type operators and $\mathcal{N}(u^n)$ denotes the nonlinear part. We then prove a direct L^∞ estimate using only the energy conservation and certain smoothing properties of the operators $\mathcal{L}_1$ and $\mathcal{L}_2$. To achieve an "optimal trade" (i.e. optimal dependence on ν) it is of some importance to use scaling-critical norms. In the end one obtains

$$\|u^{n+1}\|_\infty \le C_{\mathcal{E}(u^n)} \cdot h(\nu,\tau), \tag{1.19}$$

where $C_{\mathcal{E}(u^n)} > 0$ depends only on the energy $\mathcal{E}(u^n)$, and $h(\nu,\tau)$ typically has the form $h(\nu,\tau) = \nu^{-\alpha}\tau^{-\beta}$ for some exponents $\alpha > 0$, $\beta > 0$.

Step 3 (Identification of the optimal time step constraint). Here we work on the inequality

$$\sqrt{\frac{2\nu}{\tau}} \ge \frac{3}{2}(C_{\mathcal{E}(u^n)} \cdot h(\nu,\tau))^2 \tag{1.20}$$

and determine an optimal $\tau_{\max} = \tau_{\max}(\nu,\mathcal{E})$. A suitable induction procedure then closes the needed estimates and yields the result.

The rest of this paper is organized as follows. In Section 2 we set up the notation and collect various preliminary materials. The discrete energy inequality is proved in Lemma 2.6. Some Sobolev inequalities on $\mathbb{T}^2$ with explicit constants are also presented here. Section 3 is devoted to the TEFL proof for the general case $\nu > 0$ in dimension two. In Section 4 we give refined analysis for the special case $\nu = 1$ in dimension two. In Section 5 we carry out the error analysis. In the last section we give concluding remarks.

2. Notation and preliminaries

For any two positive quantities X and Y, we shall write $X \lesssim Y$ or $Y \gtrsim X$ if $X \le CY$ for some constant $C > 0$ whose precise value is unimportant. We shall write $X \sim Y$ if both $X \lesssim Y$ and $Y \lesssim X$ hold. We write $X \lesssim_\alpha Y$ if the constant C depends on some parameter α. We shall write $X = O(Y)$ if $|X| \lesssim Y$ and $X = O_\alpha(Y)$ if $|X| \lesssim_\alpha Y$.

We shall denote $X \ll Y$ if $X \le cY$ for some sufficiently small constant c. The smallness of the constant c is usually clear from the context. The notation $X \gg Y$ is similarly defined. Note that our use of $\ll$ and $\gg$ here is *different* from the usual Vinogradov notation in number theory or asymptotic analysis.

For any $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$, we denote $|x| = |x|_2 = \sqrt{x_1^2 + \cdots + x_d^2}$, and $|x|_\infty = \max_{1\le j\le d} |x_j|$. Also occasionally we use the Japanese bracket notation: $\langle x\rangle = (1+|x|^2)^{\frac{1}{2}}$.

For $1 \le p \le \infty$ and any function $f : x \in \mathbb{T}^d \to \mathbb{R}$, we denote the Lebesgue L^p-norm of f as

$$\|f\|_{L^p_x(\mathbb{T}^d)} = \|f\|_{L^p(\mathbb{T}^d)} = \|f\|_p.$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

If $(a_j)_{j\in I}$ is a sequence of complex numbers and I is the index set, we denote the discrete l^p-norm as

$$\|(a_j)\|_{l^p_j(j\in I)} = \|(a_j)\|_{l^p(I)} = \begin{cases} \left(\sum_{j\in I} |a_j|^p\right)^{\frac{1}{p}}, & 0<p<\infty, \\ \sup_{j\in I}|a_j|, & p=\infty. \end{cases} \tag{2.1}$$

For example, $\|\hat{f}(k)\|_{l^2_k(\mathbb{Z}^d)} = \left(\sum_{k\in\mathbb{Z}^d}|\hat{f}(k)|^2\right)^{\frac{1}{2}}$. If $f=(f_1,\cdots,f_m)$ is a vector-valued function, we denote $|f| = \sqrt{\sum_{j=1}^m |f_j|^2}$, and $\|f\|_p = \|(\sum_{j=1}^m f_j^2)^{\frac{1}{2}}\|_p$. We use similar convention for the corresponding discrete l^p norms for the vector-valued case.

We use the following convention for the Fourier transform pair:

$$\hat{f}(k) = \int_{\mathbb{T}^d} f(x)e^{-2\pi i k\cdot x}dx, \quad f(x) = \sum_{k\in\mathbb{Z}^d}\hat{f}(k)e^{2\pi i k\cdot x}, \tag{2.2}$$

and denote for $0\le s\in\mathbb{R}$,

$$\|f\|_{\dot{H}^s} = \|f\|_{\dot{H}^s(\mathbb{T}^d)} = \||\nabla|^s f\|_{L^2(\mathbb{T}^d)} = \|(2\pi|k|)^s\hat{f}(k)\|_{l^2_k(\mathbb{Z}^d)}, \tag{2.3a}$$

$$\|f\|_{H^s} = \sqrt{\|f\|_2^2 + \|f\|_{\dot{H}^s}^2} = \|\langle 2\pi|k|\rangle^s\hat{f}(k)\|_{l^2_k(\mathbb{Z}^d)}. \tag{2.3b}$$

To simplify the notation, in the later part of this paper we shall often denote

$$E_n = \mathcal{E}(u^n) = \int_{\mathbb{T}^d}(\frac{1}{2}\nu|\nabla u^n|^2 + \frac{1}{4}((u^n)^4 - 2(u^n)^2+1))dx, \tag{2.4}$$

where u^n is the discrete numerical solution computed according to the scheme (1.9).

Lemma 2.1. *Let $0\le\epsilon<1$. For any $\omega\in\mathbb{R}^2$ with $|\omega|=1$, we have*

$$F(\omega) = \int_{[-\frac{1}{2},\frac{1}{2})^2}(1+\epsilon^2|k|^2+2\epsilon\omega\cdot k)^{-1}dk \ge 1+\frac{1}{6}\epsilon^2 - \frac{7}{30}\epsilon^4.$$

In particular if $\epsilon^2\le\frac{5}{7}$, then $F(\omega)\ge 1$.

Proof. Denote the integrand as $g(k)$. By symmetry we have

$$F(\omega) = \int_{[-\frac{1}{2},\frac{1}{2})^2}\frac{1}{2}(g(k)+g(-k))dk.$$

Denote $a=\epsilon^2$. We shall slightly abuse the notation and write $k^2=|k|^2$. Clearly (note that $|k|^2\le\frac{1}{2}$)

$$\begin{aligned}\frac{1}{2}(g(k)+g(-k)) &= (1+ak^2)^{-1}\cdot(1-\frac{4a(\omega\cdot k)^2}{(1+ak^2)^2})^{-1} \\ &\ge (1+ak^2)^{-1}\cdot(1+\frac{4a(\omega\cdot k)^2}{(1+ak^2)^2}) \\ &\ge 1-ak^2+(1-ak^2)^3\cdot 4a(\omega\cdot k)^2.\end{aligned}$$

By using symmetry (under the swapping of variables $k_1\leftrightarrow k_2$) and the fact that $|\omega|=1$, we have for any f,

$$\int_{[-\frac{1}{2},\frac{1}{2})^2} f(k^2)(\omega\cdot k)^2dk_1dk_2 = \frac{1}{2}\int_{[-\frac{1}{2},\frac{1}{2})^2} f(k^2)k^2dk_1dk_2.$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Here we also used the fact

$$\int_{[-\frac{1}{2},\frac{1}{2})^2} f(k^2)k_1k_2dk_1dk_2 = 0.$$

Thus we only need to work with the integrand $1 - ak^2 + (1 - ak^2)^3 \cdot 2ak^2$. An explicit calculation then yields the result

$$1 + \frac{a}{6} - \frac{7a^2}{30} + \frac{9a^3}{140} - \frac{83a^4}{12600} \geq 1 + \frac{a}{6} - \frac{7a^2}{30}.$$

This completes the proof of the lemma. □

Let

$$I(n) = \int_{[-\frac{1}{2},\frac{1}{2})^2} \frac{1}{|k+n|^2}dk - \frac{1}{|n|^2}. \tag{2.5}$$

By symmetry, we have for any $n = (n_1, n_2)$, $I(n) = I(\pm n_1, \pm n_2)$. Furthermore, by direct numerical verification, we have the following:

$$\begin{aligned} &I(1,0) \geq 0.1731, \quad I(1,1) \geq 0.0525, \\ &I(2,0) \geq 0.0105, \quad I(2,1) \geq 0.007, \quad I(2,2) \geq 0.0027, \\ &I(3,0) \geq 0.0020, \quad I(3,1) \geq 0.0016, \quad I(3,2) \geq 0.0010, \quad I(3,3) \geq 0.0005. \end{aligned} \tag{2.6}$$

Lemma 2.2. *For any $\frac{1}{\sqrt{2}} < r_1 < r_2$, we have*

$$\sum_{\substack{r_1 \leq |n| \leq r_2 \\ n \in \mathbb{Z}^2}} \frac{1}{|n|^2} \leq 2\pi\Big(\ln(r_2 + \frac{1}{\sqrt{2}}) - \ln(r_1 - \frac{1}{\sqrt{2}})\Big). \tag{2.7}$$

