Analytical and Numerical Results for Phase Field, Implicit Free Boundary, and Fluid Models

by

Xinyu Cheng

B.Sc., The Chinese University of Hong Kong, 2015
M.Sc., The University of British Columbia, 2017

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Mathematics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Analytical and Numerical Results for Phase Field, Implicit Free Boundary, and Fluid Models

submitted by **Xinyu Cheng** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Mathematics**.

Examining Committee:

Brian Wetton, Professor, Mathematics, UBC

Supervisor

Tai-Peng Tsai, Professor, Mathematics, UBC

Supervisory Committee Member

Eldad Haber, Professor, EOAS, UBC

University Examiner

Ian Frigaard, Professor, Mathematics and Mechanics, UBC

University Examiner

Hongjie Dong, Professor, Applied Mathematics, Brown University

External Examiner

Additional Supervisory Committee Members:

Dong Li, Professor, Mathematics, SUSTech

Supervisory Committee Member

Stephen Gustafson, Professor, Mathematics, UBC

Supervisory Committee Member

Abstract

In this dissertation, we study analytical and numerical methods on three topics in the area of partial differential equations (PDE). These topics are: the Allen-Cahn dynamics (AC) in the study of phase field models for materials science problems, the Oxygen depletion model (OD) in the study of free boundary problems, and the stationary surface quasi-geostrophic equation (SQG) in the study of fluid dynamics. We first study the behaviour in the meta-stable regime of AC and show by computation evidence and asymptotic analysis that backward Euler method satisfies energy stability with large time steps. We also give a rigorous proof for the two-dimensional radially symmetric case. In the second project, we show several mathematical formulations of OD from the literature and give a new formulation based on a gradient flow with constraint. We prove the equivalence of all formulations and study the numerical approximations of the problem that arise from the different formulations. More general (vector, higher order) implicit free boundary value problems are discussed. In the final project, we develop a new framework of "convex integration scheme" and construct a non-trivial solution to the stationary SQG. We thus prove the non-uniqueness of the solutions to the stationary SQG.

Lay Summary

This thesis studies the Allen-Cahn dynamics (AC), oxygen depletion problem (OD) and surface quasi-geostrophic equation (SQG), which are interesting topics in the area of partial differential equations. These equations model phenomena in material science, biology and fluid dynamics. We study AC and OD dynamics by implementing backward Euler numerical schemes, while for SQG we construct a non-trivial solution by applying a newly developed convex integration scheme.

Preface

This thesis is based on original research projects by the author and the contents of research articles that are published or submitted for publication to research journals from this thesis are presented below. Contributions of collaborators in each research article will be clarified.

Chapter 3 is based on the paper "X. Cheng, D. Li, K. Promislow and B. Wetton. Asymptotic behaviour of time stepping methods for phase field models. *Journal of scientific computing*, 86(3), 1-34, 2021" [19], which has been published. The framework and methodology of this project was developed by the author, D. Li, K. Promislow and B. Wetton. The author contributes 25% of the research framework, implementation of the new method and detailed computations including 50% of the work in the asymptotic analysis and rigorous radial analysis of AC with BE. The author was not involved in the computational aspect of the work.

Chapter 4 is based on the paper "X. Cheng, Z. Fu and B. Wetton. Equivalent formulations of the oxygen depletion problem, other implicit free boundary value problems, and implications for numerical approximation, arXiv:2105. 03538, 2021" [17], which has been submitted and put on arXiv. The research and preparation of the manuscript were done by the author, Z. Fu and B. Wetton in equal parts. The author contributes 50% to the research framework and methodology including the rigorous analysis of different formulations and proof of the equivalence. The author was not involved in the computational aspect of this work.

Chapter 5 is based on the paper "X. Cheng, H. Kwon and D. Li. Non-uniqueness of steady-state weak solutions to the surface quasi-geostrophic equations. arXiv:2007.09591, 2020" [18], which has been submitted and under review. The development of the new systematic method and the manuscript composition were carried out together by the author, H. Kwon and D. Li. The author contributes 33.3% to the work including developing original research scheme, parameter computations and manuscript composition.

Table of Contents

List of Figures

Acknowledgements

Firstly, I would like to thank my supervisors Dr. Dong Li and Dr. Brian Wetton who helped me find suitable projects. In addition, they have spent many efforts on helping me build qualified mathematics skills and develop interests in diverse areas. Moreover, they took much time in providing me with valuable ideas and enlightening suggestions, which are crucial to accomplishment of this thesis.

Secondly, I would like to express my thanks to my collaborators Dr. Hyunju Kwon, Dr. Keith Promislow for the fruitful discussions and suggestions.

Last but not least I would like to thanks my friends, colleagues and officemates for their warmness and caring that help me enjoy doing research and come up with more ideas. I give a list of names below and I do apologize if any name is missing: A. Bulut, Z. Fu, C. Griffith, C. Lai, N. Lai, T. Rüd, L. Wang, J. Xu, T. Yang, Y. Zhou, etc.

Dedication

I dedicate this thesis:

In memory of my grandfather, Mr. Shiyan Cheng for his efforts to support the family during the difficult times.

In memory of my math teacher from high school, Mr. Jiannan Yang for showing me the beauty of mathematics.

To my firmest supporters, my parents and my fiancée: Mrs. Xiaodong Huang, Mr. Chun Cheng and Ms. Jiaying Catherine Wu, for always being there for me.

Chapter 1

Introduction

Partial differential equations (PDE) often describe mathematical models of physical, biological, or financial phenomena. For example, generalized nonlinear wave equations describe the properties of waves including sound waves, light waves, elastic waves and other waves, which help us to study physical phenomena including noise and music [26], electromagnetics [32, 61] and fluid dynamics [71].

Equations that accurately describe these physical phenomena are the basis of simulation tools that allow inexpensive virtual experiments and design optimization. It is important to understand the properties of the equations well in order to design accurate and efficient computational approximation methods. Analytic properties of the equations can themselves give insight into the behavior of the systems.

To study PDE, we are interested in the solutions to a partial differential equation in some domain under specific initial conditions and boundary conditions. For example, in the study of Tsunamis, the domain is the ocean, the boundary conditions describe how the waves interact with the shore and the initial conditions could be an approximation of the ocean's early response to an earthquake. More information about the solutions can help the physical models to be better understood. In mathematics, usually the focus of study is on the existence, uniqueness, regularity ("degree of smoothness") and the dynamical behavior of the solutions. Existence, uniqueness and regularity are often investigated with analytic tools such as fixed point theorems, the method of calculus of variation, and iteration methods, which will be introduced in Chapter 2. Although sometimes it is possible to find explicit solutions of certain simple PDEs such as Laplace's equation and heat equation, usually there are no explicit solutions. Thus, it is necessary to compute approximate solutions using computer simulations. Throughout the area of partial differential equations, it is necessary to develop well behaved numerical schemes that are guaranteed to approximate PDEs to an expected accuracy and that preserve important solution properties.

In the spirit of these ideas (analytic properties of PDEs and their relationship to their numerical approximation), three main research topics will

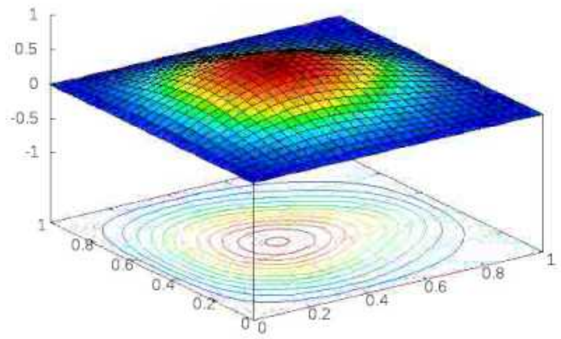

Figure 1.1: Computational Simulation of 2D Waves [94]

be presented in this dissertation. These topics are: the Allen-Cahn dynamics (AC), the oxygen depletion problem (OD) and the surface quasi-geostrophic equation (SQG). They correspond to the research areas of materials science, free boundary problems and fluid dynamics respectively.

These topics are popular research areas in the study of partial differential equations. The phase field models we consider are of nonlinear diffusion type; the free boundary problem also has a diffusion term with nonlinearity introduced by the location of the free boundary. In both cases we consider implicit time stepping of the dynamics, motivated by the numerical approximations of the problems. In general, we consider equations of the form

$$\partial_t u = \Delta u + g(u)$$

with some nonlinearity $g(u)$. The left hand side (LHS) is the rate of change of a quantity u with respect to time at a particular point. The second term on the right hand side (RHS) is a local reaction term that is generated more of the quantity u (if $g(u) > 0$) ir reduces u (if $g(u) < 0$). The first term on the RHS is a term that represents diffusion, that is, there exists a flux of u from regions with large values of u to regions of small values of u.

Fully-implicit schemes discretize time derivative and treat both diffusion and nonlinearity as an implicit problem for the next time step:

$$\frac{u_{n+1} - u_n}{k} = \Delta u_{n+1} + g(u_{n+1})$$

where k is size of each time step, u_n is the solution from the previous step and

u_{n+1} needs to be solved. This equation represents an approximation of the solution which is discretized in time. Instead of $u(t)$ with time continuous, we have $u_n \approx u(nk)$. The equation represents the map from u_n to u_{n+1}. It is an implicit equation and a nonlinear, nonlocal problem and must be solved at each time step to find u_{n+1} from u_n. Under certain conditions, it can be shown that the time discrete approximation is convergent (increasingly accurate as k gets smaller). It is known that some form of implicit time stepping is needed to allow reasonable time steps in numerical approximation of these problems. In a subsequent step, the spatially continuous solution is approximated using a finite number of values on a grid.

More specifically, in the first part of this dissertation we are interested in the behaviour in the meta-stable regime of AC (solutions spend a long time near a state that is not a stationary solution and eventually change rapidly) as the small length scale $\epsilon \to 0$, where ϵ parametrizes the width of the layer where phase changes. Various time-stepping schemes have been applied in the literature to study such behavior including energy stable schemes where stability can be attained with help of large stabilizers but at the cost of the accuracy. Fully-implicit backward Euler scheme, however, gives more accurate approximation without a guarantee of energy stability. We will show by computation evidence and asymptotic analysis that the backward Euler method satisfies the energy stability in the meta-stable regime with large time steps. We will also give a rigorous proof for the two-dimensional radially symmetric case.

In the second part of this thesis we study the OD problem, which has a moving boundary whose location in time is part of the problem to be solved.

We show the equivalence of several different formulations of the problem including an energy gradient flow, a variational inequality, and a fixed boundary mapping method. The energy gradient flow formulation is new in the literature and we show a convergence result for a numerical scheme based on it.

In the third part of this dissertation, we are interested in the uniqueness and regularity of solutions to a PDE, arising from the study of fluid dynamics. Fluids usually follow certain energy (entropy, etc.) conservation laws, however poor regularity may cause the break down of such conservation. This phenomenon is called energy cascade [41], which has been studied by experimental physics. These phenomena indicate that lack of smoothness can lead to a break down of conservation and therefore it is natural to investigate what function spaces represent the minimal amount of smoothness required for the conservation laws to hold, which is known as the Onsager conjectures [67]. In this part of the thesis we focus on a stationary two-dimensional SQG system. SQG has applications to both meteorological and oceanic flow [68] and it is known to be a simplified model for incompressible Euler equations. We then develop a new framework of the “convex integration scheme” and implement it to construct a non-trivial solution to the stationary SQG. This proof of non-uniqueness of the stationary SQG system may be relevant to the long standing open problem of uniqueness of solutions to the Euler equations as in [67] which are fundamental equations in the study of fluid mechanics.

The structure of this thesis is organized as follows. Notation and preliminaries will be introduced in Chapter 2; we study the dynamics of AC in Chapter 3 and the free boundary dynamics of OD can be seen in Chapter 4.

Finally, we present the non-uniqueness of SQG in Chapter 5. Conclusions and remarks for future work can be found in Chapter 6.

Chapter 2

Notation and preliminaries

2.1 Notation and definitions

2.1.1 $O(h(x))$ and $o(h(x))$

We say that a function $f(x)$ is $O(h(x))$ if there exists a positive absolute constant C such that $|f(x)| \leq C|h(x)|$ for any x. We say that $f(x)$ is $o(h(x))$ if $\lim_{x\to 0} \frac{f(x)}{h(x)} = 0$ or $\lim_{x\to\infty} \frac{f(x)}{h(x)} = 0$ depending on the context.

2.1.2 $\lesssim$ and $\ll$

We say two positive quantities $A \lesssim B$ if there exists a positive absolute constant C such that $A \leq C \cdot B$. Similarly, $A \gtrsim B$ means $A \geq CB$, and $A \sim B$ when $A \lesssim B$ and $A \gtrsim B$.

We say two positive quantities $A \ll B$ if A/B is "small", where "small enough" is clear from the context.

2.1.3 L^p Space

Assume the domain Ω is given. If $1 \leq p < \infty$, the space $L^p(\Omega)$ consists of all complex-valued measurable functions that satisfy

$$\int_\Omega |f(x)|^p \ dx < \infty \ .$$

For $f \in L^p(\Omega)$ we define the L^p norm of f by

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(x)|^p \ dx \right)^{1/p} \ .$$

2.1.4 Weak Derivatives and Sobolev Space

We use the notation below:

$$\begin{aligned} x &= (x_1, x_2, ..., x_n) \in \mathbb{R}^n \\ \alpha &= (\alpha_1, \alpha_2, ..., \alpha_n) \in \mathbb{Z}_+^n \\ \partial^\alpha f &= \frac{\partial^{\alpha_1 + ... + \alpha_n} f}{\partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} ... \partial_{x_n}^{\alpha_n}} \quad . \end{aligned} \tag{2.1.1}$$

We define the weak derivative in the following sense: For $u,\ v \in L^1_{loc}(\Omega)$, (i.e they are locally integrable); $\forall \phi \in C_0^\infty(\Omega)$, i.e ϕ is infinitely differentiable (smooth) and compactly supported; and

$$\int_\Omega u(x) \ \partial^\alpha \phi(x) \ dx = (-1)^{\alpha_1 + ... + \alpha_n} \int_\Omega v(x) \ \phi(x) \ dx,$$

then v is defined to be the weak partial derivative of u, denoted by $\partial^\alpha u$. If u is “smooth” enough, its weak derivative coincides with its derivative and the equation above is basically integration by parts.

Suppose $u \in L^p(\Omega)$ and all weak derivatives $\partial^\alpha u$ exist for $|\alpha| = \alpha_1 + ... + \alpha_n \le k$, such that $\partial^\alpha u \in L^p(\Omega)$ for $|\alpha| \le k$, then we say $u \in W^{k,p}(\Omega)$, and such space is called Sobolev space. The norm in $W^{k,p}(\Omega)$ is defined as :

$$\|u\|_{W^{k,p}(\Omega)} = \left(\sum_{|\alpha| \le k} \int_\Omega |\partial^\alpha u|^p \, dx \right)^{\frac{1}{p}} .$$

Throughout this dissertation, for $p = 2$ case, we use the convention $H^k(\Omega)$ denote the space $W^{k,2}(\Omega)$. For more details, we refer to chapter 5, [34].

2.1.5 Fourier Transform

In this thesis we use the following convention for Fourier expansion on $\mathbb{T}^d := (\mathbb{R}/(2\pi))^d$:

$$f(x) = \frac{1}{(2\pi)^d} \sum_{k \in \mathbb{Z}^d} \widehat{f}(k) e^{ik \cdot x} \ , \ \widehat{f}(k) = \int_\Omega f(x) e^{-ik \cdot x} \, dx \ .$$

Taking advantage of the Fourier expansion, we define the equivalent H^s-norm and $\dot{H}^s$-norm of function f by

$$\|f\|_{H^s} = \frac{1}{(2\pi)^{d/2}} \left(\sum_{k \in \mathbb{Z}^d} (1 + |k|^{2s}) |\widehat{f}(k)|^2 \right)^{\frac{1}{2}} \ , \ \|f\|_{\dot{H}^s} = \frac{1}{(2\pi)^{d/2}} \left(\sum_{k \in \mathbb{Z}^d} |k|^{2s} |\widehat{f}(k)|^2 \right)^{\frac{1}{2}} .$$

The equivalence of two norms are well known, we refer to Appendix A in [83].

2.1.6 Convergence of Fourier Series in Periodic Domains

Given f being a $L^p(\mathbb{T}^d)$ periodic function for $p > 1$, and denote the Dirichlet partial sum $D_N f := \frac{1}{(2\pi)^d} \sum_{|k|\leq N} \widehat{f}(k) e^{ik\cdot x}$, then

$$\|D_N f - f\|_{L^p(\mathbb{T}^d))} \to 0\ ,\ and\ D_N f \to f \text{ pointwise almost everywhere}\ . \quad (2.1.2)$$

This was originally proved by Carleson in [13].

2.1.7 Banach Fixed-point Theorem

Given a Banach space $(X, \|.\|)$ and a contraction map $T : X \to X$ s.t $\|T(x) - T(y)\| \leq \beta\|x - y\|$ with $0 < \beta < 1$, then there exists a fix-point x, s.t $T(x) = x$. We refer to [34] for details.

2.1.8 Duhamel's Formula

Consider a linear inhomogeneous evolution equation for a function $u(x,t) : \Omega \times (0,\infty) \to \mathbb{R}$, with a spatial domain $\Omega \subset \mathbb{R}^d$, of the form

$$\begin{cases} u_t(x,t) - Lu(x,t) = f(x,t)\ ,\ (x,t) \in \Omega \times (0,\infty) \\ u|_{\partial\Omega} = 0 \\ u(x,0) = u_0(x)\ ,\ x \in \Omega\ , \end{cases} \quad (2.1.3)$$

where L is a linear differential operator that involves no time derivatives and the boundary condition could be replaced by periodic boundary condition.

Then formally, the solution to this equation system is:

$$u(x,t) = e^{Lt}u_0 + \int_0^t e^{L(t-s)} f \; ds \tag{2.1.4}$$

where e^{Lt} is the homogeneous solution operator, or $e^{Lt}u_0$ solves the homogeneous equation with initial data u_0. In fact $e^{Lt}u_0$ is often given as a convolution between a well-defined kernel and the initial data u_0. For more details, we refer to [34].

