GENERALIZING THE CAHN-HILLIARD EQUATION WITH APPLICATIONS IN MIGRATION MODELING

By

ZACHARY JAMES HILLIARD

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
Department of Mathematics and Statistics

MAY 2020

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of ZACHARY JAMES HILLIARD find it satisfactory and recommend that it be accepted.

Lynn Schreyer, Ph.D., Chair

Nikolaos Voulgarakis, Ph.D.

Hong-Ming Yin, Ph.D.

Sergey Lapin, Ph.D.

Tiziana Giorgi, Ph.D.

ACKNOWLEDGMENTS

I would like to thank my entire committee. Your expertise and guidance has been invaluable. I am especially grateful to Dr. Schreyer as we have had many enlightening conversations and her constructive feedback has always pushed me to improve.

GENERALIZING THE CAHN-HILLIARD EQUATION WITH APPLICATIONS IN MIGRATION MODELING

Abstract

by Zachary James Hilliard, Ph.D.
Washington State University
May 2020

Chair: Lynn Schreyer

The Cahn-Hilliard Equation is a nonlinear parabolic partial differential equation that was originally developed to model phase separation of a two-phase solution. A generalization of this equation has been derived via hybrid mixture theory and phase field theory as a means to model migration by using fluid flow through a porous media as an analogue for a population migrating over a given terrain.

We examine this generalized equation both numerically and analytically. Analytically, we show that there are solutions to the generalized equation, which includes anisotropy and a forcing function. Included in the numerical analysis is a novel method of time-stepping that reduces the stiffness of the problem by making a change of variables in the continuous setting.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

We study a generalization of the Cahn-Hilliard equation. This generalization is a model for migration that is derived by making an analogy between migration of some population over a given terrain and fluid flow through a porous medium and then using the frameworks of hybrid mixture theory and phase field theory to derive governing equations [36]. In this analogue, fluid density corresponds to population density and the conductivity of the porous medium corresponds to the trafficability of the terrain (i.e., the difficulty of traversing the terrain). Before we introduce this model, let us familiarize ourselves with two variants of the classical Cahn-Hilliard equation.

The Cahn-Hilliard equation was originally developed as a model for the phase decomposition of a binary solution and is a gradient flow of the Ginzburg-Landau free energy functional

$$e[u] = \int_\Omega \left(F(u) + \frac{\alpha}{2} |\nabla u|^2 \right) \tag{1.1}$$

where u is an order parameter that denotes how much of each phase is present at a given point in the domain of interest $\Omega \subset \mathbb{R}^d$, F is a bulk energy function that penalizes the mixing of the two phases, and $\alpha > 0$ controls the interfacial energy between regions of pure phases [7, 27]. Generally F takes the form of a "double well" with minima at $u = \pm 1$, which represent the two pure phases. The fact that the

Cahn-Hilliard equation (and its cousin, the Allen-Cahn equation) is a gradient flow on (1.1) governs the dynamics of the evolution of solutions. Having F take the form of a double well makes these dynamics interesting and physically relevant, but this will be discussed in detail later on. The equation $F(u) = \frac{1}{4}(u^2 - 1)^2$ is particularly common, but other forms in which F' is singular at $u = \pm 1$ are also possible and more physically relevant in the particular application of phase decomposition [2, 10, 16, 23]. The standard Cahn-Hill1iard equation is most often posed as a coupled system of PDEs

$$\begin{cases} u_t = \nabla \cdot b(u)\nabla w \\ \\ w = F'(u) - \alpha \Delta u \end{cases} \tag{1.2}$$

with no-flux boundary conditions of the form $\partial_{\boldsymbol{\nu}} u = \partial_{\boldsymbol{\nu}} w = 0$ being the most common. Here $\boldsymbol{\nu}$ is the outward unit normal vector on $\partial\Omega$. With these boundary conditions, (1.2) is a descent of the energy functional (1.1) under the constant mass constraint $\frac{d}{dt}\int_\Omega u = 0$. The function b is referred to as the mobility. To prove existence of weak solutions, b is often assumed to be non-degenerate (i.e., $0 < b_1 \leq b(u) \leq b_2$ for some constants b_2 and b_1) or constant. However, under certain assumptions on b and F, weak solutions have been shown to exist when only requiring $0 \leq b(u)$ [12, 17].

To see that the energy given by (1.1) decreases for solutions of (1.2), assume that u and w are sufficiently smooth and solve (1.2) with the given no-flux boundary

conditions. Differentiating (1.1) yields

$$\begin{aligned}\frac{d}{dt}e[u] &= \int_\Omega \left(F'(u)u_t + \alpha\nabla u\cdot\nabla u_t\right)\\ &= \int_\Omega \left(F'(u) - \alpha\Delta u\right)u_t\\ &= \int_\Omega w\nabla\cdot b(u)\nabla w\\ &= -\int_\Omega b(u)\left|\nabla w\right|^2 \le 0.\end{aligned}$$

We see that $\frac{d}{dt}e[u(t)] = 0$ only at steady state, but may be arbitrarily small. This is our first indication that solutions may tend towards "pseudo steady states" in which u is essentially constant over a long duration before undergoing a rapid evolution. This behavior can pose some challenges numerically when discretizing the time variable.

When we eventually solve the Cahn-Hilliard equation numerically, it will be important (as for solving any PDE numerically) to use a stable method. Because the Cahn-Hilliard equation is nonlinear and a descent of the Ginzburg-Landau energy functional (1.1), the natural way to classify stability is by requiring that the energy of the numerical solution is also decreasing. When solving (1.2) directly, often a severe restriction on the size of the time-step to ensure this stability, which we will see in Chapter 4. To overcome this, one can introduce viscous relaxation [37]. This relaxation amounts to solving

$$\begin{cases} u_t = \nabla\cdot b(u)\nabla w \\ \\ w = F'(u) - \alpha\Delta u + \theta u_t \end{cases} \tag{1.3}$$

for some positive $\theta = O(\delta t)$. Equation (1.3) is known as the viscous Cahn-Hilliard equation. To see why the introduction of θ makes energy stability more feasible, let us assume that u and w form a smooth solution to (1.3) with no-flux boundary conditions $\partial_{\boldsymbol{\nu}} u = \partial_{\boldsymbol{\nu}} w = 0$. Differentiating (1.1) now yields

$$\begin{aligned}
\frac{d}{dt} e[u] &= \int_\Omega \left(F'(u) u_t + \alpha \nabla u \cdot \nabla u_t \right) \\
&= \int_\Omega \left(F'(u) - \alpha \Delta u \right) u_t \\
&= \int_\Omega \left(F'(u) - \alpha \Delta u + \theta u_t \right) u_t - \theta \int_\Omega |u_t|^2 \\
&= \int_\Omega w \nabla \cdot b(u) \nabla w - \theta \left\| u_t \right\|^2_{L^2(\Omega)} \\
&= - \int_\Omega b(u) \left| \nabla w \right|^2 - \theta \left\| u_t \right\|^2_{L^2(\Omega)} \leq 0
\end{aligned}$$

so that $e(u)$ decreases for solutions of (1.3) faster than for solutions of (1.2).

We will continue our discussion of the Cahn-Hilliard equation as follows. In Chapter 2, we introduce the migration model and informally discuss the effects that various parameters have on solutions to the model before establishing the existence of solutions to the model in Chapter 3. We will then move on to some numerical techniques and analysis in Chapter 4 before making some concluding remarks in Chapter 5.

CHAPTER 2. THE MIGRATION MODEL

Now let us turn our attention to the migration model [36]. The governing equation is given by the system of PDEs

$$\begin{cases} u_t = \nabla \cdot b(u)\mathbf{K}\nabla w \\ \\ w = f(u) - \nabla \cdot \mathbf{A}\nabla u + \phi \end{cases} \tag{2.1}$$

where $\mathbf{K}$ and $\mathbf{A}$ are symmetric, positive-definite matrices that may depend on certain parameters related to the species migrating and the terrain, but not on u or w. The function ϕ is given and may depend on x or possibly also on t, and $f = F'$ is the derivative of a double well energy function. We will also be interested in the viscous relaxation of this governing equation, which is given by

$$\begin{cases} u_t = \nabla \cdot b(u)\mathbf{K}\nabla w \\ \\ w = f(u) - \nabla \cdot \mathbf{A}\nabla u + \phi + \theta u_t \end{cases} \tag{2.2}$$

for $\theta > 0$. In either case, we will be looking for solutions over some bounded domain $\Omega \subset \mathbb{R}^d$ with no-flux boundary conditions

$$\partial_{\mathbf{A}\boldsymbol{\nu}} u = \partial_{\mathbf{K}\boldsymbol{\nu}} w = 0 \tag{2.3}$$

on $\partial\Omega$, which yield the constant mass constraint $\frac{d}{dt}\int_\Omega u = 0$.

We now look at an alternate energy functional given by

$$E[u] = \int_\Omega \left(F(u) + \frac{1}{2}\nabla u \cdot \mathbf{A}\nabla u + u\phi \right) \tag{2.4}$$

and note that both (2.1) and (2.2) are descents of this functional. Assume that $\mathbf{A}$ and ϕ are independent of t and u and w are sufficiently smooth. If u and w solve (2.1), we follow the same steps as in Chapter 1 to obtain

$$\frac{d}{dt}E[u] = -\int_\Omega b(u)\nabla w \cdot \mathbf{K}\nabla w \leq 0,$$

while on the other hand obtaining

$$\frac{d}{dt}E[u] = -\int_\Omega b(u)\nabla w \cdot \mathbf{K}\nabla w - \theta\int_\Omega |u_t|^2 \leq 0$$

if u and w solve (2.2) instead. Again, we see that $\frac{d}{dt}E[u(t)] = 0$ only at steady state but may be arbitrarily small.

The parameters above can be interpreted both in terms of migration and in terms of flow through a porous medium via equations derived using hybrid mixture theory. First, we look at flow through a porous medium. Here u is the density of a fluid flowing through a porous media Ω, $b(u)\mathbf{K}$ is the conductivity of the porous medium, w is the pressure, F is the bulk energy as a function of density, $\mathbf{A}$ governs the interfacial energy, and ϕ is a gravity potential. Now if we look at the migration context, we have u representing the density (count per area) of some population, which here we consider being a herd of ungulates (hoofed mammals) such as zebras or elk, $\mathbf{K}$ is the trafficability of the terrain over which the migration takes place, $b(u)$ relates to how efficiently a given population can move, F relates to the density that the species prefers, $\mathbf{A}$ represents the external threat a herd perceives and controls the shape of

the herd, and ϕ represents a danger potential driving the movement of the herd in a particular direction. The extra variable w does not have a direct analogue in the migration setting.

Now that we have a statement of the model, let us take a closer look as to how each of the parameters affect the solution u. We will do this by first examining how F, $\mathbf{A}$, and ϕ affect minimizers of (2.4) (steady state solutions), and then how $b(u)$ and $\mathbf{K}$ affect the evolution of some initial condition towards steady state. We then show how we can construct $\mathbf{K}$ and $\mathbf{A}$ from a given terrain.

2.1 The Double Well Function

Let us assume that a given herd has some natural density that is most preferable and let $u = 1$ represent this density (i.e., we consider u to be density normalized by this natural density). Then the states $u = 0$ and $u = 1$ should be the lowest energy states. Making the simplest possible choice, we set $F(u) \propto u^2(1-u)^2$. The particular choice of $F(u) = 5u^2(1-u)^2$ has given good initial results and was used when performing all simulations shown in this chapter.

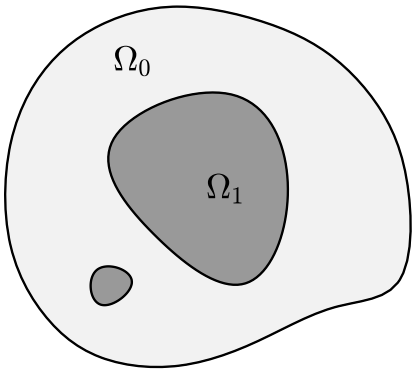

Figure 2.1: Effect of F

To see the effect of F on the minima of (2.4), choose some open $\Omega_1 \subset \Omega$ and set $\Omega_0 = \Omega \backslash \Omega_1$. Set $u = \chi_{\Omega_1}$, the characteristic function over Ω_1, and observe that this choice minimizes $\int_\Omega F(u)$. Note that if this function were to be reached as the result of some evolution of (2.1) or (2.2) (which is impossible as we must have $u \in H^1(\Omega)$ as will be seen in Chapter 3), Ω_1 would be required to satisfy $|\Omega_1| = \int_\Omega u|_{t=0}$ due to the constant mass constraint imposed by the boundary condition $\partial_{\mathbf{K}\boldsymbol{\nu}} w = 0$.

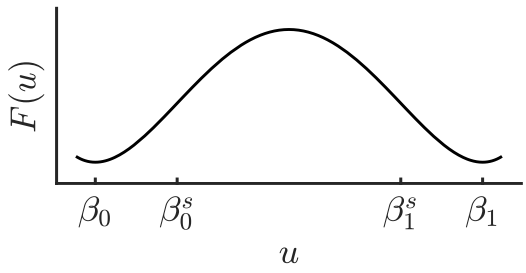

Figure 2.2: A generic double well function

In general, a double well function will have the form depicted in Figure 2.2 with two distinct minima at some β_0 and β_1 and inflection points at β_0^s and β_1^s. These points would be ordered $\beta_0 < \beta_0^s < \beta_1^s < \beta_1$ with the interval (β_0^s, β_1^s) on which F is concave known as the *spinodal* interval [15, 34]. As we will see, initial conditions for (1.2), (1.3), (2.1), and (2.2) that are contained in this spinodal interval are unstable and will undergo *spinodal decomposition* so that eventually the solution u will mostly take values $u < \beta_0^s$ or $\beta_1^s < u$ over Ω. As an indication that solutions with values only in the spinodal region are unstable, let us temporarily set $b(u)\mathbf{K} = \mathbf{I}$, $\mathbf{A} = 0$, and $\phi = 0$ in (2.1) to obtain

$$u_t = \Delta f(u).$$

If we approximate $f(u) \approx f(a) + f'(a)(u - a)$ for some constant $a \in \mathbb{R}$, we have the

heat equation

$$u_t = f'(a)\Delta u. \tag{2.5}$$

Note that if $a \in (\beta_0^s, \beta_1^s)$, then we have $f'(a) = F''(a) < 0$ so that we actually have the backwards heat equation, which is notoriously ill-posed. Alternatively, if $a < \beta_0^s$ or $\beta_1^s < a$, then our linearization would be the forward heat equation, which is well posed and generally has smooth solutions [20, 44].

2.2 Interface Energy

Now let us turn our attention to the effects of $\mathbf{A}$ on solutions to (2.1) and (2.2). We will always assume that $\mathbf{A}$ is symmetric and uniformly positive definite over $\overline{\Omega}$ and for now we assume that $\mathbf{A} = \alpha\mathbf{I}$ with $\alpha > 0$ constant. Clearly choosing u to be constant over Ω will minimize $\int_\Omega \frac{1}{2}\nabla u \cdot \mathbf{A}\nabla u$. If we respect the mass constraint, we would have to choose $u = m\chi_\Omega$, where $m = \frac{1}{|\Omega|}\int u$ is the average value of some initial condition. We will be primarily interested in the case where α is small and $\beta_0^s < m < \beta_1^s$. This is because if α is too large, then $u = m\chi_\Omega$ is a global minimizer of $\int_\Omega \left(F(u) + \frac{\alpha}{2}|\nabla u|^2\right)$, and if $F''(m) > 0$, then $u = m\chi_\Omega$ is at least a local minimizer. These are formally stated below, but the corresponding proofs are delayed until Section 2.6.

Theorem 2.1. *Let $F \in C^2(\mathbb{R};\mathbb{R})$ with $|F''(s)| \leq L$ for some $L > 0$ and all $s \in \mathbb{R}$. Let $\mathbf{A}\colon \Omega \mapsto \mathbb{R}^{d\times d}$ be uniformly positive definite on an open, connected domain*

$\Omega \subset \mathbb{R}^d$ with a Lipschitz boundary. Fix some $m \in \mathbb{R}$ and consider the affine space $U = \{u \in H^1(\Omega) \colon \frac{1}{|\Omega|}\int_\Omega u = m\}$. If $F''(m) > 0$, then $u^* = m\chi_\Omega$ is a local minimizer of the energy functional

$$E[u] = \int_\Omega \left(F(u) + \frac{1}{2}\nabla u \cdot \mathbf{A}\nabla u \right)$$

over U.

Theorem 2.2. *In Theorem 2.1, we took $\mathbf{A}$ to be uniformly positive definite. Suppose $A_1 |\xi|^2 \leq \xi \cdot \mathbf{A}(x)\xi \leq A_2 |\xi|^2$ for some A_1 and $A_2 > 0$ and all $x \in \Omega$ and $\xi \in \mathbb{R}^d$. There exists some C depending only on Ω and L such that if $A_1 > C$, then u^* is a global minimizer of E over U.*

Under the assumptions that α is small and $\beta_0^s < m < \beta_1^s$, we expect minimizers of $\int_\Omega \left(F(u) + \frac{1}{2}\nabla u \cdot \mathbf{A}\nabla u\right)$ will be quite similar to the minimizers of $\int_\Omega F(u)$ except that we require $\int_\Omega |\nabla u|^2 < \infty$ so that characteristic functions $u = \chi_{\Omega_1}$ with $\Omega_1 \subsetneq \Omega$ are not generally permissible. Due to this, let us take one further step and assume that u is continuous. This restriction causes the appearance of a transition or interface region $\Omega_{1/2}$ between Ω_0 and Ω_1. To be precise, let us define

$$\Omega_0 = \{x \in \Omega \colon u(x) < \beta_0^s\}$$

$$\Omega_{1/2} = \{x \in \Omega \colon \beta_0^s \leq u(x) \leq \beta_1^s\}$$

$$\Omega_1 = \{x \in \Omega \colon \beta_1^s < u(x)\}$$

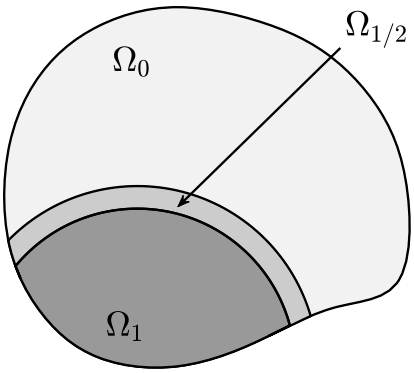

Figure 2.3: Effect of $\mathbf{A}$

and observe that these definitions are consistent with our discussion in Section 2.1. To see that these regions develop, consider the dynamic case for an initial condition $g \approx m\chi_\Omega$. Because $f'(m) < 0$, the evolution will start off similar to the backwards heat equation, and we will see a rapid movement of the values of u towards β_0^s and β_1^s. Once the values of u pass either of these values, we will have $f'(u) > 0$ on $\Omega_0 \cup \Omega_1$ and have local behavior similar to the forward heat equation so that solutions will locally approach some constant value. In order to minimize $\int_\Omega \left(F(u) + \frac{1}{2}\nabla u \cdot \mathbf{A}\nabla u\right)$, we should expect $\left|\Omega_{1/2}\right|$ to be small in comparison to $|\Omega_1| + |\Omega_2|$ as well as being circular due to the following (informal) reasoning. First, u is nearly constant on Ω_0 and Ω_1 so that $|\nabla u|$ will be largest on $\Omega_{1/2}$ which, together with $F(u)$ being largest on $\Omega_{1/2}$, would imply that $\left|\Omega_{1/2}\right|$ should be small. In fact, if $\mathbf{A} = \alpha\mathbf{I}$, then $\Omega_{1/2}$ should have a width of $O(\sqrt{\alpha})$ [34]. Second, we can write our mass constraint in terms of Ω_0, $\Omega_{1/2}$, and Ω_1 as

$$m\,|\Omega| = \int_{\Omega_0} u + \int_{\Omega_{1/2}} u + \int_{\Omega_1} u \approx \beta_0\,|\Omega_0| + \frac{\beta_0^s + \beta_1^s}{2}\left|\Omega_{1/2}\right| + \beta_1\,|\Omega_1|$$

which can be re-arranged to get

$$(m - \beta_0)\,|\Omega_0| + (m - \frac{\beta_0^s + \beta_1^s}{2})\left|\Omega_{1/2}\right| + (m - \beta_1)\,|\Omega_1| \approx 0.$$

Assuming that $\left|\Omega_{1/2}\right|$ is small gives us a system of approximations

$$\begin{cases} |\Omega_1| + |\Omega_2| \approx |\Omega| \\ (m - \beta_0)\,|\Omega_0| + (m - \beta_1)\,|\Omega_1| \approx 0 \end{cases}$$

so that $|\Omega_0|$ and $|\Omega_1|$ are essentially fixed for a given Ω and F. In fact, if we solve the approximate equations above we get $|\Omega_0| \approx \frac{\beta_1 - m}{\beta_1 - \beta_0}\,|\Omega|$ and $|\Omega_1| \approx \frac{m - \beta_0}{\beta_1 - \beta_0}\,|\Omega|$. It is interesting to observe that these approximations are quite reasonable for $m = \beta_0$, $m = (\beta_0 + \beta_1)/2$, and $m = \beta_1$ in which case we have $|\Omega_1| \approx 0$, $|\Omega_1| \approx |\Omega|/2$, and $|\Omega_1| \approx |\Omega|$. From here, we consider the isoparametric inequality (i.e., that the surface of minimal surface area enclosing a fixed volume is a sphere [33]). We conjecture that if $|\Omega_1| < |\Omega_0|$, then $\Omega_1 = B \cap \Omega$ for some ball $B \subset \mathbb{R}^d$. Further, in an analogue to the monotonicity requirement when $\Omega \subset \mathbb{R}$ [8], it is likely that $B \cap \partial\Omega \neq \emptyset$. Going back to the isoparametric inequality, this would be analogous to only counting a portion of the surface when looking for a surface of minimal area. We should note here that there is no reason to believe that this holds for general $\mathbf{A}$. However, if the spectral radius $\rho(\mathbf{A})$ is small throughout Ω, while the spectrum $\sigma(\mathbf{A})$ with the corresponding eigenvectors are nearly constant on Ω, the conjecture is still reasonable with the alteration of B being elliptic with axes and radii governed by the eigenvalues and eigenvectors of $\mathbf{A}$.