Also for $r_2 > \frac{1}{\sqrt{2}}$,

$$\sum_{\substack{|n| \geq r_2 \\ n \in \mathbb{Z}^2}} \frac{1}{|n|^4} \leq \pi\left(r_2 - \frac{1}{\sqrt{2}}\right)^{-2}, \tag{2.8}$$

and more generally for any $s > 2$,

$$\sum_{\substack{|n| \geq r_2 \\ n \in \mathbb{Z}^2}} \frac{1}{|n|^s} \leq 2\pi \cdot \frac{1}{s-2}\left(r_2 - \frac{1}{\sqrt{2}}\right)^{-(s-2)}. \tag{2.9}$$

Proof. By Lemma 2.1 and (2.5)-(2.6), we have

$$\sum_{r_1 \leq |n| \leq r_2} |n|^{-2} \leq \sum_{r_1 \leq |n| \leq r_2} \int_{[-\frac{1}{2},\frac{1}{2})^2} |n+k|^{-2}dk.$$

Now for each $n \in \mathbb{Z}^2$, define $I_n = \{z \in \mathbb{R}^2 : \ |z - n|_\infty < \frac{1}{2}\}$. Clearly I_n and $I_{n'}$ are disjoint if $n \neq n'$. Obviously

$$\bigcup_{r_1 \leq |n| \leq r_2} I_n \subset \{z \in \mathbb{R}^2 : \ r_1 - \frac{1}{\sqrt{2}} \leq |z| \leq r_2 + \frac{1}{\sqrt{2}}\}.$$

Thus

$$\sum_{r_1 \leq |n| \leq r_2} |n|^{-2} \leq \int_{r_1 - \frac{1}{\sqrt{2}} \leq |z| \leq r_2 + \frac{1}{\sqrt{2}}} |z|^{-2}dz = 2\pi\Big(\ln(r_2 + \frac{1}{\sqrt{2}}) - \ln(r_1 - \frac{1}{\sqrt{2}})\Big).$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

For the second inequality, we note that

$$|n|^{-4} = (|n|^{-2})^2 \le \left(\int_{[-\frac{1}{2},\frac{1}{2})^2} |n+k|^{-2} dk \right)^2 \le \int_{[-\frac{1}{2},\frac{1}{2})^2} |n+k|^{-4} dk.$$

Thus

$$\sum_{|n|\ge r_2} |n|^{-4} \le \int_{|z|\ge r_2 - \frac{1}{\sqrt{2}}} |z|^{-4} dz = \pi \left(r_2 - \frac{1}{\sqrt{2}} \right)^{-2}.$$

This completes the proof of the lemma. □

It follows from the above lemma that for any $r \ge 2$, we have

$$\begin{aligned} \sum_{\substack{0<|k|\le r \\ k\in\mathbb{Z}^2}} |k|^{-2} &= 6 + \sum_{2\le|k|\le r} |k|^{-2} \\ &\le 6 + 2\pi(\ln(r+\frac{1}{\sqrt{2}}) - \ln(2-\frac{1}{\sqrt{2}})) \le 4.4 + 2\pi \ln\left(r + \frac{1}{\sqrt{2}}\right). \end{aligned} \tag{2.10}$$

Lemma 2.3. *Let $f \in H^2(\mathbb{T}^2)$ with mean zero. Then for any $r \ge 2$, we have*

$$\begin{aligned} &\|f\|_{L^\infty(\mathbb{T}^2)} \\ &\le \|f\|_{\dot{H}^1} \cdot \frac{1}{2\pi} \left(4.386 + 2\pi(\ln(r+\frac{1}{\sqrt{2}})) \right)^{\frac{1}{2}} + \|f\|_{\dot{H}^2} \cdot \frac{1}{4\pi^{\frac{3}{2}}} \left(r - \frac{1}{\sqrt{2}} \right)^{-1}. \end{aligned} \tag{2.11}$$

Proof. Splitting f into low and high frequencies, we have

$$\|f\|_\infty \le \|f\|_{\dot{H}^1} \left(\sum_{0<|k|\le r} \frac{1}{(2\pi|k|)^2} \right)^{\frac{1}{2}} + \|f\|_{\dot{H}^2} \left(\sum_{|k|>r} \frac{1}{(2\pi|k|)^4} \right)^{\frac{1}{2}}. \tag{2.12}$$

By Lemma 2.2, we have

$$\begin{aligned} \sum_{0<|k|\le r} |k|^{-2} &= 6 + \sum_{2\le|k|\le r} |k|^{-2} \le 6 + 2\pi \left(\ln(r+\frac{1}{\sqrt{2}}) - \ln(2-\frac{1}{\sqrt{2}}) \right) \\ &\le 4.386 + 2\pi \ln\left(r + \frac{1}{\sqrt{2}}\right), \\ \sum_{|k|>r} |k|^{-4} &\le \pi \left(r - \frac{1}{\sqrt{2}} \right)^{-2}. \end{aligned}$$

This completes the proof of the lemma. □

Lemma 2.4. *Denote $\mathbb{T}$ the one-periodic torus on $\mathbb{R}$ which we identify as $\mathbb{T} = [-\frac{1}{2}, \frac{1}{2})$. For any smooth $f : \mathbb{T} \to \mathbb{R}$, we have*

$$\|f - \bar{f}\|_{L^\infty(\mathbb{T})} \le \frac{1}{2} \|f'\|_{L^1(\mathbb{T})},$$

where $\bar{f}$ denotes the average of f on $\mathbb{T}$. On $\mathbb{T}^2 = [-\frac{1}{2}, \frac{1}{2})^2$, we have

$$\begin{aligned} \|f - \bar{f}\|_{L^2(\mathbb{T}^2)} &\le \frac{1}{2} \|(|\partial_1 f| + |\partial_2 f|)\|_{L^1(\mathbb{T}^2)} \\ &\le \frac{1}{\sqrt{2}} \|\nabla f\|_{L^1(\mathbb{T}^2)} = \frac{1}{\sqrt{2}} \|\sqrt{|\partial_1 f|^2 + |\partial_2 f|^2}\|_{L^1(\mathbb{T}^2)}. \end{aligned}$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Note that we use the convention $|\nabla f| = \sqrt{|\partial_1 f|^2 + |\partial_2 f|^2}$.

Proof. The first inequality can be proved in two ways. Without loss of generality, one may assume $\bar{f} = 0$. One can then choose some $x_0 \in \mathbb{T}$ such that $f(x_0) = 0$. Writing $f(x) = \int_{x_0}^{x} f'(s)ds = -\int_{x}^{x_0+1} f'(s)ds$ then yields the result. Alternatively one can resort to Fourier analysis and show that

$$K(x) = \mathcal{F}^{-1}\left(\frac{1}{2\pi i k} 1_{k\neq 0}\right) = \frac{1}{2}\operatorname{sgn}(x) - x = \begin{cases} \frac{1}{2} - x, & 0 < x \leq \frac{1}{2}; \\ -\frac{1}{2} - x, & -\frac{1}{2} \leq x < 0. \end{cases}$$

Obviously $\|K\|_\infty \leq \frac{1}{2}$ and the desired inequality follows. For the second inequality we may also assume $\bar{f} = 0$. Observe

$$|f(x_1, x_2) - \underbrace{\int f(y_1, x_2) dy_1}_{=:B(x_2)}| \leq \frac{1}{2}\int |\partial_1 f(y_1, x_2)| dy_1, \tag{2.13a}$$

$$|f(x_1, x_2) - \underbrace{\int f(x_1, y_2) dy_2}_{=:A(x_1)}| \leq \frac{1}{2}\int |\partial_2 f(x_1, y_2)| dy_2. \tag{2.13b}$$

Note that since $\bar{f} = 0$ we have $\int A(x_1)dx_1 = 0 = \int B(x_2)dx_2$. Clearly then

$$\|A(x_1)\|_{L^\infty_{x_1}} \leq \frac{1}{2}\int |\partial_1 A| dx_1 \leq \frac{1}{2}\|\partial_1 f\|_{L^1}, \quad \|B(x_2)\|_{L^\infty_{x_2}} \leq \frac{1}{2}\|\partial_2 f\|_{L^1}.$$

It follows that

$$\begin{aligned} f^2 &\leq -A(x_1)B(x_2) + (A(x_1) + B(x_2))f + \frac{1}{4}\int |\partial_1 f(y_1, x_2)| dy_1 \int |\partial_2 f(x_1, y_2)| dy_2 \\ &\leq \frac{1}{2}f^2 + \frac{1}{2}(A(x_1) + B(x_2))^2 - A(x_1)B(x_2) \\ &\qquad + \frac{1}{4}\int |\partial_1 f(y_1, x_2)| dy_1 \int |\partial_2 f(x_1, y_2)| dy_2. \end{aligned}$$

We then obtain

$$\begin{aligned} \|f\|_{L^2}^2 &\leq \int A(x_1)^2 dx_1 + \int B(x_2)^2 dx_2 + \frac{1}{2}\|\partial_1 f\|_{L^1}\|\partial_2 f\|_{L^1} \\ &\leq \frac{1}{4}\left(\|\partial_1 f\|_{L^1} + \|\partial_2 f\|_{L^1}\right)^2, \end{aligned}$$

where in the last inequality we used $\|A(x_1)\|_\infty \leq \frac{1}{2}\|\partial_1 f\|_{L^1}$, $\|B(x_2)\|_\infty \leq \frac{1}{2}\|\partial_2 f\|_{L^1}$. □