2.2 Several Important Inequalities

2.2.1 Hölder's Inequality

Given $f \in L^p(\Omega)$ and $g \in L^q(\Omega)$, such that $\frac{1}{p} + \frac{1}{q} = 1$ then

$$\|fg\|_{L^1(\Omega)} \leq \|f\|_{L^p(\Omega)} \|g\|_{L^q(\Omega)}.$$

2.2.2 Young's Inequality

Given a, b, p, q positive real numbers, such that $\frac{1}{p} + \frac{1}{q} = 1$, then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

2.2.3 Morrey's Inequality

Assume Ω is a bounded Lipschitz domain in $\mathbb{R}^d$ with $d \leq 3$ and $f \in H^2(\Omega)$ then

$$\|f\|_{L^\infty(\Omega)} \lesssim \|f\|_{H^2(\Omega)} \, .$$

In fact a stronger statement can be made with the help of Hölder space as in [34].

2.2.4 Gagliardo–Nirenberg Interpolation Inequality

For functions $u : \Omega \to \mathbb{R}$ defined on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, fix $1 \leq q,\ r \leq \infty$ and a natural number m. Suppose also that a real number α and a natural number j are such that

$$\frac{1}{p} = \frac{j}{d} + \left(\frac{1}{r} - \frac{m}{d}\right)\alpha + \frac{1-\alpha}{q}$$

and

$$\frac{j}{m} \leq \alpha \leq 1\ .$$

Then

$$\|D^j u\|_{L^p} \leq C_1(\Omega)\|D^m u\|_{L^r}^{\alpha}\|u\|_{L^q}^{1-\alpha} + C_2(\Omega)\|u\|_{L^s}$$

where $s > 0$ is arbitrary. We refer to [34].

Chapter 3

Stability and accuracy of time stepping schemes on phase field models

In this chapter we will consider two phase field models: Allen-Cahn (AC) and Cahn-Hilliard (CH) equations. The (AC) model was developed in [2] by Allen and Cahn to study the competition of crystal grain orientations in an annealing process; while the (CH) was introduced in [12] by Cahn and Hilliard to describe the process of phase separation of different metals in a binary alloy. These equations are presented as:

$$\begin{cases} \partial_t u = \Delta u - \dfrac{f(u)}{\epsilon^2}, \quad (x,t) \in \Omega \times (0,\infty) \\ u(x,0) = u_0 \end{cases}, \tag{AC}$$

and

$$\begin{cases} \partial_t u = -\epsilon \Delta\Delta u + \dfrac{\Delta f(u)}{\epsilon}, \quad (x,t) \in \Omega \times (0,\infty) \\ u(x,0) = u_0 \end{cases}, \tag{CH}$$

where $u(x,t)$ is a real valued function and values of u in $(-1,1)$ represent a mixture of the two phases, with -1 representing the pure state of one phase and $+1$ representing the pure state of the other phase. Vector position x is in the spatial domain Ω, which is taken to be two or three dimensional periodic domain in this work, and t is time. Here ϵ is a small parameter representing an average distance over which phases mix. The energy term $f(u)$ is often chosen to be

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

It is well known that, as $\epsilon \to 0$, the limiting problem of (AC) is given by a mean curvature flow while the limiting problem of (CH) becomes Mullins-Sekerka problem [66]; we refer to [46] for AC and [69], [1] for CH, where asymptotic and rigorous analysis are provided. More detail can be found in section 3.3. Although the limiting behavior of AC and CH are well known, there are related materials science models that are studied only numerically and this current chapter presents ideas about how to approach these models in an appropriate way numerically.

Adaptive time stepping methods for metastable dynamics of the Allen-Cahn and Cahn-Hilliard equations are investigated in the spatially continuous, semi-discrete setting. In this chapter we analyse the performance of a number of first and second order methods, formally predicting step sizes required to satisfy specified local truncation error σ in the limit of small length scale parameter $\epsilon \to 0$ during meta-stable dynamics. The formal predictions are made under stability assumptions that include the preservation of the asymptotic structure of the diffuse interface, a concept we call profile fidelity.

In this setting, definite statements about the relative behaviour of time stepping methods can be made. Some methods, including all so-called energy stable methods but also some fully implicit methods, require asymptotically more time steps than others. The formal analysis is confirmed in computational studies. We observe that some provably energy stable methods popular in the literature perform worse than some more standard schemes. We show further that when Backward Euler is applied to meta-stable Allen-Cahn dynamics, the energy decay and profile fidelity properties for these discretizations are preserved for much larger time steps than previous analysis would suggest. The results are established asymptotically for general interfaces, with a rigorous proof for radial interfaces. It is shown analytically and computationally that for most reaction terms, Eyre type time stepping performs asymptotically worse due to loss of profile fidelity.

3.1 Discussion

The mathematical literature for computational methods for AC dynamics, and its higher order relative CH dynamics, is dominated by the proposal, use, and analysis of so-called energy stable schemes [76, 85, 90]. AC and CH dynamics are gradient flows on an energy functional, and the solution should decrease that energy in time. Energy stable schemes guarantee that decrease no matter what time step is chosen. This is a desirable property not shared by standard fully implicit, semi-implicit (IMEX), or exponential integrator time stepping methods. We will show in this work that some (but not all) fully implicit methods can outperform energy stable schemes when subject to fixed accuracy requirements. The recent article [91] gives especially clear evidence that when time steps are chosen appropriately, fully implicit methods are conditionally energy stable, and further that the large time steps allowed by energy stable schemes can come at the cost of significant loss of accuracy. We show that in the meta-stable dynamic regime of AC and CH, some fully implicit methods can take optimally sized time steps. By optimal, we mean the asymptotically largest time steps as the order parameter $\epsilon \to 0$ that satisfy a given local error tolerance. Here, ϵ represents the width of interfacial layers in metastable dynamics and, like the authors of [91], we use the form of the equations scaled so that these dynamics transpire in an $O(1)$ time scale. When the dynamics are in this metastable regime, which dominates the time of typical phenomena of interest, definite statements about the behaviour of different time stepping methods can be made. This criteria does not take into account solver efficiency. However, we can make definite statements on how efficient solvers for nonlinear implicit time stepping need to be to outperform other methods.

A combination of asymptotic analysis and careful computational work backs up our claims. In addition, we present a rigorous result for implicit time stepping for meta-stable AC dynamics in radial geometry that shows that asymptotically larger time steps can be taken than previous analysis would suggest. These time steps preserve the diffuse interface structure (a property that we call proflie fidelity) and also the energy decay property of the equations. This result is shown for a class of reaction terms. An interesting result in Section 3.6.2 shows that Eyre-type time stepping can perform asymptotically worse with most reaction terms, while implicit time stepping has uniform asymptotic behaviour over a class of reaction terms. This was predicted by the analysis and confirmed computationally.

Our study focuses on pure materials science applications rather than the use of Cahn-Hilliard equations to track interfaces in so-called *diffuse interface methods* [93] in which the CH dynamics are coupled to other physics. We consider the simplest form of AC and CH dynamics, whose Gamma limit (as $\epsilon \to 0$) is well understood and use that well known structure to gain insight into the behaviour of the schemes. The authors believe that the insight gained from these studies will also apply to schemes used for other materials science models which are less well understood.

We consider a number of first and second order time stepping schemes: the energy stable Eyre's method [35]; Backward (Implicit) Euler (BE) [43]; Trapezoidal Rule (TR) [43]; Second order Backward Differentiation Formula (BDF2) [43]; Secant [33]; standard semi-implicit (linear IMEX) methods of first and second order [4]; first and second order Scalar Auxiliary Methods (SAV) [76] for which a modified energy stability can be proved; and finally a second order Singular Diagonally Implicit Runge Kutta method with good stability properties (DIRK2) [43]. The resulting implicit systems are considered in the spatially continuous semi-discrete setting in a 2D periodic domain, with numerical validation done with a suitably refined Fourier spectral approximation. Time step schemes that result in nonlinear systems are solved with Newton's iterations using the Preconditioned Conjugate Gradient Solver (PCG) developed in [21] at each iteration. Adaptive time stepping is done based on a user-specified local error tolerance σ. The variation of the number of time steps with ϵ for fixed σ is predicted based on formal consideration of the local truncation error of the schemes in the metastable dynamics. The formal predictions are then validated in computational studies. With this criteria, first order BE performs better (asymptotically fewer time steps as $\epsilon \to 0$) than Eyre and first order IMEX and SAV. Second order TR and BDF2 perform better than Secant, DIRK2, and second order IMEX and SAV. The difference in both cases is asymptotically larger for CH than AC. These comparisons are also valid for computational time, using PCG counts as the measure, to similar accuracy. It is seen that optimal numbers of time steps are obtained when the dominant local truncation error is a higher order time derivative. This observation may have application in other systems with metastable dynamics. We observe that standard IMEX methods perform almost identically to SAV methods of the same order in the scenario we consider, at reduced computational cost.

It is observed that the global accuracy of BE is better than a naïve prediction based on the size of the local truncation error would suggest. A formal analysis of the scheme for the AC case shows that the dominant error made in one time step is asymptotically smaller than expected. This is due to a special structure of the local truncation error for BE, in which the asymptotically largest term lies in a strongly damped space.

We introduce the equations and numerical schemes in Section 3.2 with some introductory analysis. The scaling for AC and CH is chosen so that the metastable interface dynamics (approximate curvature motion for AC and Mullins-Sekerka flow for CH) occurs in $O(1)$ time. In Section 3.3 we examine the metastable dynamics of the equations and make predictions for the behaviour of the time steps with ϵ and local error tolerance σ under stability assumptions which are verified numerically in Section 3.4. We give an asymptotic analysis for the surprising accuracy and stability properties for BE with large time steps applied to AC in Section 3.5. In Section 3.6 we present the rigorous result for BE applied to AC with large time steps and also show the loss of profile fidelity for Eyre-type time stepping for most reaction terms.

3.2 Equations and Schemes

We consider the simplest form of the AC dynamics for $u(\mathbf{x}, t)$ given by

$$u_t = \Delta u - \frac{1}{\epsilon^2} f(u) \tag{3.2.1}$$

where $f(u) = u^3 - u$ is the classical form of the reaction term. More general, smooth reaction terms are considered in Section 3.6. Non-smooth reaction terms and degenerate mobility are also of interest in some material science applications and there are stability and convergence results for implicit time stepping applied to these problems in [5, 7] for example. However, the asymptotic behaviour of time stepping schemes for these problems is not clear.

CH dynamics is described by a higher order partial differential equation

$$u_t = -\epsilon \Delta\Delta u + \frac{1}{\epsilon} \Delta f(u). \tag{3.2.2}$$

For computational simplicity, we consider the two-dimensional (2D) cases of these equations in a doubly periodic cell $\mathbb{T}^2 := [0, 2\pi]^2$. The time scaling in the equations above is chosen to give sharp interface (as $\epsilon \to 0$) motion in

$O(1)$ time. The sharp interface limit yields curvature driven flow for AC and a nonlocal Mullins-Sekerka flow for CH [66]. Both types of dynamics have an associated energy functional

$$\mathscr{E} = \int \left(|\nabla u|^2/2 + W(u)/\epsilon^2\right) \tag{3.2.3}$$

where $W(u) = \frac{1}{4}(u^2-1)^2$ and the reaction term $f(u) = W'(u)$. The energy $\mathscr{E}(t)$ is monotonically decreasing due to the gradient flow nature of the dynamics. For AC the gradient is in L^2 and for CH it is H^{-1}.

3.2.1 Time stepping

Backward Euler

We consider the simplest implicit scheme, first order Backward Euler (BE), also known as Implicit Euler. Applied to (3.2.1) keeping space continuous, we have

$$\frac{u_{n+1}-u_n}{k_n} = \Delta u_{n+1} - \frac{1}{\epsilon^2} f(u_{n+1}) \, .$$

where $u_n(\mathbf{x})$ approximates the exact solution $u(\mathbf{x}, t_n)$ and $k_n = t_{n+1} - t_n$ is the time step. We use the classical $f(u) = u^3 - u$ as mentioned above. Dropping the subscript on the time step and the unknown solution at time level $n+1$ we have the nonlinear problem

$$u - k\Delta u + \frac{k}{\epsilon^2} f(u) = u_n \tag{3.2.4}$$

for u given u_n.

Definition 1. *A time stepping scheme is said to have the* energy decay property *if* $\mathscr{E}(u_{n+1}) \le \mathscr{E}(u_n)$.

This property could be conditional on the choice of time step size. Additionally, it could depend on u_n. If a scheme has the energy decay property for any u_n and k, the scheme is called *unconditionally energy stable*.

Theorem 1. *Consider* (3.2.4)*, assume that* $u_n \in H^2(\Omega)$ *and* u_n *takes values in* $[-1,1]$*, then there exists* $u \in H^2(\Omega)$ *that solves* (3.2.4) *with values in* $[-1,1]$*. Define* $f_\infty := \max\{|f'(s)|, s \in [-1,1]\}$*, then if* $k \le 2\epsilon^2/f_\infty$ *the solution* u *is unique and satisfies the energy decay property. Note that the energy stability result was established earlier in [91] with a different proof.*

Proof. The existence of u follows from the standard method of sub-/super-solutions using comparison functions -1 and $+1$. To establish uniqueness, we assume u_1 and u_2 are solutions. Then their difference $w = u_1 - u_2$ is a solution of

$$(1 - k\Delta)w = -k \cdot \frac{f(u_1) - f(u_2)}{\epsilon^2} = -\frac{k}{\epsilon^2} \cdot f'(s(x))w \ ,$$

where s takes values between u_1 and u_2, and hence in $[-1,1]$. Isolating w leads to the elliptic problem

$$\left[1 + \frac{kf'(s)}{\epsilon^2} - k\Delta\right] w = 0,$$

and if $k < \epsilon^2/f_\infty$ then the corresponding elliptic operator is strictly positive and w is zero by the maximum principle. To establish energy decay, we take the inner product of (3.2.4) with the test function $u - u_n$:

$$\frac{1}{k}\int |u-u_n|^2 + \frac{1}{2}\int \left(|\nabla u|^2 - |\nabla u_n|^2 + |\nabla u - \nabla u_n|^2\right) = -\frac{1}{\epsilon^2}(f(u),(u-u_n)) \ .$$

From the Fundamental Theorem of Calculus we develop the expansion,

$$|F(u)-F(u_n)-f(u)(u-u_n)| = \left|\int_u^{u_n} f'(s)(s-u_n)\,ds\right| \le \frac{f_\infty}{2}(u-u_n)^2.$$

Using this relation to eliminate $f(u)$ yields the equality,

$$(\frac{1}{k} - \frac{f_\infty}{2\epsilon^2})\int |u-u_n|^2 + E[u] - E[u_n] \le 0.$$

This implies the desired energy decay for $k < 2\epsilon^2/f_\infty$. The theorem is also true when homogeneous Neumann boundary conditions are specified. □

Thus we have existence of solutions to (3.2.4) for any time step size, and uniqueness and energy stability under the resitriction $k \le 2\epsilon^2/f_\infty$. This is true for any u_n under the restrictions of the Theorem. We shall see in Section 3.6 that asympoticaly larger time steps $k = o(\epsilon)$ can be taken when the dynamics are slow (interface motion) with locally unique, energy stable solutions. This is verified in computational tests.

Eyre's Method

An alternative first order scheme to fully implicit BE was proposed by Eyre [35]:

$$u - k\Delta u + \frac{k}{\epsilon^2}u^3 = u_n + \frac{k}{\epsilon^2}u_n \tag{3.2.5}$$

The scheme is derived conceptually by keeping a convex part of the reaction term $f(u) = u^3 - u$ implicit and a concave part explicit. In this sense, it is an IMEX method but an unusual one since a nonlinear term is kept implicit and a linear term is handled explicitly. The method has appealing properties:

Theorem 2 (from [35])**.** *The time step (3.2.5) has a unique solution u for any u_n and k that is unconditionally energy stable.*

Additional first order schemes considered are the SAV scheme [76] and a linear IMEX method [4]:

$$u - k\Delta u + \frac{Sk}{\epsilon^2} u = u_n - \frac{k}{\epsilon^2}\left(u_n^3 - (S+1)u_n\right) \tag{3.2.6}$$

with $S > 0$, sometimes called a stabilization term. We take $S = 2$ since that makes the left hand side a linearization about the far field values, but computational performance is relatively insensitive to S. The SAV scheme is energy stable with a modified energy. We use the same stabilization coefficient as above in the SAV scheme. There is a class of linearly implicit energy stable schemes [16, 54, 55] that require an asymptotically large stabilization term $O(\epsilon^{-p})$ with p large and increasing from AC to CH and 2D to 3D for the analysis. These methods are theoretically interesting but are extremely inaccurate and not useful for practical applications. We have further discussion of these schemes in Remark 3.

All time stepping schemes can be applied to CH (3.2.2), with BE and Eyre shown below:

$$u + k\Delta\Delta u - \frac{k}{\epsilon^2}\Delta f(u) = u_n \qquad \text{BE}$$

$$u + k\Delta\Delta u - \frac{k}{\epsilon^2}\Delta u^3 = u_n - \frac{k}{\epsilon^2}\Delta u_n \qquad \text{Eyre}$$

In this case, BE is known to have unique solutions with the energy decay property when $k < \epsilon^3$ [91] and Eyre is unconditionally energy stable [35].

Second Order Schemes

We also consider the second order methods Trapezoidal Rule (TR), Secant (S) [33], Second Order Backward Differencing (BDF2), and Second Order Singular Diagonal Implicit Runge Kutta (DIRK2) [43] methods. These are described below for $u_t = \mathscr{F}(u)$ with

$$\mathscr{F}(u) = \Delta u - f(u)/\epsilon^2 \qquad \text{for AC}$$

$$\text{and} \quad \mathscr{F}(u) = -\epsilon\Delta\Delta u + \Delta f(u)/\epsilon \qquad \text{for CH}$$

With this notation:

$$\text{(TR)} \quad u - \frac{k}{2}\mathscr{F}(u) = u_n + \frac{k}{2}\mathscr{F}(u_n)$$

$$\text{(BDF2)} \quad \frac{3u}{2} - k\mathscr{F}(u) = 2u_n - \frac{1}{2}u_{n-1}.$$

Secant is a variant of TR with the term $f(u) - f(u_n)$ replaced by

$$(W(u) - W(u_n))/(u - u_n)$$

where W is the energy term from (3.2.3). It is known to be conditionally energy stable [33]. For the simple form of W we have taken, the expression above can be factored explicitly. DIRK2 is a two stage method

$$\begin{aligned} u_* - \alpha k \mathscr{F}(u_*) &= u_n \\ u - \alpha k \mathscr{F}(u) &= u_n + (1-\alpha) k \mathscr{F}(u_*) \end{aligned}$$

with $\alpha = 1 - 1/\sqrt{2}$. Both DIRK2 and BDF2 are A-stable, and so preferable to TR and Secant from the perspective of stiff ODE solver theory [43]. A second order linear IMEX method (SBDF2 [4]) and two variants of second order SAV methods based on BDF2 are also considered.