2.3 Gravity/Danger Potential

The effect of ϕ on minimizers of

$$E[u] = \int_\Omega \left(F(u) + \frac{1}{2}\nabla u \cdot \mathbf{A}\nabla u + u\phi \right) \tag{2.4}$$

is quite simple. First, suppose $\mathbf{A}$ is constant over Ω and note that, *with the exception of boundary effects* discussed above, the value of $\int_\Omega \left(F(u) + \frac{1}{2}\nabla \cdot \mathbf{A}\nabla u \right)$ is invariant under translations of u so that the term $\int_\Omega u\phi$ is the only portion with explicit spacial components. Taking the liberty of speaking loosely, this means that the largest values of u (i.e., Ω_1) should be located where ϕ is small, especially if $\nabla\phi$ is constant or nearly so. In the context of migration, this means that the herd wants to move where ϕ is smallest, i.e., where the danger potential is minimum. If $\mathbf{A}$ is not constant, there may be competition between locating Ω_1 where ϕ is small and locating the interface $\Omega_{1/2}$ where $\rho(\mathbf{A})$ is small. If $|\nabla\phi|$ is large, there is no reason to believe that either Ω_1 or Ω_0 should be spherical or elliptic.

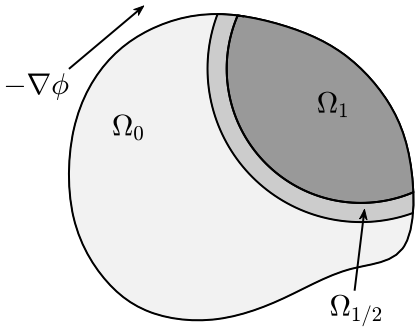

Figure 2.4: Effect of ϕ

2.4 Dynamic Behavior

Let us consider how the characterization the minimizers of $E[u]$ affect the dynamic behavior of (2.1) or (2.2) starting from some initial condition $g(x)$. To see the most interesting behavior, let us take $F(u) = (u-\beta_0)^2(u-\beta_1)^2$ and constrain $\left|g(x) - \frac{1}{2}(\beta_0+\beta_1)\right| \leq \varepsilon$ for some $0 < \varepsilon << \frac{1}{2}(\beta_1 - \beta_0)$. For simplicity, take $\mathbf{A} = \alpha(x)\mathbf{I}$ with $0 < \alpha(x) << 1$ and $\nabla\phi$ some nonzero constant vector with $|\nabla\phi|$ moderate (say $|\nabla\phi| \leq 1$).

First, we will see a *decomposition* period in which the regions Ω_0 and Ω_1 will develop (see Figures 2.5 and 2.6 for $0 \leq t \leq 0.01$). At this point, each set will be very disconnected and will be distributed over most of Ω. Next, we will see a *coarsening* period over which Ω_0 and Ω_1 will become more connected (see Figures 2.5 and 2.6 for $0.01 \leq t \leq 5$). Finally, we will see a *migration* period over which Ω_1 will be pushed in the $-\nabla\phi$ direction (see Figure 2.6 for $0.1 \leq t \leq 5$). Observe that u will nearly be constant on both Ω_0 and Ω_1 so that we will have $\nabla w \approx \nabla\phi$ and $u_t \approx \nabla \cdot b(u)\mathbf{K}\nabla\phi$, which is consistent with flow being pushed in the general direction of $-\nabla\phi$. The effects of $b(u)$, $\mathbf{K}$, and $\mathbf{A}$ are most apparent during this region. $\mathbf{K}$ and $\mathbf{A}$ govern local changes in u, which manifest as controlling the interface $\Omega_{1/2}$, while $b(u)$ controls how quickly the interface can travel in the general direction of $-\nabla\phi$. When $b(u)$ is taken to be degenerate, it is most often the case that $b(u) \approx 0$ on at least one of Ω_0 and Ω_1,

which can increase the required time for Ω_0 and Ω_1 to become connected.

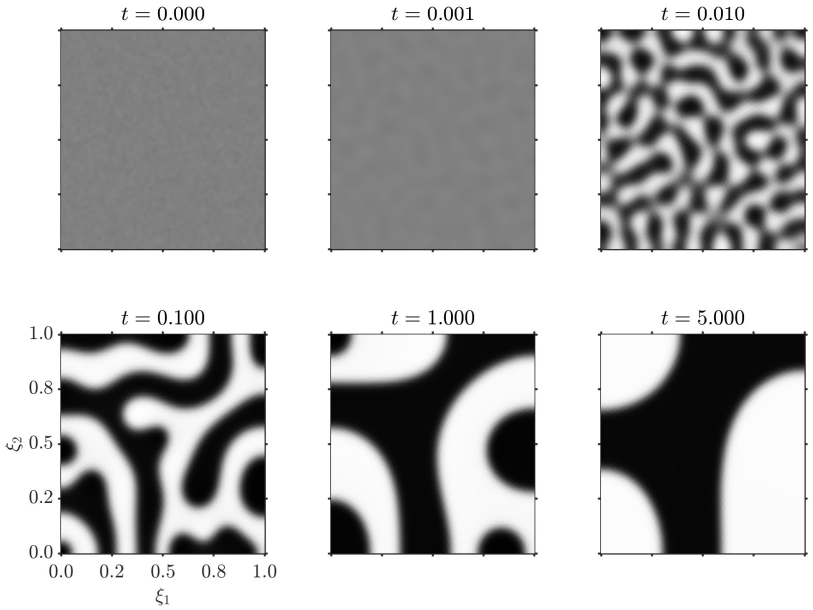

Figure 2.5: A solution with $\phi = 0$. At $t = 5$, the solution is near a pseudo steady state.

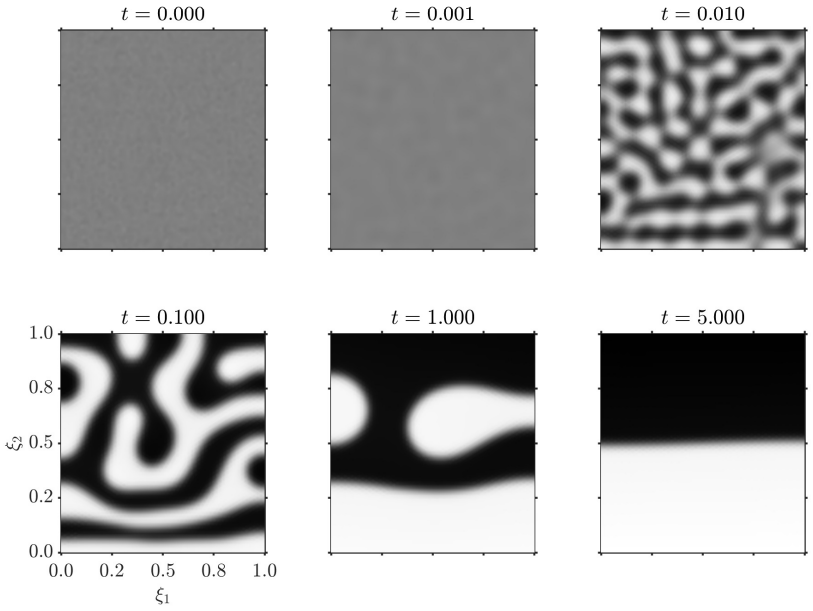

Figure 2.6: A solution with $\phi(\xi_1, \xi_2) = \xi_2$. At $t = 5$, the solution is near steady state.

2.5 Terrain Effects

Perhaps the most distinctive feature of this model is its ability to incorporate the effects of terrain. For now, we $\mathbf{K}$ and $\mathbf{A}$ to directly depend only on the terrain over which the migration takes place. We believe that elevation and surface type (i.e., forest, paved, plains, etc.) will be the two most important factors in these terms. Also assume that the elevation will determine the tensor portion of $\mathbf{K}$ as this will introduce preferable directions (e.g., directions in which the terrain is level) to travel that we use as eigenvectors of $\mathbf{K}$. This is the source of anisotropy in the model. We assume that the effect of surface type is isotropic so that we can capture this by a scalar multiple. We define $\mathbf{A}$ in terms of $\mathbf{K}$.

Suppose our elevation is given by some $z \in C^1(\overline{\Omega}; \mathbb{R})$ where $\Omega \subset \mathbb{R}^2$ for our purposes in this section. Choose some $x_0 \in \Omega$ and consider the level set $z(x) = z(x_0)$. It will be easiest to travel along this curve as opposed to perpendicular to the curve, especially if the slope is large. From this reasoning, we introduce a local change of coordinates given by $\xi_1' \propto \nabla z(x_0)$ perpendicular to the level set and ξ_2' tangent to the level set. Observe that we can take $\{\xi_1', \xi_2'\}$ to be an orthonormal basis for $\mathbb{R}^2$ (if

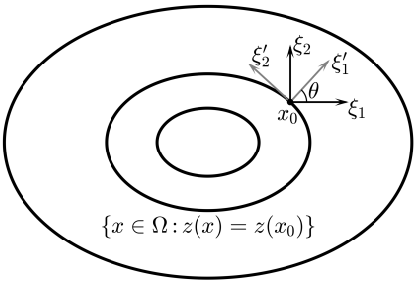

Figure 2.7: Change of coordinates

$\nabla z(x_0) = 0$, simply use the standard basis vectors). From here, we define $\mathbf{K}$ as

$$\mathbf{K} = \alpha_{\text{surf}} \mathbf{R}^{\mathsf{T}} \begin{bmatrix} \eta_1(|\nabla z|) & 0 \\ 0 & \eta_2(|\nabla z|) \end{bmatrix} \mathbf{R} \tag{2.6}$$

where $\mathbf{R} = [\xi_1' \ \xi_2']$ is the orthogonal rotation matrix from the standard coordinate system $\{\xi_1, \xi_2\}$ into the rotated system $\{\xi_1', \xi_2'\}$, α_{surf} is the surface conductivity, and η_1 and η_2 are functions which determine the effect of the steepness of the terrain on trafficability. Each η_i should be a decreasing function with $\eta_i(0) = 1$ so that if the terrain is flat, we recover $\mathbf{K} = \alpha_{\text{surf}}\mathbf{I}$. The choice $\eta_i(m) = \frac{1}{1+c_i m}$ with $c_1 = 10$ and $c_2 = 1$ or $c_2 = 0$ have given good results. Note that we should always have $\eta_1(m) \leq \eta_2(m)$ with equality only holding at $m = 0$. Moreover, if each η_i is a continuous positive function, then $\mathbf{K}$ will be uniformly positive definite over $\overline{\Omega}$, provided that $\alpha_{\text{surf}} > 0$ and $|\nabla z|$ are bounded on $\overline{\Omega}$.

The choice of $\mathbf{A} \propto \mathbf{K}^{-1}$ has given quite reasonable results. Observe that this choice gives us $\rho(\mathbf{A}) \propto \frac{1+c_1|\nabla z|}{\alpha_{\text{surf}}}$ so that the interface energy is highest when the terrain is steep and the surface is difficult to traverse. This has the effect of causing a herd to stay away from regions that are difficult to traverse. Treating $\mathbf{A} \propto \mathbf{K}^{-1}$ has a distinct advantage over simply setting $\mathbf{A} = \alpha\mathbf{I}$ for some constant $\alpha > 0$. This is due to the latter case only penalizing the speed at which the herd can move across difficult terrain, but does not give any motivation for the herd to stay away from that terrain. In the simulations below, the curves in blue are level sets of the terrain.

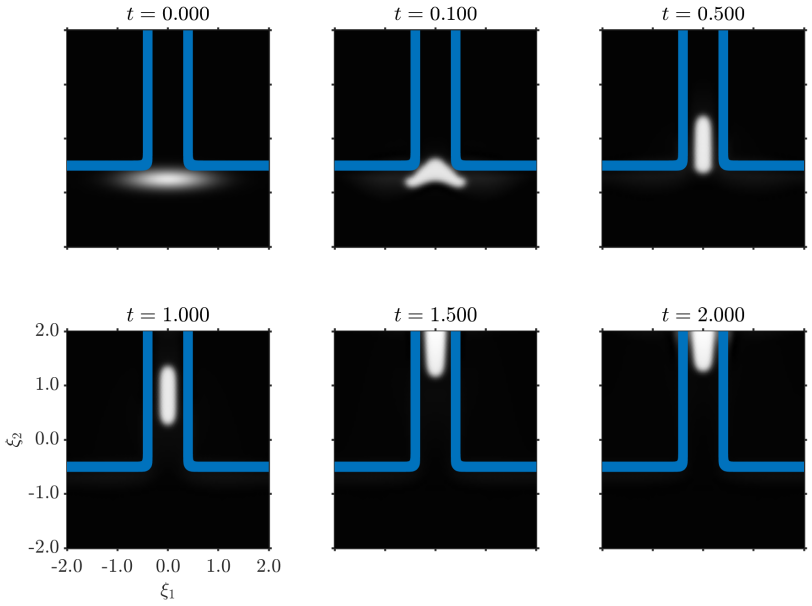

Figure 2.8: A numerical solution showing a herd moving through a canyon

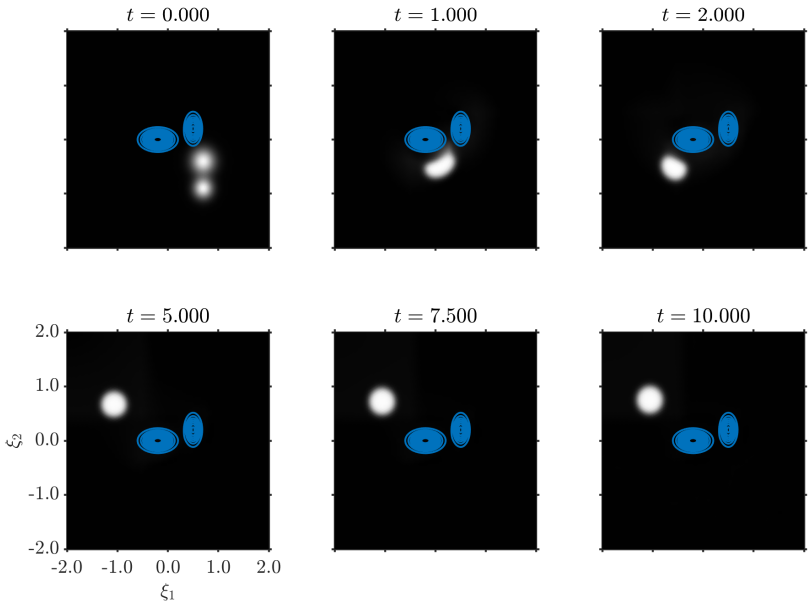

Figure 2.9: A numerical solution showing a herd moving around two hills

2.6 Proofs Regarding Energy Minimizing Functions

Proof of Theorem 2.1. Let $V = \{v \in H^1(\Omega) \colon \int_\Omega v = 0\}$. Due to Poincare's inequality and our assumption on $\mathbf{A}$, we know that V is a Banach space with norm given by $\|v\|_V^2 = \int_\Omega \nabla v \cdot \mathbf{A} \nabla v$. We show that for any sequence $v_j \to 0$ in V, there is some $N \in \mathbb{N}$ such that $E[u^*] \leq E[u^* + v_j]$ for all $j > N$. Using the Taylor expansion of F centered at m, we see that

$$\begin{aligned} E[u^* + v_j] &= \int_\Omega \left(F(m) + F'(m)v_j + \frac{1}{2}F''(c_j)v_j^2 + \frac{1}{2}\nabla v_j \cdot \mathbf{A}\nabla v_j \right) \\ &= E[u^*] + \frac{1}{2}\int_\Omega F''(c_j)v_j^2 + \frac{1}{2}\|v_j\|_V^2 \end{aligned}$$

for some function c_j bounded pointwise almost everywhere between m and $m + v_j$. Using Egoroff's theorem [21, 32], we have that $v_j \to 0$ uniformly, except on sets of arbitrarily small measure. Thus for any $k \in \mathbb{N}$ there is some $\Omega_k \subset \Omega$ with $|\Omega \backslash \Omega_k| < 1/k$ on which $v_j \to 0$ uniformly. Because F'' is continuous and $F''(m) > 0$, for each k, there is some $N_k \in \mathbb{N}$ such that $F''(c_j) > 0$ over Ω_k for all $j > N_k$.