Lemma 2.5. *For any smooth* $f : \mathbb{T}^2 \to \mathbb{R}$, *we have*

$$\|f - \bar{f}\|_{L^\infty(\mathbb{T}^2)} \leq \frac{\sqrt{6.05}}{(2\pi)^2}\|f\|_{\dot{H}^2(\mathbb{T}^2)}. \tag{2.14}$$

Proof. Note that by Lemma 2.2, for any $r > \frac{1}{\sqrt{2}}$, we have

$$\sum_{0\neq k\in\mathbb{Z}^2} \frac{1}{(2\pi|k|)^4} \leq \frac{1}{(2\pi)^4}\left(\sum_{0<|k|\leq r} \frac{1}{|k|^4} + \pi(r - \frac{1}{\sqrt{2}})^{-2}\right).$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Choosing $r = 10$ then yields the result. One should note that $\pi(10 - \frac{1}{\sqrt{2}})^{-2} \approx 0.0363788 < 0.037$ and

$$\sum_{0<|k|\le 10} |k|^{-4} \le \sum_{0<|k|_\infty \le 10} |k|^{-4} \approx 6.00355 < 6.0036.$$

This completes the proof. □

Lemma 2.6 (Discrete energy estimate)**.** *For any $n \ge 0$,*

$$\begin{aligned} & E_{n+1} - E_n + \left(\frac{1}{2} + \sqrt{\frac{2\nu}{\tau}}\right) \|u^{n+1} - u^n\|_2^2 \\ \le\ & \|u^{n+1} - u^n\|_2^2 \cdot \frac{3}{2} \max\{\|u^n\|_\infty^2,\ \|u^{n+1}\|_\infty^2\}, \end{aligned} \tag{2.15}$$

where E_n was defined in (2.4)*.*

Proof. In this proof we denote by $(\cdot,\cdot)$ the usual L^2 inner product. Recall

$$\frac{u^{n+1} - u^n}{\tau} = -\nu\Delta^2 u^{n+1} + \Delta\Big(f(u^n)\Big).$$

Taking the L^2 inner product with $(-\Delta)^{-1}(u^{n+1} - u^n)$ on both sides and applying the identity $b \cdot (b - a) = \frac{1}{2}(|b|^2 - |a|^2 + |b - a|^2)$, we get

$$\begin{aligned} & \frac{1}{\tau}\||\nabla|^{-1}(u^{n+1} - u^n)\|_2^2 + \frac{\nu}{2}\Big(\|\nabla u^{n+1}\|_2^2 - \|\nabla u^n\|_2^2 + \|\nabla(u^{n+1} - u^n)\|_2^2\Big) \\ & = (\Delta\Big(f(u^n)\Big), (-\Delta)^{-1}(u^{n+1} - u^n)). \end{aligned} \tag{2.16}$$

Clearly

$$(\Delta\Big(f(u^n)\Big), (-\Delta)^{-1}(u^{n+1} - u^n)) = -(f(u^n), u^{n+1} - u^n).$$

By using the auxiliary function $g(s) = F(u^n + s(u^{n+1} - u^n))$ (recall $f = F'$) and the Taylor expansion $g(1) = g(0) + g'(0) + \int_0^1 g''(s)(1 - s)ds$, we get

$$\begin{aligned} F(u^{n+1}) =& F(u^n) + f(u^n)(u^{n+1} - u^n) - \frac{1}{2}(u^{n+1} - u^n)^2 \\ & + (u^{n+1} - u^n)^2 \int_0^1 \tilde{f}'(u^n + s(u^{n+1} - u^n))(1 - s)ds, \end{aligned} \tag{2.17}$$

where $\tilde{f}(z) = z^3$ and $\tilde{f}'(z) = 3z^2$ (for $z \in \mathbb{R}$). From this it is easy to see that

$$\begin{aligned} & -(f(u^n), u^{n+1} - u^n) \\ \le & F(u^n) - F(u^{n+1}) - \frac{1}{2}\|u^{n+1} - u^n\|_2^2 + \|u^{n+1} - u^n\|_2^2 \cdot \frac{3}{2} \max\{\|u^n\|_\infty^2, \|u^{n+1}\|_\infty^2\}. \end{aligned} \tag{2.18}$$

Thus

$$\begin{aligned} & E_{n+1} - E_n + \frac{1}{\tau}\||\nabla|^{-1}(u^{n+1} - u^n)\|_2^2 + \frac{\nu}{2}\|\nabla(u^{n+1} - u^n)\|_2^2 + \frac{1}{2}\|u^{n+1} - u^n\|_2^2 \\ \le & \|u^{n+1} - u^n\|_2^2 \cdot \frac{3}{2} \max\{\|u^n\|_\infty^2, \|u^{n+1}\|_\infty^2\}. \end{aligned} \tag{2.19}$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Finally, observe

$$\begin{aligned}&\frac{1}{\tau}\||\nabla|^{-1}(u^{n+1}-u^n)\|_2^2+\frac{\nu}{2}\|\nabla(u^{n+1}-u^n)\|_2^2\\ \geq&\sqrt{\frac{2\nu}{\tau}}\||\nabla|^{-1}(u^{n+1}-u^n)\|_2\|\nabla(u^{n+1}-u^n)\|_2\geq\sqrt{\frac{2\nu}{\tau}}\|u^{n+1}-u^n\|_2^2.\end{aligned}$$

The desired inequality then follows easily. □

Lemma 2.7. *Let $d\leq 3$ and $\beta>0$. Consider on the torus $\mathbb{T}^d$,*

$$K(x)=\mathcal{F}^{-1}((1+\beta(2\pi|k|)^4)^{-1})=(1+\beta\Delta^2)^{-1}\delta_0, \tag{2.20}$$

where δ_0 is the periodic Dirac comb. Then for any $1\leq p\leq\infty$,

$$\|K\|_{L^p(\mathbb{T}^d)}\leq c_{d,p}\,(1+\beta^{-d(\frac{1}{4}-\frac{1}{4p})}), \tag{2.21}$$

where $c_{d,p}>0$ depends only on d and p. Define

$$\widetilde{K}=\mathcal{F}^{-1}((1+\beta(2\pi|k|)^4)^{-1}1_{k\neq 0}). \tag{2.22}$$

Then

$$\|\widetilde{K}\|_{L^p(\mathbb{T}^d)}\leq\tilde{c}_{d,p}\,\beta^{-d(\frac{1}{4}-\frac{1}{4p})}, \tag{2.23}$$

where $\tilde{c}_{d,p}>0$ depends only on d and p.

Proof. Define

$$K_1(x)=\int_{\mathbb{R}^d}e^{2\pi i\xi\cdot x}\cdot\frac{1}{1+(2\pi|\xi|)^4}d\xi.$$

It is easy to check that $|K_1(x)|\lesssim\langle x\rangle^{-10}$ and $K_1\in L^1_x(\mathbb{R}^d)$ for $d\leq 3$. Now note that for $d\leq 3$, if $|x|_\infty\leq\frac{1}{2}$, then $|x|\leq\sqrt{d}\cdot\frac{1}{2}\leq\frac{\sqrt{3}}{2}$. Thus if $|l|\geq 4$, then

$$\frac{1}{2}|l|\leq|x+l|\leq 2|l|,\qquad\forall\,|x|_\infty\leq\frac{1}{2}.$$

It follows that for all $1\leq p\leq\infty$ and $|l|\geq 4$,

$$\|\langle\beta^{-\frac{1}{4}}(x+l)\rangle^{-10}\|_{L^p_x(|x|_\infty<\frac{1}{2})}\leq\langle\beta^{-\frac{1}{4}}\frac{1}{2}|l|\rangle^{-10}\ \leq\|\langle\beta^{-\frac{1}{4}}\frac{1}{4}|x+l|\rangle^{-10}\|_{L^1_x(|x|_\infty<\frac{1}{2})}.$$

Clearly then

$$\begin{aligned}\|K\|_{L^p(\mathbb{T}^d)}&\leq\beta^{-\frac{d}{4}}\sum_{l\in\mathbb{Z}^d}\|K_1(\beta^{-\frac{1}{4}}(x+l))\|_{L^p_x(|x|_\infty<\frac{1}{2})}\\ &\lesssim\ \beta^{-\frac{d}{4}}\sum_{|l|\leq 4}\|K_1(\beta^{-\frac{1}{4}}(x+l))\|_p+\beta^{-\frac{d}{4}}\sum_{|l|>4}\|\langle\beta^{-\frac{1}{4}}\frac{1}{4}|x+l|\rangle^{-10}\|_{L^1_x(|x|_\infty<\frac{1}{2})}\\ &\lesssim\ \beta^{-d(\frac{1}{4}-\frac{1}{4p})}+1.\end{aligned} \tag{2.24}$$

Now we consider the estimate for $\widetilde{K}(x)=K(x)-1$. Obviously by using the previous bound we have $\|\widetilde{K}\|_1\lesssim\|K\|_1+1\lesssim 1$. Alternatively one can compute

$$\|\tilde{K}\|_{L^1_x(\mathbb{T}^d)}\leq 1+\|\sum_{l\in\mathbb{Z}^d}\beta^{-\frac{d}{4}}|K_1(\beta^{-\frac{1}{4}}(x+l))|\|_{L^1_x(\mathbb{T}^d)}\leq 1+\|K_1\|_{L^1_x(\mathbb{R}^d)}\lesssim 1.$$