There are many other specialized schemes in the literature and we mention two second order unconditionally energy stable concave-convex splitting schemes here. They are nonlinear two step schemes but the nonlinear problem at each time step is convex. One is based on TR, with the cubic term handled implicitly and the linear reaction term extrapolated from two previous time steps [31]. The second is a variant of SBDF2, again with the cubic term handled implicitly and an additional moderate stabilizing term [92]. Both schemes have the same asymptotic error behaviour as the Secant, DIRK2, and SBDF2 methods shown in detail below.

3.2.2 Spatial discretization and solution procedure

The current work concentrates on the time stepping errors, and it is convenient to consider the semi-discrete, spatially continuous approximation. This idealization is approximated well by the Fourier spectral spatial discretization. The computational results shown have sufficient spatial resolution that spatial errors do not affect the results in the digits shown.

We use the Preconditioned Conjugate Gradient (PCG) solvers developed in [21] for the schemes involving nonlinear implicit problems. We note that there has been recent promising work in the use of preconditioned steepest descent with approximate line search in solving these nonlinear problems [14, 38]. Another approach has been to recast the implicit step as a minimization problem [91]. Both these techniques have the advantage that they look for local solutions which can be unique and have energy decay even for large time steps, as shown rigorously in Section 3.6.

The computations in this work are done in a full 2D setting, rather than in a reduced dimensional radial setting as could be done, in order to give PCG iteration counts for the nonlinear time stepping methods that have meaning for more general computations. Note that the PCG counts are independent of spatial resolution when the problem is resolved.

3.2.3 Error estimation and adaptive time stepping

We perform two time steps of the same size k in order to use a specialized predictor u_p for u_{n+2}.

$$u_p = u_n + \frac{k}{3}(\mathscr{F}(u_n) + 4\mathscr{F}(u_{n+1}) + \mathscr{F}(u_{n+2})) \tag{3.2.7}$$

where $\mathscr{F}(u) = \Delta u - f(u)/\epsilon^2$ for AC and $-\epsilon\Delta\Delta u + \Delta f(u)/\epsilon$ for CH as above. Time step sizes are adjusted so that

$$\|u_{n+2} - u_p\|_\infty \leq \sigma.$$

The predictor u_p is formally one order more accurate than the numerical approximation u_{n+2} from time stepping, up to fifth order. The predictor has an inherent dominant local error $k^5 u_{ttttt}/90$ that is a pure time derivative of u, which is shown below in Section 3.3.1 to be a desirable property.

For the one step methods, the time step is adjusted adaptively to maintain a local error below σ as described in [21]. For BDF2 and its linear variants, time steps are only adjusted by a factor of two. When time steps are reduced (using Hermite cubic interpolation for the restart value) or increased, four time steps are taken before checking the local error to allow relaxation of the initial error layer.

3.3 Local Truncation Errors in Metastable Dynamics

3.3.1 Metastable dynamics

In our formulation, it is known that after a short time $O(\epsilon^2)$ solutions to AC tend to interfaces between regions of solution near the equilibrium values,

$$u \approx \pm 1.$$

These interfaces have width ϵ and move approximately with curvature motion. We refer to this dynamics as metastable or slow, even though with the particular time scaling we have chosen it occurs in in $O(1)$ time. For the majority of the time, the solution will be in this regime, so we concentrate now on the expected and observed behaviour of time stepping in this setting. With the choice of $f(u) = u^3 - u$, we have

$$u(x,t) \approx g(z) \tag{3.3.1}$$

with $g(z) := \tanh(z/\sqrt{2})$ and $z = \mathrm{dist}(x,\Gamma)/\epsilon$, where Γ is the approximate interface with arc length parameter s moving with curvature motion (normal velocity equal to curvature). We fix its location at the $u = 0$ level set. The local coordinates (s,z) are shown in Figure 3.1. This structural result on the metastable solution can be obtained with formal asymptotics. In the outer asymptotic region for AC the solution takes the form $u = \pm 1$ to all orders. Curvature motion as the limit $\epsilon \to 0$ has been proven rigorously [1, 69].

CH has the same metastable solution structure (3.3.1) with normal interface velocity given by Mullins-Sekerka flow, in $O(1)$ time in our scaling (3.2.2). We refer the reader to the review article [74] for details. It has been shown that numerical schemes can accurately approximate this limit with implicit time stepping with appropriate scaling of the time step with ϵ [39]. We will show that this limit can be taken with asymptotically larger time steps than in that analysis.

From (3.3.1), we see that time and space derivatives are large near the

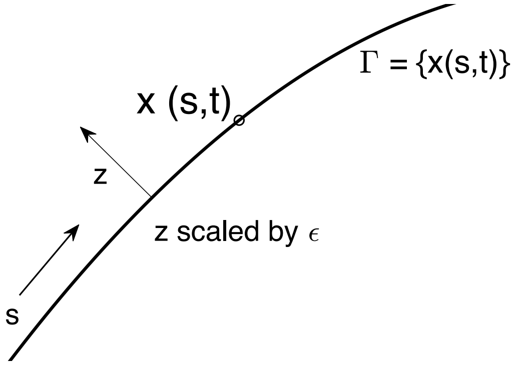

Figure 3.1: Sketch of the local coordinates of the metastable solution

interface. Starting with

$$u(x,t) \approx g(\mathrm{dist}(x,\Gamma)/\epsilon)$$

we can take a time derivative to obtain:

$$u_t \approx g'(\mathrm{dist}(x,\Gamma))V/\epsilon$$

where V is the normal velocity at the point on Γ closest to x. Formally taking higher derivatives in this pattern yields:

$$\frac{\partial^n u}{\partial t^n} = O(\epsilon^{-n}). \tag{3.3.2}$$

This is used to analyze the truncation error of the time stepping schemes.

Predicted time step sizes for AC

A standard strategy for adaptive time stepping is to have a user specified local error tolerance of σ. The error for each time step is estimated and the time step adjusted so that there is an estimated error in that single time step less than σ. It is known that the dominant local truncation error for BE is $k^2 u_{tt}/2$ which in metastable dynamics is $O(k^2/\epsilon^2)$ from (3.3.2). The local truncation error restriction then requires time steps of size

$$k = O(\sqrt{\sigma}\epsilon) \quad \text{(BE)}$$

We now proceed to determine the expected behaviour of time steps with ϵ and σ from the other schemes. We can write the BE scheme (3.2.4) and Eyre's scheme (3.2.5) for AC in an instructive way

$$\begin{aligned} u - u_n - k\Delta u + k\left[u^3 - u\right]/\epsilon^2 &= 0 \text{ (BE)} \\ u - u_n - k\Delta u + k\left[u^3 - u\right]/\epsilon^2 + k(u - u_n)/\epsilon^2 &= 0 \text{ (Eyre).} \end{aligned}$$

Knowing that the truncation error for BE is $O(k^2/\epsilon^2)$ we see that the truncation error for the Eyre scheme is dominated by the last term in its expression above, which has leading order $k^2 u_t/\epsilon^2 = O(k^2/\epsilon^3)$. Our time step prediction in this case is

$$k = O(\sqrt{\sigma}\epsilon^{3/2}) \quad \text{(Eyre)}$$

Thus, the advantage of the Eyre scheme to be able to take large time steps and remain energy stable is never realized if accurate computational results are required. Reference [91] has an alternate way to view the loss of accuracy that does not highlight this asymptotic difference. The first order IMEX and SAV schemes have the same asymptotic behaviour as Eyre.

Remark 1. *It is well known that when large time steps are taken with Eyre's* *method, the dynamics occur in a slower time scale. This is an exact result for AC [91], qualitative for CH [22]. In this work, time steps are restricted by a specified local error tolerance. Thus, we do not see a change in time scale for the results of Eyre's method, rather we see decreased time step size.*

Remark 2. *The formal local error analysis above relies on the stability of the schemes in metastable dynamics under the resulting time step restrictions. More than simple stability, the analysis requires that the time stepping preserves the asymptotic structure of the diffuse interface. This is the concept we have named* profile fidelity. *All predicitions described in this section lead to time stepping that preserves profile fidelity for the classical choice of $f(u) = u^3 - u$. We observe the predicted time step behaviour in ϵ and σ computationally. In Section 3.6.2 we show that for (most) other reaction terms, Eyre time stepping loses profile fidelity for time steps $k = O(\epsilon^{3/2})$ and in these cases, $k = O(\epsilon^2)$ is needed for accuracy.*

Remark 3. *The first order, linearly implicit energy stable scheme for 2D AC is analyzed in [16]. The analysis requires a stabilization term of order $\epsilon^{-2}|\ln\epsilon|$. If such a scheme were implemented, the time steps required for a local error tolerance of σ would be $k = O(\sqrt{\sigma}\epsilon^{5/2}/|\ln\epsilon|)$, prohibitively small for practical computation.*

We can determine the dominant term in the local truncation errors of the second order schemes applied to AC:

$$\begin{aligned}
&\text{(TR)} && k^3 u_{ttt}/12 = O(k^3/\epsilon^3)\\
&\text{(S)} && k^3\left(u_{ttt}/12 + u u_t^2/(2\epsilon^2)\right) = O(k^3/\epsilon^4)\\
&\text{(DIRK2)} && k^3\left((\alpha^2(1-\alpha)+\alpha/2-1/6)u_{ttt} - 3\alpha^2(1-\alpha)u u_t^2/(2\epsilon^2)\right) = O(k^3/\epsilon^4)\\
&\text{(BDF2)} && -k^3 u_{ttt}/3 = O(k^3/\epsilon^3)\\
&\text{(SBDF2)} && k^3\left(3u^2+(M+1)\right)u_{tt}/\epsilon^2 = O(k^3/\epsilon^4)
\end{aligned}$$

We consider two second order SAV variants based on how an extrapolated approximation is computed. If the extrapolated value of u_{n+1} is taken as $2u_n - u_{n-1}$ the scheme (referred to as SAV2-A) behaves similarly to SBDF2. If the extrapolated value is computed with a first order linear IMEX scheme as suggested in [76] (referred to as SAV2-B), the scheme has a local truncation error of order k^3/ϵ^5. The results are summarized in Table 3.1. It is clear that BE takes asymptotically (as $\epsilon \to 0$) fewer time steps than Eyre, although they are both first order in time step size. TR and BDF2 take asymptotically fewer time steps than Secant, DIRK2, SBDF2, SAV2-A and SAV2-B although they are all second order methods. The computations in Section 3.4 below show that these time step estimates correspond to real computational behaviour.

Remark 4. *We predict the number M of time steps in Tables 3.1 and 3.2 and how it varies with ϵ and σ. As shown in Figure 3.2 we are also predicting how a profile of time steps $k(t)$ behaves with ϵ and σ.*

Method (AC)	L	k	$M = O(1/k)$
BE	k^2/ϵ^2	$\sqrt{\sigma\epsilon}$	$1/(\sqrt{\sigma\epsilon})$
Eyre, IMEX1, SAV1	k^2/ϵ^3	$\sqrt{\sigma}\epsilon^{3/2}$	$1/(\sqrt{\sigma}\epsilon^{3/2})$
TR, BDF2	k^3/ϵ^3	$\sqrt[3]{\sigma\epsilon}$	$1/(\sqrt[3]{\sigma\epsilon})$
S, DIRK2, SBDF2, SAV2-A	k^3/ϵ^4	$\sqrt[3]{\sigma}\epsilon^{4/3}$	$1/(\sqrt[3]{\sigma}\epsilon^{4/3})$
SAV2-B	k^3/ϵ^5	$\sqrt[3]{\sigma}\epsilon^{5/3}$	$1/(\sqrt[3]{\sigma}\epsilon^{5/3})$

Table 3.1: Order predictions for the behaviour of the numerical schemes with local error tolerance σ in the metastable regime of AC dynamics. Here, L is the local error, k is the time step size, and M is the number of time steps to reach a fixed end time.

Predicted time step sizes for CH

The same local truncation analysis can be done for the CH in the metastable regime where the solution has the same interface structure (3.3.1) with the interface Γ moving approximately with Mullins-Sekerka flow in $O(1)$ time. BE, TR, BDF2, and SBDF2 have the same error expressions as above, but Eyre, Secant and DIRK2 have local truncation errors when applied to CH listed below:

$$
\begin{aligned}
\text{(Eyre)}\quad & k^2(u_{tt}/2 - \Delta u_t/\epsilon) = O(k^2/\epsilon^4) \\
\text{(S)}\quad & k^3\left(u_{ttt}/12 - \Delta(u u_t^2)/(2\epsilon)\right) = O(k^3/\epsilon^5) \\
\text{(DIRK2)}\quad & k^3\left((\alpha^2(1-\alpha) + \alpha/2 - 1/6)u_{ttt} + 3\alpha^2(1-\alpha)\Delta(u u_t^2)/(2\epsilon)\right) = O(k^3/\epsilon^5) \\
\text{(SBDF2)}\quad & k^3\left(3u^2 + (M+1)\right)\Delta u_{tt}/\epsilon = O(k^3/\epsilon^5)
\end{aligned}
$$

where we have used the fact that the Laplacian Δ increases the size of terms by $1/\epsilon^2$ near the interface. The first order IMEX and SAV schemes have the same asymptotic behaviour as Eyre. SAV2-A behaves similarly to SBDF2 as before, with SAV2-B worse by a power of ϵ as for the AC case above. The re-

Method (CH)	L	k	$M = O(1/k)$
BE	k^2/ϵ^2	$\sqrt{\sigma}\epsilon$	$1/(\sqrt{\sigma}\epsilon)$
Eyre, IMEX1, SAV1	k^2/ϵ^4	$\sqrt{\sigma}\epsilon^2$	$1/(\sqrt{\sigma}\epsilon^2)$
TR, BDF2	k^3/ϵ^3	$\sqrt[3]{\sigma}\epsilon$	$1/(\sqrt[3]{\sigma}\epsilon)$
S, DIRK2, SBDF2, SAV2-A	k^3/ϵ^5	$\sqrt[3]{\sigma}\epsilon^{5/3}$	$1/(\sqrt[3]{\sigma}\epsilon^{5/3})$

Table 3.2: Order predictions for the behaviour of the numerical schemes with local error tolerance σ in the metastable regime of CH dynamics. Here, L is the local error, k is the time step size, and M is the number of time steps to reach a fixed end time.

sults are summarized in Table 3.2. The predictions in this table are validated in the numerical experiments in the next section. Although the methods all have the formal order of accuracy in terms of time step size, the behaviour as $\epsilon \to 0$ varies significantly. Note that the gap between BE and the other first order schemes, and between TR/BDF2 and Secant/DIRK2/SBDF2/SAV2-A is wider for CH dynamics than it was for AC.

Discussion: the source of increased local error

In the metastable regime, the two terms in AC and CH (diffusion and nonlinear reaction) are both large but approximately cancel to give the slow dynamics. The methods with asymptotically (as $\epsilon \to 0$) small local errors (BE, TR, BDF2) have dominant truncation errors that are pure time derivatives of the solution, which inherit this high order cancellation. The other methods which have large local errors have truncation errors that involve the reaction term individually. This imbalance amplifies the size of the error. As an example, DIRK2 applied to $u_t = \mathscr{F}(u)$ has an error proportional to $\mathscr{F}''u_t^2$. From this discussion, we believe the ranking of the schemes in this work will also apply to other nonlinear problems with metastable dynamics.

3.4 Computational Results

3.4.1 Allen-Cahn

We take initial conditions in the form of a radial front

$$\tanh \frac{\sqrt{(x-\pi)^2+(y-\pi)^2}-2}{\epsilon\sqrt{2}}$$

and compute with $\epsilon = 0.2$, 0.1, 0,05 and 0.025. The benchmark for accuracy is the time T at which the value at the domain centre (π,π) changes from negative to positive. Except for the exponentially small (in ϵ) derivative discontinuities at the periodic boundaries, the dynamics approximate the sharp interface limit of curvature motion of a circle shrinking to a point at the domain centre. The expectation from asymptotic analysis of the sharp interface limit is that

$$T = 2 + O(\epsilon^2).$$

This is confirmed by the numerical solutions below. A video of the dynamics is available [88].

First order methods

The PCG approach is known to have bounded condition number under the scaling $k = C\epsilon^2$ for BE with $C < 1$ [91] and we observe good behaviour in the example below even with $C > 1$ in the metastable regime. It is observed computationally in this work that the PCG for Eyre's method is independent of k and ϵ although the authors are not aware of a proof in the literature. PCG counts can be used as a proxy for computational time when comparing

	BE			Eyre		
σ	M	CG	E	M	CG	E
1e-4	717	5,348 [7.46]	0.003	2,350	14,856 [6.32]	0.047
1e-5	2,225 (3.10)	9,448 [4.24]	0.001	7,351 (3.12)	28,263 [3.85]	0.014
1e-6	7,010 (3.15)	23,017 [3.28]	0.001	23,172 (3.15)	68,148 [2.94]	0.004

Table 3.3: Computational results for the AC benchmark problem with fixed $\epsilon = 0.2$ and local error tolerance σ varied. BE results are on the left, Eyre on the right. Here, M is the total number of time steps taken (with the ratio to the value above in brackets), CG is the number of conjugate iterations (with the ratio to the number of time steps in brackets), E is the error in the benchmark time.

	BE			Eyre		
ϵ	M	CG	E	M	CG	E
0.2	717	5,348 [7.46]	0.003	2,350	14,856 [6.32]	0.047
0.1	1,291 (1.80)	12,354 [9.57]	0.001	6,463 (2.75)	44,717 [6.92]	0.069
0.05	2,412 (1.87)	27,782 [11.52]	0.001	18,218 (2.83)	143,416 [7.87]	0.099
0.025	4,630 (1.92)	64,884 [14.01]	$*$	52,595 (2.89)	497,846 [9.47]	0.141

Table 3.4: Computational results for the AC benchmark problem with fixed local error tolerance $\sigma = 10^{-4}$ and ϵ varied. Here, M is the total number of time steps taken (with the ratio to the value above in brackets) and CG is the number of preconditioned conjugate gradient iterations (with the ratio to the number of time steps in brackets), E is the error in the benchmark time with $*$ denoting a result correct to three decimal places.

methods.