Due to Sobolev embedding ([20, 21, 32], especially [32] Theorem 8.8), we have that there is some $p > 2$ depending on the ambient dimension d and a constant $C > 0$ such that $C\|v\|_{L^p(\Omega)}^2 \leq \|v\|_V^2$ for all $v \in V$. Next, set $q = 2p/(p-2)$ and $r = 2$ and note that $1/r = 1/p + 1/q$ so that by Hölder's inequality, we have that

$$\|v_j\|_{L^2(\Omega\backslash\Omega_k)} \leq \|v_j\|_{L^p(\Omega\backslash\Omega_k)} \left\|\chi_{\Omega\backslash\Omega_k}\right\|_{L^q(\Omega\backslash\Omega_k)} \leq \|v_j\|_{L^p(\Omega\backslash\Omega_k)} \left(\frac{1}{k}\right)^{1/q}.$$

Upon noting that

$$\int_\Omega F''(c_j)v_j^2 \geq \int_{\Omega_k} F''(c_j)v_j^2 - L\left\|v_j\right\|^2_{L^2(\Omega\setminus\Omega_k)},$$

we return to our previous inequality to see that for all $j > N_k$

$$\begin{aligned}\int_\Omega F''(c_j)v_j^2 + \left\|v_j\right\|_V^2 &\geq \int_{\Omega_k} F''(c_j)v_j^2 - L\left\|v_j\right\|^2_{L^2(\Omega\setminus\Omega_k)} + \left\|v\right\|_V^2 \\ &\geq \int_{\Omega_k} F''(c_j)v_j^2 - L\left\|v_j\right\|^2_{L^p(\Omega)}\left(\frac{1}{k}\right)^{\frac{2}{q}} + C\left\|v_j\right\|^2_{L^p(\Omega)} \\ &\geq 0\end{aligned}$$

after choosing k large enough so that $C - L/k^{2/q} > 0$. Fixing $N = N_k$ ensures that $E[u^* + v_j] \geq E[u^*]$ for all $j > N$. □

Proof of Theorem 2.2. Following the same steps as above, we see that for any $v \in V$, we have

$$\begin{aligned}E[u^* + v] &= \int_\Omega \left(F(m) + F'(m)v + \frac{1}{2}F''(c)v^2 + \frac{1}{2}\nabla v \cdot \mathbf{A}\nabla v\right) \\ &= E[u^*] + \frac{1}{2}\int_\Omega F''(c)v^2 + \frac{1}{2}\left\|v\right\|_V^2 \\ &\geq E[u^*] + \frac{1}{2}\int_\Omega F''(c)v^2 + \frac{A_1}{2}\left\|\nabla v\right\|^2_{L^2(\Omega)} \\ &\geq E[u^*] + \frac{1}{2}\int_\Omega F''(c)v^2 + \frac{1}{2}A_1C_\Omega\left\|v\right\|^2_{L^2(\Omega)} \\ &\geq E[u^*] + \frac{1}{2}\left(A_1C_\Omega - L\right)\left\|v\right\|^2_{L^2(\Omega)}\end{aligned}$$

where $C_\Omega > 0$ depends only on Ω by Poincare's inequality and c is some function

which lies pointwise almost everywhere between m and $m+v$. Restricting $A_1 \geq L/C_\Omega$ completes the proof. $\square$

CHAPTER 3. EXISTENCE OF SOLUTIONS

Now that we have an intuitive understanding of how solutions should behave, let us turn our attention to establishing the existence of solutions. First, let us review some of the progress in the standard Cahn-Hilliard case, particularly when variable mobility is taken. In the isotropic case with no potential function ϕ, existence has been shown even when the mobility $b(u)$ is allowed to be degenerate [12, 17], but uniqueness is still an open problem. Taking the mobility to be variable but non-degenerate allows for some uniqueness results for $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, but these results are few and may only extend up to some finite time T^* when $d = 3$ [3, 35]. Some anisotropic equations have been analyzed for constant mobility [14] as particular cases of the Cahn-Hilliard-Gurtin equations first proposed by Gurtin [27]. The potential term ϕ comes up quite infrequently in the literature under any context, but we see Bernoff and Topaz use this term when modeling biological aggregation [5] and Schimperna includes this term as a possible way to couple a Cahn-Hilliard system to other physical processes [35]. All of the results above are obtained by some variant of using compactness after obtaining bounds from Galerkin approximations.

For the rest of this chapter, we will focus on the standard problem

$$\begin{cases} u_t = \nabla \cdot b(u)\mathbf{K}\nabla w & (x,t) \in \Omega_T \\ w = f(u) - \nabla \cdot \mathbf{A}\nabla u + \phi & (x,t) \in \Omega_T \\ \partial_{\mathbf{A}\boldsymbol{\nu}} u = \partial_{\mathbf{K}\boldsymbol{\nu}} w = 0 & (x,t) \in \partial\Omega \times [0,T] \\ u = g & (x,t) \in \Omega \times \{0\} \end{cases} \tag{3.1}$$

and its viscous relaxation

$$\begin{cases} u_t = \nabla \cdot b(u)\mathbf{K}\nabla w & (x,t) \in \Omega_T \\ w = f(u) - \nabla \cdot \mathbf{A}\nabla u + \phi + \theta u_t & (x,t) \in \Omega_T \\ \partial_{\mathbf{A}\boldsymbol{\nu}} u = \partial_{\mathbf{K}\boldsymbol{\nu}} w = 0 & (x,t) \in \partial\Omega \times [0,T] \\ u = g & (x,t) \in \Omega \times \{0\} \end{cases} \tag{3.2}$$

where $\Omega_T = \Omega \times (0,T]$ with $T > 0$ arbitrary. In Section 3.1, we use a compactness argument similar to what is done by Elliot and Garcke in [17] to establish the existence of solutions to (3.1). In Section 3.2, we make a similar argument to establish the existence of solutions to (3.2), but allow the parameters $\mathbf{A}$ and ϕ to vary with respect to t, so that more care is required when obtaining bounds for the sequence of approximate solutions. To the best of our knowledge, these are the first results to incorporate anisotropic terms that are allowed to vary in space, especially when also considering variable mobility and a potential term ϕ. Additionally, these are the first results to allow any parameters to directly depend on t.

3.1 Standard Case

Let us now make the following assumptions:

$$\Omega \subset \mathbb{R}^d \text{ is a bounded domain with } \partial\Omega \text{ Lipschitz} \tag{3.3a}$$

$$b \in C(\mathbb{R};\mathbb{R}) \text{ and } \exists\, b_1, b_2 > 0 \; s.t. \; b_1 \leq b(s) \leq b_2 \; \forall \; s \in \mathbb{R} \tag{3.3b}$$

$$\mathbf{K}\colon \overline{\Omega_T} \mapsto \mathbb{R}^{d\times d} \; s.t. \; K_{ij} = K_{ji}, \text{ and } \exists\, K_1, K_2 > 0 \; s.t. \tag{3.3c}$$

$$K_1 \left|\xi\right|^2 \leq \xi \cdot \mathbf{K}(x,t)\xi \leq \theta \left|y\right|^2 \; \forall \; \xi \in \mathbb{R}^d \text{ and } (x,t) \in \overline{\Omega_T}$$

$$K_{ij} \in L^\infty(\Omega_T)$$

$$\mathbf{A}\colon \overline{\Omega} \mapsto \mathbb{R}^{d\times d} \; s.t. \; A_{ij} = A_{ji}, \text{ and } \exists\, A_1, A_2 > 0 \; s.t. \tag{3.3d}$$

$$A_1 \left|\xi\right|^2 \leq \xi \cdot \mathbf{A}(x)\xi \leq A_2 \left|\xi\right|^2 \; \forall \; \xi \in \mathbb{R}^d \text{ and } x \in \overline{\Omega}$$

$$A_{ij} \in L^\infty(\Omega)$$

$$F \in C^2(\mathbb{R};\mathbb{R}^+) \text{ and } \exists\, c_1, c_2, L \; s.t. \tag{3.3e}$$

$$c_1 s^2 - c_2 \leq F(s) \text{ and } \left|F''(s)\right| \leq L \; \forall \; s \in \mathbb{R}$$

$$\phi \in L^2(\Omega) \tag{3.3f}$$

where and K_{ij} and A_{ij} refer to the components of the matrices $\mathbf{K}$ and $\mathbf{A}$ respectively. Note that it is not strictly necessary to have $\mathbf{K}$ and $\mathbf{A}$ defined pointwise everywhere, but it is necessary that at almost all $t \in (0,T)$, that $\mathbf{K}$ and $\mathbf{A}$ are defined almost everywhere both in Ω and on $\partial\Omega$.

Theorem 3.1. *Under the assumptions listed above,* (3.1) *has a solution for any* $g \in H^1(\Omega)$ *in the sense that there are functions* u, w *such that*

$$\langle u_t, \psi \rangle_{H^{-1}(\Omega), H^1(\Omega)} + \int_\Omega b(u) \nabla w \cdot \mathbf{K} \nabla \psi = 0 \tag{3.4}$$

$$\int_\Omega w\zeta = \int_\Omega \left(f(u) + \phi\right) \zeta + \int_\Omega \nabla u \cdot \mathbf{A} \nabla \zeta \tag{3.5}$$

for all ψ *and* $\zeta \in H^1(\Omega)$ *and almost all* $t \in (0, T)$*. The functions* u *and* w *have regularity* $u \in C([0,T]; L^2(\Omega)) \cap L^\infty(0, T; H^1(\Omega))$*,* $u_t \in L^2(0, T; H^{-1}(\Omega))$*,* $w \in L^2(0, T; H^1(\Omega))$*, and* $u(0) = g$ *in* $L^2(\Omega)$*.*

Here $H^{-1}(\Omega)$ is the dual of $H^1(\Omega)$ with $\langle \cdot, \cdot \rangle_{H^{-1}(\Omega), H^1(\Omega)}$ the standard dual pairing.

Proof. As $\partial\Omega$ is Lipschitz, $C^\infty(\overline{\Omega})$ is dense in $H^1(\Omega)$. Let $\{\psi_i\}_{i=1}^\infty \subset C^\infty(\overline{\Omega})$ be a basis for $H^1(\Omega)$ which is orthogonal in $H^1(\Omega)$ and orthonormal in $L^2(\Omega)$ and set $V_n = \text{span}\{\psi_i\}_{i=1}^n$. We first establish a sequence of functions $\{u^n, w^n\}_{n=1}^\infty$ that satisfy (3.4) and (3.5) over the subset of test functions ψ and $\zeta \in V^n$ and establish uniform bounds to use compactness to pass to the limit.

For each n, set $u^n = c_1^n \psi_1 + \ldots + c_n^n \psi_n$ and $w^n = d_1^n \psi_1 + \ldots + d_n^n \psi_n$ for functions $c_i(t), d_i(t)$ to be determined. Substituting into (3.4) and (3.5) yields the Galerkin ansatz

$$\frac{d}{dt} c_j^n = -\sum_i d_i^n \int_\Omega b\left(\sum_k c_k^n \psi_k\right) \nabla \psi_i \cdot \mathbf{K} \nabla \psi_j \tag{3.6}$$

$$d_i^n = \int_\Omega \left[f\left(\sum_k c_k^n \psi_k\right) + \phi \right] \psi_i + \sum_k c_k^n \int_\Omega \nabla \psi_k \cdot \mathbf{A} \nabla \psi_i \tag{3.7}$$

for all $i, j = 1, 2, \ldots, n$. Substituting (3.7) into (3.6) and applying the initial condition

$$c_j^n(0) = (g, \psi_j)_{H^1(\Omega)}, \quad j = 1, 2, \ldots, n \tag{3.8}$$

yields a system of ordinary differential equations that has a unique, absolutely continuous solution $\{c_i^n(t)\}_{i=1}^n$ for $0 \leq t \leq T$. (3.7) gives us the corresponding functions $\{d_i^n(t)\}_{i=1}^n$ so that we obtain our sequence of functions $\{u^n, w^n\}_{n=1}^\infty$. To more clearly see that these solutions exist, note that there are positive constants C_1 and C_2 which are independent of c_j^n and satisfy

$$\sum_{j=1}^n \left| \frac{d}{dt} c_j^n \right| \leq C_1 + C_2 \sum_{j=1}^n \left| c_j^n \right|$$

(see e.g. [38]).

To obtain appropriate bounds, first note that differentiating (2.4) yields

$$\frac{d}{dt} E[u^n] = \int_\Omega (f(u^n) u_t^n + \nabla u^n \cdot \mathbf{A} \nabla u_t^n + \phi u_t^n) \tag{3.9}$$

for almost all t so that choosing test functions $\psi = w^n$ and $\xi = u_t^n$ in (3.4) and (3.5) (substituting $u = u^n$ and $w = w^n$) yields

$$\int_\Omega u_t^n w^n = -\int_\Omega b(u^n) \nabla w^n \cdot \mathbf{K} \nabla w^n \tag{3.10}$$

$$\int_\Omega u_t^n w^n = \int_\Omega f(u^n) u_t^n + \nabla u^n \cdot \mathbf{A} \nabla u_t^n + \phi u_t^n \tag{3.11}$$

so that

$$\begin{aligned}
&\frac{d}{dt}E[u^n] + \int_\Omega b(u^n)\nabla w^n \cdot \mathbf{K}\nabla w^n = 0 \\
\implies & E[u^n] + \int_0^t \int_\Omega b(u^n)\nabla w^n \cdot \mathbf{K}\nabla w^n = E[u^n(0)] \\
\implies & \int_\Omega \left(F(u^n) + \frac{1}{2}\nabla u^n \cdot \mathbf{A}\nabla u^n + u^n\phi \right) + \int_0^t \int_\Omega b(u^n)\nabla w^n \cdot \mathbf{K}\nabla w^n \leq C \\
\implies & \int_\Omega \left(c_1 |u^n|^2 - c_2 + \frac{A_1}{2}|\nabla u^n|^2 + u^n\phi \right) + \int_0^t \int_\Omega b_1 K_1 |\nabla w^n|^2 \leq C \\
\implies & \|u^n\|_{H^1(\Omega)} + \int_0^t \int_\Omega |\nabla w^n|^2 \leq C
\end{aligned} \tag{3.12}$$

where $C > 0$ is some positive constant which may vary from line to line. Note that $E[u^n(0)] \leq C$ follows from (3.8), (3.3d) and (3.3e). Because the inequality holds for almost all t, we have that $\{u^n\}$ is bounded in $L^\infty(0,T;H^1(\Omega))$. Now choosing $\xi = w^n$ in (3.5) yields

$$\begin{aligned}
&\int_\Omega |w^n|^2 = \int_\Omega \left(f(u^n) + \phi\right) w^n + \int_\Omega \nabla u^n \cdot \mathbf{A}\nabla w^n \\
\implies & \|w^n\|^2_{L^2(\Omega)} \leq \|f(u^n) + \phi\|_{L^2(\Omega)} \|w^n\|_{L^2(\Omega)} + A_2 \int_\Omega |\nabla u^n| \, |\nabla w^n| \\
\implies & \|w^n\|^2_{L^2(\Omega)} \leq \eta \|w^n\|^2_{L^2(\Omega)} + \frac{1}{4\eta}\|f(u^n) + \phi\|^2_{L^2(\Omega)} + \\
& \qquad\qquad \eta \int_\Omega |\nabla w^n|^2 + \frac{A_2^2}{4\eta}\int_\Omega |\nabla u^n|^2 \\
\implies & \|w^n\|^2_{L^2(\Omega)} - \frac{\eta}{1-\eta}\int_\Omega |\nabla w^n|^2 \leq C_\eta \left(1 + \|u^n\|^2_{H^1(\Omega)} + \|\phi\|^2_{L^2(\Omega)}\right)
\end{aligned} \tag{3.13}$$

where $0 < \eta < 1$ is arbitrary and C_η depends on η but is independent of u^n and w^n. Fixing η while integrating (3.13) over $(0,t)$ and adding to (3.12) gives us that $\{w^n\}$

is bounded in $L^2(0,T;H^1(\Omega))$. To see that $\{u_t^n\}$ is bounded, choose any $\psi \in H^1(\Omega)$ and decompose it into $\psi = \psi_1^n + \psi_2^n$ where $\psi_1^n \in V_n$ and $\psi_2^n \in V_n^\perp$ are orthogonal in $L^2(\Omega)$. Then we have

$$\begin{aligned}
\left|\langle u_t^n, \psi\rangle_{H^{-1}(\Omega),H^1(\Omega)}\right| &= \left|\int_\Omega u_t^n \psi_1^n\right| \\
&= \left|\int_\Omega b(u)\nabla w^n \cdot \mathbf{K}\nabla\psi_1^n\right| \\
&\le K_2 b_2 \int_\Omega |\nabla w^n| \left|\nabla\psi^1\right| \\
&\le C \|\psi\|_{H^1(\Omega)} \int_\Omega |\nabla w^n|
\end{aligned}$$

so that

$$\|u_t^n\|_{H^{-1}(\Omega)} \le C \int_\Omega |\nabla w^n| \tag{3.14}$$

which gives $\{u_t^n\}$ bounded in $L^2(0,T;H^{-1}(\Omega))$ after noting

$$\left(\int_\Omega |\nabla w^n|\right)^2 \le |\Omega| \, \|\nabla w^n\|_{L^2(\Omega)}^2$$

and integrating over $(0,T)$.

Due to weak compactness, we have the existence of limit functions u, w and a (not relabeled) subsequence such that

$$u^n \rightharpoonup u \text{ weak} - * \text{ in } L^\infty(0,T;H^1(\Omega)) \tag{3.15a}$$

$$u_t^n \rightharpoonup u_t \text{ weakly in } L^2(0,T;H^{-1}(\Omega)) \tag{3.15b}$$

$$w^n \rightharpoonup w \text{ weakly in } L^2(0,T;H^1(\Omega)). \tag{3.15c}$$

Due to the embeddings $H^1(\Omega) \hookrightarrow L^2(\Omega)$ (compact) and $L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$ (continuous), we have, again passing to a further subsequence, that

$$u^n \to u \text{ in } C([0,T]; L^2(\Omega)) \tag{3.15d}$$

by the Aubin-Lions lemma so that $u(0) = g$ in $L^2(\Omega)$. It may be prudent to recall here that if $u^n \rightharpoonup u$ weakly in $L^2(0,T; H^1(\Omega))$ and $u_t^n \rightharpoonup v$ weakly in $L^2(0,T; H^{-1}(\Omega))$, then $v = u_t$.

We conclude our proof by a standard method of passing to the limit in (3.4) and (3.5). First, choose any interval $(t_1, t_2) \subset (0,T)$ and functions $\psi, \xi \in H^1(\Omega)$. Then construct sequences $\psi^m \to \psi$ and $\xi^m \to \xi$ in $H^1(\Omega)$ where ψ^m, $\xi^m \in V^m$. Extend these to $L^2(0,T; H^1(\Omega))$ by considering the characteristic function $\chi_{(t_1,t_2)}$ to get $\chi_{(t_1,t_2)}\psi^m \to \chi_{(t_1,t_2)}\psi$ in $L^2(0,T; H^1(\Omega))$ and similarly for ξ^m. Then for each fixed m, we pass to the limit in n to get

$$\int_0^T \left\langle u_t, \chi_{(t_1,t_2)}\psi^m \right\rangle_{H^{-1}(\Omega), H^1(\Omega)} = -\int_0^T \int_\Omega b(u)\nabla w \cdot \mathbf{K}\nabla\chi_{(t_1,t_2)}\psi^m$$
$$\int_0^T \int_\Omega w\chi_{(t_1,t_2)}\xi^m = \int_0^T \int_\Omega \left(f(u) + \phi\right)\chi_{(t_1,t_2)}\xi^m + \int_0^T \int_\Omega \nabla u \cdot \mathbf{A}\nabla\chi_{(t_1,t_2)}\xi^m$$

before passing to the limit in m and noting that the interval (t_1, t_2) is arbitrary to recover (3.4) and (3.5) for almost all t. □

Corollary 3.1. *If we make the additional assumptions that $\partial\Omega \in C^{2,1}$, $\phi \in H^1(\Omega)$, and $A_{ij} \in C^{1,1}(\overline{\Omega}; \mathbb{R})$ for almost all $t \in (0,T)$, then we have that $u \in L^2(0,T; H^3(\Omega))$ and $w = f(u) - \nabla \cdot (\mathbf{A}\nabla u) + \phi$ in $L^2(0,T; H^1(\Omega))$.*

Proof. This follows immediately from elliptic regularity upon noticing that, for almost all t, u is a weak solution to the boundary value problem

$$\begin{cases} \nabla \cdot \mathbf{A}\nabla u + \lambda u = h, & x \in \Omega \\ \partial_{\mathbf{A}\boldsymbol{\nu}} u = 0, & x \in \partial\Omega \end{cases}$$

for some $\lambda > 0$ where $h = f(u) + \phi - w + \lambda u \in L^2(0,T;H^1(\Omega))$ [26]. □

3.2 Viscosity Relaxation

Now we turn our attention to (3.2). As we will see, this relaxation provides enough regularity to allow $\mathbf{A}$ and ϕ to depend on t in addition to $\mathbf{K}$. For this section, we keep the same assumptions as in Section 3.1, with the exception that we alter (3.3d) and (3.3f) to

$$\mathbf{A}\colon \overline{\Omega_T} \to \mathbb{R}^{d\times d} \; s.t. \; A_{ij} = A_{ji}, \text{ and } \exists \; A_1, A_2 > 0 \; s.t. \tag{3.16a}$$

$$A_1 |\xi|^2 \leq \xi \cdot \mathbf{A}(x,t)\xi \leq A_2 |\xi|^2 \;\; \forall \; \xi \in \mathbb{R}^d \text{ and } (x,t) \in \overline{\Omega_T}$$

$$A_{ij} \in L^\infty(\Omega_T) \text{ and differentiable in } t \; s.t. \; \|\rho(\mathbf{A}_t)\|_{L^\infty(\Omega)} \in L^1(0,T)$$

$$\phi \in L^2(\Omega_T) \tag{3.16b}$$

where $\rho(\mathbf{A}_t)$ is the spectral radius of $\mathbf{A}_t$.