We then bound the L^2 norm as

$$\|\widetilde{K}\|_{L^2_x(\mathbb{T}^d)}=\|\frac{1}{1+\beta(2\pi|k|)^4}\|_{l^2_k(0\neq k\in\mathbb{Z}^d)}\lesssim\begin{cases}\beta^{-1}, & \text{if } \beta\geq 1;\\ \beta^{-\frac{d}{8}}, & \text{if } 0<\beta<1.\end{cases}$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Since $d \le 3$, we have the uniform bound for all $\beta > 0$ as $\|\widetilde{K}\|_{L^2_x(\mathbb{T}^d)} \lesssim \beta^{-\frac{d}{8}}$. Similarly,

$$\|\widetilde{K}\|_{L^\infty_x(\mathbb{T}^d)} \le \|\frac{1}{1+\beta(2\pi|k|)^4}\|_{l^1_k(0\neq k\in\mathbb{Z}^d)} \lesssim \begin{cases}\beta^{-1}, & \text{if } \beta \ge 1;\\ \beta^{-\frac{d}{4}}, & \text{if } 0<\beta<1;\end{cases}$$
$$\lesssim \beta^{-\frac{d}{4}}, \quad \forall \beta > 0 \text{ (since } d \le 3 \Rightarrow \frac{d}{4} < 1).$$

By using interpolation we then get the L^p estimate. □

3. Proof for general $\nu > 0$

In this section, we consider the semi-implicit scheme (1.9) in the representative two-dimensional case with $\nu > 0$.

Lemma 3.1. *Let $d = 2$ and $\nu > 0$. Let $\tau > 0$. Then for any $g \in L^4(\mathbb{T}^2)$ with zero mean, we have*

$$\|(1+\nu\tau\Delta^2)^{-1}g\|_\infty \le C_1(\nu\tau)^{-\frac{1}{8}}\|g\|_4. \tag{3.1}$$

For any $g_1 \in L^{\frac{4}{3}}(\mathbb{T}^2)$, we have

$$\|\tau\Delta(1+\nu\tau\Delta^2)^{-1}g_1\|_\infty \le C_2\tau(\nu\tau)^{-\frac{7}{8}}\|g_1\|_{\frac{4}{3}}. \tag{3.2}$$

In the above $C_1 > 0$, $C_2 > 0$ are absolute constants.

Proof. Denote $\beta = \nu\tau$. The first inequality follows from Lemma 2.7 (see the bound for $\tilde{K}$ therein). For the second inequality denote

$$K_\beta = \mathcal{F}^{-1}\left(\frac{\beta^{\frac{1}{2}}(2\pi|k|)^2}{1+\beta(2\pi|k|)^4}\right). \tag{3.3}$$

We then have $\|K_\beta\|_4 \le \|\widehat{K_\beta}\|_{l^{\frac{4}{3}}_k} \lesssim \beta^{-\frac{3}{8}}$. □

Lemma 3.2. *Let $d \ge 1$. If $E_p = \int_{\mathbb{T}^d} \frac{1}{4}(v^2-1)^2 dx$, then*

$$\|v\|^2_{L^4(\mathbb{T}^d)} \le 1 + 2\sqrt{E_p}, \qquad \|v^3 - v\|^2_{L^{\frac{4}{3}}(\mathbb{T}^d)} \le 4E_p(1+2\sqrt{E_p}). \tag{3.4}$$

Proof. Obvious. For the second inequality, note that $\|(v^2-1)v\|_{\frac{4}{3}} \le \|v^2-1\|_2\|v\|_4$. □

Theorem 3.1 (Conditional energy stability for 2D $\nu > 0$)**.** *Let $d = 2$ and $\nu > 0$. Assume $u_0 \in H^1(\mathbb{T}^2)$ and has zero mean. Assume $\|u^0\|_\infty \le L_0$ and $0 < L_0 < \infty$. Take*

$$\tau_{\max} = \min\Big\{\frac{8\nu}{9L_0^4}, \tau^{(1)}_{\max}\Big\}, \tag{3.5}$$

where

$$\tau^{(1)}_{\max} = 0.04\nu^3\min\{\alpha_1^{-8}, \alpha_2^{-\frac{8}{3}}\}, \alpha_1 = C_1\sqrt{1+2\sqrt{E_0}}, \alpha_2 = 2C_2\sqrt{E_0}\cdot\sqrt{1+2\sqrt{E_0}}. \tag{3.6}$$

In the above C_1, C_2 are the same constants in Lemma 3.1. Then for any $0 < \tau \le \tau_{\max}$, the scheme (1.9) is energy stable, i.e. $\mathcal{E}(u^{n+1}) \le \mathcal{E}(u^n)$ for $n \ge 0$.

Proof. We shall use induction.

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Step 1 (The base step $n = 0$). Thanks to our choice of τ, we clearly have for $n = 0$, $\sqrt{2\nu/\tau} \geq \frac{3}{2}\|u^0\|_\infty^2$. To ensure $E_1 \leq E_0$, we need to check $\sqrt{2\nu/\tau} \geq \frac{3}{2}\|u^1\|_\infty^2$. By Lemmas 3.1 and 3.2, we have

$$\|u^1\|_\infty \leq \alpha_1(\nu\tau)^{-\frac{1}{8}} + \alpha_2\tau(\nu\tau)^{-\frac{7}{8}},$$

where α_1 and α_2 are given by (3.6). By Lemma 2.6, to show $E_1 \leq E_0$, it suffices for us to check the inequality

$$\Big(\frac{2}{3}\sqrt{\frac{2\nu}{\tau}}\Big)^{\frac{1}{2}} \geq \alpha_1(\nu\tau)^{-\frac{1}{8}} + \alpha_2\tau(\nu\tau)^{-\frac{7}{8}}.$$

Now set $\tau = \nu^3 r$ and for r we need to check $2^{\frac{3}{4}}3^{-\frac{1}{2}} \geq \alpha_1 r^{\frac{1}{8}} + \alpha_2 r^{\frac{3}{8}}$. Let $r = \alpha_1^4\alpha_2^{-4}z$. Then we need $2^{\frac{3}{4}}3^{-\frac{1}{2}} \geq \alpha_1^{\frac{3}{2}}\alpha_2^{-\frac{1}{2}}(z^{\frac{1}{8}} + z^{\frac{3}{8}})$. Now we choose $\eta \in (0, 1)$, such that

$$z^{\frac{1}{8}} \leq \eta 2^{\frac{3}{4}}3^{-\frac{1}{2}}\alpha_1^{-\frac{3}{2}}\alpha_2^{\frac{1}{2}}, \quad z^{\frac{3}{8}} \leq (1-\eta)2^{\frac{3}{4}}3^{-\frac{1}{2}}\alpha_1^{-\frac{3}{2}}\alpha_2^{\frac{1}{2}}.$$

A nearly optimal choice is $\eta = 0.690119$ for which

$$(\eta 2^{\frac{3}{4}}3^{-\frac{1}{2}})^8 \approx 0.0406525 > 0.04, \quad ((1-\eta)2^{\frac{3}{4}}3^{-\frac{1}{2}})^{\frac{8}{3}} \approx 0.0406523 > 0.04.$$

Thus it suffices to require $z \leq 0.04\min\{\alpha_1^{-12}\alpha_2^4,\, \alpha_1^{-4}\alpha_2^{\frac{4}{3}}\}$. Thus for τ we need

$$0 < \tau \leq \tau_{\max}^{(1)} = 0.04\nu^3\min\{\alpha_1^{-8},\, \alpha_2^{-\frac{8}{3}}\}.$$

This in turn guarantees that $E_1 \leq E_0$.

Step 2 (Induction). For $n \geq 1$, the induction hypothesis is that

$$E_n \leq E_{n-1}, \quad \|u^n\|_\infty \leq \alpha_1(\nu\tau)^{-\frac{1}{8}} + \alpha_2\tau(\nu\tau)^{-\frac{7}{8}}. \tag{3.7}$$

Clearly by using similar estimates as in Step 1 for u^1, one can check that

$$\|u^{n+1}\|_\infty \leq \alpha_1(\nu\tau)^{-\frac{1}{8}} + \alpha_2\tau(\nu\tau)^{-\frac{7}{8}}, \quad \sqrt{\frac{2\nu}{\tau}} \geq \frac{3}{2}\max\{\|u^n\|_\infty^2,\, \|u^{n+1}\|_\infty^2\}. \tag{3.8}$$

Then by Lemma 2.6 we obtain $E_{n+1} \leq E_n$ which completes the induction step.

□

Remark 3.1. Despite our rigorous theoretic analysis, the role of the time step still merits further investigation especially regarding the nonlinear evolution. Here we provide a simple numerical example where $\nu = 0.01$, the initial condition is $u_0 = 0.1(\sin(3Kx)\sin(2Ky) + \sin(5Kx)\sin(5Ky))$ $(K = 2\pi)$ and the number of Fourier modes $N_x = N_y = 128$. The time step is taken to be $\tau = 0.1,\ 0.01,\ 0.001$, and 0.0001 respectively. Figure 1 illustrates the energy evolutions for different choices of time step. It seems that the time step has a significant influence on the profile of energy evolution, despite that the energy dissipation property holds for all cases.