Results of the numerical experiments in which σ and ϵ were varied for BE and Eyre are shown in Tables 3.3 and 3.4. Spatial errors do not affect the digits shown in any of the computational results in this chapter.

Table 3.3 validates the second order $O(k^2)$ local truncation error since the number of time steps was predicted to be $M = O(1/\sqrt{\sigma})$ for both methods with ϵ constant, noting that $\sqrt{10} \approx 3.16$. Such results for other schemes and for the CH benchmark problem below are not shown, but verify the formal

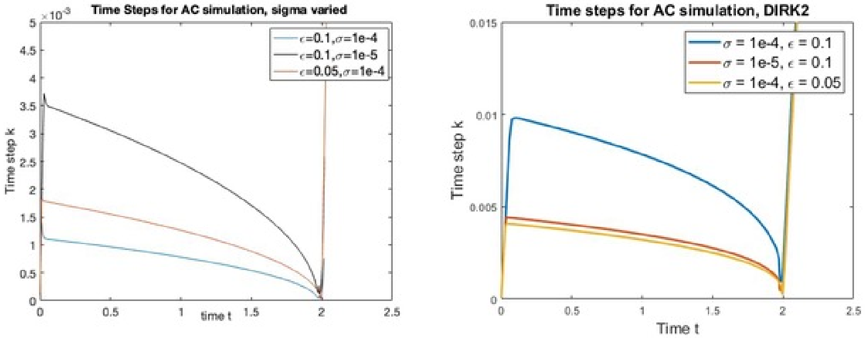

Figure 3.2: Time steps k for Allen-Cahn dynamics with ϵ and σ varied using BE (left) and DIRK2 (right). The time steps decrease in size as the simulation approaches the topological singularity at $t \approx 2$. Note that for each method, the profiles $k(t)$ have the same *shape* as σ and ϵ are varied and scale with these quantities according to our theoretical predictions. In particular, note that time steps decrease more quickly for BE as σ is decreased but more quickly for DIRK2 as ϵ is decreased, as we predict.

accuracy of the schemes. Table 3.4 validates the prediction of $M = O(1/\epsilon)$ for BE and $M = O(1/\epsilon^{3/2})$ for Eyre with σ constant, noting that $2^{3/2} \approx 2.83$. Both tables validate the prediction that for the same local tolerance σ, Eyre involves more computational work than BE and gives less accurate answers. CG counts for both methods are small as expected. You see (unexpectedly) that the final accuracy of BE does not seem to degrade as $\epsilon \to 0$ for fixed σ. This is discussed in Section 3.5 below. Although BE does not guarantee energy stability, no step accepted by the local error tolerance exhibited an energy increase.

For completeness, we show the time step sizes as a function of time for BE in Figure 3.2 with ϵ and σ varied. As mentioned in Remark 4 our predictions for the behaviour of the time steps sizes k as ϵ and σ are varied describe a profile $k(t)$.

	IMEX1		SAV1	
ϵ	M	E	M	E
0.2	3,932	0.067	3,936	0.067
0.1	11,110 (2.83)	0.096	11,112 (2.82)	0.096
0.05	31,676 (2.85)	0.138	31,682 (2.85)	0.138
0.025	90,748 (2.86)	0.198	90,760 (2.86)	0.198

Table 3.5: Computational results for the AC benchmark problem with fixed local error tolerance $\sigma = 10^{-4}$ and ϵ varied. Here, M is the total number of time steps taken (with the ratio to the value above in brackets) and E is the error in the benchmark time.

We repeat the $\epsilon \to 0$ study for IMEX1 and SAV1 in Table 3.5. These methods require a fixed number of FFT calculations per time step to invert the constant coefficient linear implicit aspect of the schemes, with SAV1 requiring four times as many solves as IMEX1. It is seen that IMEX1 behaves almost identically to SAV1 and both are superior to Eyre's method when computational cost is considered. In the context of this study, there is no benefit from the theoretical guarantees of energy stable schemes and BE is the optimal (with our asymptotic definition) first order scheme with IMEX1 the runner up. This will remain true for other nonlinear solver strategies for BE as long as they require fewer than $O(1/\sqrt{\epsilon})$ iterations when adaptive time steps are taken.

Remark 5. *Note that for the BE computation for $\epsilon = 0.025$ we can still get reasonable accuracy taking $\sigma = 10^{-2}$. In this case, the maximum value of k/ϵ^2 is 14.6. Clearly, the theory which guarantees existence of solutions and energy decay for $k < \epsilon^2$ [91] can be improved for metastable dynamics. This is explored in the analysis in Sections 3.5 and 3.6 below.*

Second Order Methods

The CG counts of all the nonlinear second order methods are relatively insensitive to ϵ, similar to the first order methods shown above. We show the number of time steps used for the seven methods in Table 3.6, for $\sigma = 10^{-4}$ fixed and ϵ varied. All second order methods give at least three digits of accuracy to the benchmark time with this tolerance σ. The superiority of TR and BDF2 is clearly seen with $M = O(1/\epsilon)$, compared to $M = O(1/\epsilon^{4/3})$ (noting that $2^{4/3} \approx 2.52$) for Secant, DIRK2, SBDF2, SAV2-A and $M = O(1/\epsilon^{5/3})$ (noting that $2^{5/3} \approx 3.18$) for SAV-B as predicted above. The pattern in the number of time steps for the multi-step methods is a bit rougher due to the strict criteria we have used for adaptive time step change. As above, we see no benefit from the theoretical guarantees of energy stable schemes. Fully implicit methods TR and BDF2 are asymptotically optimal in terms of the number of time steps and are computationally optimal if the solvers require fewer than $O(1/\sqrt[3]{\epsilon})$ iterations when adaptive time steps are taken (which appears to be the case with the Newton PCG solver we used). SBDF2 is the runner up and notably it is comparable to the fully implicit DIRK2 method but does not have the overhead of a nonlinear solve.

It is interesting to note that the slight change in the extrapolation procedure in the SAV2 schemes makes such a difference to their asymptotic performance. It is confirmation that merely considering the order of time stepping scheme and its theoretical energy stability properties is not the whole story.

ϵ	TR	S	BDF2	DIRK2	SBDF2	SAV2-A	SAV2-B
0.2	170	236	280	180	588	768	1,572
0.1	278 (1.64)	512 (2.16)	472 (1.69)	364 (2.02)	1,384 (2.35)	1,572 (2.04)	5,436 (3.46)
0.05	492 (1.77)	1,208 (2.36)	860 (1.82)	814 (2.24)	3,260 (2.35)	3,392 (2.16)	15,088 (2.78)
0.025	916 (1.86)	2,960 (2.45)	1,632 (1.90)	1,894 (2.33)	7,600 (2.33)	7,980 (2.35)	48,048 (3.18)

Table 3.6: Computational results for the second order methods applied to the AC benchmark problem with fixed local error tolerance $\sigma = 10^{-4}$ and ϵ varied. Shown are the total number of time steps taken (with the ratio to the value above in brackets)

3.4.2 Cahn-Hilliard

For the initial conditions we take

$$\tanh\left(\frac{r-5/2}{\epsilon\sqrt{2}}\right)+\tanh\left(\frac{3/2-r}{\epsilon\sqrt{2}}\right)+1$$

with $r = \sqrt{(x-\pi)^2+(y-\pi)^2}$ and compute with ϵ = 0.2, 0.1, 0,05 and 0.025. The dynamics approximate the sharp interface limit of two concentric circles, with the inner circle shrinking. As before, the benchmark is the time T at which the value at the domain centre (π,π) changes from negative to positive. A video of the dynamics is available [89] .

First order methods

Results of the numerical experiments in which ϵ is varied for the first order methods are shown in Table 3.7. These validate the prediction of $M = O(1/\epsilon)$ for BE and $M = O(1/\epsilon^2)$ for Eyre and IMEX1 with σ constant. As for the AC case, SAV1 behaves similarly to IMEX1 at increased computational cost. For CH, the implicit problem for BE is more difficult to solve as $\epsilon \to 0$ with fixed σ, but it is still more accurate than Eyre stepping for equivalent

	BE			Eyre			IMEX1	
ϵ	M	CG	E	M	CG	E	M	E
0.2	730	5,348 [7.33]	$*$	3,055	36,684 [12.0]	0.019	9,982	0.014
0.1	1,184 (1.62)	24,778 [20.9]	0.001	12,751 (4.17)	190,204 [14.0]	0.021	43,332 (0.015)	0.015
0.05	2,068 (1.75)	66,307 [32.1]	$*$	52,753 (4.13)	937,774 [17.8]	0.022	181,234 (4.18)	0.015
0.025	3,768 (1.82)	198,771 [52.8]	$*$	215,443 (4.08)	4,504,278 [20.9]	0.022	740,366 (4.09)	0.015

Table 3.7: Computational results for the first order methods applied to the CH benchmark problem with fixed local error tolerance $\sigma = 10^{-4}$ and ϵ varied. Here, M is the total number of time steps taken (with the ratio to the value above in brackets), CG is the number of conjugate iterations (with the ratio to the number of time steps in brackets), and E is the error in the benchmark time with $*$ denoting a result correct to three decimal places.

computational cost. It will be asymptotically more efficient as long as the solution strategy for the nonlinear problem requires fewer than $O(1/\epsilon)$ iterations with adaptive time stepping. As with AC, we see that BE does not suffer from global accuracy decrease as $\epsilon \to 0$.

Second order methods

The CG counts for the second order methods behave like those of BE with ϵ as shown above. We show the number of time steps used for the four methods in Table 3.8, for $\sigma = 10^{-4}$ fixed and ϵ varied. The superiority of TR and BDF2 is clearly seen, consistent with $M = O(1/\epsilon)$, compared to $M = O(1/\epsilon^{5/3})$ (noting that $2^{5/3} \approx 3.17$) for Secant, DIRK2, and SBDF2 as predicted above. Results for SAV2-A are comparable to those for SBDF2. Again, the implications for the asymptotic computational superiority of fully implicit TR and BDF2 under the assumption of sufficient solver efficiency are clear.

ϵ	TR	S	BDF2	DIRK2	SBDF2
0.2	230	534	320	378	1,388
0.1	314 (1.36)	1,530 (2.87)	468 (1.46)	788 (2.08)	4,108 (2.96)
0.05	474 (1.51)	4,722 (3.08)	748 (1.60)	1,906 (2.42)	12,352 (3.01)
0.025	792 (1.67)	14,924 (3.16)	1,312 (1.75)	6,048 (3.17)	44,060 (3.57)

Table 3.8: Computational results for the second order methods applied to the CH benchmark problem with fixed local error tolerance $\sigma = 10^{-4}$ and ϵ varied. Shown are the total number of time steps taken (with the ratio to the value above in brackets)

3.5 Asymptotic Analysis of Properties of BE AC Solutions

The results in Table 3.4 present the accuracy for BE applied to AC with fixed local error tolerance $\sigma = 10^{-4}$ under various values of ϵ. It is remarkable the accuracy in the benchmark time does not degrade as $\epsilon \to 0$. This is unexpected, as a naïve prediction would be that the final accuracy scaled like $M\sigma = O(\sqrt{\sigma}/\epsilon)$ where M is the number of time steps. It is clear that the resulting solution accuracy for the schemes under specified local error tolerance is a nontrivial question.

We present below the asymptotic analysis of a fully implicit BE time step (3.2.4) in two dimensions assuming the solution is in the meta-stable regime. That is, u_n is approximately described as a curve $\mathbf{x}_n(s)$ parametrized by arc length with normal $\hat{n}$, dressed with the heteroclinic profile (3.3.1). We take the scaling $k = c\epsilon$ with c independent of ϵ, both sufficiently small depending only on the curve $\mathbf{x}_n$. We consider the formal asymptotics for the implicit time step u of (3.2.4) in this setting, anticipating that u will have the same local dependence $u(s,z)$. Using

$$\Delta \approx \frac{1}{\epsilon^2}\frac{\partial^2}{\partial z^2}+\frac{\kappa}{\epsilon}\frac{\partial}{\partial z}$$

where κ is the curvature of the interface, we find at leading order $O(\epsilon^{-1})$ that u has the same homoclinic structure around a new curve $\mathbf{x}(s)$. That is,

$$u_{n+1} \approx g(z)+\epsilon v(z,s) \tag{3.5.1}$$

with $g(z)=\tanh(z/\sqrt{2})$ and where we have changed coordinates to (s,z) with

$$(x,y)=\mathbf{x}(s)+\epsilon z\hat{n}$$

based on the curve $\mathbf{x}(s)$ after the implicit time step. In the language of Remark 2 we predict that the scheme preserves profile fidelity and show below that this is asymptotically consistent. In (3.5.1), $v(z,s)$ is the correction to the leading order solution. We will identify the size and structure of this term below.

We take

$$\mathbf{x}_n=\mathbf{x}-k\rho(s)\hat{n} \tag{3.5.2}$$

where ρ is the average normal speed through the time step. Recalling that $k=c\epsilon$ and the spatial scaling of z, we have

$$u_n \approx g(z-c\rho(s)). \tag{3.5.3}$$

A diagram is shown in Figure 3.3. Note that the variation in normal direction

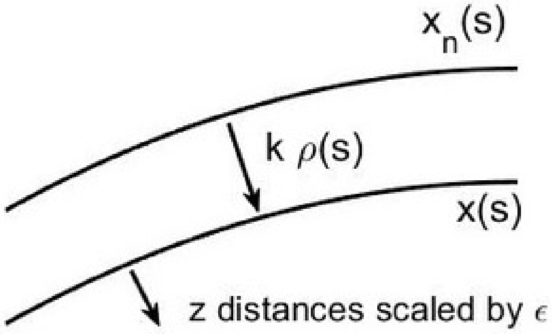

Figure 3.3: Sketch of the asymptotic analysis of the fully implicit problem. Here, $\rho(s)$ the average normal speed of the interface between time steps.

appears in higher order asymptotic terms, so it is consistent in what follows to use the same $\hat{n}$ as normal direction for both curves, *i.e.* the same "z".

Considering now the next order term $O(1)$ in (3.2.4) with the forms (3.5.1) and (3.5.3):

$$g' c\rho + \frac{1}{2} g'' c^2 \rho^2 + \frac{1}{6} g''' c^3 \rho^3 \approx c\kappa g' - c\mathscr{L} v \tag{3.5.4}$$

where $\mathscr{L} := \partial^2/\partial z^2 + f'(g)\cdot$ and we have used the smallness of c for the cubic Taylor approximation of $g(z) - g_n(z)$ on the right hand side. We consider (3.5.4) at each s in the $L^2(\mathbb{R})$ orthogonal decomposition of $G := \text{span}\{g'(z)\}$ and $G^\perp$. Note that $g'' \in G^\perp$ (this does not depend on the specific reaction term $f = u^3 - u$ chosen here) and $\mathscr{L}$ has G as its kernel and has bounded inverse on $G^\perp$ from standard Fredholm theory [52]. Thus we have $\rho = \kappa + O(c^2)$ and $v = O(c)$ in $G^\perp$. Careful examination of these results shows that the errors in $G^\perp$ do not accumulate and are globally of size $O(c\epsilon) = O(k)$ and so decrease as $\epsilon \to 0$. Global errors in interface *position* after $O(1/k)$ time steps are of size

c^2, independent of ϵ. Global solution value errors due to the position error have size $O(c^2/\epsilon) = O(k^2/\epsilon^3)$ and so it is seen that BE behaves like a second order method in this scaling. This explains the unexpected accuracy in AC BE computations as $\epsilon \to 0$.

Remark 6. *Note that the error estimator (3.2.7) uses $\mathscr{F}(u)$ which sees the undamped dominant truncation error term, which is why the number of time steps behaves with ϵ in the manner predicted in Section 3.3.1. Thus for BE applied to AC in the metastable regime, the estimator asymptotically overestimates the local errors actually made.*

The formal asymptotic results can also be used to show that the implicit time steps in this scaling lead to energy decrease. Neglecting the $O(c^2)$ terms in the interface motion, we have from (3.5.2)

$$\mathbf{x}_n = \mathbf{x} - k\kappa\hat{n}.$$

Using the identities for arc length parametrized curves $|\mathbf{x}_s| = 1$, $\kappa\hat{n} = \mathbf{x}_{ss}$ and $\mathbf{x}_s \cdot \mathbf{x}_{sss} = -\kappa^2$ it follows by taking the s derivative of the equation above and the dot product with $\mathbf{x}_s$ at each s that

$$|\mathbf{x}_{n,s}| \geq 1 + k\kappa^2 \geq 1 = |\mathbf{x}_s|.$$

This shows that the metastable curve at time n is longer than at the next step $n+1$. Since the energy $\mathscr{E}$ is proportional to curve length to highest order in the metastable regime [64], we have shown formally that implicit time stepping for AC has the energy decay property under this time step scaling. Large, accurate, fully implicit time steps can be taken in computations validated in Section 3.4.

In the next section we show a closely related rigorous result in a radial geometry. The main result is in Proposition 3.6.1. A key ingredient is an identification of a dominant term in the space G that represents the interface motion, separate from heavily damped terms in the perpendicular space, as shown here. Care must be taken to control the size of terms which are formally neglected in this asymptotic analysis.