Theorem 3.2. *Under the assumptions given above,* (2.2) *has a solution for any* $g \in H^1(\Omega)$ *when* $\theta > 0$ *in the sense that there are functions* u, w *such that*

$$\int_\Omega u_t \psi + \int_\Omega b(u) \nabla w \cdot \mathbf{K} \nabla \psi = 0, \quad \psi \in H^1(\Omega) \tag{3.17}$$

$$\int_\Omega w\zeta = \int_\Omega \left(f(u) + \phi\right) \zeta + \int_\Omega \nabla u \cdot \mathbf{A} \nabla \zeta + \theta \int_\Omega u_t \zeta \tag{3.18}$$

for all ψ *and* $\zeta \in H^1(\Omega)$ *and almost all* $t \in (0, T)$. *The functions* u *and* w *have regularity* $u \in C([0,T]; L^2(\Omega)) \cap L^\infty(0, T; H^1(\Omega))$, $u_t \in L^2(\Omega_T)$, $w \in L^2(0, T; H^1(\Omega))$, *and* $u(0) = g$ *in* $L^2(\Omega)$.

Proof. Using the same method as in the proof of (3.1), we arrive at the Galerkin ansatz

$$\frac{d}{dt} c_j^n = -\sum_i d_i^n \int_\Omega b \left(\sum_k c_k^n \psi_k \right) \nabla \psi_i \cdot \mathbf{K} \nabla \psi_j \tag{3.19}$$

$$d_i^n = \int_\Omega \left[f \left(\sum_k c_k^n \psi_k \right) + \phi \right] \psi_i + \sum_k c_k^n \int_\Omega \nabla \psi_k \cdot \mathbf{A} \nabla \psi_i + \theta \frac{d}{dt} c_i^n \tag{3.20}$$

from which we eliminate d_i^n to get the system of ODEs

$$\left(\mathbf{I} + \theta \mathbf{M}\right) \frac{d}{dt} \boldsymbol{c}^n = -\mathbf{M} \boldsymbol{y} \tag{3.21}$$

where $\boldsymbol{c}^n = [c_1^n, \ldots, c_n^n]^\mathsf{T}$ and $\boldsymbol{y}^n = [y_1^n, \ldots, y_n^n]^\mathsf{T}$ with

$$M_{ij}(\boldsymbol{c}^n) = \int_\Omega b \left(\sum_k c_k^n \psi_k \right) \nabla \psi_i \cdot \mathbf{K} \nabla \psi_j,$$

and

$$y_i(\boldsymbol{c}^n) = \int_\Omega \left[f \left(\sum_k c_k^n \psi_k \right) + \phi \right] \psi_i + \sum_k c_k^n \int_\Omega \nabla \psi_k \cdot \mathbf{A} \nabla \psi_i.$$

Noting that $\mathbf{M}$ is symmetric positive semi-definite and $|y_i| \leq C_1(t) + C_2(t)\sum_k |c_k^n|$ with C_1 and C_2 integrable, we apply the initial condition (3.8) and use the absolutely continuous solution of the initial value problem to define the functions u^n and w^n.

To show that the sequences $\{u^n\}$ and $\{w^n\}$ are bounded, we first choose test functions $\psi = w^n$, $\zeta = u_t^n$ to get

$$\begin{aligned}
&\int_\Omega (f(u^n)u_t^n + \nabla u^n \cdot \mathbf{A}\nabla u_t^n + \phi u_t^n) + \theta \int_\Omega |u_t^n|^2 + \int_\Omega b(u^n)\nabla w^n \cdot \mathbf{K}\nabla w^n = 0 \\
\implies &\frac{d}{dt}\int_\Omega \left(F(u^n) + \frac{1}{2}\nabla u^n \cdot \mathbf{A}\nabla u^n\right) + \theta \left\|u_t^n\right\|_{L^2(\Omega)}^2 + \int_\Omega b(u^n)\nabla w^n \cdot \mathbf{K}\nabla w^n \\
&\qquad = \int_\Omega (\nabla u^n \cdot \mathbf{A}_t \nabla u^n - \phi u_t^n) \\
\implies &\frac{d}{dt}\int_\Omega \left(F(u^n) + \frac{1}{2}\nabla u^n \cdot \mathbf{A}\nabla u^n\right) + \frac{\theta}{2}\left\|u_t^n\right\|_{L^2(\Omega)}^2 + \int_\Omega b(u^n)\nabla w^n \cdot \mathbf{K}\nabla w^n \quad (3.22) \\
&\qquad \leq 2\frac{\|\rho(\mathbf{A}_t)\|_{L^\infty(\Omega)}}{A_1}\int_\Omega \left(F(u^n) + \frac{1}{2}\nabla u^n \cdot \mathbf{A}\nabla u^n\right) + \frac{1}{2\theta}\left\|\phi\right\|_{L^2(\Omega)}^2 \\
\implies &\int_\Omega \left(F(u^n) + \frac{1}{2}\nabla u^n \cdot \mathbf{A}\nabla u^n\right) \quad (3.23) \\
&\qquad \leq C\left[\int_\Omega \left(F(u^n) + \frac{1}{2}\nabla u^n \cdot \mathbf{A}\nabla u^n\right)\Big|_{t=0} + \frac{1}{2\theta}\int_0^t \left\|\phi\right\|_{L^2(\Omega)}^2\right]
\end{aligned}$$

where we added the term $2\frac{\|\rho(\mathbf{A}_t)\|_{L^\infty(\Omega)}}{A_1}\int_\Omega F(u^n)$ to the right-hand side in (3.22) so that we can use Gronwall's inequality to obtain (3.23), which gives us that $\{u^n\}$ is bounded in $L^\infty(0,T;H^1(\Omega))$. From here, we integrate (3.22) to obtain that $\{u_t^n\}$ and $\{|\nabla w^n|\}$ are bounded in $L^2(\Omega_T)$. We now choose the test function $\xi = w^n$ in (3.18) to obtain

$$\left\|w^n\right\|_{L^2(\Omega)}^2 \leq C\left(1 + \left\|u^n\right\|_{H^1(\Omega)}^2 + \left\|u_t^n\right\|_{L^2(\Omega)}^2 + \left\|\nabla w^n\right\|_{L^2(\Omega)}^2\right)$$

so that after integrating, we see that $\{w^n\}$ is bounded in $L^2(0,T;H^1(\Omega))$. Applying the same compactness argument as for the Theorem 3.1 concludes the proof. $\square$

CHAPTER 4. NUMERICAL METHODS

Now that we have established that solutions to our generalization of the Cahn-Hilliard equation exist, let us discuss how we can approximate these solutions numerically. There are a few properties of the Cahn-Hilliard equation that one should be aware of when finding numerical solutions. The first is that solutions can alternate between a short timescale and a long timescale. In the context of our discussion in Chapter 2, if $\mathbf{A} = \alpha\mathbf{I}$, $b(u)\mathbf{K} = \mathbf{I}$, and ϕ is constant, then the development of the interfacial region $\Omega_{1/2}$ is on an $O(\alpha)$ timescale while the coarsening of regions Ω_0 and Ω_1 occur on an $O(1/\sqrt{\alpha})$ timescale [34]. Because of this, a common practice is to use an adaptive time-step where either an error estimate is obtained by either using two different Runge-Kutta schemes or by using the energy functional (2.4) [31, 42]. In addition to the existence of multiple timescales, there are multiple length scales as well. For example, under the assumptions above, the width of $\Omega_{1/2}$ will be $O(\sqrt{\alpha})$ [34], while the diameters of Ω_0 and Ω_1 will be more comparable to the diameter of the entire domain Ω (e.g., see Figures 2.5 and 2.6). When using a finite element method, unless one uses an adaptive mesh [22], this generally necessitates the use of a uniformly fine mesh. The most common choices for the spacial discretization are (continuous) Galerkin finite elements [2, 3, 4, 13, 15, 18] and discontinuous Galerkin finite elements [22, 28, 43], but finite difference methods are still used sometimes

[9]. We will approach this problem by using Galerkin finite elements for our spacial approximation and an implicit-explicit (IMEX) scheme for our time discretization.

In Section 4.1, we review Galerkin finite elements as applied to our problem [24, 29]. In Section 4.2, we adapt a first order IMEX scheme from the standard Cahn-Hilliard equation with constant mobility to our problem and discus its stability [37, 39]. In Section 4.3, we propose a new formulation of the problem that is framed as a change of variables in the continuous setting. Working in this framework allows us to effectively captures both the short term and long term behavior of solutions, without the need of using an adaptive time-step, which generally requires the tuning of parameters [31, 39]. In Section 4.4, we conclude by discussing a few practical details that should be considered when implementing the methods proposed in the previous sections.

4.1 Spacial Discretization

In this section, we use Galerkin finite elements to form the spacial discretization of our problem. This methodology is quite standard and general references can be found in [24, 29] while examples relating to the Cahn-Hilliard equation can be found

in [2, 37], Recall that the strong form of our problem is given by

$$\begin{cases} u_t = \nabla \cdot b(u)\mathbf{K}\nabla w & (x,t) \in \Omega_T \\ w = f(u) - \nabla \cdot \mathbf{A}\nabla u + \phi & (x,t) \in \Omega_T \\ \partial_{\mathbf{A}\boldsymbol{\nu}} u = \partial_{\mathbf{K}\boldsymbol{\nu}} w = 0 & (x,t) \in \partial\Omega \times [0,T] \\ u = g & (x,t) \in \Omega \times \{0\} \end{cases} \tag{3.1}$$

where $T > 0$ is arbitrary and $\Omega_T = [0,T) \times \Omega$ for some bounded domain $\Omega \subset \mathbb{R}^d$ with $\partial\Omega$ at least Lipschitz. Instead of attempting to solve (3.1) directly, we again turn to the weak form. That is, we will approximate u, $w \in L^2([0,T]; H^1(\Omega))$ such that

$$\langle u_t, \psi \rangle_{H^{-1}(\Omega), H^1(\Omega)} + \int_\Omega b(u)\nabla w \cdot \mathbf{K}\nabla\psi = 0 \qquad \forall \psi \in H^1(\Omega) \tag{4.1a}$$

$$\int_\Omega w\zeta = \int_\Omega (f(u) + \phi)\zeta + \int_\Omega \nabla u \cdot \mathbf{A}\nabla\zeta \qquad \forall \zeta \in H^1(\Omega) \tag{4.1b}$$

hold for almost all t and that $u(0) = g$ in $H^1(\Omega)$. Observe that for (almost every) fixed time t, we have that both the unknown functions (u and w) and the test functions (ψ and ζ) are elements of $H^1(\Omega)$. Before we can solve these equations, we must have a way to express these functions in a usable manner. In order to do this, we construct a finite dimensional subspace $V_h \subset H^1(\Omega)$ by choosing *continuous* basis functions $\{\psi_i\}_{i=1}^M$ for V_h which have certain properties to be specified later. The choice that each ψ_i is continuous is not strictly necessary but would complicate the application of the divergence theorem later on and lead to a different formulation (i.e., discontinuous Galerkin finite elements for the most common choice of discontinuous ψ_i) [1, 22]. Until

V_h is more concretely defined, the subscript h can simply be thought of as some sort of tolerance on how well functions in $H^1(\Omega)$ should be approximated by functions in V_h.

After the ψ_i are chosen, we solve (4.1) over V_h instead of $H^1(\Omega)$. To do this, first set $u_h = c_1\psi_1 + \ldots + c_M\psi_M$ and $w_h = d_1\psi_1 + \ldots + d_M\psi_M$ where c_i and d_i are unknown functions of t. Recall that $H^1(\Omega)$ is a Hilbert space so that there exists a subspace $V_h^\perp$ orthogonal to V_h such that $H^1(\Omega) = V_h \oplus V_h^\perp$ so that given some function $v \in H^1(\Omega)$, there are unique functions $v^{(1)} \in V_h$ and $v^{(2)} \in V_h^\perp$ such that $v = v^{(1)} + v^{(2)}$ [30]. We will use this decomposition to define the operator $\Pi\colon H^1(\Omega) \to V_h$ as $\Pi v = v^{(1)}$ from above. Once V_h is fixed, our goal is to choose c_i and d_i so that $u_h = \Pi u$ and $w_h = \Pi w$. Unfortunately, due to the nonlinear nature of the problem, this is not necessarily possible. However, as was seen in Chapter 3, the error $\|u - u_h\|_{L^2(\Omega_T)}$ can be arbitrarily small, depending on the choice of V_h.

Replacing u and w by u_h and w_h in (4.1) and projecting into V_h produces the system of equations

$$\langle u_{h,t}, \psi_j \rangle_{H^{-1}(\Omega), H^1(\Omega)} + \int_\Omega \Pi b(u_h) \nabla w_h \cdot \mathbf{K} \nabla \psi_j = 0$$
$$\int_\Omega w_h \psi_j = \int_\Omega \Pi(f(u_h) + \phi)\psi_j + \int_\Omega \nabla u_h \cdot \mathbf{A} \nabla \psi_j$$

for each $j = 1, 2, \ldots, M$ where $u_{h,t}$ is the time derivative of u_h. These equations can

be expanded in terms of c_i and d_i as

$$\sum_{i=1}^{M} c_i' \int_\Omega \psi_i \psi_j + \sum_{i=1}^{M} d_i \int_\Omega \Pi b(u_h) \nabla \psi_i \cdot \mathbf{K} \nabla \psi_j = 0$$
$$\sum_{i=1}^{M} d_i \int_\Omega \psi_i \psi_j = \int_\Omega \Pi(f(u_h) + \phi)\psi_j + \sum_{i=1}^{M} c_i \int_\Omega \nabla \psi_i \cdot \mathbf{A} \nabla \psi_j$$

and then (after recalling $\mathbf{A}$ and $\mathbf{K}$ are symmetric) re-written in matrix form. This is given by

$$\mathcal{M}\boldsymbol{c}' + \boldsymbol{b} \cdot \mathcal{B}\boldsymbol{d} = 0 \tag{4.2a}$$

$$\mathcal{M}\boldsymbol{d} = \mathcal{M}(\boldsymbol{f} + \boldsymbol{\phi}) + \mathcal{A}\boldsymbol{c} \tag{4.2b}$$

where at each t, $\boldsymbol{c}$ and $\boldsymbol{d}$ are vectors in $\mathbb{R}^M$ with components c_i and d_i respectively; $\boldsymbol{b}$, $\boldsymbol{f}$, and $\boldsymbol{\phi}$ are vectors which encode $\Pi b(u_h)$, $\Pi f(u_h)$, and $\Pi\phi$ respectively; and $\mathcal{M}$, $\mathcal{A}$ are symmetric matrices with elements

$$\mathcal{M}_{ij} = \int_\Omega \psi_i \psi_j \tag{4.3a}$$

$$\mathcal{A}_{ij} = \int_\Omega \nabla \psi_i \cdot \mathbf{A} \nabla \psi_j \tag{4.3b}$$

and $\mathcal{B}$ is a tensor with elements

$$\mathcal{B}_{lij} = \int_\Omega \psi_l \nabla \psi_i \cdot \mathbf{K} \nabla \psi_j. \tag{4.3c}$$

The encoding of ϕ into $\boldsymbol{\phi}$ is as follows: if $\Pi\phi = \alpha_1\psi_1 + \ldots + \alpha_M\psi_M$, then $\boldsymbol{\phi} = [\alpha_1, \ldots, \alpha_M]^\mathsf{T}$. This methodology is the same for all representations of elements of V_h. We will return to the original weak formulation before choosing a method for

solving the ordinary differential equation (4.2) but note here that we will impose the initial condition $\boldsymbol{c}(0) = \boldsymbol{g}$. Although the result may be obvious, it is important to realize that these matrices are very structured.

Proposition 4.1. *Let $\mathcal{M}$, $\mathcal{A}$, and $\mathcal{B}$ be defined as in* (4.3). *that $\mathcal{M}$ is positive definite, $\mathcal{A}$ is positive-semidefinite, and the matrix $\boldsymbol{b} \cdot \mathcal{B}$ with components $\sum_l b_l \mathcal{B}_{lij}$ is positive-semidefinite for any choice of $\boldsymbol{b} \in \mathbb{R}^N$ with $b_l \geq 0$.*

Proof. Choose any nonzero $\boldsymbol{v} \in \mathbb{R}^N$ and define $v = v_1\psi_1 + \ldots + v_N\psi_N \in V_h$. Then

$$\boldsymbol{v} \cdot \mathcal{M}\boldsymbol{v} = \int_\Omega \left(\sum_{i=1}^N v_i \psi_i \right) \left(\sum_{j=1}^N v_j \psi_j \right) = \|v\|^2_{L^2(\Omega)} > 0.$$

Similarly, we see that

$$\boldsymbol{v} \cdot \mathcal{M}\boldsymbol{v} = \int_\Omega \nabla v \cdot \mathbf{A} \nabla v \geq 0$$

and

$$\boldsymbol{v} \cdot (\boldsymbol{b} \cdot \mathcal{B})\, \boldsymbol{v} = \sum_{l=1}^N b_l \int_\Omega \nabla v \cdot \mathbf{K} \nabla v \geq 0$$

where we note that v being a constant function is allowed. □

Now let us turn our attention to the choice of basis functions ψ_i. We have already mentioned that each ψ_i must be continuous in order for the current formulation to be valid. Beyond that, we should choose $\{\psi_i\}$ so that $\mathcal{M}$, $\mathcal{A}$, and $\mathcal{B}$ can be either calculated exactly or easily approximated using some numerical integration such as Gauss quadrature. If we were to choose $\{\psi_i\}$ such that the support of each ψ_i is all

of Ω, then $\mathcal{M}$, $\mathcal{A}$, and $\mathcal{B}$ would all be densely populated and we would move into the territory of spectral methods [25]. On the other hand, if the support of each ψ_i is disjoint from most other ψ_j, then $\mathcal{M}$, $\mathcal{A}$, and $\mathcal{B}$ would all be sparsely populated. For our implementation, we chose $\{\psi_i\}$ to consist of piecewise linear basis functions. We will return to this topic in Section 4.4, but conclude our discussion here by noting that if we were to include the viscous relaxation as in (3.2), (4.2b) would become $\mathcal{M}\boldsymbol{d} = \mathcal{M}(\boldsymbol{f} + \boldsymbol{\phi}) + \mathcal{A}\boldsymbol{c} + \theta\mathcal{M}\boldsymbol{c}'$.

4.2 Temporal Discretization

Before stating our discretization, let us first consider some preliminaries. As is quite common, we will approximate u and w at discrete time levels $\{t_n\}_{n=0}^N$ where $t_0 = 0$, $t_N = T$, and $t_{n+1} - t_n = \delta t_{n+1} > 0$ for $n = 0, 1, \ldots, N-1$. It is possible to choose δt_{n+1} based on the solution at t_n or other parameters [31]. This is known as adaptive time-stepping and, when done well, can reduce the computation time to estimate the solution at time T but may increase the time required to step from each time t_n to t_{n+1} as will be seen shortly. While on the topic of adaptivity, we note here that it is possible to change the spacial discretization scheme at each time-step. That is, instead of keeping V_h fixed at each time-step, we can use a sequence of subspaces $\{V_h^n\}_{n=0}^N$ where each V_h^n is an M_n-dimensional subspace of $H^1(\Omega)$ and is only used to

approximate the solution at time t_n. The main idea being that if a single space V_h is used for all time-steps, then we will have to use a large number of basis functions so that V_h approximates $H^1(\Omega)$ well for arbitrary functions, as we do not know how u and w will evolve *a priori*. The potential benefit of updating the approximation spaces is that we can reduce the number of necessary basis functions by tailoring V_h^{n+1} to the solution at t_{n+1}, which should be similar to the solution at t_n. The reduction in degrees of freedom does come at a cost. No matter what, $\mathcal{M}$, $\mathcal{A}$, and $\mathcal{B}$ will all have to be re-computed whenever $V_h^{n+1} \neq V_h^n$ and, depending on how V_h^{n+1} is chosen, there may be an iterative process in which a guess for a good V_h^{n+1} is taken, a tentative solution at t_{n+1} is found and either accepted or used to make a new guess for V_h^{n+1} based on some *a posteriori* analysis [6, 11, 40]. Adaptive mesh techniques as well as *hp*-refinement are examples of methods where V_h is not held fixed. We will generally keep $\delta t \equiv \delta t_n$ and $V_h \equiv V_h^n$ fixed for our implementation, but include this general discussion as these are ideas that one should consider when choosing how best to solve a given PDE.