4. Refined results for $\nu = 1$

Consider the semi-implicit scheme (1.9) with $\nu = 1$:

$$\frac{u^{n+1} - u^n}{\tau} = -\Delta^2 u^{n+1} + \Delta(f(u^n)), \quad n \geq 0, \tag{4.1}$$

where $\tau > 0$ is the time step. Then $(1 + \tau\Delta^2)u^{n+1} = u^n + \tau\Delta(f(u^n))$, which gives

$$u^{n+1} = \frac{1 + 2\tau\Delta}{1 + \tau\Delta^2}u^n + \frac{\tau\Delta}{1 + \tau\Delta^2}((u^n)^3 - 3u^n). \tag{4.2}$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

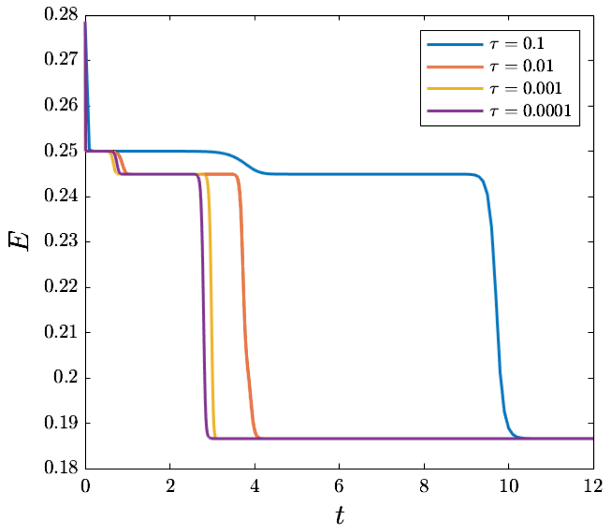

FIGURE 1. Energy profiles of 2D Cahn-Hilliard equation for different choices of time step $\tau = 0.1,\ 0.01,\ 0.001$, and 0.0001

Lemma 4.1. *We have*

(4.3a)

$$\Big(\sum_{0\neq k\in\mathbb{Z}^2}\big(\frac{1}{1+\tau(2\pi|k|)^4}\big)^2\frac{1}{(2\pi|k|)^2}\Big)^{\frac{1}{2}} \le \begin{cases}(-\frac{1}{8\pi}\ln\tau+0.52)^{\frac{1}{2}}, & \textit{if } 0<\tau\le(4\pi)^{-4};\\ 0.8803, & \textit{if } \tau>(4\pi)^{-4},\end{cases}$$

(4.3b)

$$\Big(\sum_{0\neq k\in\mathbb{Z}^2}\big(\frac{1-2\tau(2\pi|k|)^2}{1+\tau(2\pi|k|)^4}\big)^2\frac{1}{(2\pi|k|)^2}\Big)^{\frac{1}{2}} \le\ h(\tau),$$

where

$$h(\tau)=\begin{cases}(-\frac{1}{8\pi}\ln\tau+0.52)^{\frac{1}{2}}+0.00872, & \textit{if } 0<\tau\le(4\pi)^{-4};\\ 0.8988, & \textit{if } \tau>(4\pi)^{-4}.\end{cases} \tag{4.3c}$$

Proof. The natural cut-off is $k_0=k_0(\tau)=\frac{1}{2\pi}\tau^{-\frac{1}{4}}$.

Case 1 ($k_0\ge 2$). This is equivalent to $0<\tau\le(4\pi)^{-4}<4.02\times10^{-5}$. Then by (2.10), we have

$$\sum_{0<|k|\le k_0}(1+\tau(2\pi|k|)^4)^{-2}|k|^{-2}\le\sum_{0<|k|\le k_0}|k|^{-2}\le 4.4+2\pi\ln(k_0+\frac{1}{\sqrt{2}})$$
$$=4.4+2\pi\ln k_0+2\pi\ln(1+\frac{1}{k_0\sqrt{2}})\le-\frac{\pi}{2}\ln\tau-5.24.$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Also by Lemma 2.2, we have

$$\sum_{|k|>k_0}(1+\tau(2\pi|k|)^4)^{-2}|k|^{-2} \le \frac{(2\pi)^{-8}}{\tau^2}\sum_{|k|>k_0}|k|^{-10}$$
$$\le \tau^{-2}(2\pi)^{-7}\cdot\frac{1}{8}\left(k_0-\frac{1}{\sqrt{2}}\right)^{-8} \le 25.76.$$

Thus

$$\left(\sum_{0\neq k\in\mathbb{Z}^2}(1+\tau(2\pi|k|)^4)^{-2}(2\pi|k|)^{-2}\right)^{\frac{1}{2}} \le \frac{1}{2\pi}\left(-\frac{\pi}{2}\ln\tau+20.52\right)^{\frac{1}{2}}$$
$$\le \left(-\frac{1}{8\pi}\ln\tau+0.52\right)^{\frac{1}{2}}.$$

Now observe (for the first inequality we use $1+\tau(2\pi|k|)^4 \ge 2\sqrt{\tau}(2\pi|k|)^2$)

$$\sum_{0<|k|\le k_0}\left(\frac{2\tau(2\pi|k|)^2}{1+\tau(2\pi|k|)^4}\right)^2|k|^{-2} \le \tau\left(-\frac{\pi}{2}\ln\tau-5.24\right)$$
$$< 0.0015, \quad \forall\, 0<\tau<4.02\times10^{-5},$$
$$\sum_{|k|>k_0}\left(\frac{2\tau(2\pi|k|)^2}{1+\tau(2\pi|k|)^4}\right)^2|k|^{-2} \le 4\cdot(2\pi)^{-4}\sum_{|k|>k_0}|k|^{-6}$$
$$\le (2\pi)^{-3}\cdot\left(2-\frac{1}{\sqrt{2}}\right)^{-4} < 0.0015.$$

Thus

$$\Big(\sum_{0\neq k\in\mathbb{Z}^2}\left(\frac{2\tau(2\pi|k|)^2}{1+\tau(2\pi|k|)^4}\right)^2(2\pi|k|)^{-2}\Big)^{\frac{1}{2}} \le \frac{1}{2\pi}(0.0015+0.0015)^{\frac{1}{2}} < 0.00872.$$

This gives the desired result (4.3) when $k_0 \ge 2$.

Case 2 ($0<k_0<2$). In this regime $\tau>(4\pi)^{-4}$. In this case,

$$\sum_{0<|k|<2}(1+\tau(2\pi|k|)^4)^{-2}|k|^{-2} = 4(1+\tau(2\pi)^4)^{-2}+4\cdot\frac{1}{2}\cdot(1+\tau(2\pi\sqrt{2})^4)^{-2} \le 4.83,$$
$$\sum_{|k|\ge2}(1+\tau(2\pi|k|)^4)^{-2}|k|^{-2} \le \tau^{-2}(2\pi)^{-8}\sum_{|k|\ge2}|k|^{-10}$$
$$\le (4\pi)^8(2\pi)^{-7}\cdot\frac{1}{8}(2-\frac{1}{\sqrt{2}})^{-8} \le 25.76.$$

Thus

$$\Big(\sum_{0\neq k\in\mathbb{Z}^2}(1+\tau(2\pi|k|)^4)^{-2}(2\pi|k|)^{-2}\Big)^{\frac{1}{2}} \le \frac{1}{2\pi}\sqrt{4.83+25.76} < 0.8803.$$

Now observe

$$\sum_{0<|k|<2}\left(\frac{2\tau(2\pi|k|)^2}{1+\tau(2\pi|k|)^4}\right)^2|k|^{-2} = 4\tau^2(2\pi)^4\Big(4(1+\tau(2\pi)^4)^{-2}+8(1+\tau(2\pi\sqrt{2})^4)^{-2}\Big)$$
$$= \pi^{-4}x^2((1+x)^{-2}+2(1+4x)^{-2}) < 0.012, \quad \forall\, x>\frac{1}{16},$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

where we have denoted $x = (2\pi)^4\tau > \frac{1}{16}$. On the other hand, we have

$$\sum_{|k|\geq 2}\Big(\frac{2\tau(2\pi|k|)^2}{1+\tau(2\pi|k|)^4}\Big)^2|k|^{-2} \leq 4\cdot(2\pi)^{-4}\sum_{|k|\geq 2}|k|^{-6} \leq (2\pi)^{-3}\cdot(2-\frac{1}{\sqrt{2}})^{-4} < 0.0015.$$

Thus

$$\Big(\sum_{0\neq k\in\mathbb{Z}^2}\Big(\frac{2\tau(2\pi|k|)^2}{1+\tau(2\pi|k|)^4}\Big)^2(2\pi|k|)^{-2}\Big)^{\frac{1}{2}} \leq \frac{1}{2\pi}(0.012+0.0015)^{\frac{1}{2}} < 0.0185.$$

Finally to estimate the left-hand side of (4.3b), we just use the triangle inequality $\|A+B\|_{l^2_k} \leq \|A\|_{l^2_k} + \|B\|_{l^2_k}$.