3.6 Rigorous Radial Analysis of AC With BE and Eyre Time Stepping

We derive rigorous asymptotic evolution of a radially symmetric profile for BE and first order Eyre-type methods for the Allen-Cahn equation in $\mathbb{R}^2$. Extensions to radial profiles in $\mathbb{R}^d$ is immediate. More precisely we consider a splitting $f = f_+ - f_-$ and study the iterative scheme

$$\begin{aligned} &\frac{u-u_n}{k} = u_{rr} + \frac{1}{r}u_r - \frac{1}{\varepsilon^2}(f_+(u) - f_-(u_n))\,, \qquad\qquad r \in [0,\infty) \\ &u_r(0) = 0, u(\infty) = 1. \end{aligned}$$

For simplicity, we assume that f is smooth, odd about $u = 0$, has precisely three simple zeros at $u = \pm 1$ and at $u = 0$, and tends to $\pm\infty$ as $u \to \pm\infty$. This includes the classical choice of $f(u) = u^3 - u$ but we consider other reaction terms in this class since Eyre's method can have quite different behaviour as shown in Section 3.6.2. The BE scheme corresponds to the choice $f_- \equiv 0$ while

Eyre-type schemes take $f'_+, f'_- \geq 0$. We pose the problem on the affine space

$$Y := \{u + 1 \in H^1_R(0,\infty) \,\big|\, \partial_r u(0) = 0.\},$$

with $u_n \in Y$ as a given. The assumptions on f imply that the continuous 1D Allen-Cahn equation has a steady state solution

$$g_{zz} = f(u), \tag{3.6.1}$$

which is heteroclinic to ±1; that is $g \to \pm 1$ as $z \to \pm\infty$. Considering $R > 1$, we modify this g at the order $O(e^{-1/\epsilon})$ so that $g' = 0$ on $(-\infty, -1/\epsilon)$ for $\epsilon \ll 1$. This introduces exponentially small residuals in the sequel that have no impact upon the salient results of our analysis.

We introduce $z = \frac{r-R}{\varepsilon}$, the weighted inner product

$$\langle u, v \rangle_R := \int_{-R/\epsilon}^{\infty} u(z)v(z)(R + \epsilon z)dz,$$

and the associated spaces L^2_R and H^1_R with $B_{H^1_R}(\delta)$ the ball that is centered at the origin with radius δ in the space H^1_R. We rewrite the iterative equation as

$$\frac{u - u_n}{k} = \epsilon^{-2}(u_{zz} - (f_+(u) - f_-(u_n))) + \frac{\epsilon^{-1}u_z}{R + \epsilon z}, \tag{3.6.2}$$

on the domain $z \in (-R/\epsilon, \infty)$. We decompose u_n and u as

$$u_n = g\left(z + \frac{R - R_n}{\epsilon}\right) + v_n,$$

$$u = g(z) + v,$$

where R_n and v_n are taken as given and R and v are to be determined. The profile associated to u_n is denoted g_n and observe that it admits the expansion

$$g_n = g\left(z + \frac{R - R_n}{\epsilon}\right) = g + g' \frac{R - R_n}{\epsilon} + O\left(\left(\frac{R - R_n}{\epsilon}\right)^2\right).$$

In the sequel we will enforce the orthogonality conditions

$$\langle v, g' \rangle_R = 0, \quad \langle v_n, g_n' \rangle_R = 0, \tag{3.6.3}$$

and denote the corresponding subspaces of L^2_R by $X^\perp$ and $X_n^\perp$ respectively with the associated orthogonal projections Π and Π_n.

At this point the analysis of the implicit and Eyre-type schemes diverges sufficiently that we approach them distinctly.

3.6.1 Backward Euler estimates

For BE we take $f'_- \equiv 0$, $f = f_+$, and write the iterative map as

$$v + \frac{k}{\epsilon^2} L v = v_n - (g - g_n) + \frac{k g'}{\epsilon(R + \epsilon z)} - \frac{k}{\epsilon^2} \mathcal{N}, \tag{3.6.4}$$

where we have introduced the linear operator

$$L := -\left(\partial_z^2 + \frac{\epsilon}{R + \epsilon z} \partial_z\right) + f'(g) = -\frac{1}{R + \epsilon z} \partial_z \left((R + \epsilon z) \partial_z\right) + f'(g), \tag{3.6.5}$$

and the nonlinearity

$$\mathcal{N}(v) := f(g + v) - (f(g) + f'(g) v).$$

The operator L is self-adjoint in the weighted inner product for which the eigenvalue problem takes the form

$$L\psi = \frac{\lambda}{R+\epsilon z}\psi,$$

subject to $\partial_z \psi(-R/\epsilon) = 0$ and $\psi \to 0$ as $z \to \infty$. Since the profile g solves (3.6.1), it will be useful to compare L to the simpler operator

$$L_0 := -\partial_z^2 + f'(g), \tag{3.6.6}$$

arising as the linearization of (3.6.1) about g in $L^2(\mathbb{R})$. The operator L_0 is self-adjoint on $L^2(\mathbb{R})$, and since g is heteroclinic with $g' > 0$, the Sturm-Liouville theory on $L^2(\mathbb{R})$ implies that L_0 has a simple, ground-state eigenvalue at $\lambda = 0$ with eigenfunction g' and the remainder of the spectrum of L_0 is strictly positive, in particular L_0 is uniformly coercive on the space $\{g'\}^{\perp}_{L^2(\mathbb{R})}$. While L does not generically have a kernel, it does have an eigenspace with a small associated eigenvalue. However, for ϵ sufficiently small, it inherits the coercivity of L_0.

Lemma 1. *Fix $\epsilon_0 > 0$ sufficiently small, then there exists $\alpha > 0$, independent of $R \geq 1$ and of $\epsilon \in (0, \epsilon_0)$, such that*

$$\langle Lv, v\rangle_R \geq \alpha \|v\|^2_{H^1_R}. \tag{3.6.7}$$

for all $v \in H^1_R$ satisfying $\langle v, g'\rangle_R = 0$.

Proof. We defer the proof of L^2_R coercivity to Section 3.7. To extend coercivity to H^1_R we observe that

$$\langle Lv, v\rangle_R = \int_{-R/\epsilon}^{\infty} (R+\epsilon z)\left(|v'|^2 + f'(g)|v|^2\right)dz,$$

so that for any $t \in (0,1)$ we may write

$$\begin{aligned}\langle Lv, v\rangle_R &= t\langle Lv, v\rangle_R + (1-t)\langle Lv, v\rangle_R,\\ &\geq \int_{-R/\epsilon}^{\infty} (R+\epsilon z)(t|v'|^2 + ((1-t)\alpha - t\|f'(g)\|_\infty)|v|^2)dz,\\ &\geq \tilde{\alpha}\|v\|^2_{H^1_R},\end{aligned}$$

where we have introduced $\tilde{\alpha} := \alpha/(1+\alpha+\|f'(g)\|_\infty) > 0$. Dropping the tilde, we have (3.6.7) with α independent of $\epsilon > 0$ and $R > 1$. □

We assume throughout our analysis that $\|v\|_{H^1_R}$ and $\|v_n\|_{H^1_R}$ are uniformly bounded by $\delta \ll 1$. Returning to (3.6.4), we denote its right-hand side as $\mathscr{F}_{\mathrm{BE}}$. To have the inversion of the operator on the left-hand side be contractive the term $\mathscr{F}_{\mathrm{BE}}$ must be approximately orthogonal to the small eigenspace of L. As Lemma 1 shows it is sufficient to be L_R-orthogonal to g', the kernel of L_0. To this end we determine $R = \hat{R}_{\mathrm{BE}}(v, v_n, R_n)$ such that $\mathscr{F}_{\mathrm{BE}} \in X^\perp$, or equivalently

$$\langle \mathscr{F}_{\mathrm{BE}}, g'\rangle_R = 0. \tag{3.6.8}$$

Assuming this condition has been enforced we introduce

$$M := I + \frac{k}{\epsilon^2}L,$$

and may rewrite the BE iteration in the equivalent formulation

$$v = \mathscr{G}_{\mathrm{BE}}(v, v_n, R - R_n) := M^{-1}\Pi\mathscr{F}_{\mathrm{BE}}(v, v_n, R - R_n). \tag{3.6.9}$$

The key step is the introduction of the operator Π, the orthogonal projection onto $X^\perp$. This is redundant when $\mathscr{F}_{\mathrm{BE}} \in X^\perp$, but preserves contractivity for choices of (v, v_n, R_R_n) when it is not. Our goal is to show the function $\mathscr{G}_{\mathrm{BE}}$ is a contraction mapping and to develop asymptotic formula for R and v.

Lemma 2. *The function $R = \hat{R}_{\mathrm{BE}}$ satisfies the implicit relation*

$$\frac{R - R_n}{k} = -\frac{1}{R} + \frac{k}{4R^3} - \frac{b_1 k^2}{\epsilon^2 R^3} + O\left(\delta, \frac{k^3}{\epsilon^2}, \frac{\delta^2}{\epsilon}\right). \tag{3.6.10}$$

where

$$b_1 := \frac{\|g''\|_R^2}{6\|g'\|_R^2} > 0. \tag{3.6.11}$$

Moreover we have the Lipshitz estimate

$$|\hat{R}_{\mathrm{BE}}(v; v_n, R_n) - \hat{R}_{\mathrm{BE}}(\tilde{v}; v_n, R_n)| \le c\frac{k\delta}{\epsilon}\|v - \tilde{v}\|_R, \tag{3.6.12}$$

so long as $k\delta^2 \ll \epsilon^2$.

Proof. Due to parity considerations, we remark that $\|g'\|_R^2 = R\|g'\|_{L^2(\mathbb{R})}^2$, up to exponentially small terms. For brevity, and as an element of foreshadowing, we approximate $(R - R_n)$ by k in the O-error terms. We address the terms in

$\mathscr{F}_{\mathrm{BE}}$ and derive the following elementary estimates,

$$\langle v_n, g'\rangle_R = \langle v_n, (g'-g_n')\rangle_R = -\langle v_n, g''\rangle_R \frac{R-R_n}{\epsilon} + O\left(\delta\frac{k^2}{\epsilon^2}\right), \tag{3.6.13}$$

$$\begin{aligned}\langle g-g_n, g'\rangle_R = -\|g'\|_R^2\left(\frac{(R-R_n)}{\epsilon} - \frac{(R-R_n)^2}{4R\epsilon}\right) + \\ + \frac{\|g''\|_R^2}{6}\frac{(R-R_n)^3}{\epsilon^3} + O\left(\frac{k^4}{\epsilon^3}\right),\end{aligned} \tag{3.6.14}$$

$$\left\langle \frac{g'}{R+\epsilon z}, g'\right\rangle_R = \|g'\|_{L^2(\mathbb{R})}^2 = \frac{\|g'\|_R^2}{R}, \tag{3.6.15}$$

$$|\langle \mathscr{N}, g'\rangle_R| \le c\delta^2. \tag{3.6.16}$$

For this scheme, $\mathscr{F}_{\mathrm{BE}}$ depends upon v only through $\mathscr{N}$. Collecting terms in the orthogonality condition that are linear in $R-R_n$ and identifying relevant higher order terms yields the relation

$$\frac{R-R_n}{k} = -\frac{1}{R} + \frac{(R-R_n)^2}{4Rk} + \frac{b_1(R-R_n)^3}{k\epsilon^2} + O\left(\delta, \frac{k^3}{\epsilon^2}, \frac{\delta^2}{\epsilon}\right) \tag{3.6.17}$$

where b_1 is given in (3.6.11). Under the assumptions on k and δ we have the leading order result $R-R_n = -k/R$. Substituting this relation into (3.6.17) yields the result (3.6.10).

To obtain the Lipschitz estimate we observe from the estimates above that

$$|\hat{R}_{\mathrm{BE}}(v) - \hat{R}_{\mathrm{BE}}(\tilde{v})| \le c\frac{k}{\epsilon\|g'\|_R^2}\left|\langle \mathscr{N}(v), g'\rangle_R - \langle \mathscr{N}(\tilde{v}), \tilde{g}'\rangle_R\right|.$$

The nonlinearity satisfies the Lipschitz properties

$$\|\mathscr{N}(v) - \mathscr{N}(\tilde{v})\|_R \le c\delta\|v-\tilde{v}\|_R,$$

while

$$\|g' - \tilde{g}'\|_R \le c \frac{|\hat{R}_{\mathrm{BE}}(v) - \hat{R}_{\mathrm{BE}}(\tilde{v})|}{\epsilon}.$$

Adding and subtracting $\langle \mathcal{N}(\tilde{v}), g' \rangle_R$ and using (3.6.16), we arrive at the estimates

$$|\hat{R}_{\mathrm{BE}}(v) - \hat{R}_{\mathrm{BE}}(\tilde{v})| \le c \left(\frac{k\delta}{\epsilon} \|v - \tilde{v}\|_R + \frac{k\delta^2}{\epsilon^2} |\hat{R}_{\mathrm{BE}}(v) - \hat{R}_{\mathrm{BE}}(\tilde{v})| \right).$$

Imposing the condition $k\delta^2 \ll \epsilon^2$ yields (3.6.12). □

To establish bounds on the map $\mathcal{G}_{\mathrm{BE}}$ defined in (3.6.9) we apply M to both sides of the relation and take the L^2_R inner product with respect to $\mathcal{G}_{\mathrm{BE}}$. Using the coercivity estimate (3.6.7) we find

$$\|\mathcal{G}_{\mathrm{BE}}\|_R^2 + \alpha \frac{k}{\epsilon^2} \|\mathcal{G}_{\mathrm{BE}}\|_{H^1_R}^2 \le \|\Pi \mathcal{F}_{\mathrm{BE}}\|_R \|\mathcal{G}_{\mathrm{BE}}\|_R.$$

Taking $v, v_n \in B_{H^1_R}(\delta)$ for $\delta \ll 1$ and recalling that the projection Π crucially cancels the leading order term in $g - g_n$, we estimate

$$\|\mathcal{G}_{\mathrm{BE}}\|_R + \alpha \frac{k}{\epsilon^2} \|\mathcal{G}_{\mathrm{BE}}\|_{H^1_R} \le c \left(\delta + \frac{k^2}{\epsilon^2} + k + \frac{k}{\epsilon^2} \delta^2 \right). \tag{3.6.18}$$

For the BE system we examine distinguished limits $k = \epsilon^s$, for $s \in (1,2)$, which we call the large time-step regime, for which the H^1_R term is dominant on the left-hand side of (3.6.18). We drop the L^2_R term to find,

$$\|\mathcal{G}_{\mathrm{BE}}\|_{H^1_R} \le c \left(\delta \epsilon^{2-s} + \epsilon^s + \delta^2 \right).$$

Taking $\delta = \epsilon^{s'}$ for any $s' > \max\{s/2, 2(s-1)\}$ then we determine that

$$\|\mathcal{G}_{\mathrm{BE}}\|_{H^1_R} \le c(\epsilon^{2-s+s'} + \epsilon^s + \epsilon^{2s'}) \le \delta,$$

for ϵ sufficiently small. In particular, since $s > \max\{s/2, 2(s-1)\}$ in the large time-stepping regime, we may take $\delta = k = \epsilon^s$, so that, viewing $\mathcal{G}_{\mathrm{BE}}$ as a map on (v, v_n), we have $\mathcal{G}_{\mathrm{BE}} : B_{H^1_R}(k) \times B_{H^1_R}(k) \mapsto B_{H^1_R}(k)$, for all s in the large time-step regime.

Proposition 3.6.1. *Fix $1 < s < 2$, then in the distinguished limit $k = \epsilon^s$ the function $\mathcal{G}_{\mathrm{BE}}$ defined in (3.6.9) with $R := R_{n+1} = \hat{R}_{\mathrm{BE}}(v; v_n, R_n)$ maps $B_{H^1_R}(k) \times B_{H^1_R}(k)$ into $B_{H^1_R}(k)$ and is a strict contraction. In particular it has a unique solution $v \in B_{H^1_R}(k)$, denoted by v_{n+1} which satisfies*

$$\left\| v_{n+1} - \frac{k}{R^2} L\Pi g'' \right\|_{H^1_R} \le c\frac{\epsilon^2}{k}\|v_n\|_R + O(k^2).$$

In particular there exists $c > 0$ such that for all $v_0 \in B_{H^1_R}(ck)$ and $R_0 > 1$ the sequence $\{(v_n, R_n)\}_{n=1}^N$ satisfies $v_n \in B_{H^1_R}(ck)$ while $\{R_n\}_{n=0}^N$ satisfies the backwards Euler iteration

$$\frac{R_{n+1} - R_n}{k} = -\frac{1}{R} - \frac{b_1 k^2}{\epsilon^2 R^3} + O(k), \tag{3.6.19}$$

where $b_1 > 0$ is given by (3.6.11). Here N is the iteration number such that $R_N > 1$ and $R_{N+1} < 1$.