In order to determine a time-stepping scheme to solve the system of ODEs given in the previous section, let us recall that the weak form for our problem is given by

$$\langle u_t, \psi \rangle_{H^{-1}(\Omega), H^1(\Omega)} + \int_\Omega b(u) \nabla w \cdot \mathbf{K} \nabla \psi = 0 \qquad \forall \psi \in H^1(\Omega) \tag{4.1a}$$

$$\int_\Omega w\zeta = \int_\Omega (f(u) + \phi)\zeta + \int_\Omega \nabla u \cdot \mathbf{A} \nabla \zeta \qquad \forall \zeta \in H^1(\Omega) \tag{4.1b}$$

and we have an initial condition $u(0) = g$ in $H^1(\Omega)$. As a compromise between speed and stability, we chose to use the first order implicit-explicit (IMEX) scheme

$$\frac{1}{\delta t}\int_\Omega (u^{n+1} - u^n)\psi + \int_\Omega b(u^n)\nabla w^{n+1} \cdot \mathbf{K}\nabla\psi = 0 \qquad \forall \psi \in H^1(\Omega) \tag{4.5a}$$

$$\int_\Omega w^{n+1}\zeta = \int_\Omega (f(u^n) + \phi)\zeta + \int_\Omega \nabla u^{n+1} \cdot \mathbf{A}\nabla\zeta \qquad \forall \zeta \in H^1(\Omega) \tag{4.5b}$$

where u^n and w^n denote the approximate solution at t_n. Note that setting

$$f(u^n) = f(u^{n+1}) + f'(\xi)(u^n - u^{n+1})$$

in (4.5b) gives us

$$\int_\Omega f'(\xi)(u^{n+1} - u^n)\zeta = \int_\Omega (f(u^{n+1}) + \phi)\zeta + \int_\Omega \nabla u^{n+1} \cdot \mathbf{A}\nabla\zeta - \int_\Omega w^{n+1}\zeta$$

where and ξ is some function that lies between u^n and u^{n+1} pointwise. Note that if u^{n+1} and w^{n+1} were exact, then the right-hand side would be zero. This can be used as justification to include the viscous term in (4.5b) to get an alternate equation

$$\int_\Omega w^{n+1}\zeta = \int_\Omega (f(u^n) + \phi)\zeta + \int_\Omega \nabla u^{n+1} \cdot \mathbf{A}\nabla\zeta + S\int_\Omega (u^{n+1} - u^n)\zeta$$

which amounts to solving (2.2) with $\theta = S\delta t$. Note that the consistency error introduced by this extra term is of the same order as the error caused by treating $f(u)$ explicitly [37]. Now let us examine the effect of this relaxation on the stability of our scheme.

Theorem 4.1. *Let* $\mathbf{A}$, $\mathbf{K}$, b, f, *and* ϕ *be as in Theorem 3.1. Suppose* u^{n+1}, u^n, w^{n+1}, *and* w^n *are all* $H^1(\Omega)$ *functions which satisfy* (4.5) *for some* $\delta t > 0$. *If* $\delta t \leq \dfrac{8b_1A_1}{L^2b_2^2K_2}$, *then* $E[u^{n+1}] \leq E[u^n]$ *where the functional* E *is given by*

$$E[u] = \int_\Omega \left(F(u) + \frac{1}{2}\nabla u \cdot \mathbf{A}\nabla u + u\phi \right). \tag{2.4}$$

Proof. Set $\psi = w^{n+1}$ and $\zeta = u^{n+1} - u^n$ to get

$$\int_\Omega \left(f(u^n) + \phi\right)\left(u^{n+1} - u^n\right) + \int_\Omega \nabla u^{n+1} \cdot \mathbf{A}\nabla(u^{n+1} - u^n) = -\delta t \int_\Omega b(u^n)\nabla w^n \cdot \mathbf{K}\nabla w^n.$$

Expanding F about u^n gives us

$$F(u^{n+1}) - F(u^n) = f(u^n)(u^{n+1} - u^n) + \frac{1}{2}f'(\xi)(u^{n+1} - u^n)^2$$

for some function ξ so, after recalling that $2p \cdot (p - q) = |p - q|^2 + |p|^2 - |q|^2$ where $|p|^2 = p \cdot p$ holds for any bilinear operator $\cdot$, we have

$$\begin{aligned} E[u^{n+1}] - E[u^n] \leq \frac{L}{2}\left\|u^{n+1} - u^n\right\|_{L^2(\Omega)}^2 - \frac{A_1}{2}\left\|\nabla u^{n+1} - \nabla u^n\right\|_{L^2(\Omega)}^2 \\ - \delta t b_1 \int_\Omega \nabla w^{n+1} \cdot \mathbf{K}\nabla w^{n+1} \end{aligned} \tag{4.6}$$

where we have used $|f'(\xi)| \leq L$, $A_1 \left|\nabla(u^{n+1} - u^n)\right|^2 \leq \nabla(u^{n+1} - u^n) \cdot \mathbf{A}\nabla(u^{n+1} - u^n)$, and $b_1 \leq b(u^n)$. In order to dominate $\|u^{n+1} - u^n\|_{L^2(\Omega)}$ by $\|\nabla u^{n+1} - \nabla u^n\|_{L^2(\Omega)}$, we set $\zeta = u^{n+1} - u^n$ in (4.5a) and use $b(u^n) \leq b_2$ along with Cauchy-Schwarz and Young's inequality to obtain

$$\left\|u^{n+1} - u^n\right\|_{L^2(\Omega)}^2 \leq \delta t b_2 \left(\eta \int_\Omega \nabla w^{n+1} \cdot \mathbf{K}\nabla w^{n+1} + \frac{K_2}{4\eta}\left\|\nabla u^{n+1} - \nabla u^n\right\|_{L^2(\Omega)}^2 \right) \tag{4.7}$$

for all $\eta > 0$. Substituting (4.7) into (4.6) yields

$$E[u^{n+1}] - E[u^n] \leq \frac{1}{2}\left(\frac{L\delta t b_2 K_2}{4\eta} - A_1\right)\left\|\nabla u^{n+1} + \nabla u^n\right\|^2_{L^2(\Omega)} + \delta t\left(\frac{Lb_2\eta}{2} - b_1\right)\int_\Omega \nabla w^{n+1}\cdot\mathbf{K}\nabla w^{n+1}$$

so that $E[u^{n+1}] - E[u^n] \leq 0$ if $\delta t \leq \dfrac{4A_1\eta}{Lb_2K_2}$ and $\eta \leq \dfrac{2b_1}{Lb_2}$. Choosing the largest value of η completes the proof. □

Observe that the restriction $\delta t \leq \dfrac{8b_1A_1}{L^2b_2^2K_2}$ makes sense intuitively, as rapid evolution of u (i.e., large b_2 and K_2) as well as sharp fronts (i.e., small A_1 and large L) should both require a small time-step. However, as will be seen in the next theorem, if we introduce a large enough relaxation, we will not have any stability restrictions on δt. This is essentially due to the viscous relaxation causing $E[u]$ to decrease faster in the continuous setting as seen in Chapter 2.

Theorem 4.2. *Let* $\mathbf{A}$, $\mathbf{K}$, b, f, *and* ϕ *be as in Theorem 3.1. Suppose* u^{n+1}, u^n, w^{n+1}, *and* w^n *are all* $H^1(\Omega)$ *functions which satisfy*

$$\frac{1}{\delta t}\int_\Omega (u^{n+1} - u^n)\psi + \int_\Omega b(u^n)\nabla w^{n+1}\cdot\mathbf{K}\nabla\psi = 0 \tag{4.8a}$$

$$\int_\Omega w^{n+1}\zeta = \int_\Omega (f(u^n) + \phi)\zeta + \int_\Omega \nabla u^{n+1}\cdot\mathbf{A}\nabla\zeta + S\int_\Omega (u^{n+1} - u^n)\zeta \tag{4.8b}$$

for all $\zeta \in H^1(\Omega)$. *If* $S \geq L/2$, *then* $E[u^{n+1}] \leq E[u^n]$ *for any* $\delta t > 0$.

Proof. Upon setting $\psi = w^{n+1}$, $\zeta = u^{n+1} - u^n$, and following the same procedure as

in the proof of Theorem 4.1, we see that

$$E[u^{n+1}] - E[u^n] \leq \frac{L}{2}\left\|u^{n+1} - u^n\right\|^2_{L^2(\Omega)} - \frac{1}{2}\int_\Omega \nabla(u^{n+1} - u^n)\cdot \mathbf{A}\nabla(u^{n+1} - u^n) - \delta t \int_\Omega b(u^n)\nabla w^{n+1}\cdot \mathbf{K}\nabla w^{n+1} - S\left\|u^{n+1} - u^n\right\|^2_{L^2(\Omega)}.$$

Recalling that $\mathbf{A}$ and $\mathbf{K}$ are positive definite and $b(u^n)$ is positive completes the proof. □

It should be noted that the stability theorems given above allow the flexibility of allowing δt to vary so that the theorems remain valid even if we were to use an adaptive time-step method.

To get an idea of the size of the consistency error introduced by several choices for the relaxation parameter S, we applied the method outlined above to the case when $\Omega = (0,1) \subset \mathbb{R}$, $\phi = 0$, $b(u) = 0.01$, $F(u) = 5u^2(1-u)^2$, $\mathbf{K} = 1$, and $\mathbf{A} = 0.001$. The initial condition g was a linear interpolation over 500 equally spaced nodes with values sampled from the normal distribution $\mathcal{N}(0.5, 0.01)$. Assuming that the solution u remains bounded in the interval $[0,1]$, we can set $L = 10$ which gives the stability restriction of $\delta t \leq 8\cdot 10^{-3}$, but in order to focus on the consistency error, computations were done using $\delta t = 10^{-4}$. The error at the finial simulation time $T = 10$ is summarized in Figure 4.1.

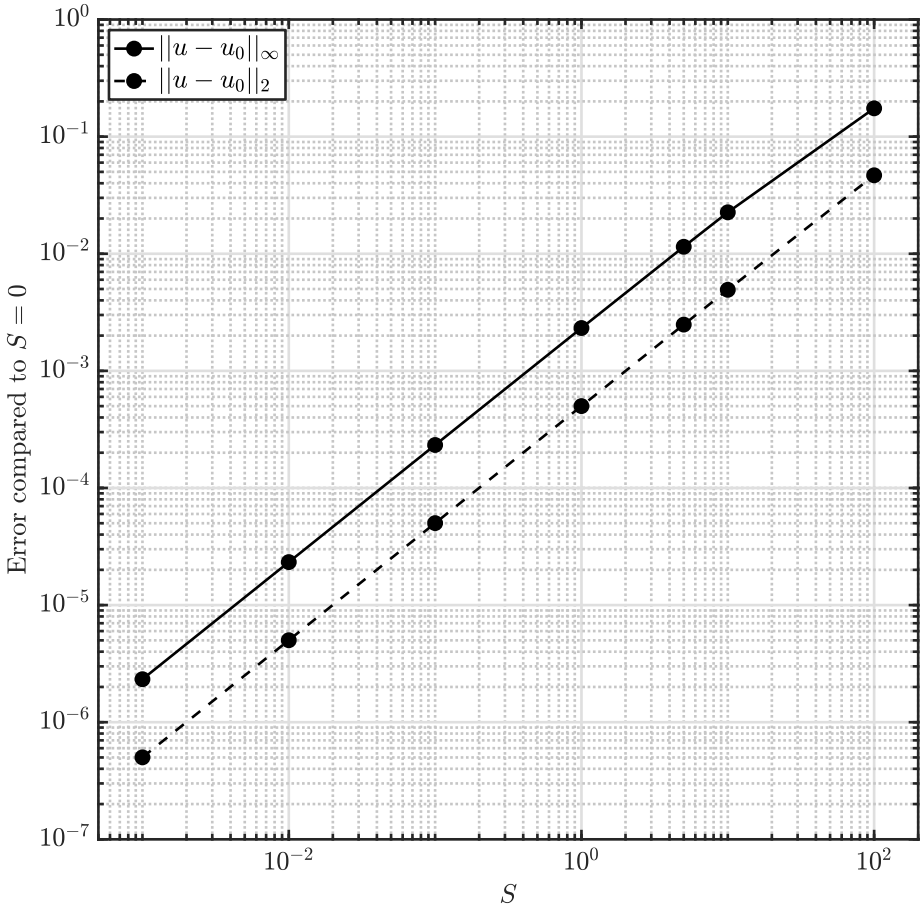

Figure 4.1: Consistency error at $T = 10$

4.3 Arc-length Coordinates

We have mentioned several times that solutions can alternate between periods of nearly steady-state and extremely rapid evolution. In order to make a simulation as efficient as possible, we would like to use a large time-step during periods where u

is nearly constant and a small time-step during periods where u is rapidly changing. These periods can be seen by monitoring the value of $E[u]$ as shown in Figure 4.2, which was generated using the same parameters as given at the end of the previous section with the exception that 200 computational nodes were used and $S = 5$.

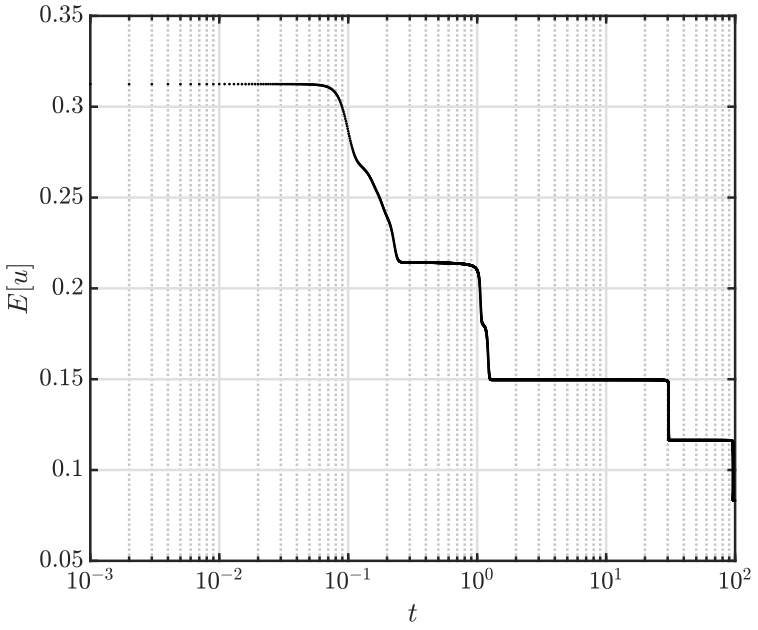

Figure 4.2: A typical energy profile

Because we know that

$$\frac{d}{dt}E[u] = -\int_{\Omega} b(u)\nabla w \cdot \mathbf{K}\nabla w,$$

we can (for example) compare $-\delta t \int_{\Omega} b(u^{n+1})\nabla w^{n+1} \cdot \mathbf{K}\nabla w^{n+1}$ to $E[u^{n+1} - E[u^n]$ as the basis for an adaptive time-step technique, which would necessitate the use and tuning of parameters related to both the tolerance of the comparison and the method

used to increase or decrease δt [31, 39]. We propose instead to perform a change of variables to achieve the desired effect. Let

$$s(t) = \int_0^t \sqrt{1 + \left| \frac{d}{d\tau} E[u(\tau)] \right|^2} \, d\tau \tag{4.9}$$

be the arc-length of the graph of $E[u]$ and note that s is a strictly increasing function and thus invertible. Note that the time-derivative of s is given by

$$s_t(u, w) = \sqrt{1 + \left(\int_\Omega b(u) \nabla w \cdot \mathbf{K} \nabla w \right)^2} \tag{4.10}$$

and can be expressed entirely in terms of the primary unknowns u and w without requiring additional regularity. We now make a change of variables from t to s to get the weak form

$$s_t(u, w) \left\langle u_s, \psi \right\rangle_{H^{-1}(\Omega), H^1(\Omega)} + \int_\Omega b(u) \nabla w \cdot \mathbf{K} \nabla \psi = 0 \qquad \forall \psi \in H^1(\Omega) \tag{4.11a}$$

$$\int_\Omega w \zeta = \int_\Omega (f(u) + \phi) \zeta + \int_\Omega \nabla u \cdot \mathbf{A} \nabla \zeta \qquad \forall \zeta \in H^1(\Omega) \tag{4.11b}$$

$$t_s = 1 / s_t(u, w) \tag{4.11c}$$

which we can use to approximate u, w, and t at various values of s. The analogous IMEX scheme to (4.5) is

$$\frac{s_t^n}{\delta s} \int_\Omega (u^{n+1} - u^n) \psi + \int_\Omega b(u^n) \nabla w^{n+1} \cdot \mathbf{K} \nabla \psi = 0 \qquad \forall \psi \in H^1(\Omega) \tag{4.12a}$$

$$\int_\Omega w^{n+1} \zeta = \int_\Omega (f(u^n) + \phi) \zeta + \int_\Omega \nabla u^{n+1} \cdot \mathbf{A} \nabla \zeta \qquad \forall \zeta \in H^1(\Omega) \tag{4.12b}$$

$$\frac{1}{\delta s} \left(t^{n+1} - t^n \right) = \frac{1}{s_t^n} \tag{4.12c}$$

where $\delta s > 0$ and $s_t^n = s_t(u^n, w^n)$. This particular scheme is effectively an adaptive time-step technique for (4.5) with

$$\delta t_n = \frac{\delta s}{\sqrt{1 + \left(\int_\Omega b(u^n)\nabla w^n \cdot \mathbf{K}\nabla w^n\right)^2}},$$

but we have the additional option of adapting δs. For example, we could compare δs to $\sqrt{(t^{n+1} - t^n)^2 + \left(E[u^{n+1}] - E[u^n]\right)^2}$ and increase or decrease δs based on some tolerance between these two values. If we wish to still use the viscous relaxation, we propose to use the scheme

$$\frac{s_t^n}{\delta s}\int_\Omega (u^{n+1} - u^n)\psi + \int_\Omega b(u^n)\nabla w^{n+1} \cdot \mathbf{K}\nabla\psi = 0 \qquad \forall \psi \in H^1(\Omega) \tag{4.13a}$$

$$\int_\Omega w^{n+1}\zeta = \int_\Omega (f(u^n) + \phi)\zeta + \int_\Omega \nabla u^{n+1} \cdot \mathbf{A}\nabla\zeta + s_t^n S\int_\Omega (u^{n+1} - u^n)\zeta \qquad \forall \zeta \in H^1(\Omega) \tag{4.13b}$$

$$\frac{1}{\delta s}\left(t^{n+1} - t^n\right) = \frac{1}{s_t^n} \tag{4.13c}$$

so that, in the continuous setting, we are still approximating a viscous relaxation parameter $\theta = S\delta s$ constant over all iterations. Observe that the forward Euler scheme in (4.12c) could be easily replaced with backwards Euler or Crank-Nicolson as computing s_t^{n+1} from (4.12a) and (4.12b) can be done before computing t^{n+1} and the same can be said for (4.8). We should note that Theorems 4.1 and 4.2 (with δs in place of δt) hold for (4.12) and (4.13) respectively, which can be seen by substituting $\delta t = \delta s / s_t^n$ and noting that $s_t^n \geq 1$.