□

Lemma 4.2. *Denote $\mathcal{E}(u^n) = E_n$. Then*

$$\begin{aligned}
&\Big\|\Delta\frac{\tau\Delta}{1+\tau\Delta^2}((u^n)^3-3u^n)\Big\|_{L^2(\mathbb{T}^2)} \leq 3E_n;\\
&\|u^{n+1}\|_{\dot{H}^2(\mathbb{T}^2)} \leq (2+\tau^{-1})\sqrt{E_n}+3E_n;\\
&\|u^{n+1}\|_{\dot{H}^1(\mathbb{T}^2)} \leq 2\sqrt{2}\sqrt{E_n}+3E_n.
\end{aligned}$$

Proof. For simplicity of notation we denote $v = u^{n+1}$ and $u = u^n$. We shall prove the first two inequalities at one stroke. Using (4.2) and by Lemma 2.4, we have (below $\overline{u^3}$ denotes the average of u^3 on $\mathbb{T}^2$)

$$\begin{aligned}
\|v\|_{\dot{H}^2} &\leq (2+\tau^{-1})\|u\|_2 + \|u^3-\overline{u^3}-3u\|_2\\
&\leq (2+\tau^{-1})\frac{1}{\sqrt{2}}\|\nabla u\|_1 + \frac{1}{\sqrt{2}}\|3\nabla u(u^2-1)\|_1\\
&\leq (2+\tau^{-1})\frac{1}{\sqrt{2}}\|\nabla u\|_2 + 3\|2\cdot\frac{1}{\sqrt{2}}\nabla u\cdot\frac{1}{2}(u^2-1)\|_1\\
&\leq (2+\tau^{-1})\sqrt{E_n}+3E_n.
\end{aligned}$$

The third inequality is similarly proved. We omit details. □

Theorem 4.1 (Conditional energy stability for 2D $\nu = 1$, second approach)**.** *Let $d = 2$ and $\nu = 1$. Assume $u_0 \in H^1(\mathbb{T}^2)$ and has zero mean. Let $u^0 = u_0$ and assume $\|u^0\|_\infty \leq 1.25$. Take*

$$\tau_{\max} = \min\Big\{\frac{1}{2},\ \frac{1}{B_1(\ln B_1)^2},\ \frac{8}{9}\frac{1}{B_2}\Big\}, \tag{4.5}$$

where $B_1 = \frac{3.1}{99.8}(\sqrt{2E_0}+0.19676E_0)^4$, $B_2 = (0.8988\sqrt{2E_0}+0.18692E_0)^4$. Then for any $0 < \tau \leq \tau_{\max}$, the scheme (4.1) is energy stable, i.e. $\mathcal{E}(u^{n+1}) \leq \mathcal{E}(u^n)$, for all $n \geq 0$.

Proof. Note that since by assumption $0 < \tau \leq \frac{1}{2}$, we have

$$\frac{1}{2}+\sqrt{\frac{2}{\tau}} \geq 2.5 \geq \frac{3}{2}\|u^0\|_\infty^2.$$

We shall focus on the induction step (since the estimate for u^1 is the same as the estimate for u^{n+1} below). Our inductive hypothesis is for $n \geq 1$,

$$E_n \leq E_{n-1}, \quad \|u^n\|_\infty \leq \sqrt{2E_0}h(\tau)+0.18692E_0, \tag{4.6}$$

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

where $h(\tau)$ is given by (4.3c). By Lemma 4.1, we have

$$\left\|\frac{1+2\tau\Delta}{1+\tau\Delta^2}u^n\right\|_\infty \le h(\tau)\|u^n\|_{\dot{H}^1} \le \sqrt{2E_0}h(\tau).$$

On the other hand by Lemmas 2.5 and 4.2, we have

$$\left\|\frac{\tau\Delta}{1+\tau\Delta^2}\Pi_N((u^n)^3-3u^n)\right\|_\infty \le \frac{\sqrt{6.05}}{(2\pi)^2}\cdot 3E_0 \le 0.18692E_0.$$

Clearly then $\|u^{n+1}\|_\infty \le \sqrt{2E_0}h(\tau)+0.18692E_0$. It remains for us to verify $E_{n+1} \le E_n$. This amounts to check (1.17). It suffices for us to prove

$$\sqrt{\left(\frac{1}{2}+\sqrt{\frac{2}{\tau}}\right)\cdot\frac{2}{3}} \ge \sqrt{2E_0}h(\tau)+0.18692E_0.$$

Case 1 ($\tau > (4\pi)^{-4}$). We only need to show $(\frac{1}{2}+\sqrt{\frac{2}{\tau}})\cdot\frac{2}{3} \ge (0.8988\sqrt{2E_0}+0.18692E_0)^2$, or in a slightly simpler form $\sqrt{2/\tau} \ge (0.8988\sqrt{2E_0}+0.18692E_0)^2\cdot\frac{3}{2}-\frac{1}{2}$. Thus it is sufficient to require $\tau \le \frac{8}{9}\frac{1}{A_1}$, where $A_1 = (0.8988\sqrt{2E_0}+0.18692E_0)^4$.

Case 2 ($0 < \tau \le (4\pi)^{-4}$). Then we need to show

$$\sqrt{(\frac{1}{2}+\sqrt{\frac{2}{\tau}})\cdot\frac{2}{3}} \ge \sqrt{2E_0}\underbrace{\left((-\frac{1}{8\pi}\ln\tau+0.52)^{\frac{1}{2}}+0.00872\right)}_{=:f_0(\tau)}+0.18692E_0.$$

It is not difficult to check that $f_0(\tau) \ge 0.95$ for $0 < \tau \le (4\pi)^{-4}$. Note that $0.18692/0.95 < 0.19676$. Thus it suffices for us to show

$$\sqrt{(\frac{1}{2}+\sqrt{\frac{2}{\tau}})\cdot\frac{2}{3}}\cdot\frac{1}{f_0(\tau)} \ge \sqrt{2E_0}+0.19676E_0.$$

We only need to show $\frac{8}{9}\cdot(\sqrt{2E_0}+0.19676E_0)^{-4} \ge \tau f_0(\tau)^4$. It is not difficult to check that $f_0(\tau) \le 1.54(-\ln\tau/(8\pi))^{\frac{1}{2}}$ for $0 < \tau \le (4\pi)^{-4}$. We then only need to prove

$$99.8(\sqrt{2E_0}+0.19676E_0)^{-4} \ge \tau(\ln\tau)^2. \tag{4.7}$$

Now discuss two cases.

Case (a) ($E_0 \ge 4.57916$). In this case we have

$$99.8(\sqrt{2E_0}+0.19676E_0)^{-4} \le 0.419538 < 3.1e^{-2} \approx 0.419539,$$

Consider $h(x) = x(\ln x)^2$ for $0 < x \le 1$. If $0 < x \le e^{-2}$, then $h'(x) \ge 0$. If $A \ge e^2/3.1 \approx 2.38357$ and $0 < x \le (\ln B)^{-2}/B$ for $B = 3.1A$, then $h(x) \le 1/A$. By using this fact, it suffices to require $0 < \tau \le (\ln B_1)^{-2}/B_1$, where $B_1 = (3.1/99.8)(\sqrt{2E_0}+0.19676E_0)^4$.

Case (b) ($E_0 < 4.57916$). In this case we have $99.8(\sqrt{2E_0}+0.19676E_0)^{-4} \ge 0.418$. It suffices to require $0 < \tau \le 0.04$ since $\sup_{0<\tau\le 0.04}\tau(\ln\tau)^2 < 0.41445 < 0.418$. Recall that we are in the sub-case $0 < \tau < (2\pi)^{-4} < 0.04$, so (4.7) is satisfied.

□

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

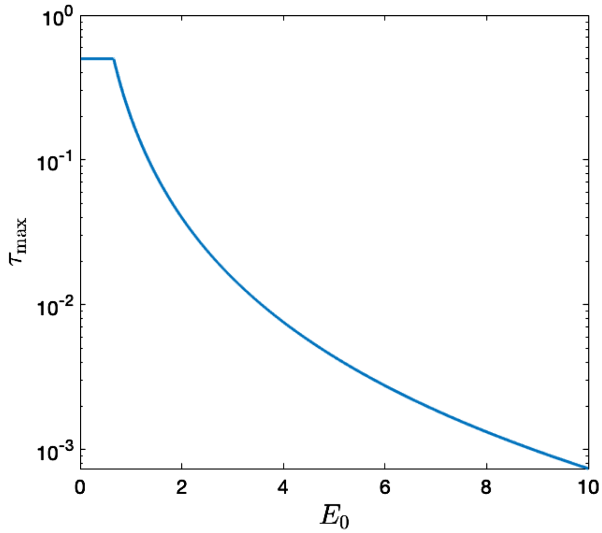

FIGURE 2. $\tau_{\max}$ with respect to E_0 given by Theorem 4.1 (in the case of $\nu = 1$)

Remark 4.1. Note that Theorem 4.1 provides explicit relationship between E_0 and $\tau_{\max}$, as illustrated in Figure 2. It can be observed that in the case of $\nu = 1$, when $E_0 = 8$, the energy dissipation can be ensured theoretically if $\tau < \tau_{\max} \approx 0.001$.

5. Error estimates

In this section, we carry out the L^2-error estimates for the Cahn-Hilliard equation.