Proof. We have established the mapping property. To establish the contractivity we must control the impact of f upon the projection Π through the motion of the front R. We assume that $v,\tilde{v},v_n \in B_{H^1_R}(k)$ and denote $R = R(v)$ and $\tilde{R} = R(\tilde{v})$, with the associated front profiles denoted by g and $\tilde{g}$. The estimate (3.6.12) establishes that $\hat{R}_{\rm BE}$ is Lipschitz with constant $ck\delta/\epsilon$, which in the the large time-step regime reduces to ck^2/ϵ. Following the proof of (3.6.12) we find that

$$\|\mathscr{F}_{\rm BE}(v) - \mathscr{F}_{\rm BE}(\tilde{v})\|_{H^1_R} \leq c\frac{k^2}{\epsilon^2}\|v-\tilde{v}\|_{H^1_R}. \tag{3.6.20}$$

In the large time-step regime, using (3.6.7) we deduce the bound

$$\|M^{-1}\Pi f\|_{H^1_R} \leq \alpha^{-1}\frac{\epsilon^2}{k}\|f\|_R \tag{3.6.21}$$

We wish to obtain a bound on the difference of $\mathscr{G}_{\rm BE}$ at two values of v:

$$\mathscr{G}_{\rm BE}(v,v_n) - \mathscr{G}_{\rm BE}(\tilde{v},v_n) = M^{-1}\Pi\mathscr{F}_{\rm BE} - \tilde{M}^{-1}\tilde{\Pi}\tilde{\mathscr{F}}_{\rm BE}.$$

We first bound the difference

$$\mathrm{g}_{\rm BE} := (M^{-1}\Pi - \tilde{M}^{-1}\tilde{\Pi})\mathscr{F}_{\rm BE}. \tag{3.6.22}$$

The analysis is complicated by the fact that M is only uniformly invertible on the range of Π. To factor these projected inverses we act with M, observing

$$M\mathrm{g}_{\rm BE} = (\Pi - M\tilde{M}^{-1}\tilde{\Pi})\mathscr{F}_{\rm BE} = (\Pi\tilde{M} - M)\tilde{M}^{-1}\tilde{\Pi}\mathscr{F}_{\rm BE} + \Pi(I-\tilde{\Pi})\mathscr{F}_{\rm BE}, \tag{3.6.23}$$

where we used that fact that $\tilde{M}\tilde{M}^{-1}\tilde{\Pi} = \tilde{\Pi}$ and hence $\tilde{M}\tilde{M}^{-1}\tilde{\Pi} + (I-\tilde{\Pi}) = I$. Since the right-hand side of (3.6.23) lies in the range of Π we may invert boundedly,

$$\Pi g_{\mathrm{BE}} = M^{-1}\Pi(\tilde{M}-M)\tilde{M}^{-1}\tilde{\Pi}\mathscr{F}_{\mathrm{BE}} + M^{-1}\Pi(I-\tilde{\Pi})\mathscr{F}_{\mathrm{BE}}. \tag{3.6.24}$$

To recover the whole g_{BE} we act with $(I-\Pi)$ on (3.6.22) obtaining

$$(I-\Pi)g_{\mathrm{BE}} = -(I-\Pi)\tilde{M}^{-1}\tilde{\Pi}\mathscr{F}_{\mathrm{BE}} = -(\tilde{\Pi}-\Pi)\tilde{M}^{-1}\tilde{\Pi}\mathscr{F}_{\mathrm{BE}}. \tag{3.6.25}$$

Adding (3.6.25) to (3.6.24) yields a regularized expression that accounts for the shifts in the projections

$$g_{\mathrm{BE}} = M^{-1}\Pi(\tilde{M}-M)\tilde{M}^{-1}\tilde{\Pi}\mathscr{F}_{\mathrm{BE}} + M^{-1}\Pi(\Pi-\tilde{\Pi})\mathscr{F}_{\mathrm{BE}} + (\Pi-\tilde{\Pi})\tilde{M}^{-1}\tilde{\Pi}\mathscr{F}_{\mathrm{BE}}. \tag{3.6.26}$$

The operators $M^{-1}\Pi$ and $\tilde{M}^{-1}\tilde{\Pi}$ are bounded using (3.6.21), while

$$\begin{aligned}\|\tilde{M}-M\|_{R*} = \frac{k}{\epsilon}\|f'(g)-f'(\tilde{g})\|_{R*} \le c\|g-\tilde{g}\|_{R*},\\ \le c\frac{|R-\tilde{R}|}{\epsilon} \le c\frac{k^2}{\epsilon^2}\|v-\tilde{v}\|_{H^1_R},\end{aligned}$$

where $\|\cdot\|_{R*}$ denotes the operator norm from L^2_R into itself. The projections satisfy

$$\begin{aligned}\|(\Pi-\tilde{\Pi})f\|_R &= \|g'\langle g',f\rangle_R - \tilde{g}'\langle\tilde{g}',f\rangle\|_R,\\ &\le c\frac{k^2}{\epsilon^2}\|v-\tilde{v}\|_{H^1_R}\|f\|_R + c\frac{\epsilon^2}{k}\frac{k^2}{\epsilon^2}\|v-\tilde{v}\|_{H^1_R}\end{aligned}$$

Applying these estimates to (3.6.26) and using (3.6.18) to estimate $\Pi\mathscr{F}_{\mathrm{BE}}$ we obtain

$$\begin{aligned}\|g_{\mathrm{BE}}\|_{H^1_R} \leq& c\left(\frac{\epsilon^4}{k^2}\frac{k^2}{\epsilon^2}+\frac{\epsilon^2}{k}\frac{k^2}{\epsilon^2}\right)\|v-\tilde{v}\|_{H^1_R}\|\Pi\mathscr{F}_{\mathrm{BE}}\|_{L^2},\\ \leq& c\left(\epsilon^2+k\right)\frac{k^2}{\epsilon^2}\|v-\tilde{v}\|_{H^1_R}.\end{aligned} \tag{3.6.27}$$

Finally we write

$$\mathscr{G}_{\mathrm{BE}}(v,v_n)-\mathscr{G}_{\mathrm{BE}}(\tilde{v},v_n)=g_{\mathrm{BE}}+\tilde{M}^{-1}\tilde{\Pi}(\mathscr{F}_{\mathrm{BE}}-\tilde{\mathscr{F}}_{\mathrm{BE}}),$$

and using (3.6.20), (3.6.27) estimate

$$\|\mathscr{G}_{\mathrm{BE}}(v,v_n)-\mathscr{G}_{\mathrm{BE}}(\tilde{v},v_n)\|_{H^1_R} \leq c\left(\frac{k^3}{\epsilon^2}+k\right)\|v-\tilde{v}\|_{H^1_R},$$

which is contractive so long as $k \ll \epsilon^{\frac{2}{3}}$ which holds with the large time-step regime.

Within the large time-step regime the leading order iteration (3.6.17) simplifies as $k \ll k^2/\epsilon^2$ and the dominant correction is given by the b_1 term. To compare to standard notation we rewrite the regime as $\epsilon^2 \ll k = \delta \ll 1$ and replace the internal parameter δ with k, the result is the large time-step interation (3.6.19). □

3.6.2 Eyre-type iterations

For an Eyre iteration the map (3.6.2) takes the form

$$v+\frac{k}{\epsilon^2}L_+v=v_n-(g-g_n)+\frac{kg'}{\epsilon(R+\epsilon z)}+\frac{k}{\epsilon^2}(\mathscr{R}-\mathscr{N}), \tag{3.6.28}$$

where we have introduced the Eyre linear operator

$$L_+ := -\left(\partial_z^2 + \frac{\epsilon}{R+\epsilon z}\partial_z\right) + f_+'(g) = -\frac{1}{R+\epsilon z}\partial_z\left((R+\epsilon z)\partial_z\right) + f_+'(g), \tag{3.6.29}$$

the explicit-term residual

$$\mathscr{R}(v, v_n) := f_-(g_n) - f_-(g) + f_-'(g_n)v_n,$$

and the nonlinearity

$$\mathcal{N}(v, v_n) := \mathcal{N}_+(v) - \mathcal{N}_-(v_n),$$

which we further decompose into implicit and explicit parts

$$\begin{aligned} \mathcal{N}_+(v) &:= f_+(g+v) - (f_+(g) + f_+'(g)v), \\ \mathcal{N}_-(v_n) &:= f_-(g_n+v_n) - (f_-(g_n) + f_-'(g_n)v_n). \end{aligned}$$

The operator L_+ is self-adjoint in the weighted inner product for which the eigenvalue problem takes the form

$$L_+\psi = \frac{\lambda}{R+\epsilon z}\psi,$$

subject to $\partial_z\psi(-R/\epsilon) = 0$ and $\psi \to 0$ as $z \to \infty$. The coercivity estimate is substantially simpler than for BE as the operator L_+ is strictly positive without constraint.

Lemma 3. *There exists $\alpha_+ > 0$, independent of $R \geq 1$, such that*

$$\langle L_+ v, v\rangle_R \geq \alpha_+ \|v\|^2_{H^1_R}. \tag{3.6.30}$$

for all $v \in H^1_R$.

Proof. Since $f'_+ \geq 0$ the normalized ground-state eigenfunction ψ_0 of L_+, satisfies

$$\lambda_0^+ = \langle L_+\psi_0, \psi_0\rangle_R = \int_{-R/\epsilon}^{\infty} \left((\partial_z \psi_0)^2 + f'_+(g)\psi_0^2\right)(R+\epsilon z)dz > 0.$$

Since the ground-state eigenvalue is strictly positive, this establishes the L^2_R coercivity of L_+ with $\alpha_+ = \lambda_0^+$. The H^1_R coercivity follows as in Lemma 1. □

We assume throughout our analysis that $\|v\|_{H^1_R}$ and $\|v_n\|_{H^1_R}$ are uniformly bounded by $\delta \ll 1$. We denote the right-hand side of (3.6.28) by $\mathscr{F}_{\mathrm{E}}$ and introduce

$$M_+ := I + \frac{k}{\epsilon^2} L_+,$$

which is strictly contractive on the full space L^2_R, and re-write the Eyre iteration as

$$v = \mathscr{G}_{\mathrm{E}}(v, v_n, R - R_n) := M_+^{-1}\Pi\mathscr{F}_{\mathrm{E}}(v, v_n, R - R_n). \tag{3.6.31}$$

For the Eyre iteration the role of the projection Π is diminished as M_+ is contractive without it. Our goal is to show the existence of a map $R = \hat{R}_{\mathrm{E}}(v, v_n, R_n)$, for which

$$\langle \mathscr{F}_{\mathrm{E}}, g'\rangle_R = 0, \tag{3.6.32}$$

to establish the contractive mapping properties of $\mathscr{G}_{\mathrm{E}}$, and to develop asymptotic formula for R and v. We do this in the long time-stepping regime, $k \gg \epsilon^2$, which has no lower bound for the Eyre scheme.

Lemma 4. *Assume $k \gg \epsilon^2$. There exists a smooth function $\hat{R}_{\mathrm{E}} : B_{H^1_R}(\delta) \times B_{H^1_R}(\delta) \times \mathbb{R} \mapsto \mathbb{R}$ such that the profile $g = g(z;R)$ satisfies (3.6.32). The function $R = \hat{R}_{\mathrm{E}}$ satisfies the implicit relation*

$$R - R_n = -\frac{\epsilon^2}{c_- R} + O\left(\epsilon^3, \delta\epsilon, \frac{\epsilon^4}{k}\right). \tag{3.6.33}$$

where we have introduced the leading order Eyre time constant

$$c_- := \frac{\langle f_-'(g) g', g' \rangle_R}{\|g'\|_R^2} > 0 \tag{3.6.34}$$

when $f_-' \not\equiv 0$. Moreover we have the Lipshitz estimate

$$|\hat{R}_{\mathrm{E}}(v; v_n, R_n) - \hat{R}_{\mathrm{E}}(\tilde{v}; v_n, R_n)| \leq c\epsilon\delta \|v - \tilde{v}\|_R, \tag{3.6.35}$$

so long as $\delta \ll 1$.

Proof. Due to parity considerations, we remark that $\|g'\|_R^2 = R\|g'\|^2_{L^2(\mathbb{R})}$, up to exponentially small terms. For brevity, and as an element of foreshadowing, we approximate $(R - R_n)$ by ϵ^2 in the O-error terms. Addressing the terms in

$\mathscr{F}_{\rm BE}$ one by one, we record

$$\langle v_n, g'\rangle_R = \langle v_n, (g' - g'_n)\rangle_R = O(\delta\epsilon) \tag{3.6.36}$$

$$\langle g - g_n, g'\rangle_R = -\|g'\|_R^2 \frac{(R - R_n)}{\epsilon} + O\left(\epsilon^3\right), \tag{3.6.37}$$

$$\left\langle \frac{g'}{R + \epsilon z}, g' \right\rangle_R = \|g'\|^2_{L^2(\mathbb{R})} = \frac{\|g'\|_R^2}{R}, \tag{3.6.38}$$

$$\langle \mathscr{R}, g'\rangle_R = \frac{\langle f'_-(g)g', g'\rangle_R (R - R_n)}{\epsilon} + \langle f'_-(g) v_n, g'\rangle_R + O(\epsilon^2, \epsilon\delta), \tag{3.6.39}$$

$$|\langle \mathscr{N}, g'\rangle_R| \le c\delta^2. \tag{3.6.40}$$

With these reductions we can simplify the orthgonality condition, identifying terms that are linear in $R - R_n$ and most relevant higher order terms. The result is the balance

$$\frac{R - R_n}{k}\left(1 + \frac{c_- k}{\epsilon^2}\right) = -\frac{1}{R} - \frac{\langle f'_-(g)g', v_n\rangle_R}{\epsilon \|g'\|_R^2} + O\left(\epsilon, \delta, \frac{\delta^2}{\epsilon}\right), \tag{3.6.41}$$

where c_-, introduced in (3.6.34) is positive since $f'_- \ge 0$ by assumption. The largest terms and error terms come from the residual, and we kept the lower order constant on the left-hand side to emphasize that in the long time-stepping regime, the residual dominates the natural time-step term. Indeed, the iteration is independent of step size, k, given at leading order by (3.6.41).

To obtain the Lipshitz estimate we observe from the bounds above that dependence of $\hat{R}_{\rm E}$ on v arises from the balance of the linear $R - R_n$ term in the residual against the nonlinearity. Since both these terms are multiplied by k/ϵ^2 this factor cancels and we have the balance

$$|\hat{R}_{\mathrm{E}}(v)-\hat{R}_{\mathrm{E}}(\tilde{v})| \le \frac{c\epsilon}{\langle f'_-(g)g',g'\rangle_R}\left|\langle \mathcal{N}(v),g'\rangle_R - \langle \mathcal{N}(\tilde{v}),\tilde{g}'\rangle_R\right|.$$

The nonlinearity satisfies the Lipshitz properties

$$\|\mathcal{N}(v,v_n)-\mathcal{N}(\tilde{v},v_n)\|_R \le c\delta\|v-\tilde{v}\|_R,$$

while

$$\|g'-\tilde{g}'\|_{L^2} \le c\frac{|\hat{R}_{\mathrm{E}}(v)-\hat{R}_{\mathrm{E}}(\tilde{v})|}{\epsilon}.$$

Adding and subtracting $\langle \mathcal{N}(\tilde{v},v_n),g'\rangle_R$ and using (3.6.40), we arrive at the estimates

$$|\hat{R}_{\mathrm{E}}(v)-\hat{R}_{\mathrm{E}}(\tilde{v})| \le c\left(\epsilon\delta\|v-\tilde{v}\|_R + \delta^2|\hat{R}_{\mathrm{BE}}(v)-\hat{R}_{\mathrm{BE}}(\tilde{v})|\right).$$

Imposing the condition $\delta \ll 1$ yields (3.6.35). □

We may now establish the main result on the Eyre sequence.

Proposition 3.6.2. *There exists $c>0$ such that for any $k \gg \epsilon^2$ the function $\mathcal{G}_{\mathrm{E}}$ defined in (3.6.31) with $R := \hat{R}_{\mathrm{E}}(v;v_n,R_n)$ maps $B_{H^1_R}(c\epsilon)\times B_{H^1_R}(c\epsilon)$ into $B_{H^1_R}(\mathcal{G}_{\mathrm{E}}(0,v_n),\epsilon^2)$ and is a strict contraction, satisfying*

$$\|\mathcal{G}_{\mathrm{E}}(v,v_n)-\mathcal{G}_{\mathrm{E}}(\tilde{v},v_n)\|_{H^1_R} \le c\epsilon^2\|v-\tilde{v}\|_{H^1_R}. \tag{3.6.42}$$

In particular $\mathcal{G}_{\mathrm{E}}$ has a unique fixed point in that set, which we denote v_{n+1}.

Moreover, if the Eyre balance parameter

$$\gamma := \|L_+^{-1}\Pi \circ f_-'(g)g'\|_{H^1_R{}^*} < 1 \tag{3.6.43}$$

then for any $\rho \in (0,1)$ *there exists* $c > 0$ *such that for all* $v_0 \in B_{H^1_R}(c\epsilon)$ *and* $R_0 > 1$*, the sequence* $\{(v_n, R_n)\}_{n=1}^N$ *satisfies* $v_n \in B_{H^1_R}(c\epsilon)$ *and*

$$\frac{R_{n+1} - R_n}{\epsilon^2} = -\frac{c_E}{R_{n+1}} + O\left(\epsilon^{1-\rho}\right).$$

where the Eyre number, c_E*, is defined by*

$$c_E := \frac{\|g'\|_R^2}{\langle f_-'(g)g', g'\rangle_R + \langle K_+ L_+^{-1}\Pi f_-'(g)g', f_-'(g)g'\rangle_R} > 0, \tag{3.6.44}$$

where $K_+ > 0$ *is defined in (3.6.52) and* $K_+L_+^{-1}\Pi > 0$ *is self-adjoint.*

Proof. To establish the contractivity of $\mathscr{G}_{\mathrm{E}}$ we follow the arguments for backward Euler, sketching only the differences. We introduce

$$\mathrm{g_E} := (M_+^{-1}\Pi - \tilde{M}_+^{-1}\tilde{\Pi})\mathscr{F}_{\mathrm{E}}; \tag{3.6.45}$$

and derive the expression

$$\mathrm{g_E} = M_+^{-1}\Pi(\tilde{M}_+ - M_+)\tilde{M}_+^{-1}\tilde{\Pi}\mathscr{F}_{\mathrm{E}} + \\ M_+^{-1}\Pi(\Pi - \tilde{\Pi})\mathscr{F}_{\mathrm{E}} + (\Pi - \tilde{\Pi})\tilde{M}_+^{-1}\tilde{\Pi}\mathscr{F}_{\mathrm{E}}. \tag{3.6.46}$$

The operators $M_+^{-1}\Pi$ and $\tilde{M}_+^{-1}\tilde{\Pi}$ are bounded as L_+ has no small eigenvalues.

Using (3.6.35) we estimate

$$\begin{aligned}\|\tilde{M}_+ - M_+\|_{R*} = \frac{k}{\epsilon}\|f_+'(g) - f_+'(\tilde{g})\|_{R*} &\le c\|g - \tilde{g}\|_{R*},\\ \le c\frac{|R - \tilde{R}|}{\epsilon} &\le c\delta\|v - \tilde{v}\|_{H^1_R}.\end{aligned}$$

Similarly the projections satisfy

$$\begin{aligned}\|(\Pi - \tilde{\Pi})\|_{R*} &= \|g'\langle g', \cdot\rangle_R - \tilde{g}'\langle \tilde{g}', \cdot\rangle\|_R,\\ &\le c\delta\|v - \tilde{v}\|_{H^1_R}\end{aligned}$$

Applying these estimates to (3.6.45,3.6.46) and following the proof of (3.6.35) to estimate $\Pi\mathscr{F}_{\mathrm{E}}$ we obtain

$$\begin{aligned}\|\mathrm{g}_{\mathrm{E}}\|_{H^1_R} \le & c\frac{\epsilon^2}{k}\delta\|v - \tilde{v}\|_{H^1_R}\|\Pi\mathscr{F}_{\mathrm{BE}}\|_{L^2},\\ &\le c\left(\frac{\epsilon^2\delta^2}{k} + \frac{\epsilon^4\delta}{k} + \epsilon\delta + \delta^2\right)\|v - \tilde{v}\|_{H^1_R}.\end{aligned} \tag{3.6.47}$$

Finally we write

$$\mathscr{G}_{\mathrm{E}}(v, v_n) - \mathscr{G}_{\mathrm{E}}(\tilde{v}, v_n) = \mathrm{g}_{\mathrm{E}} + \tilde{M}^{-1}\tilde{\Pi}(\mathscr{F}_{\mathrm{E}} - \tilde{\mathscr{F}}_{\mathrm{E}}), \tag{3.6.48}$$

and estimate the $\mathscr{F}_{\mathrm{E}}$ term from which the dominant contribution comes from the residual

$$\begin{aligned}\|\mathscr{F}_{\mathrm{E}} - \tilde{\mathscr{F}}_{\mathrm{E}}\|_R \le c\frac{k}{\epsilon^2}\|f_-(g) - f_-(\tilde{g})\|_R &\le c\frac{k}{\epsilon^2}\frac{|R - \tilde{R}|}{\epsilon},\\ \le c\frac{k\delta}{\epsilon}\|v - \tilde{v}\|_R,&\end{aligned}$$

where we used (3.6.35) in the last inequality. In particular we deduce that

$$\|\tilde{M}_+^{-1}\tilde{\Pi}(\mathscr{F}_{\rm E}-\tilde{\mathscr{F}}_{\rm E})\|_{H^1_R} \leq c\epsilon\delta\|v-\tilde{v}\|_R. \tag{3.6.49}$$

Combining (3.6.47), (3.6.49) and (3.6.48), imposing $\delta=\epsilon$, and using $k \gg \epsilon^2$ we arrive at strict contractivity on $B_{H^1_R}(c\delta)$ for any fixed $c>0$.