Let us now compare the scheme given in (4.5) to (4.12). For this comparison, we move to $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$ while keeping $\phi = 0$, $b(u) = 10^{-2}$, $\mathbf{K} = \mathbf{I}$, $\mathbf{A} = 10^{-3}\mathbf{I}$, and $F(u) = 5u^2(1-u)^2$. For our spacial discretization, we set V_h to be the space of continuous piecewise linear functions defined on a triangulation of Ω (see Section 4.4 for more information). This mesh had 1936 nodes 3710 triangles, and in order to put the stability of these methods to the test, we sampled the uniform distribution $\mathcal{U}(0,1)$ at each node to define our initial condition (one sample and mesh held constant through all simulations). This introduces very large gradients in the initial condition and causes u_t to be quite large for small t. Note that we are not using the viscous relaxation here so our stability condition is again $\delta t \leq 8 \cdot 10^{-3}$ (taking $L = 10$), but we tested δt and δs ranging from 10^{-2} to $5 \cdot 10^{-4}$. The simulations using $\delta t = 5 \cdot 10^{-4}$ and $\delta s = 5 \cdot 10^{-4}$ were taken to be the exact solution for the corresponding methods and used to compute the error when using a larger δt or δs at times $T = 1$, $T = 10$, and $T = 100$. Figures which show the solutions, energy profiles, and error can be seen in Appendix 5, but we consistently see that changing to the arc-length format increases accuracy for comparable number of time-steps while capturing solution behavior that occurs on a time-scale orders of magnitude smaller than δs.

We conclude by emphasizing that our technique takes place in the continuous setting, not in the discrete. This method essentially reduces the stiffness of the

problem at the cost of introducing a nonlinear, nonlocal operator s_t into the problem. One should note that the computational cost of evaluating s_t is not large, and can still lead to a system of linear equations to be solved at each step if s_t is treated explicitly.

4.4 Notes on Implementation

Now that we have methods to fully discretize our problem, let us combine the temporal and spacial approximations. For the sake of being specific, let us restrict $\Omega \subset \mathbb{R}^2$. In order to define the basis functions for V_h, we first approximate Ω by a polygonal domain Ω_h and create a triangulation $\mathcal{T}^h$ of Ω_h. We follow [29] and require that $\mathcal{T}^h = \{\mathcal{T}_k\}$ is a partition of $\overline{\Omega_h}$ into closed triangles $\mathcal{T}_k$ such that $\cup_k \mathcal{T}_k = \overline{\Omega_h}$ and the interiors of $\mathcal{T}_k$ and $\mathcal{T}_l$ are disjoint if $k \neq l$. Moreover, we require that the triangulation is conforming in the sense that if $k \neq l$ and $\mathcal{T}_k \cap \mathcal{T}_l \neq \{\}$, then $\mathcal{T}_k \cap \mathcal{T}_l$ is either a common edge or vertex of the two triangles.

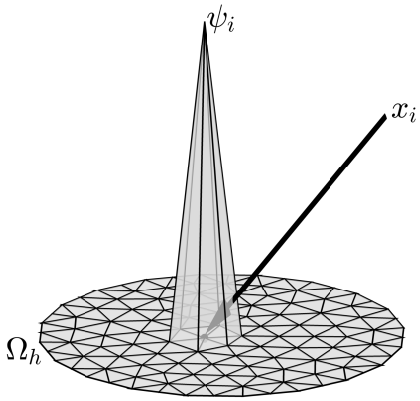

Figure 4.3: A basis function

From here, we set

$$V_h = \{\psi \in C(\overline{\Omega_h}; \mathbb{R}) \colon \psi|_{\mathcal{T}_k} \in \mathcal{P}_1(\mathcal{T}_k)\} \tag{4.14}$$

where $\mathcal{P}_1(\mathcal{T}_k)$ is the space of polynomials of degree at most 1 whose domains are restricted to $\mathcal{T}_k$. Note that V_h has a convenient basis $\{\psi_i\}_{i=1}^N$ given by $\psi_i(x_j) = \delta_{ij}$ where δ_{ij} is the Kronecker delta and $\{x_i\}_{i=1}^N$ are the vertices of the triangles in $\mathcal{T}^h$. We should note here that other choices for V_h as well as the mesh $\mathcal{T}^h$ are quite common and lead to different finite element methods [1, 29, 41]. Moreover, we can now specify that the parameter h as $h = \max_k \operatorname{diam}(\mathcal{T}_k)$ where $\operatorname{diam}(\mathcal{T}_k) = \sup_{x,y\in\mathcal{T}_k} |x - y|$ is the diameter of $\mathcal{T}_k$.

Now that we have our basis functions defined, our numerical integration to evaluate (4.3) becomes quite simple. In order to approximate the integrals

$$\mathcal{M}_{ij} = \int_\Omega \psi_i \psi_j \tag{4.3a}$$

$$\mathcal{A}_{ij} = \int_\Omega \nabla\psi_i \cdot \mathbf{A}\nabla\psi_j \tag{4.3b}$$

$$\mathcal{B}_{lij} = \int_\Omega \psi_l \nabla\psi_i \cdot \mathbf{K}\nabla\psi_j, \tag{4.3c}$$

we note that $\int_\Omega \psi_i\psi_j \approx \int_{\Omega_h} \psi_i\psi_j = \sum_{k=1}^{M''} \int_{\mathcal{T}_k} \psi_i\psi_j$ so that we need only approximate these integrals over each $\mathcal{T}_k$.

Proposition 4.2. *Let $\mathcal{T}_k \subset \mathbb{R}^2$ be a triangle with vertices x_1, x_2, x_3 and edge midpoints y_1, y_2 and y_3. Then if p is a polynomial of degree at most 1 over $\overline{\mathcal{T}_k}$, we*

have

$$\int_{\mathcal{T}_k} p = \frac{|\mathcal{T}_k|}{3}(p(x_1) + p(x_2) + p(x_3)),$$

and if p is a polynomial of degree at most 2 over $\overline{\mathcal{T}_k}$, then

$$\int_{\mathcal{T}_k} p = \frac{|\mathcal{T}_k|}{3}(p(y_1) + p(y_2) + p(y_3)).$$

Proof. First, note that we can make an affine change of variables from $\mathcal{T}_0$ to $\mathcal{T}_k$, where $\mathcal{T}_0$ is the triangle with vertices $(0,1)$, $(1,0)$, and $(0,-1)$. Represent this change of variables by $\xi = A\eta + b$ where $\xi \in \mathcal{T}_k$ and $\eta \in \mathcal{T}_0$. Then

$$\int_{\mathcal{T}_k} p(\xi)\, d\xi = \int_{\mathcal{T}_0} p(\eta)\, |\det(A)|\, d\eta.$$

Upon setting $p = 1$ and noting $|\mathcal{T}_0| = 1$, we see that $|\mathcal{T}_k| = |\det(A)|$ so that it is sufficient to show our theorem holds on $\mathcal{T}_0$. Further, it is sufficient to show that the theorem holds for $p(\xi) = 1, \xi_1, \xi_2, \xi_1^2, \xi_2^2$, and $\xi_1\xi_2$ as these functions form a basis for $\mathcal{P}_2$. These are easily verified by computing the appropriate double integrals. □

Now in order to calculate the integrals in (4.3), we only need to compute ψ_i on each midpoint of edges of $\mathcal{T}_k$, $\nabla\psi_i$ on each $\mathcal{T}_k$, and $|\mathcal{T}_k|$ in terms of the node set $\{x_i\}_{i=1}^M$. Additionally, we will need $\mathbf{K}$ and $\mathbf{A}$ on each $\mathcal{T}_k$. Due to the nature of how we will compute $\mathbf{K}$ from a given terrain function, we will take both $\mathbf{K}$ and $\mathbf{A}$ to be constant on each $\mathcal{T}_k$ individually. Although we are only working with the case that $\mathcal{T}_k$ is a triangle in $\mathbb{R}^2$, we show how to compute $|\mathcal{T}_k|$ and $\nabla\psi_i$ for the general case when $\mathcal{T}_k$ is a simplex in $\mathbb{R}^d$ (recall that a simplex in $\mathbb{R}^d$ is the convex hull of $d+1$ vertices).

Proposition 4.3. *Let $\mathcal{T}_k \subset \mathbb{R}^d$ be a simplex with vertices $\{x_i\}_{i=1}^{d+1}$ and $|\mathcal{T}_k| > 0$. Define $\mathbf{X} \in \mathbb{R}^{(d+1)\times(d+1)}$ as the matrix whose i-th column is $[x_i,\ 1]^\mathsf{T}$. Let ψ_i be the first degree polynomial with $\psi_i(x_j) = \delta_{ij}$ for all $i, j = 1, 2, \ldots, d+1$. Suppose that $\psi_i(\xi_1, \xi_2, \ldots, \xi_d) = \sum_{l=1}^{d} \lambda_l^{(i)} \xi_l + \lambda_{d+1}^{(i)}$ for constants $\lambda_l^{(i)}$ and let $\mathbf{\Lambda} \in \mathbb{R}^{(d+1)\times(d+1)}$ have components $\Lambda_{il} = \lambda_l^{(i)}$. Then*

(i) $|\mathcal{T}_k| = \frac{1}{d!}\,|det(\mathbf{X})|$ *and*

(ii) $\mathbf{\Lambda} = \mathbf{X}^{-1}$.

Proof. (i) Let $\hat{\mathbf{X}} \in \mathbb{R}^{d\times d}$ be the matrix with columns $x_i - x_{d+1}$ for $i = 1, 2, \ldots, d$ and $\mathcal{T}_0$ be the simplex with vertices 0, $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$ where $\boldsymbol{e}_i$ is the i-th standard basis function for $\mathbb{R}^d$. Set $A(\eta) = \hat{\mathbf{X}}\eta + x_{d+1}$ and note that $A(\mathcal{T}_0) = \mathcal{T}_k$ so that $|\mathcal{T}_k| = \int_{\mathcal{T}_k} 1 = \int_{\mathcal{T}_0} \left|\det(\hat{\mathbf{X}})\right|$. Subtracting column $d+1$ from all other columns in $\mathbf{X}$ reveals that $|\det(\mathbf{X})| = \left|\det(\hat{\mathbf{X}})\right|$, so all that remains is to show that $|\mathcal{T}_0| = \frac{1}{d!}$. This can be demonstrated inductively. The base case $d = 1$ is clearly true as $\mathcal{T}_0$ is just the interval $(0, 1)$ in this case. Assume our claim is true up to $\mathbb{R}^{d-1}$ and let $\hat{\mathcal{T}}_0$ be the projection of $\mathcal{T}_0 \subset \mathbb{R}^d$ onto the span of $\boldsymbol{e}_1$ through $\boldsymbol{e}_{d-1}$. Observing that $\mathcal{T}_0$ is the convex hull of $\boldsymbol{e}_d$ and $\hat{\mathcal{T}}_0 \times \{0\}$, we represent $\mathcal{T}_0$ as

$$\mathcal{T}_0 = \{(\theta y, 1-\theta) \in \mathbb{R}^{d-1} \times \mathbb{R} \colon y \in \hat{\mathcal{T}}_0,\ 0 \le \theta \le 1\},$$

so that

$$|\mathcal{T}_0| = \int_0^1 \mathcal{H}_{d-1}(\theta\hat{\mathcal{T}}_0)\, d\theta = \int_0^1 \theta^{d-1}\mathcal{H}_{d-1}(\hat{\mathcal{T}}_0)\, d\theta = \frac{1}{d}\mathcal{H}_{d-1}(\hat{\mathcal{T}}_0) = \frac{1}{d!},$$

where $\mathcal{H}_{d-1}$ is the $(d-1)$-surface measure.

(ii) Because $\mathbf{X}$ is invertible, it is sufficient to note that

$$\delta_{ij} = \psi_i(x_j) = \sum_{l=1}^{d} \lambda_l^{(i)} \xi_l^{(j)} + \lambda_{d+1}^{(i)} = (\mathbf{\Lambda X})_{ij}$$

so that $\mathbf{\Lambda X} = \mathbf{I}$. □

Note that Proposition 4.3 allows us to find $\nabla\psi_i$ on $\mathcal{T}_k$, we can compute $\mathcal{M}$, $\mathcal{A}$, and $\mathcal{B}$ by Algorithm 1 once we know how to find $\mathbf{K}|_{\mathcal{T}_k}$ and $\mathbf{A}|_{\mathcal{T}_k}$ from a given (approximate) elevation function $z_h \in V_h$ and surface conductivity α_{surf}. If $\mathcal{T}_k$ has vertices x_1, x_2, and x_3, then $z_h|_{\mathcal{T}_k} = z_h(x_1)\psi_1 + z_h(x_2)\psi_2 + z_h(x_3)\psi_3$ so that ∇z_h can be easily found and (2.6) used directly. In our simulations thus far, we have taken α_{surf} to be constant so that $\mathbf{K}|_{\mathcal{T}_k}$ is constant. Setting $\mathbf{A} = \alpha\mathbf{K}^{-1}$ for a fixed value of $0 < \alpha << 1$ has given promising results so far. We note here that in this algorithm, we used the notation $\mathcal{M}(i,j)$ rather than $\mathcal{M}_{ij}$ for clarity. For the best performance, this algorithm should be used to store $\mathcal{M}$, $\mathcal{A}$, and $\mathcal{B}$ in compressed coordinate (cco) format before converting to either compressed sparse row (csr) or compressed sparse column (csc) format.

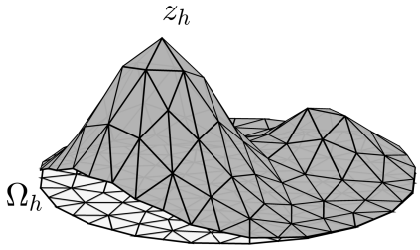

Figure 4.4: An elevation function

Now that we have a method to generate $\mathcal{M}$, $\mathcal{A}$, and $\mathcal{B}$, let us turn our attention

Algorithm 1: Compute $\mathcal{M}$, $\mathcal{A}$, $\mathcal{B}$

$\mathcal{M}(i,j) \leftarrow 0$, $\mathcal{A}(i,j) \leftarrow 0$, $\mathcal{B}(l,i,j) \leftarrow 0$;

for $\mathcal{T}_k \in \mathcal{T}^h$ **do**

look up vertices x_{i_1}, x_{i_2}, x_{i_3} for $\mathcal{T}_k$ and compute $|\mathcal{T}_k|$, $\nabla\psi_{i_m}$ for $m = 1, 2, 3$;

compute or look up $\mathbf{K}|_{\mathcal{T}_k}$ and $\mathbf{A}|_{\mathcal{T}_k}$;

for $m \leftarrow 1$ **to** 3 **do**

$\mathcal{M}(i_m, i_m) \mathrel{+}= \frac{1}{6} |\mathcal{T}_k|$;

$\mathcal{A}(i_m, i_m) \mathrel{+}= |\mathcal{T}_k| \nabla\psi_{i_m} \cdot \mathbf{A}|_{\mathcal{T}_k} \nabla\psi_{i_m}$;

for $l \leftarrow 1$ **to** 3 **do**

$\mathcal{B}(i_l, i_m, i_m) \mathrel{+}= \frac{1}{3} |\mathcal{T}_k| \nabla\psi_{i_m} \cdot \mathbf{K}|_{\mathcal{T}_k} \nabla\psi_{i_m}$;

end

for $n \leftarrow 1$ **to** $m - 1$ **do**

$\mathcal{M}(i_m, i_n)$, $\mathcal{M}(i_n, i_m) \mathrel{+}= \frac{1}{12} |\mathcal{T}_k|$;

$\mathcal{A}(i_m, i_n)$, $\mathcal{A}(i_n, i_m) \mathrel{+}= |\mathcal{T}_k| \nabla\psi_{i_m} \cdot \mathbf{A}|_{\mathcal{T}_k} \nabla\psi_{i_n}$;

for $l \leftarrow 1$ **to** 3 **do**

$\mathcal{B}(i_l, i_m, i_n)$, $\mathcal{B}(i_l, i_n, i_m) \mathrel{+}= \frac{1}{3} |\mathcal{T}_k| \nabla\psi_{i_m} \cdot \mathbf{K}|_{\mathcal{T}_k} \nabla\psi_{i_n}$;

end

end

end

end

to finally solving the discrete problem. Using (4.8) for the standard formulation gives us

$$\begin{bmatrix} \mathcal{M} & \delta t \boldsymbol{b}^n \cdot \mathcal{B} \\ -(\mathcal{A} + S\mathcal{M}) & \mathcal{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}^{n+1} \\ \boldsymbol{w}^{n+1} \end{bmatrix} = \begin{bmatrix} \mathcal{M}\boldsymbol{u}^n \\ \mathcal{M}(\boldsymbol{f}^n + \boldsymbol{\phi} - S\boldsymbol{u}^n) \end{bmatrix} \tag{4.15}$$

while using (4.13) for the arc-length formulation gives us

$$\begin{bmatrix} s_t^n \mathcal{M} & \delta s \boldsymbol{b}^n \cdot \mathcal{B} \\ -(\mathcal{A} + s_t^n S\mathcal{M}) & \mathcal{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}^{n+1} \\ \boldsymbol{w}^{n+1} \end{bmatrix} = \begin{bmatrix} \mathcal{M}\boldsymbol{u}^n \\ \mathcal{M}(\boldsymbol{f}^n + \boldsymbol{\phi} - s_t^n S\boldsymbol{u}^n) \end{bmatrix} \tag{4.16}$$

where we use the convention $\boldsymbol{u} = [u(x_1), u(x_2), \ldots, u(x_M)]^\mathsf{T}$ for $\boldsymbol{u}$, $\boldsymbol{w}$, and $\boldsymbol{\phi}$ and the convention $\boldsymbol{b} = [b(u(x_1)), b(u(x_2)), \ldots, b(u(x_M))]$ for $\boldsymbol{b}$ and $\boldsymbol{f}$. The superscript n refers to the approximate solution at $t = n\delta t$ for (4.15) or at $s = n\delta s$ for (4.16). We note here that the product $\boldsymbol{b}^n \cdot \mathcal{B}$ is a matrix with elements $(\boldsymbol{b}^n \cdot \mathcal{B})_{ij} = \sum_{l=1}^{M} b(u(x_l))\mathcal{B}_{lij}$. Upon examining the left-hand sides of (4.15) and (4.16), we see that the disadvantage to using the arc-length formulation comes when the function b is constant.

CHAPTER 5. CONCLUSION

We have covered a wide variety of topics concerning this generalized Cahn-Hilliard equation, ranging from a qualitative description in Chapter 2 to establishing existence of solutions in Chapter 3 and discussing how best to compute solutions and introducing a new framework to reduce the stiffness of the problem in Chapter 4. The topics of variable mobility, potential terms, anisotropy, and direct dependence of parameters on t are all either sparse or nonexistent in the current body of work on the Cahn-Hilliard equation so that our analysis brings new results on each of these topics. Additionally, the framework of solving the Cahn-Hilliard equation using the arc-length of the energy profile is a new approach that appears to work quite well and warrants further exploration such as incorporating an adaptive scheme or applying it to other gradient flow problems. Our generalized equation also prompts further problems such as a deeper examination of the critical points of the energy functional (2.4) and establishing the existence of solutions when the mobility is allowed to be degenerate.

BIBLIOGRAPHY

[1] Douglas N. Arnold, Franco Brezzi, Bernardo Cockburn, and Donatella Marini. Discontinuous Galerkin Methods for Elliptic Problems. In Cockburn B., Karniadakis G.E., and Shu CW., editors, *Discontinuous Galerkin Methods. Lecture Notes in Computational Science and Engineering*, volume 11. Springer, Berlin, Heidelberg, 2000.

[2] John W. Barrett and James F. Blowey. Finite element approximation of the Cahn-Hilliard equation with concentration dependent mobility. *Mathematics of Computation*, 68(226):487–517, April 1999.

[3] John W. Barrett, James F. Blowey, and Harald Garcke. Finite Element Approximation of the Cahn-Hilliard Equation with Degenerate Mobility. *SIAM J. Numer. Anal.*, 37(1):286–318, 1999.

[4] John W. Barrett, James F. Blowey, and Harald Garcke. On fully practical finite element approximations of degenerate Cahn-Hilliard systems. *ESAIM: M2AN*, 35(4):713–748, 2001.