5.1. L^2 **error estimate for an auxiliary system.** We consider the following auxiliary system on $\mathbb{T}^2$.

$$\begin{cases} \dfrac{v^{n+1} - v^n}{\tau} = -\nu\Delta^2 v^{n+1} + \Delta\Big(f(v^n)\Big) + \Delta\tilde{G}_n^1, \quad n \geq 0, \\ \dfrac{\tilde{v}^{n+1} - \tilde{v}^n}{\tau} = -\nu\Delta^2 \tilde{v}^{n+1} + \Delta\Big(f(\tilde{v}^n)\Big) + \Delta\tilde{G}_2^n, \qquad n \geq 0, \\ v^0 = v_0, \quad \tilde{v}^0 = \tilde{v}_0, \end{cases} \tag{5.1}$$

where v_0 and $\tilde{v}_0$ have mean zero, and $f(z) = z^3 - z$. Denote $\tilde{G}^n = \tilde{G}_1^n - \tilde{G}_2^n$.

Proposition 5.1. *Consider* (5.1) *on* $\mathbb{T}^2$. *Assume for some* $N_1 > 0$, $N_2 > 0$,

$$\sup_{n\geq 0} \|\tilde{v}^n\|_\infty \leq N_1, \quad \sup_{n\geq 0} \|\nabla v^n\|_2 \leq N_2, \quad \sup_{n\geq 0} \|\nabla \tilde{v}^n\|_2 \leq N_2. \tag{5.2}$$

Then for any $m \geq 1$,

$$\|v^m - \tilde{v}^m\|_2^2 \leq \exp\Big(m\tau \cdot \frac{C_1 \cdot (1 + N_1^4 + N_2^4)}{\nu}\Big) \cdot \Big(\|v_0 - \tilde{v}_0\|_2^2 + \frac{4\tau}{\nu} \sum_{n=0}^{m-1} \|\tilde{G}^n\|_2^2\Big),$$

where $C_1 > 0$ *is an absolute constant.*

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Proof. Denote $e^n = v^n - \tilde{v}^n$. Then

$$\frac{e^{n+1} - e^n}{\tau} = -\nu\Delta^2 e^{n+1} + \Delta(f(v^n) - f(\tilde{v}^n)) + \Delta\tilde{G}^n. \tag{5.3}$$

Taking the L^2 inner product with e^{n+1} on both sides, we get

$$\begin{aligned} &\frac{1}{2\tau}(\|e^{n+1}\|_2^2 - \|e^n\|_2^2 + \|e^{n+1} - e^n\|_2^2) + \nu\|\Delta e^{n+1}\|_2^2 \\ &= (\tilde{G}^n, \Delta e^{n+1}) + (f(v^n) - f(\tilde{v}^n), \Delta e^{n+1}). \end{aligned} \tag{5.4}$$

Obviously

$$|(\tilde{G}^n, \Delta e^{n+1})| \le \frac{2\|\tilde{G}^n\|_2^2}{\nu} + \frac{\nu}{8}\|\Delta e^{n+1}\|_2^2.$$

On the other hand, recalling $f(z) = z^3 - z$, we get

$$f(v^n) - f(\tilde{v}^n) = (3(\tilde{v}^2)^2 + 1)e^n + 3\tilde{v}^n(e^n)^2 + (e^n)^3.$$

We now estimate the contribution of each term. In the rest of this proof, to ease the notation, we shall denote by C an absolute constant whose value may change from line to line. Clearly

$$\begin{aligned} &|((3(\tilde{v}^n)^2 + 1)e^n, \Delta e^{n+1})| \\ &\le C \cdot (1 + \|\tilde{v}^n\|_\infty^2)\|e^n\|_2 \cdot \|\Delta e^{n+1}\|_2 \le \frac{C}{\nu}(1 + N_1^4)\|e^n\|_2^2 + \frac{\nu}{8}\|\Delta e^{n+1}\|_2^2. \end{aligned} \tag{5.5}$$

By using the interpolation inequality $\|f\|_4 \lesssim \|f\|_2^{\frac{1}{2}}\|\nabla f\|_2^{\frac{1}{2}}$, we get

$$\begin{aligned} &|(3\tilde{v}^n(e^n)^2, \Delta e^{n+1})| \\ &\le C\|\tilde{v}^n\|_\infty \cdot \|e^n\|_4^2 \cdot \|\Delta e^{n+1}\|_2 \le C \cdot \frac{N_1^2 \cdot \|e^n\|_4^4}{\nu} + \frac{\nu}{8}\|\Delta e^{n+1}\|_2^2 \\ &\le C \cdot \frac{N_1^2}{\nu}\|\nabla e^n\|_2^2\|e^n\|_2^2 + \frac{\nu}{8}\|\Delta e^{n+1}\|_2^2 \le \frac{C}{\nu}N_1^2N_2^2\|e^n\|_2^2 + \frac{\nu}{8}\|\Delta e^{n+1}\|_2^2. \end{aligned} \tag{5.6}$$

Similarly,

$$\begin{aligned} \left|((e^n)^3, \Delta e^{n+1})\right| &\le \frac{C}{\nu}\|e^n\|_6^6 + \frac{\nu}{8}\|\Delta e^{n+1}\|_2^2 \\ &\le \frac{C}{\nu}\|e^n\|_2^2\|\nabla e^n\|_2^4 + \frac{\nu}{8}\|\Delta e^{n+1}\|_2^2 \le C\frac{N_2^4}{\nu}\|e^n\|_2^2 + \frac{\nu}{8}\|\Delta e^{n+1}\|_2^2. \end{aligned} \tag{5.7}$$

Collecting the estimates, we get

$$\frac{\|e^{n+1}\|_2^2 - \|e^n\|_2^2}{\tau} \le \frac{4}{\nu}\|\tilde{G}^n\|_2^2 + \frac{C}{\nu}(1 + N_1^4 + N_2^4)\|e^n\|_2^2. \tag{5.8}$$

Define $y_n = \|e^n\|_2^2, \tilde{\alpha} = C(1 + N_1^4 + N_2^4)/\nu$ and $\tilde{\beta}_n = 4\|\tilde{G}^n\|_2^2/\nu$. Then obviously

$$\frac{y_{n+1} - y_n}{\tau} \le \tilde{\alpha}y_n + \tilde{\beta}_n.$$

The desired result then follows from a standard discrete Gronwall estimate. □

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

5.2. L^2 **error estimate for CH (proof of Theorem 1.2).** In this proof, to simplify the notation we shall denote by C a constant depending only on (ν, u_0). The value of C may vary from line to line. For any two quantities X and Y, we shall write $X \lesssim Y$ if $X \le CY$.

We need to consider the CH equation (1.8) with $u(0) = u_0$ and the semi-implicit scheme (1.9) with $\tilde{u}^0 = u_0$.

We first rewrite the PDE solution u in the discretized form. Note that for a one-variable function $h = h(t)$, we have the formula

$$\frac{1}{\tau}\int_{t_n}^{t_{n+1}} h(t)dt = h(t_n) + \frac{1}{\tau}\int_{t_n}^{t_{n+1}} h'(t)\cdot(t_{n+1}-t)dt, \tag{5.9a}$$

$$\frac{1}{\tau}\int_{t_n}^{t_{n+1}} h(t)dt = h(t_{n+1}) + \frac{1}{\tau}\int_{t_n}^{t_{n+1}} h'(t)\cdot(t_n-t)dt. \tag{5.9b}$$

By using the CH equation (1.8) and integrating the PDE for u on the time interval $[t_n, t_{n+1}]$, we get

$$\frac{u(t_{n+1})-u(t_n)}{\tau} = -\nu\Delta^2 u(t_{n+1}) + \Delta\Big(f(u(t_n))\Big) + \Delta\tilde{G}^n, \tag{5.10}$$

where

$$\tilde{G}^n = -\frac{\nu}{\tau}\int_{t_n}^{t_{n+1}} \partial_t\Delta u\cdot(t_n-t)dt + \frac{1}{\tau}\int_{t_n}^{t_{n+1}} \partial_t(f(u))\cdot(t_{n+1}-t)dt. \tag{5.11}$$

Now

$$\|\tilde{G}^n\|_2 \le \nu\int_{t_n}^{t_{n+1}} \|\partial_t\Delta u\|_2 dt + \int_{t_n}^{t_{n+1}} \|\partial_t u\|_2 dt(\|f'(u)\|_{L_t^\infty L_x^\infty}). \tag{5.12}$$

For the PDE solution we have $\|u\|_\infty \lesssim 1$. Since $\|\partial_t u\|_2 \lesssim \|\Delta\partial_t u\|_2$ (recall $\partial_t u$ has mean zero), we get

$$\|\tilde{G}^n\|_2 \lesssim \int_{t_n}^{t_{n+1}} \|\partial_t\Delta u\|_2 dt \lesssim \left(\int_{t_n}^{t_{n+1}} \|\partial_t\Delta u\|_2^2 dt\right)^{\frac{1}{2}}\cdot\sqrt{\tau}.$$

Therefore

$$\sum_{n=0}^{m-1}\|\tilde{G}^n\|_2^2 \lesssim \tau\int_0^{t_m}\|\partial_t\Delta u\|_2^2 dt \lesssim \tau\cdot(1+t_m).$$

It is easy to check that $\sup_{t\ge0}\|u(t)\|_{H^s} \lesssim 1$ and $\sup_{t\ge0}\|f(u(t))\|_{H^s} \lesssim 1$. Clearly,

$$\tau\sum_{n=0}^{m-1}\left(\|\tilde{G}^n\|_2^2\right) \lesssim (1+t_m)\tau^2. \tag{5.13}$$

By Theorem 3.1 and discrete smoothing estimates (cf. [18]), we have $\sup_{n\ge0}\|\nabla u^n\|_2 \lesssim 1$. By Proposition 5.1, we get

$$\|u^m - u(t_m)\|_2^2 \lesssim_s e^{Ct_m}(1+t_m)\tau^2.$$

This implies (1.15).