To establish bounds on the the fixed point v_{n+1} of $\mathscr{G}_{\rm E}(\cdot;v_n)$ we observe from (3.6.30) that in the large time-stepping regime

$$\|M_+^{-1}\Pi\|_{H^1_R*} \leq c\frac{\epsilon^2}{k}.$$

Using this result we expand

$$\Pi\mathscr{F}_{\rm E} = \frac{k}{\epsilon^2}\Pi\left(f_-'(g)g'\frac{R-R_n}{\epsilon}+f_-'(g)v_n\right)+O\left(\delta,\epsilon^2,k\right).$$

Inverting M_+ we find, at leading order

$$v_{n+1} = \frac{R-R_n}{\epsilon}L_+^{-1}\Pi f'(g)g' + L_+^{-1}\Pi f_-'(g)v_n + O(\epsilon^2,\delta^2)$$

In particular we deduce that

$$\left\|v_{n+1}-\frac{R-R_n}{\epsilon}L_+^{-1}\Pi f'(g)g'\right\|_{H^1_R} \leq \gamma\|v_n\|_{H^1_R}+O(\epsilon^2,\delta^2).$$

Arguing inductively, since the Eyre balance parameter $\gamma<1$ and the functions $\|L_+^{-1}\Pi f_-'(g)g'\|_{H^1_R}$ are uniformly bounded for all $R\geq 1$, we deduce that if $\delta := \|v_0\|_{H^1_R} = O(\epsilon)$ then the sequences $\{(R-R_n)\epsilon^{-2}\}_0^N$ and $\{\epsilon^{-1}\|v_n\|_{H^1_R}\}_0^N$ are uniformly bounded, independent of $\epsilon \ll 1$ and $k \gg \epsilon^2$ for all $n \leq N$ so long as

$R_n > 1$ for all $n = 0, \dots, N$.

To improve this bound we require Lipschitz estimates on the v_n component of $\mathscr{G}_{\mathrm{E}}$. To this end we find

$$\begin{aligned}\|\mathscr{G}_{\mathrm{E}}(v; v_n) - \mathscr{G}_{\mathrm{E}}(v; \tilde{v}_n)\|_{H^1_R} &\leq \|M_+^{-1}\Pi(I + \frac{k}{\epsilon^2} f'_-(g))\|_{H^1_R *} \|v_n - \tilde{v}_n\|_{H^1_R}, \\ &\leq \left(\gamma + O\left(\frac{\epsilon^2}{k}\right)\right) \|v_n - \tilde{v}_n\|_{H^1_R}. \end{aligned} \tag{3.6.50}$$

Here we introduce the quasi-steady parameter $\rho \in (0,1)$. Since $|R_n - R_m| = O(\epsilon^{2-\rho})$ for $|n - m| \leq N_\rho :=\leq \epsilon^{-\rho}$ we infer that

$$\left\| L_{+,n}^{-1} \Pi_n f'(g_n) f'_n - L_{+,m}^{-1} \Pi_m f'_-(g_m) g'_m \right\|_{H^1_R} \leq c\sqrt{\epsilon},$$

for all such n and m. For $n > N_\rho$ we define the quasi-equilibrium

$$v_{n*} := \frac{R_n - R_{n-1}}{\epsilon} E_n(z)$$

where E_n is the $R = R_n$ translate of

$$E := K_+ L_+^{-1} \Pi f'_-(g) g'. \tag{3.6.51}$$

Here the self-adjoint operator

$$K_+ := \left(I - L_+ \Pi \circ f'_-(g)\right)^{-1} > 0, \tag{3.6.52}$$

is well defined since $\|L_+ \Pi \circ f'_-(g)\|_{H^1_R *} = \gamma < 1$ by assumption. Using the Lipschitz property (3.6.50) of $\mathscr{G}_{\mathrm{E}}$ and the quasi equilibrium relation

$$\|v_{n*} - \mathscr{G}_{\mathrm{E}}(v_{n*}; v_{n*})\|_{H^1_R} = O(\epsilon^2),$$

we deduce that

$$\|v_{k+1} - v_{n*}\|_{H^1_R} \leq \gamma \|v_k - v_{n*}\|_{H^1_R} + O\left(\epsilon^{2-\rho}\right).$$

for $k = n - m_\epsilon, \ldots, n$. Since $\gamma^{N_\rho} \ll \epsilon$ we deduce from an inductive argument that

$$\left\| v_n - \frac{R_n - R_{n-1}}{\epsilon} E_n \right\|_{H^1_R} = O(\epsilon^{2-\rho}),$$

for all $n > N_s$. Inserting this result in (3.6.33) we arrive at the leading order Eyre iteration (3.6.44). □

Remark 7. *There are two examples of particular relevance*

$$f(u) = u^3 - u,$$

with the decomposition $f_+ = (1+\beta)u^3$ and $f_- = u + \beta u^3$ for $\beta > 0$. The choice $\beta = 0$ is classical and very degenerate, as in this case $f_-'(u) = 1$ and the corresponding Eyre balance parameter γ, defined in (3.6.43) is zero, and the Eyre number, (3.6.44) is 1. In this case it is possible to rewrite Eyre's method as backward Euler with a rescaled time. In particular the slow convergence to equilibrium will not be in evidence. For larger values of β the balance parameter increases from zero and the Eyre number decreases from 1. As the balance parameter increases through 1 we anticipate enhanced slowing of the *front profile as the Eyre number tends to zero. The choice of non-zero β can be viewed as spurious, a deliberate attempt to foul the method. A more robust example of non-zero balance arises naturally through the model*

$$f(u) = u^5 - \beta u^3,$$

with $\beta \geq 1$. This suggests the optimal decomposition $f_+ = u^5$ and $f_- = \beta u^3$. Here, unambiguously, increasing β increases the balance parameter and will lead to non-trivial enhanced slow-down with potential instability as γ increases through 1. *These analytic predictions are validated in a computational study below.*

Remark 8. *To leading order, in the large time-stepping regime $k \gg \epsilon^2$, the Eyre iteration recovers backward Euler with the substitution $k \mapsto c_E \epsilon^2$. This reduces to the exact result for the case $f(u) = u^3 - u$ and $f_-(u) = u$, for which $f_-' = 1$, as the Eyre constant reduces to* 1 *since $\Pi f_-(g) g' = \Pi g' = 0$.*

The strong contractivity of $\mathscr{G}_{\mathrm{E}}$ with respect to v, given in (3.6.42), arises from the strong convexity with respect to v, but the slow evolution and marginal convergence to the quasi-equilibrium, given in (3.6.50) arises from the balance between the implicit and explicit terms. The parameter γ measures this balance, with the quasi-equilibrium structure lost as γ increases towards 1. Indeed, since $\|K_+\|_{H^1_R{}^*} \sim (1-\gamma)^{-1}$, the Eyre constant will generically tend to zero as $\gamma \to 1$.

	Eyre with $f(u) = u^5 - u^3$	
ϵ	M	E
0.2	5,726	0.001
0.1	21,947 (3.83)	0.005
0.05	86,499 (3.94)	0.007
0.025	343,525(3.97)	0.007

Table 3.9: Computational results for the AC benchmark problem with fixed local error tolerance $\sigma = 10^{-4}$ and ϵ varied, using Eyre's method with reaction term $f(u) = u^5 - u^3$. Here, M is the total number of time steps taken (with the ratio to the value above in brackets) and E is the error in the benchmark time.

Computational Validation of Remark 7

We perform computations for AC with the non-classical $f(u) = u^5 - u^3$ (which also leads to meta-stable dynamics of curvature motion) using the same initial conditions and accuracy criteria as described in Section 3.4.1. BE performs almost identically to the results shown in Tables 3.3 and 3.4 for the classical $f(u) = u^3 - u$ in terms of accuracy and variation of time steps with ϵ and σ. This matches the theory in Section 3.6.1 which can be summarized as BE has profile fidelity when $k = o(\epsilon)$.

When Eyre's method is applied to the dynamics with $f(u) = u^5 - u^3$, with the natural splitting suggested in Remark 7, profile fidelity is lost as predicted. The formal prediction of $k = O(\epsilon^{3/2})$ which was seen computationally for $f(u) = u^3 - u$ in Table 3.4 is not observed for $f(u) = u^5 - u^3$. Rather, we see $k = O(\epsilon^2)$ as predicted by the theory in the previous section. The numerical results are shown in Table 3.9.

3.7 L^2_R coercivity of L

Here we show the technical argument for Lemma 1. For $u,v \in H^1(\mathbb{R})$ we define the inner product

$$\langle u,v\rangle_\ell := \int_{-\ell}^{\ell} u(s)v(s)\,ds,$$

with the standard norms L^2_ℓ and H^1_ℓ while $L^2_{\ell^c}$ is defined in $\mathbb{R}\backslash[-\ell,\ell]$. Let L_0 be as defined in (3.6.6).

Lemma 5. *Fix $\ell_0 > 0$ sufficiently large there exists $\alpha > 0$ such that for all $\ell > \ell_0$*

$$\langle L_0 u, u\rangle_\ell \geq \alpha, \tag{3.7.1}$$

for all $u \in H^1_\ell \cap L^2(\mathbb{R})$ satisfying $\langle u, g'\rangle_{L^2(\mathbb{R})} = 0$ and $1 = \|u\|_{L^2_\ell} \geq \|u\|_{L^2_{\ell^c}}$.

Proof. Let ϕ be the minimizer of $\langle L_0 u, u\rangle_\ell$ over $H^1(\mathbb{R})$ subject to $\|u\|_{L^2_\ell} = 1$ and the full-line orthogonality $\langle u, \psi_0\rangle_{L^2(\mathbb{R})} = 0$. By scaling, the minima is attained with $\|\phi\|_{L^2_\ell} = \frac{1}{2}$ and satisfies

$$L_0\phi = \lambda\phi, \qquad \text{on}\,[-\ell,\ell],$$

subject to Neumann boundary conditions $\phi_x(\pm\ell) = 0$, in addition to the full line orthogonality condition. The operator L_0 on the truncated domain has eigenvalues $\lambda^\ell_0 < \lambda^\ell_1 < ...$ which are $O(e^{-d\ell})$ far away from the eigenvalues of L_0 on the full line. In particular λ^ℓ_0 may be negative, but the rest are uniformly positive. In L^2_ℓ we partition $\phi = \beta\psi^\ell_0 + \phi^\perp$, where ψ^ℓ_0 is the L^2_ℓ

ground state of L_0 and $\phi^\perp \in L^2_\ell$ is L^2_ℓ orthogonal to ψ_0^ℓ. Then we have

$$\langle L_0\phi,\phi\rangle_\ell \geq \lambda_0\beta^2 + \lambda_1^\ell \|\phi^\perp\|^2_{L^2_\ell}. \tag{3.7.2}$$

On the other hand the orthogonality condition implies that

$$\langle \phi, g'\rangle_{\mathbb{R}} = 0 = \beta + \langle \phi, g'\rangle_{\ell^c}.$$

where the subscript ℓ^c denotes integration over $\mathbb{R}\backslash[-\ell,\ell]$ with the corresponding norms. In particular we deduce that

$$|\beta| \leq \|\phi\|_{L^2_{\ell^c}} \|g'\|_{L^2_{\ell^c}} \leq \|g'\|_{L^2_{\ell^c}} \|\phi\|_{L^2_\ell}.$$

Since g' decays exponentially at $\pm\infty$, is complementary norm is exponentially small in ℓ. From orthogonality of ψ_0^ℓ and $\phi^\perp$ we have

$$\|\phi\|^2_{L^2_\ell} = \beta^2 + \|\phi^\perp\|^2_{L^2_\ell} \leq \|\phi^\perp\|^2_{L^2_\ell} + \|g'\|^2_{L^2_{\ell^c}} \|\phi\|^2_{L^2_\ell}.$$

or equivalently

$$1 = \|\phi\|^2_{L^2_\ell} \leq \frac{1}{1 - \|g'\|^2_{L^2_{\ell^c}}} \|\phi^\perp\|^2_{L^2_\ell},$$

and taking ℓ large enough we use these bound in (3.7.2) to show that α is exponentially close to $\lambda_1^\ell > 0$. □

To complete the proof of Lemma 1 we take ℓ sufficiently large to apply Lemma 5 and then take ϵ sufficiently small that $\epsilon|z| \leq \epsilon\ell \ll 1$. Under these conditions L^2_ℓ and $L^2_R(-\ell,\ell)$ are equivalent norms, uniformly in ϵ, and we have uniform $L^2_R(-\ell,\ell)$ coercivity of L. Conversely, L is clearly L^2_R coercive on $[-R/\epsilon,\infty)\backslash[-\ell,\ell]$ since $f'(g)$ is strictly positive there. Clearly L is uniformly coercive on function with more than half their L^2_R mass in $[-R/\epsilon,\infty)\backslash[-\ell,\ell]$. The g' orthogonality condition implies approximate orthogonality to ψ_0 for ℓ large. From these we deduce the full L^2_R-coercivity of L over X.

Chapter 4

Various formulations and their numerical results of oxygen depletion problems

The Oxygen Depletion problem is an implicit free boundary value problem. The dynamics allow topological changes in the free boundary. We show several mathematical formulations of this model from the literature and give a new formulation based on a gradient flow with constraint. All formulations are shown to be equivalent. We explore the possibilities for the numerical approximation of the problem that arise from the different formulations. We show a convergence result for an approximation based on the gradient flow with constraint formulation that applies to the general dynamics including topological changes. More general (vector, higher order) implicit free boundary problems are discussed. Several open problems are described.

4.1 Discussion

The Oxygen Depletion (OD) problem is a free boundary value problem of implicit type. Implicit here means that the free boundary is specified implicitly by an extra boundary condition rather than explicitly as an interface normal velocity as for a Stefan problem [72, 79, 86]. The OD problem was introduced as a model of oxygen consumption and diffusion in living tissue but closely related problems have similar problem structure. Some of the early work is described in [25] with a great deal of subsequent interest from the analysis and numerical research communities in [6, 36, 63, 75, 79]. In the current work, we pursue an understanding of the analysis of the OD problem as the simplest example of an implicit free boundary value problem. We are motivated by an interest in a general class of implicit free boundary value problems of which some examples are given at the end of this work. We point out several open problems which are summarized in the final section.

By way of introduction, we present the OD problem in 1D for an unknown $u(x,t)$ for $x \in [0,s(t)]$ with a single free boundary $x = s(t)$ and a no flux condition $u_x = 0$ at $x = 0$. At the free boundary, $u = 0$ and additionally $u_x = 0$. These two conditions implicitly define the free boundary $x = s(t)$. In higher dimensions it is the vanishing of the solution value and is normal derivative at the boundary. The solution obeys

$$u_t = u_{xx} - 1 \tag{4.1.1}$$

for $x \in [0,s(t)]$ and it is natural to extend $u \equiv 0$ for $x > s(t)$ in a C^1 continuous

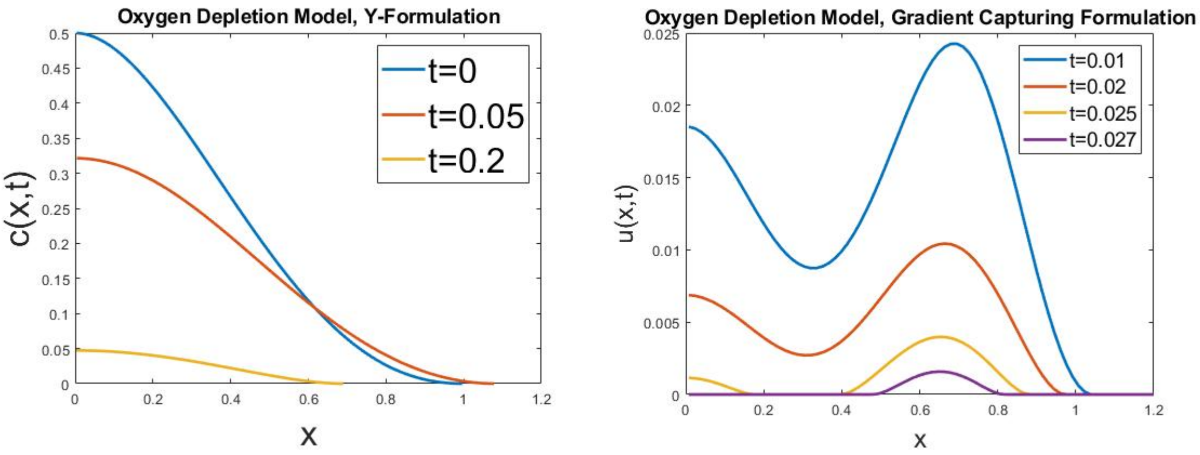

Figure 4.1: A high accuracy 1D solution of the OD problem without topological change (left). 1D solutions of the OD problem with topological changes with a capturing method (right).

way. We consider positive initial conditions for u in $[0, s(0))$. This is one of the forms of the OD problem forshadowed by the title. In another OD formulation, the time derivative u_t satisfies an explicit free boundary value problem that can be described as a one sided Stefan problem from which short time solution existence and regularity can be inferred under certain conditions. A similar reformulation can be made with an explicit velocity as a higher derivative of the solution. A further formulation, suitable only in 1D and specialized geometries in higher dimensions, results when $x \in [0, s(t)]$ is mapped linearly to $y \in [0,1]$. The numerical approximation of the resulting fixed boundary problem is of Differential Algebraic Equation (DAE) type and can achieve high accuracy. Our numerical approximation is inspired by [75], where the author shows that $s(t)$ is smooth using the idea that maps the free domain to a fixed domain.