[5] Andrew J. Bernoff and Chad M. Topaz. Biological Aggregation Driven by Social and Environmental Factors: A Nonlocal Model and Its Degenerate Cahn-Hilliard Approximation. *SIAM J. Applied Dynamical Systems*, 15(3):1528–1562, 2016.

[6] Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*. Springer, third edition, 2008.

[7] John W. Cahn and John E. Hilliard. Free Energy of a Nonuniform System. i. Interfacial Free Energy. *J. Chem. Phys.*, 28(2):258–267, July 1957.

[8] Jack Carr and Morton E. Gurtin. Structured phase transitions on a finite interval. *Arch. Rational Mech. Anal.*, 86:317–351, 1984.

[9] Wenbin Chen, Cheng Wang, Xiaoming Wang, and Steven M. Wise. Positivity-preserving, energy stable numerical schemes for the Cahn-Hilliard equation with logarithmic potential. *Journal of Computational Physics: X*, 3, 2019.

[10] M. I. M. Copetti and Charles M. Elliott. Numerical analysis of the Cahn-Hilliard equation with a logarithmic free energy. *Numer. Math.*, 63:39–65, 1992.

[11] L. Beirão da Veiga, G. Manzini, and L. Mascotto. A posteriori error estimation and adaptivity in *hp* virtual elements. *Numerische Mathematik*, 143:139–175, 2019.

[12] Shibin Dai and Qiang Du. Weak Solutions for the Cahn-Hilliard Equation with Degenerate Mobility. *Arch. Rational Mech. Anal.*, 2016.

[13] Qiang Du and R. A. Nicolaides. Numerical analysis of a continuum model of phase transition. *SIAM J. Numer. Anal.*, 28(5):1310–1322, October 1991.

[14] M. Efendiev and A. Miranville. New models of Cahn-Hilliard-Gurtin equations. *Continuum Mech. Thermodn.*, 16:441–451, 2004.

[15] Charles M. Elliott. Numerical Studies of the Cahn-Hilliard Equation for Phase Separation. *IMA Journal of Applied Mathematics*, 38:97–128, 1987.

[16] Charles M. Elliott and Donald A. French. A Nonconforming Finite-Element Method for the Two-Dimensional Cahn-Hilliard Equation. *SIAM J. Numer. Anal.*, 26(4):884–903, 1989.

[17] Charles M. Elliott and Harald Garcke. On the Cahn-Hilliard Equation with Degenerate Mobility. *SIAM J. Math. Anal.*, 27(2):402–423, 1996.

[18] C.M. Elliott, D.A. French, and F.A. Milner. A second order splitting method for the Cahn-Hilliard equation. *Numer. Math.*, 54:575–590, 1989.

[19] D. Engwirda. *Locally-optimal Delaunay-refinement and optimisation-based mesh generation*. PhD thesis, School of Mathematics and Statistics, The University of Sydney, http://hdl.handle.net/2123/13148, 2014.

[20] Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society, 2nd edition, 2010.

[21] Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 1992.

[22] Xiaobing Feng and Ohannes A. Karakashian. Fully discrete dynamic mesh discontinuous Galerkin methods for the Cahn-Hilliard equation of phase transition. *Mathematics of Computation*, 76(259):1093–1117, March 2007.

[23] Xiaobing Feng, Yukun Li, and Yulong Xing. Analysis of Mixed Interior Penalty Discontinuous Galerkin Methods for the Cahn-Hilliard Equation and the Hele-Shaw Flow. *SIAM J. Numer. Anal.*, 54(2):825–847, 2016.

[24] Mark S. Gockenbach. *Understanding and Implementing the Finite Element Method*. Society for Industrial and Applied Mathematics, 2006.

[25] David Gottlieb and Steven A. Orszag. *Numerical Analysis of Spectral Methods: Theory and Applications*. Society for Industrial and Applied Mathematics, 1977.

[26] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman Publishing, 1985.

[27] Morton E. Gurtin. Generalized Ginzburg-Landau and Cahn-Hilliard equations based on a microforce balance. *Physica D*, 92:178–192, 1996.

[28] David Kay, Vanessa Styles, and Endre Süli. Discontinuous Galerkin Finite Element Approximation of the Cahn-Hilliard Equation with Convection. *SIAM J. Numer. Anal.*, 47(4):2660–2685, 2009.

[29] Peter Knabner and Lutz Angermann. *Numerical Methods for Elliptic and Parabolic Partial Differential Equations*. Springer, 2003.

[30] Erwin Kreyszig. *Introductory Functional Analysis with Applications*. Wiley, 1989.

[31] Yibao Li, Yongho Choi, and Junseok Kim. Computationally efficient adaptive time step method for the Cahn-Hilliard equation. *Computers and Mathematics with Applications*, 73:1855–1864, 2017.

[32] Elliott H. Lieb and Michael Loss. *Analysis*, volume 14 of *Graduate Studies in Mathematics*. American Mathematical Society, 2001.

[33] Robert Osserman. The isoparametric inequality. *Bulletin of the American Mathematical Society*, 84(6), 1978.

[34] R. L. Pego. Front migration in the nonlinear Cahn-Hilliard equation. *Proc. R. Soc. Lond. A*, 422:261–278, 1989.

[35] Giulio Schimperna. Global attractors for Cahn-Hilliard equations with nonconstant mobility. *Nonlinearity*, 20(10):2365–2387, 2007.

[36] Lynn Schreyer, Nikos Voulgarakis, Zachary Hilliard, Sergey Lapin, and Loren Cobb. Modeling Refugee Flow and Mammal Migration Based on a Continuum

Mechanics Phase-Field Approach of Porous Media. *SIAM Journal on Applied Mathematics (in review)*, 2020.

[37] Jie Shen and Xiaofeng Yang. Numerical approximations of allen-cahn and cahn-hilliard equations. *Discrete and Continuous Dynamical Systems - A*, 28(4):1669–1691, 2010.

[38] Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, 2011.

[39] G. Tierra and F. Guillén-González. Numerical Methods for Solving the Cahn-Hilliard equation and its applicability to related Energy-based models. *Arch. Computat. Methods Eng.*, 22:269–289, 2015.

[40] R. Verfürth. A posteriori error estimation and adaptive mesh-refinement techniques. *Journal of Computational and Applied Mathematics*, 50:67–83, 1994.

[41] Junping Wang and Xiu Ye. The Basics of Weak Galerkin Finite Element Methods. *arXiv:1901.10035v1*, 2019.

[42] Olga Wodo and Baskar Ganapathysubramanian. Computationally efficient solution to the Cahn-Hilliard equation: Adaptive implicit time schemes, mesh sensitivity analysis and the 3D isoperimetric problem. *Journal of Computational Physics*, 230:6037–6060, 2011.

[43] Yinhua Xia, Yan Xu, and Chi-Wang Shu. Local discontinuous Galerkin methods for the Cahn-Hilliard type equations. *Journal of Computational Physics*, 227:472–491, 2007.

[44] E. C. Zachmanoglou and Dale W. Thoe. *Introduction to Partial Differential Equations with Applications*. Dover, 1986.

APPENDIX

Here we present figures to complete the comparison between two numerical methods as discussed at the end of Section 4.3. The symbols u^t and u^s refer to solutions generated by using $\delta t = 5 \cdot 10^{-4}$ and $\delta s = 5 \cdot 10^{-4}$ respectively. The solutions here were all solved on the same mesh, which was generated using `MESH2D` in MATLAB with a maximum diameter of 0.025 [19]. The initial condition u^0 was sampled from the uniform distribution $\mathcal{U}(0, 1)$ at each computational node and used for all simulations. The double well function was $F(u) = 5u^2(1-u)^2$, the potential function was $\phi = 0$, $b(u)\mathbf{K} = 10^{-2}\mathbf{I}$, and $\mathbf{A} = 10^{-3}\mathbf{I}$.

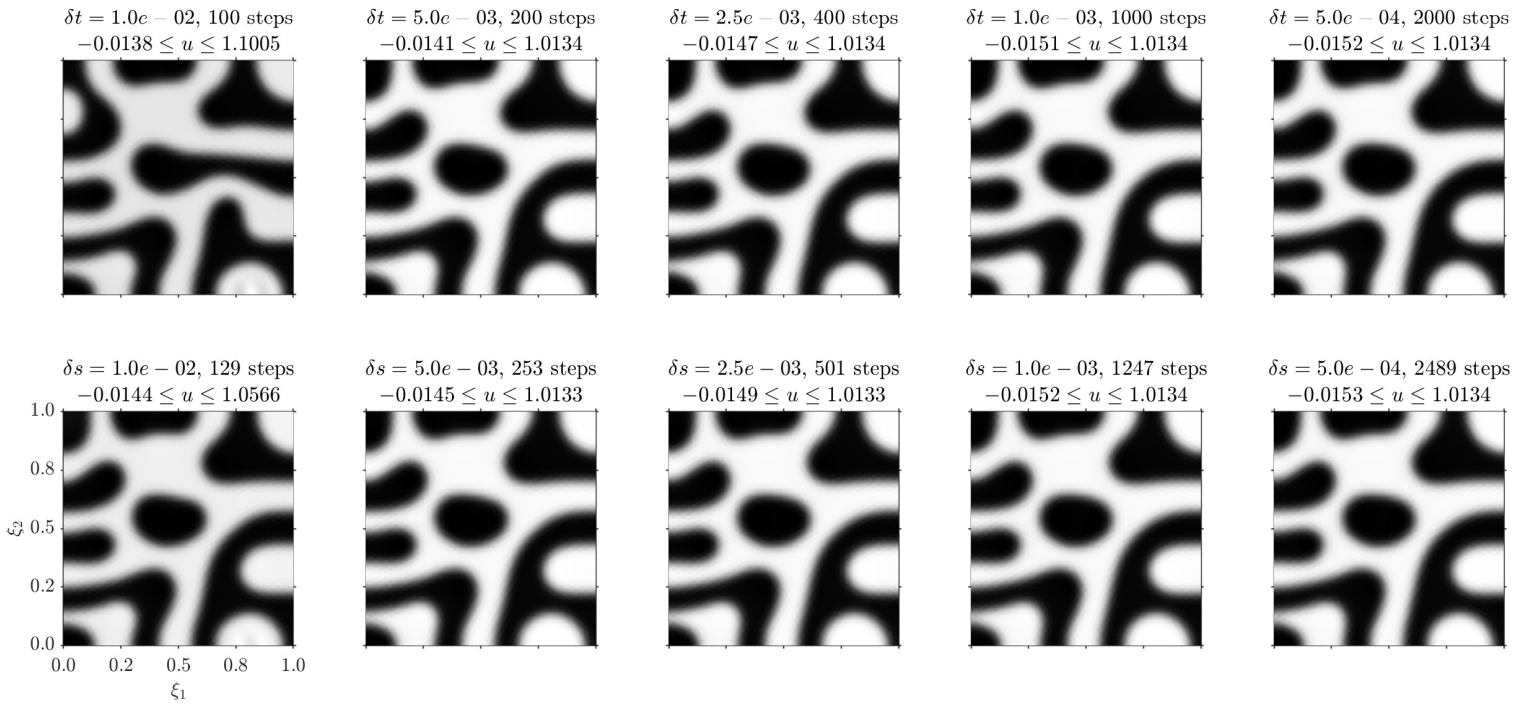

Figure A.1: Top: solutions at $t = 1$ using constant δt. Bottom: solutions at $t = 1$ using constant δs.

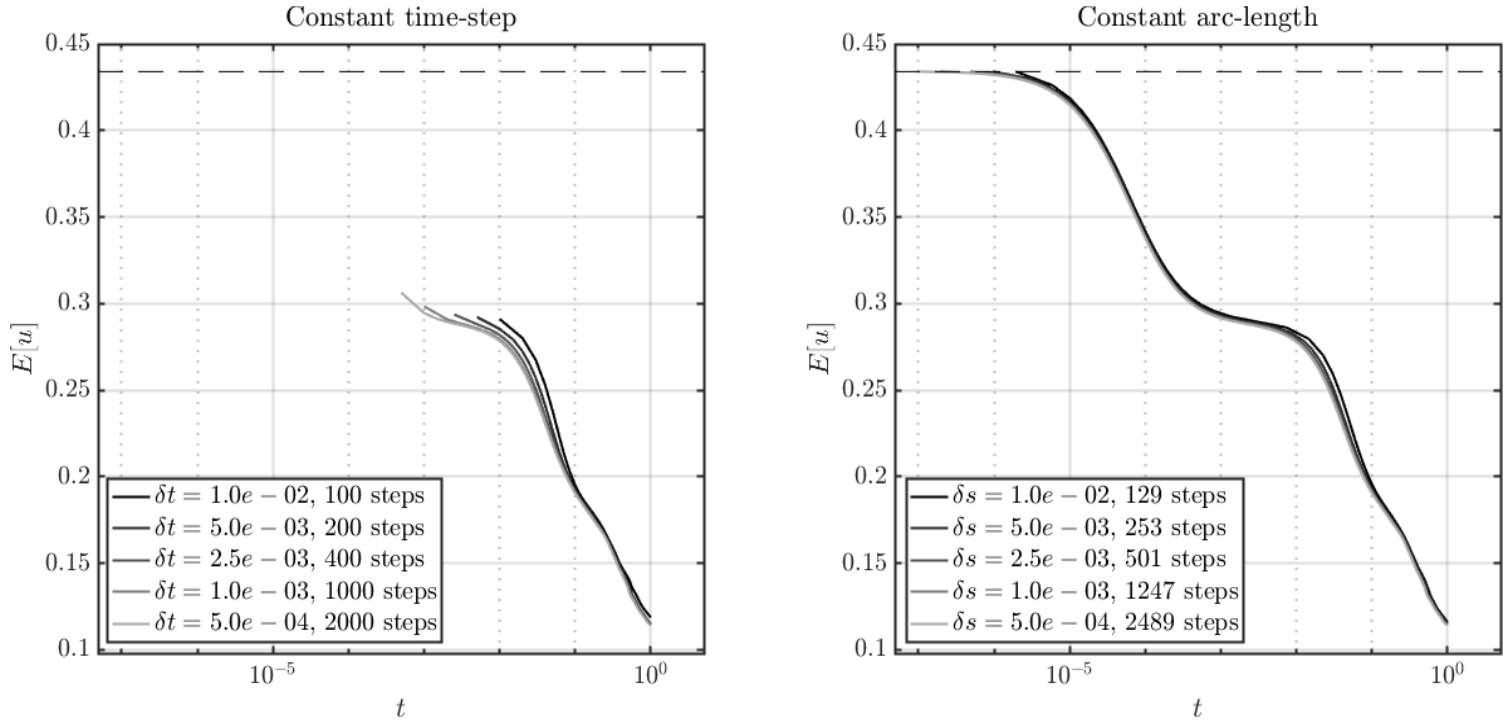

Figure A.2: Left: energy profiles when δt is constant. Right: energy profiles when δs is constant.

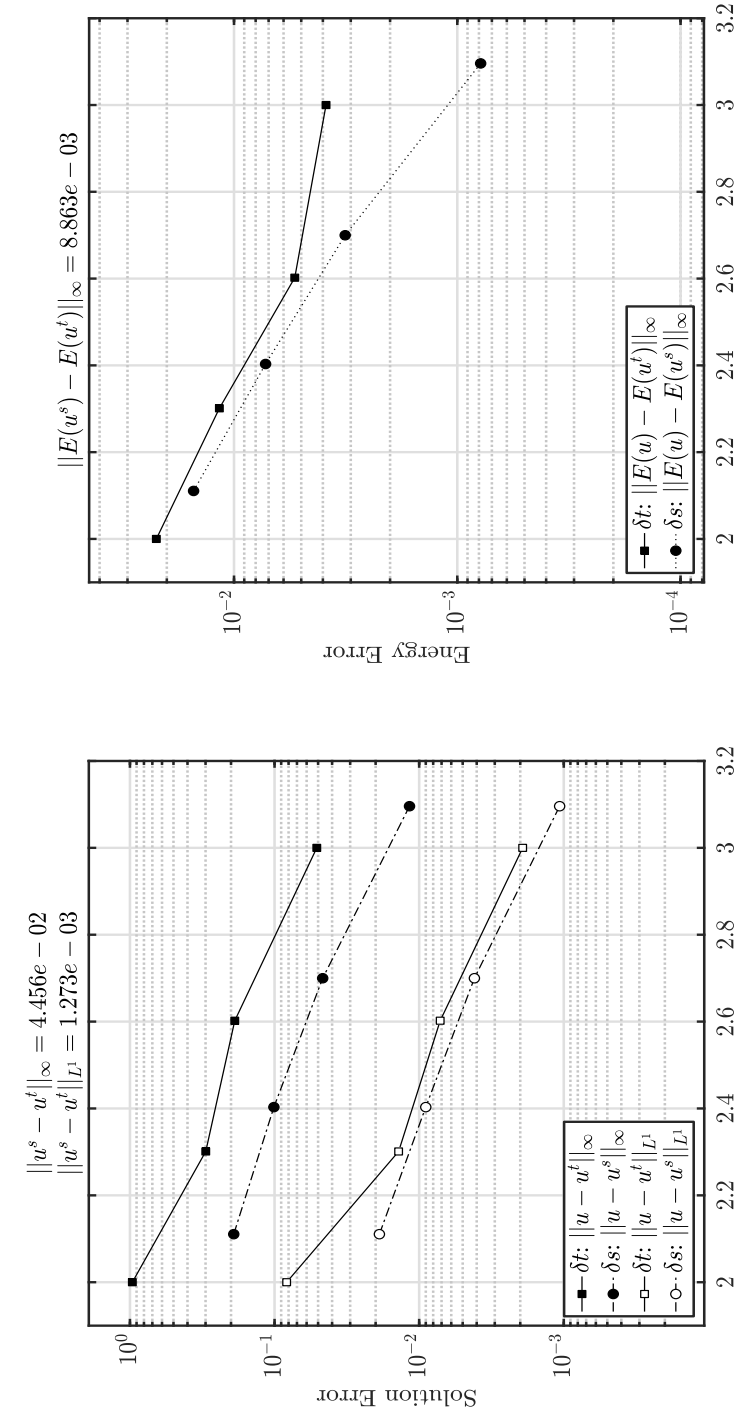

Figure A.3: Left: error in solutions at $t = 1$. Right: maximum error in energy profile over $[0, 1]$.