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

6. Concluding remarks

Implicit-explicit (IMEX) methods can simulate efficiently many phase field models such as the Cahn-Hilliard equation or thin film type equations. Compared with pure explicit methods, IMEX is more stable with larger allowable time steps whilst being efficient and accurate. In contrast with implicit methods and partially implicit methods, IMEX does not require solving a nonlinear system at each time step and is much more efficient. We are trying to balance between stability and accuracy, and look for some sort of conditional stability with affordable time step constraints, and more importantly without sacrificing accuracy too much. In numerical experiments, IMEX is often observed to be energy stable provided the time step is not taken too large.

Due to the difficulties caused by the lack of maximum principle and stiffness caused by the effect of small viscosity, the rigorous stability and convergence analysis of IMEX methods was a long standing open problem. In this work, we analyzed a model IMEX scheme successfully employed by Chen and Shen [3] for the Cahn-Hilliard equation and gave a first rigorous proof of *conditional* energy stability and rate of convergence with mild time step constraints. Our analysis does not rely on adding additional stabilization terms, truncating the nonlinearity or introducing auxiliary variables. To deal with the aforementioned difficulties caused by the lack of maximum principle and stiffness of small viscosity, we introduce a Trade-Energy-For-L^∞ (TEFL) method which is a refinement of our earlier work [15–17]. In the course of the proof we computed explicitly time step constraints in several model cases which seem to be the first done in the literature. All these developments are pivotal for future refined analysis on these algorithms.

It is expected that this new streamlined TEFL proof can be further refined and adapted to higher order cases and generalized to many other phase field models and settings.

References

[1] A. Baskaran, J. S. Lowengrub, C. Wang, and S. M. Wise, *Convergence analysis of a second order convex splitting scheme for the modified phase field crystal equation*, SIAM J. Numer. Anal. **51** (2013), no. 5, 2851–2873, DOI 10.1137/120880677. MR3118257

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

[3] L. Q. Chen and J. Shen, *Applications of semi-implicit Fourier-spectral method to phase field equations*, Comput. Phys. Commun. **108** (1998), 147–158.

[4] A. Christlieb, J. Jones, K. Promislow, B. Wetton, and M. Willoughby, *High accuracy solutions to energy gradient flows from material science models.* part A, J. Comput. Phys. **257** (2014), no. part A, 193–215, DOI 10.1016/j.jcp.2013.09.049. MR3129531

[5] C. M. Elliott, *The Cahn-Hilliard model for the kinetics of phase separation*, Mathematical models for phase change problems (Óbidos, 1988), Internat. Ser. Numer. Math., vol. 88, Birkhäuser, Basel, 1989, pp. 35–73. MR1038064

[6] C. M. Elliott and A. M. Stuart, *The global dynamics of discrete semilinear parabolic equations*, SIAM J. Numer. Anal. **30** (1993), no. 6, 1622–1663, DOI 10.1137/0730084. MR1249036

[7] D. J. Eyre, *Unconditionally gradient stable time marching the Cahn-Hilliard equation*, Computational and mathematical models of microstructural evolution (San Francisco, CA, 1998), Mater. Res. Soc. Sympos. Proc., vol. 529, MRS, Warrendale, PA, 1998, pp. 39–46, DOI 10.1557/PROC-529-39. MR1676409

[8] X. Feng, T. Tang, and J. Yang, *Long time numerical simulations for phase-field problems using p-adaptive spectral deferred correction methods*, SIAM J. Sci. Comput. **37** (2015), no. 1, A271–A294, DOI 10.1137/130928662. MR3304267

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

[9] H. Gomez and T. J. R. Hughes, *Provably unconditionally stable, second-order time-accurate, mixed variational methods for phase-field models*, J. Comput. Phys. **230** (2011), no. 13, 5310–5327, DOI 10.1016/j.jcp.2011.03.033. MR2799512

[10] Z. Guan, C. Wang, and S. M. Wise, *A convergent convex splitting scheme for the periodic nonlocal Cahn-Hilliard equation*, Numer. Math. **128** (2014), no. 2, 377–406, DOI 10.1007/s00211-014-0608-2. MR3259494

[11] J. Guo, C. Wang, S. M. Wise, and X. Yue, *An H^2 convergence of a second-order convex-splitting, finite difference scheme for the three-dimensional Cahn-Hilliard equation*, Commun. Math. Sci. **14** (2016), no. 2, 489–515, DOI 10.4310/CMS.2016.v14.n2.a8. MR3436249

[12] Y. He, Y. Liu, and T. Tang, *On large time-stepping methods for the Cahn-Hilliard equation*, Appl. Numer. Math. **57** (2007), no. 5-7, 616–628, DOI 10.1016/j.apnum.2006.07.026. MR2322435

[13] F. Liu and J. Shen, *Stabilized semi-implicit spectral deferred correction methods for Allen-Cahn and Cahn-Hilliard equations*, Math. Methods Appl. Sci. **38** (2015), no. 18, 4564–4575, DOI 10.1002/mma.2869. MR3449617

[14] D. Li, Z. Qiao, and T. Tang, *Characterizing the stabilization size for semi-implicit Fourier-spectral method to phase field equations*, SIAM J. Numer. Anal. **54** (2016), no. 3, 1653–1681, DOI 10.1137/140993193. MR3507555

[15] D. Li, Z. Qiao, and T. Tang, *Gradient bounds for a thin film epitaxy equation*, J. Differential Equations **262** (2017), no. 3, 1720–1746, DOI 10.1016/j.jde.2016.10.025. MR3582210

[16] D. Li and Z. Qiao, *On second order semi-implicit Fourier spectral methods for 2D Cahn-Hilliard equations*, J. Sci. Comput. **70** (2017), no. 1, 301–341, DOI 10.1007/s10915-016-0251-4. MR3592143

[17] D. Li and Z. Qiao, *On the stabilization size of semi-implicit Fourier-spectral methods for 3D Cahn-Hilliard equations*, Commun. Math. Sci. **15** (2017), no. 6, 1489–1506, DOI 10.4310/CMS.2017.v15.n6.a1. MR3668944

[18] D. Li and T. Tang, *Stability of the semi-implicit method for the Cahn-Hilliard equation with logarithmic potentials*, Ann. Appl. Math. **37** (2021), no. 1, 31–60, DOI 10.4208/aam.OA-2020-0003. MR4284064

[19] D. Li, *Effective maximum principles for spectral methods*, Ann. Appl. Math. **37** (2021), no. 2, 131–290, DOI 10.4208/aam.oa-2021-0003. MR4294331

[20] C.-B. Schönlieb and A. Bertozzi, *Unconditionally stable schemes for higher order inpainting*, Commun. Math. Sci. **9** (2011), no. 2, 413–457. MR2815678

[21] J. Shen and X. Yang, *Numerical approximations of Allen-Cahn and Cahn-Hilliard equations*, Discrete Contin. Dyn. Syst. **28** (2010), no. 4, 1669–1691, DOI 10.3934/dcds.2010.28.1669. MR2679727

[22] J. Shen, J. Xu, and J. Yang, *A new class of efficient and robust energy stable schemes for gradient flows*, SIAM Rev. **61** (2019), no. 3, 474–506, DOI 10.1137/17M1150153. MR3989239

[23] H. Song and C.-W. Shu, *Unconditional energy stability analysis of a second order implicit-explicit local discontinuous Galerkin method for the Cahn-Hilliard equation*, J. Sci. Comput. **73** (2017), no. 2-3, 1178–1203, DOI 10.1007/s10915-017-0497-5. MR3719623

[24] C. Xu and T. Tang, *Stability analysis of large time-stepping methods for epitaxial growth models*, SIAM J. Numer. Anal. **44** (2006), no. 4, 1759–1779, DOI 10.1137/050628143. MR2257126

[25] J. Xu, Y. Li, S. Wu, and A. Bousquet, *On the stability and accuracy of partially and fully implicit schemes for phase field modeling*, Comput. Methods Appl. Mech. Engrg. **345** (2019), 826–853, DOI 10.1016/j.cma.2018.09.017. MR3892022

[26] J. Zhu, L.-Q. Chen, J. Shen, and V. Tikare, *Coarsening kinetics from a variable-mobility Cahn-Hilliard equation: application of a semi-implicit Fourier spectral method*, Phys. Rev. E (3) **60** (1999), 3564–3572.

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

SUSTech International Center for Mathematics; and Department of Mathematics, Southern University of Science and Technology, Shenzhen 518055, People's Republic of China

Email address: lid@sustech.edu.cn

SUSTech International Center for Mathematics, Southern University of Science and Technology, Shenzhen 518055, People's Republic of China

Email address: quancy@sustech.edu.cn

SUSTech International Center for Mathematics, Shenzhen, China; and Division of Science and Technology, BNU-HKBU United International College, Zhuhai, Guangdong Province, People's Republic of China

Email address: tangt@sustech.edu.cn

Licensed to South Univ of Sci & Tech. Prepared on Wed Mar 16 08:47:13 EDT 2022 for download from IP 116.7.234.241.
License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use