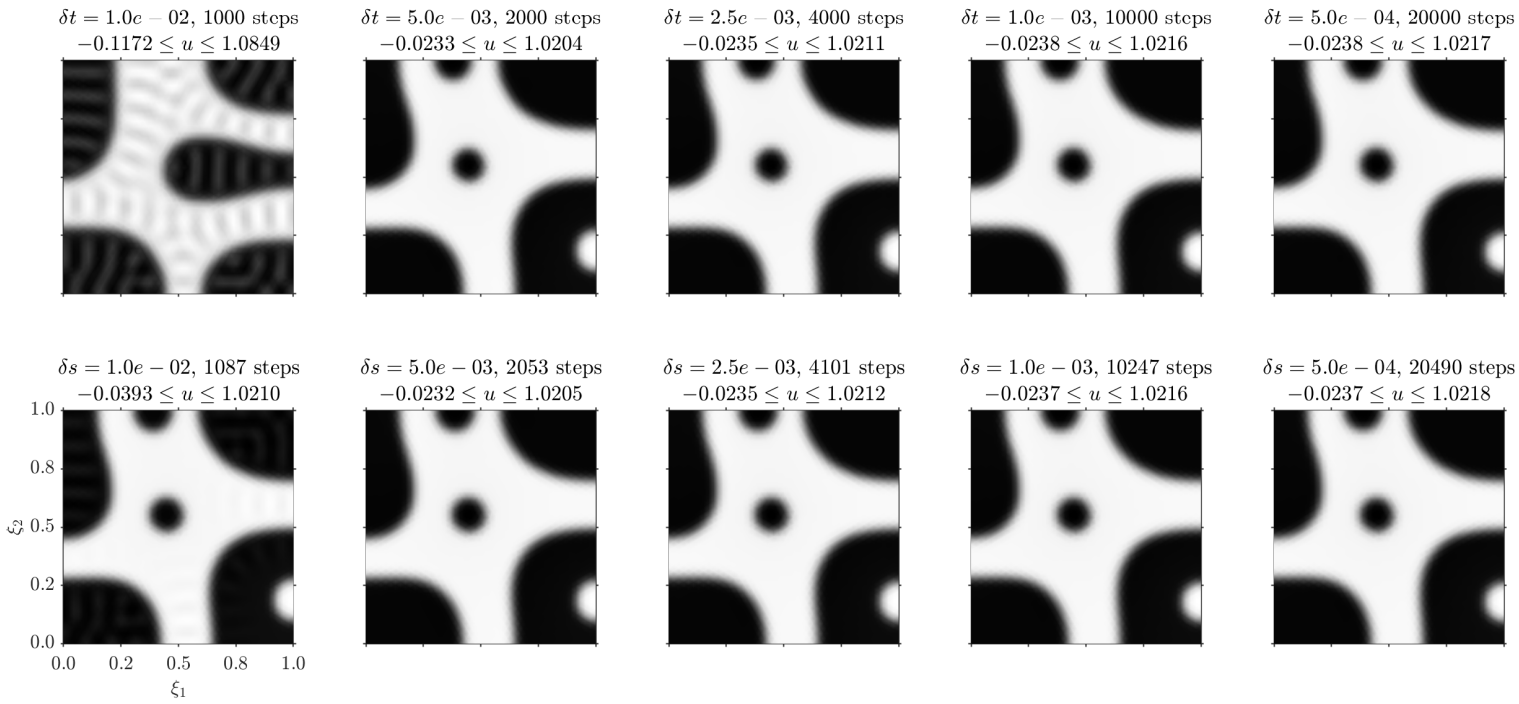

Figure A.4: Top: solutions at $t = 10$ using constant δt. Bottom: solutions at $t = 1$ using constant δs.

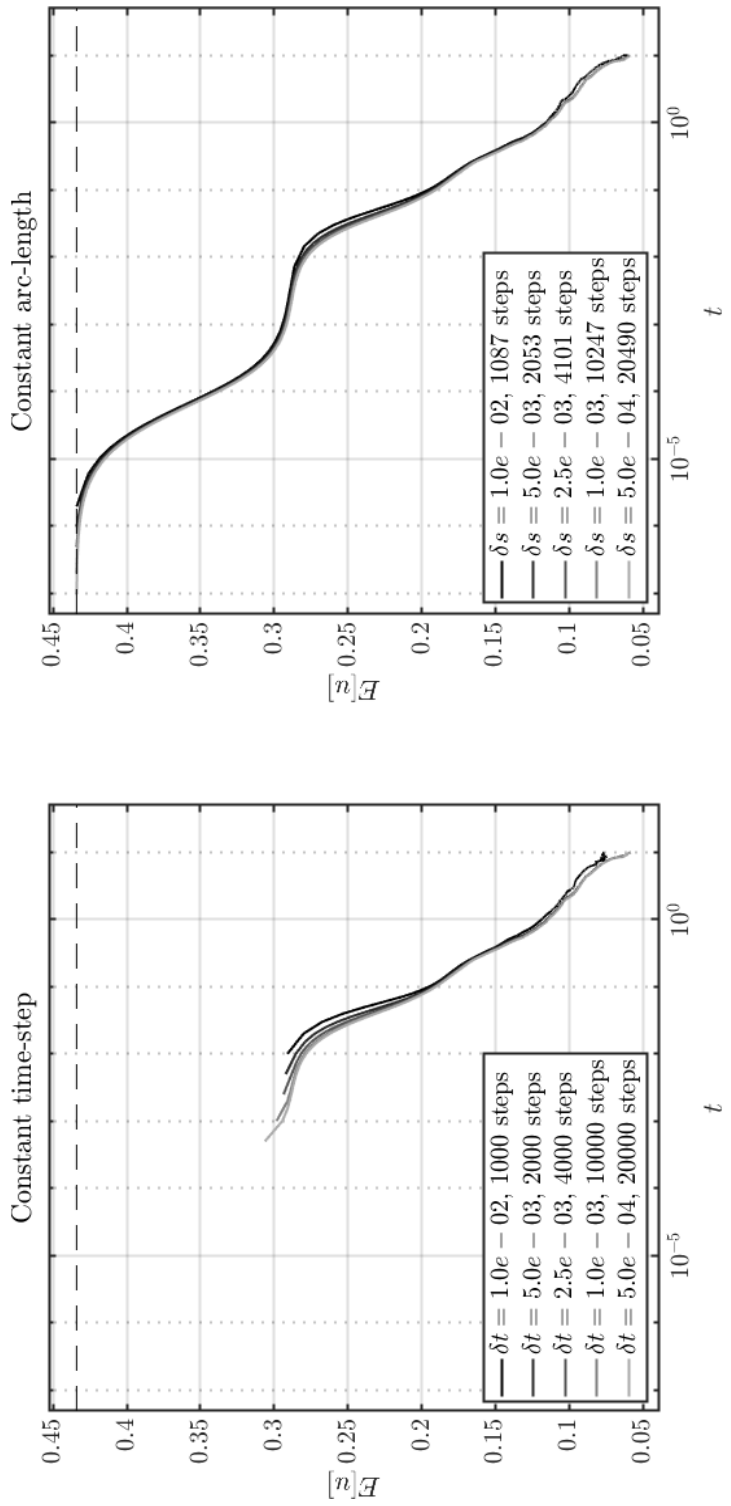

Figure A.5: Left: energy profiles when δt is constant. Right: energy profiles when δs is constant.

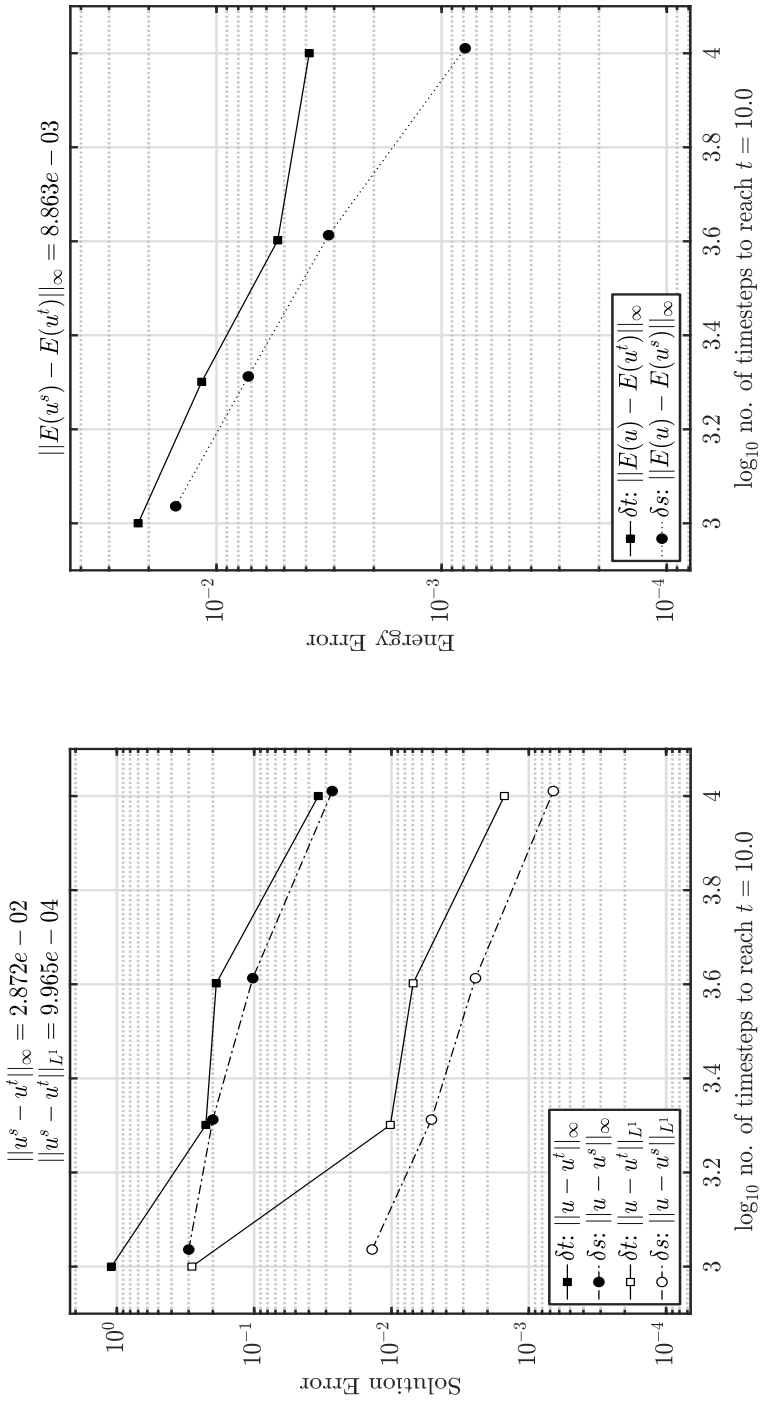

Figure A.6: Left: error in solutions at $t = 10$. Right: maximum error in energy profile over $[0, 10]$.

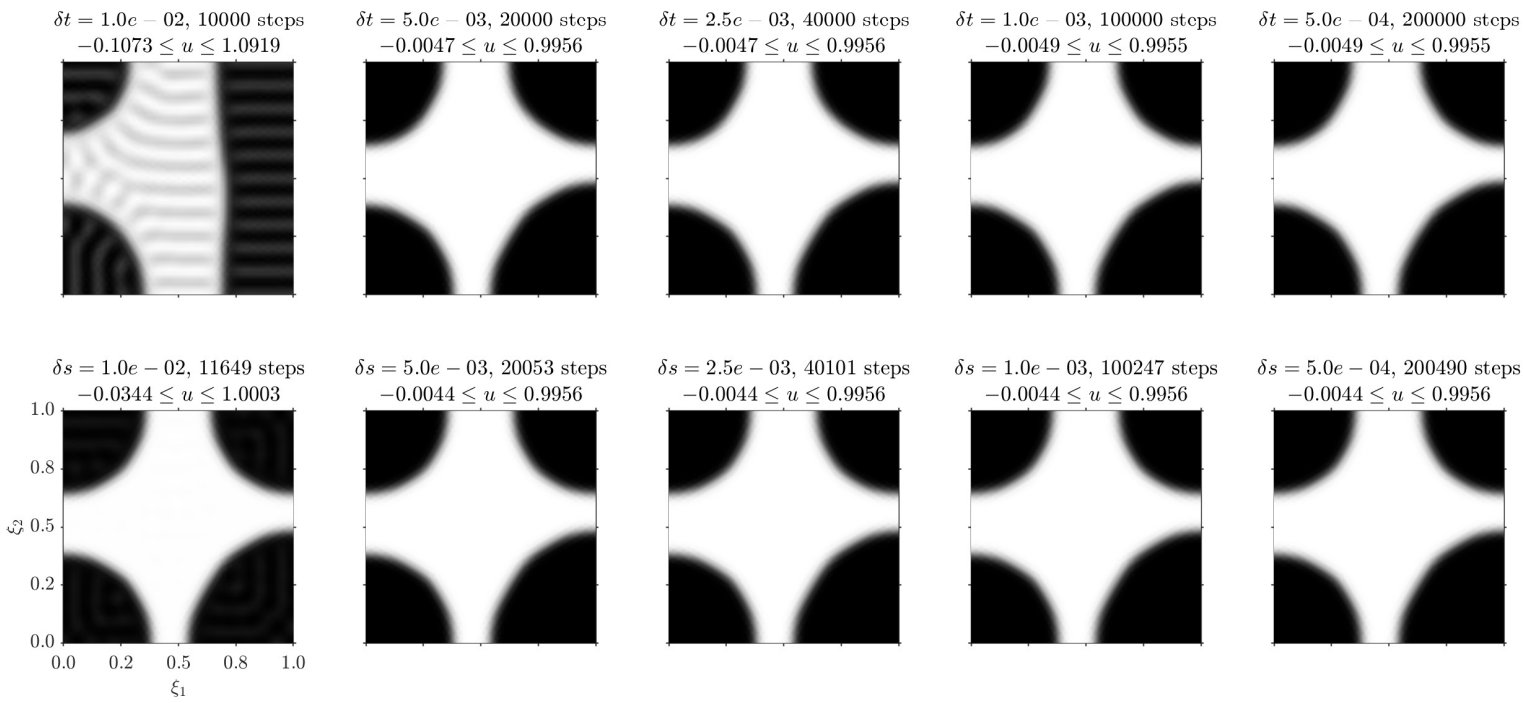

Figure A.7: Top: solutions at $t = 100$ using constant δt. Bottom: solutions at $t = 1$ using constant δs.

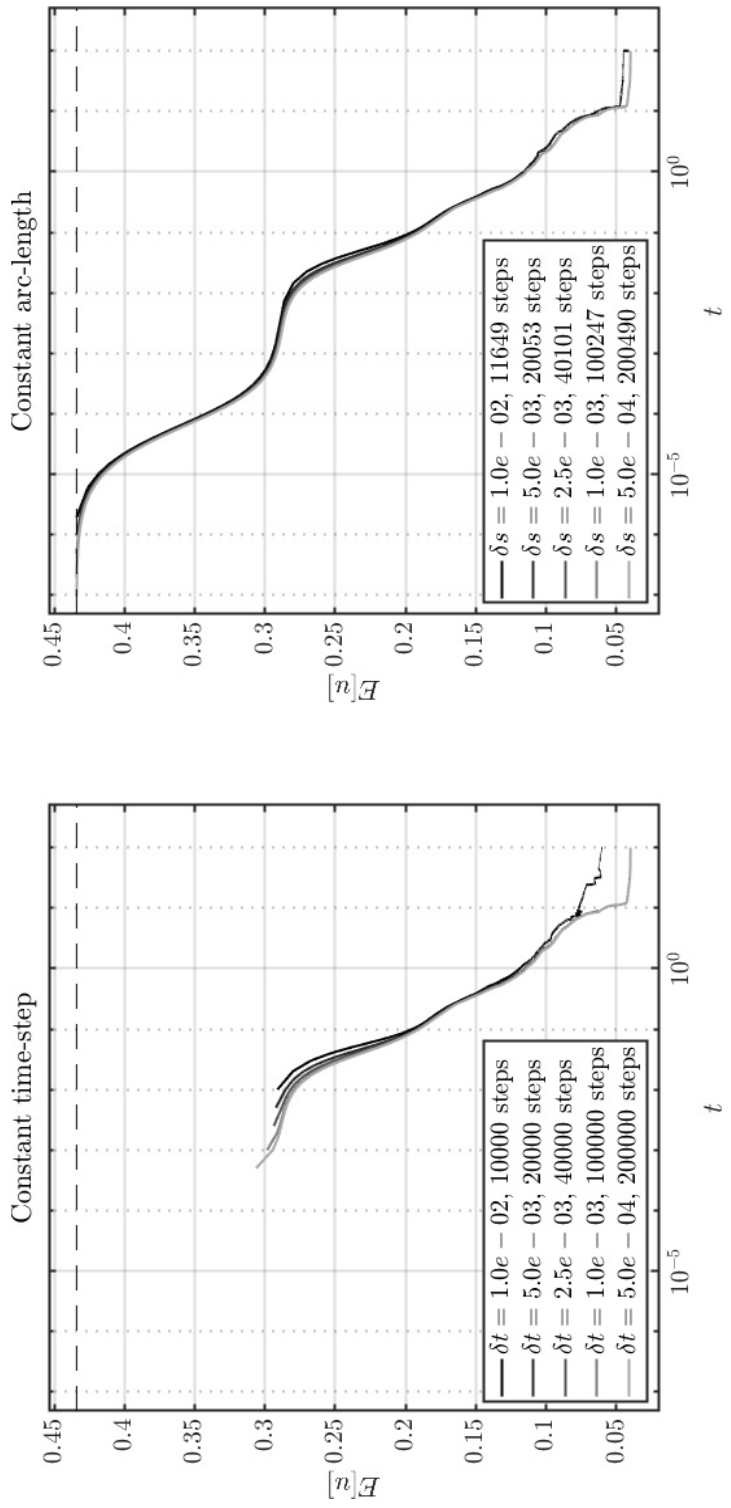

Figure A.8: Left: energy profiles when δt is constant. Right: energy profiles when δs is constant.

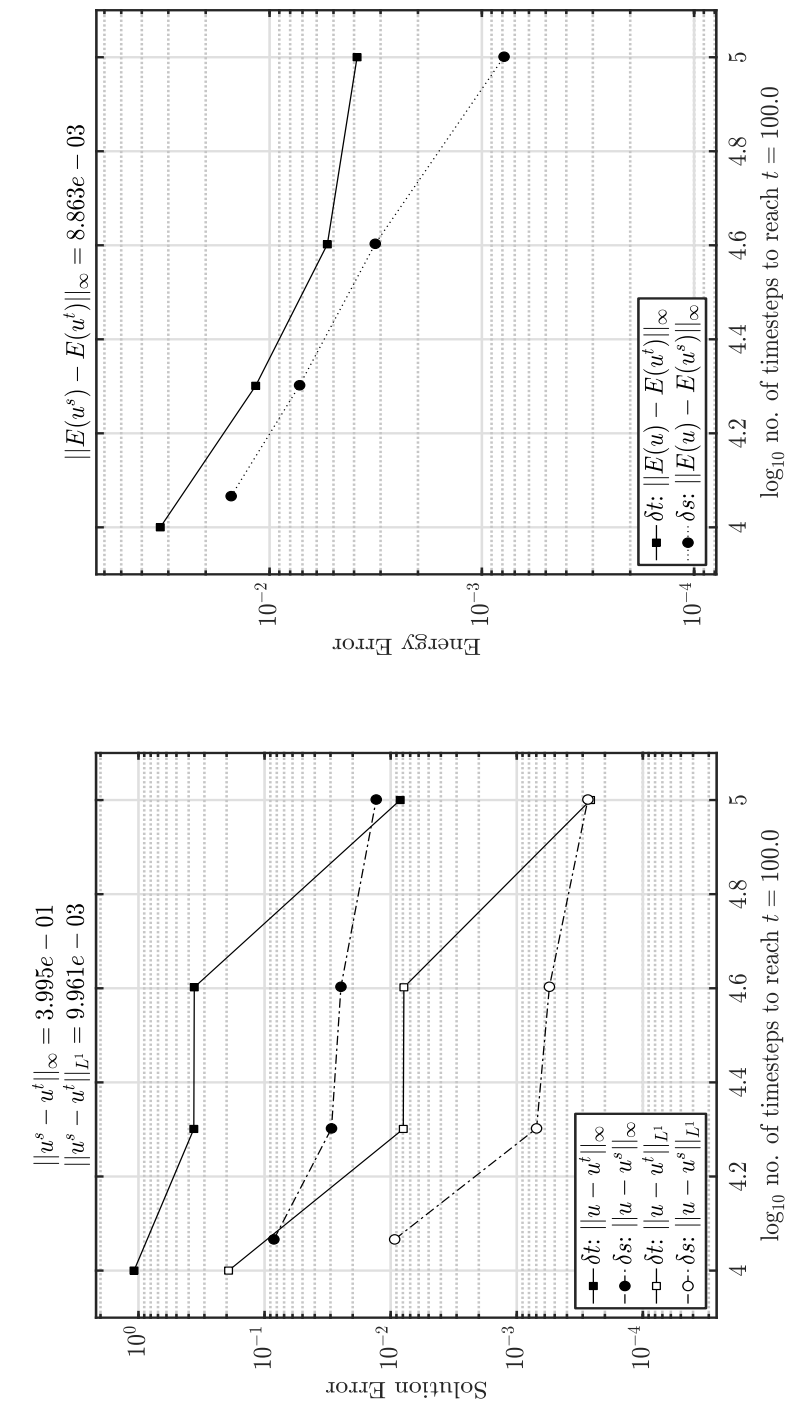

Figure A.9: Left: error in solutions at $t = 100$. Right: maximum error in energy profile over $[0, 100]$.