An example of the dynamics computed on the DAE formulation in the

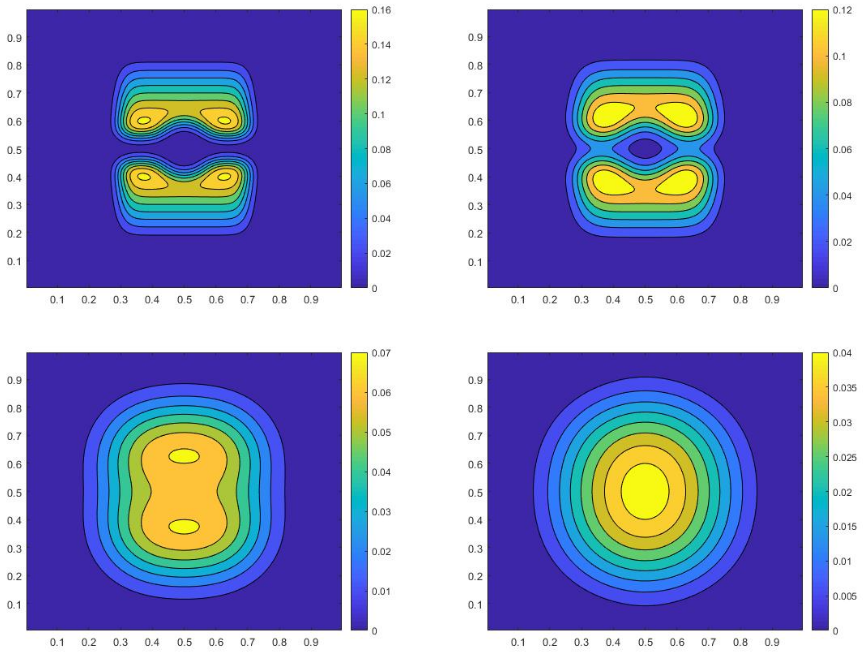

Figure 4.2: A captured 2D solution of the OD problem with topological change.

mapped region is shown in Figure 4.1 (left). Here, $s(t)$ initially moves to the right driven by diffusion and then to the left as u values decrease due to the consumption term. The solution in this formulation ends when $s(T) = 0$ ($u \equiv 0$). A specialized method in this general framework was developed in [63] to accurately compute both the solution and the end time of the dynamics. The formulations discussed so far are suitable when the solution does not undergo any topological change. Solutions of (4.1.1) can go negative, but physically relevant values of concentration have $u \geq 0$. In the 1D case, preserving nonnegativity results in the break up or merger of intervals where $u > 0$ as shown in Figure 4.1 (right). Topological change can be more complex in higher dimensions as seen in Figure 4.2.

A weak form of the solution can be introduced using a variational inequality approach (4.2.3)[62, 73]. In what follows, we use this formulation as the basis for equivalence to the others. This formulation is amenable to approximation using the Augmented Lagragian Method [49, 50, 53]. The computations in Figures 4.1(right) and 4.2 are done with a method based on the formulation of the OD as L_2 gradient flow with constraint on the energy from the elliptic obstacle problem. The obstacle problem has had considerable interest in the literature [11, 40, 56, 65, 87]. We also introduce a regularized approach with parameter ϵ, similar to the approach in [6], where we approximate the non-linearity by a family of Lipschitz monotone nonlinear terms. We show that the limit as $\epsilon \to 0$, the approximated solutions u_ϵ converges to the desired solution to OD. Throughout this chapter we show the equivalence of different formulations discussed above, each of which opens up possibilities for numerical approximation. We pursue some of the numerical approaches in more detail.

This chapter is organized as follows. In Section 4.2 we present the different formulations with technical details and show their equivalence. In Section 4.3 we present the numerical schemes and provide some analytical convergence results in some cases, numerical evidence of convergence in others. In Section 4.4 we show some additional results on the dynamics. In Section 4.5 we present some other implicit free boundary value problems of interest and indicate how our results can be extended to them, with some open questions. We end with a short Summary that includes a list of open problems.

Notation in this chapter

We define the space $H^1_+(\Omega) := \{u \in H^1(\Omega) : u \geq 0\}$. For bounded domains Ω we add homogeneous Neumann boundary conditions on $\partial\Omega$. We further denote $\mathscr{J}$ to be collection of functions $v \in L^2(0,T;H^1(\Omega))$ such that $v(t) \in H^1_+$ for a.e. $t \in (0,T)$. In some instances, we denote space derivative in 1D case by $u'(x)$ and time derivative by $\dot{u}(t)$. We also abuse the notation Ω to denote $(0,1)$ in 1D case.

4.2 Equivalent Formulations

4.2.1 Standard formulation in 1D

The one-dimensional oxygen depletion problem with associated free boundary and initial conditions is as follows:

$$\begin{cases} u_t = u_{xx} - 1, & 0 \leq x \leq s(t) \\ u(x,t) = 0, & x > s(t) \\ u_x(0,t) = 0, & t > 0 \\ u(s(t),t) = u_x(s(t),t) = 0, & t > 0 \\ u(x,0) = u_0(x), & 0 \leq x \leq 1 \\ s(0) = 1. \end{cases} \tag{4.2.1}$$

We assume here that u_0 satisfies all necessary smoothness and compatibility assumptions needed in the analyses cited below. By literature convention, we consider here a problem with a fixed, no-flux boundary condition at $x = 0$ and only one free boundary $s(t) > 0$. Uniqueness and lack of topological change when $u_0' \le 0$ follows from a modified maximum principle argument [36].

Existence can be seen by considering $v = u_t$ which satisfies a standard Stefan problem [25] with explicit interface velocity:

$$\begin{cases} v_t = v_{xx}, & 0 \le x \le s(t) \\ v_x(0,t) = 0, & t > 0 \\ v(s(t),t) = 0,\ v_x(s(t),t) = -\dot{s}(t), & t > 0 \\ s(0) = 1. \end{cases}$$

One can check the function $u = \int_0^t v \, d\tau$ solves the oxygen depletion problem. To prove existence and uniqueness of Stefan problem, one can verify that the map

$$\mathscr{T}(s)(t) := 1 - \int_0^t v_x(s(\tau),\tau))\, d\tau,\ T \ge t \ge 0,$$

defines a contraction map [59].

Remark 9. *The reformulation in $v = u_t$ to an explicit free boundary problem with interface velocity equal to $-v_x$ can be reinterpreted as a normal velocity for the problem for u with velocity equal to $-v_x = -u_{tx} = -u_{xxx}$. The authors are not aware of any analysis or computational methods based on this velocity expression with higher order spatial derivatives.*

We make the following plausible conjecture for the dynamics of the Cauchy problem in 1D with initial conditions $u_0(x) \in H^1_+$ with compact support:

Conjecture 4.2.1. *Assume u_0 has a finite $\mathscr{S}(0)$ where $\mathscr{S}(t)$ counts the number* *of free boundary points:*

$$\mathscr{S}(t) = \{x : u(x,t) = 0 \textit{ and } u(y,t) > 0 \textit{ for some } y \textit{ in every neighbourhood of } x\}.$$

Then

(i) *$\mathscr{S}(t)$ is finite for every $t > 0$.*

(ii) *There exits a finite increasing sequence of times t_j, $j = 0 \dots M$ with $t_0 = 0$ and card $\mathscr{S}(t) := n_j$ constant on every interval (t_j, t_{j+1}) and $u \equiv 0$ for $t \geq t_M$.*

(iii) *$\mathscr{S}(t) = \{s_1(t), s_2(t), \dots s_{n_j}(t)\}$ for $s_l(t)$ smooth on (t_j, t_{j+1}).*

(iii) *$u(x,t)$ is C_1 for $t > 0$ and C_∞ except at free boundary points.*

Recent related results have been shown for the Stefan problem [37]. Similar analysis of the OD problem is complicated by the reaction term that allows the formation of new zones of constraint.

4.2.2 Mapped domain formulation in 1D

Considering the same smooth solutions in 1D discussed in the previous section, we consider $s(t) > 0$ in $t \in [0,T]$, take $y = x/s(t)$, and reformulate oxygen depletion problem as

$$u_{yy} + \dot{s}syu_y - s^2u_t - s^2 = 0 \tag{4.2.2}$$

with boundary conditions $u_y(0,t) = u(1,t) = u_y(1,t) = 0$. Over a short time period, we assume that $\dot{s}(t)$ and $s(t)$ are uniformly bounded, thus the linear operator is parabolic. Assuming $s(t)$ is known, uniqueness of u is not an issue; however, to prove uniqueness of the solution pair $(s,\tilde{u})$, we introduce the map $\mathscr{G}\colon X \to Y$, where X is the closed subspace of $H^1(H^2([0,s(t)]);[0,T])\times C^1([0,T])$ that solves OD system and Y is the closed subspace of $H^1(H^2([0,1]);[0,T])\times C^1([0,T])$ that solves the reformulated system:

$$\mathscr{G}((u(x,t),s(t)) = (u(y,t),s(t)) := (U(y\cdot s(t),t),s(t)).$$

where $U(x,t)$ is the solution from the previous section. One can check that the map $\mathscr{G}$ is a bijection and so all solutions of (4.2.2) are equivalent to the solutions in the standard formulation of Section 4.2.1. A numerical method based on this formulation is presented in Section 4.3.1.

Remark 10. *A direct analysis of this formulation would be useful as a stepping stone to a convergence proof for the numerical approximation in Section 4.3.1 and an analysis of the general class of problems in Section 4.5. We have not been able to make progress on such an analysis. There are subtleties in the problem: note that changing $-s^2$ to $+s^2$ makes the problem ill defined as $s(t) = +\infty$ for $t > 0$ in that case.*

4.2.3 A parabolic variational inequality formulation

To proceed with the discussion of the problem in higher dimensions with topological changes, we consider the standard approach to weak solutions in this setting: a variational inequality formulation [57, 62]. We consider the following problem: find a function $u \in \mathscr{J}$ with $u(0) = u_0 \in H^1_+$ that solves

$$\int_0^t \int_\Omega u_t \cdot (v-u) + \int_0^t \int_\Omega \nabla u \cdot \nabla(v-u) \geq \int_0^t \int_\Omega u - v; \text{ for all } v \in \mathscr{J}, \text{ a.e. } t \in (0,T). \tag{4.2.3}$$

Proposition 4.2.2. *The variational inequality* (4.2.3) *has at most one solution and in fact suppose u_1 and u_2 solve* (4.2.3) *with distinct initial conditions u_{1_0} and u_{2_0} then*

$$\|u_1 - u_2\|_{L^\infty(0,T;L^2(\Omega))} \leq \|u_{1_0} - u_{2_0}\|_{L^2(\Omega)}. \tag{4.2.4}$$

Proof. Note $u_j \in \mathscr{J}$ satisfies (4.2.3) for j=1,2, in particular

$$\int_0^t \int_\Omega \partial_t u_1 \cdot (u_2 - u_1) + \int_0^t \int_\Omega \nabla u_1 \cdot \nabla(u_2 - u_1) \geq \int_0^t \int_\Omega u_1 - u_2,$$
$$\int_0^t \int_\Omega \partial_t u_2 \cdot (u_1 - u_2) + \int_0^t \int_\Omega \nabla u_2 \cdot \nabla(u_1 - u_2) \geq \int_0^t \int_\Omega u_2 - u_1.$$

Summing two inequalities above and denote $w = u_1 - u_2$, one has

$$\int_0^t \int_\Omega \partial_t w \cdot w + \int_0^t \int_\Omega \nabla w \cdot \nabla w \leq 0$$
$$\implies \int_0^t \int_\Omega (w^2)_t \leq 0 \implies \|w\|_{L^\infty(0,T;L^2(\Omega))} \leq \|w_0\|_{L^2(\Omega)}.$$

□

Theorem 3. *There exists a unique solution to the variational inequality* (4.2.3).

Note that this can be done by a standard monotone operator argument and we refer to [62].

Note that any smooth solution u to (4.2.1) must solve (4.2.3) and by uniqueness the solution to (4.2.3) therefore solves OD. To see this, we first observe that $u \geq 0$ and therefore $u \in \mathscr{J}$. Resulting from that, we obtain that for any $v \in \mathscr{J}$ and for a.e. $t \in (0,T)$

$$\begin{aligned}\int_0^t \int_\Omega u_t \cdot (v-u) + \int_0^t \int_\Omega \nabla u \cdot \nabla(v-u) &= \int_0^t \int_0^{s(\tau)} (u_t - u_{xx})(v-u)\, dx d\tau \\ &= \int_0^t \int_0^{s(\tau)} u - v\; dx d\tau \geq \int_0^t \int_\Omega u - v.\end{aligned}$$

4.2.4 A gradient flow formulation

In this section, we formulate the OD problem as the L_2 gradient of the energy from the elliptic obstacle problem. A formal calculation with

$$\mathscr{E}(t) := \int \frac{1}{2} |\nabla u|^2 + u$$

leads to

$$\frac{d\mathscr{E}}{dt} = -\int (\Delta u - 1)^2.$$

It is convenient to present the equivalence of the gradient flow formulation as the limit of implicit time steps as this gets us half way to the convergence result for the fully discrete method described in Section 4.3.2. The spatially continuous, time discrete solutions u_n approximate $u(\cdot, nk)$, where k is a time step. We consider the following minimization problem for $u = u_{n+1}$ to the following energy functional:

$$E[u] = \int_\Omega \frac{1}{2} |\nabla u|^2 + \frac{1}{2k} (u - u_n)^2 + u, \tag{4.2.5}$$

for $u \in H^1_+$. Existence and uniqueness of the minimizer is guaranteed by the standard calculus of variation technique and convexity of the energy functional [81].

Remark 11. *By defining the discrete energy* $\mathscr{E}^n := \int \frac{1}{2}|\nabla u_n|^2 + u_n$, *we can see that* $\mathscr{E}^{n+1} \le \mathscr{E}^n$. *This can be derived by considering* $E[u] = \int \frac{1}{2}|\nabla u|^2 + u + \frac{(u-u_n)^2}{2k}$ *with* $E[u_{n+1}] \le E[u_n]$. *This gives the discrete gradient flow structure.*

Euler-Lagrange equation of the energy minimizer

We will derive the corresponding Euler-Lagrange equation for the minimizing problem following the idea from [40]:

Theorem 4. *Suppose u is the unique minimizer to the energy minimizing problem* (4.2.5)*, then u is the (weak) solution to the following modified backward Euler scheme:*

$$\frac{u - u_n \cdot \chi_{\{u>0\}}}{k} = \Delta u - \chi_{\{u>0\}}.$$

To begin with, we consider an equivalent energy minimizing problem:

$$\tilde{E}[u] := \int \frac{1}{2}|\nabla u|^2 + \frac{1}{2k}u^2 + (1 - \frac{u_n}{k})u^+,$$

subject to

$$\tilde{\mathscr{K}} := \{v \in H^1(\Omega) : \frac{\partial v}{\partial n}|_{\partial\Omega} = 0\}$$

where $u^+ = \max(u, 0)$.

Lemma 6. *There exists a unique* $\tilde{u} \in \tilde{\mathscr{K}}$ *such that*

$$\tilde{E}[\tilde{u}] = \min_{v \in \tilde{\mathscr{K}}} \tilde{E}[v];$$

moreover such $\tilde{u}$ is the unique minimizer to (4.2.5).

Proof. Firstly, by similar argument, the existence and uniqueness of this energy minimizing problem can be proved.

Now to show the equivalence of these two minimizing problem, we recall that the minimizer $u \geq 0$, so

$$\min_{v\in\mathscr{K}} E[v] = E[u] = \tilde{E}[u] \geq \min_{v\in\tilde{\mathscr{K}}} \tilde{E}[v];$$

On the other hand, to show

$$\min_{v\in\mathscr{K}} E[v] \leq \min_{v\in\tilde{\mathscr{K}}} \tilde{E}[v],$$

we note that for any $v \in \tilde{\mathscr{K}}$, the corresponding $v^+ \in H^1_+$ under the assumption that $u_n \geq 0$. As a result,

$$E[u] \leq E[v^+] \leq \tilde{E}[v]$$

for any $v \in \tilde{\mathscr{K}}$, hence

$$E[u] \leq \inf_v \tilde{E}[v].$$

Now since $E[u] = \tilde{E}[u] = \min \tilde{E}[v]$, we have $\tilde{u} = u$ by the uniqueness.

□

It remains to derive the Euler-Lagrange equation for this new energy minimizing scheme.

Proposition 4.2.3. *Suppose u is the unique minimizer to Lemma 6, then u* *is the (weak) solution to the following modified backward Euler scheme:*

$$\frac{u - u_n \cdot \chi_{\{u>0\}}}{k} = \Delta u - \chi_{\{u>0\}}.$$

The proof of this result is found in Section 4.6.

Regularity of the minimizer

We follow the idea in [8] to formulate our problem in a variational inequality:

$$u \in H^1_+ : \int_\Omega \nabla u \cdot \nabla(v-u) + \frac{u}{k}(v-u)\,dx \geq \int_\Omega \left(\frac{u_n}{k} - 1\right)\cdot(v-u) \text{for all } v \in H^1_+. \quad (4.2.6)$$

To see the equivalence of the energy minimization and elliptic variational inequality we now state the proposition.

Proposition 4.2.4. *Any solution to the minimization problem* (4.2.5) *is also a solution to the variational inequality* (4.2.6) *and vice versa.*

Proof. Suppose u is an energy minimizer to (4.2.5). Let $v \in H^1_+$, note that H^1_+ is convex then $(1-\lambda)u + \lambda v \in H^1_+$ for any $\lambda \in [0,1]$. Using $(1-\lambda)u + \lambda v$ as a competitor in $E[u] \leq E[(1-\lambda)u + \lambda v]$, we can derive from the order $O(\lambda)$:

$$\int_\Omega \nabla u \cdot \nabla(v-u) + \frac{u}{k}(v-u)\,dx \geq \int_\Omega \left(\frac{u_n}{k} - 1\right)\cdot(v-u), \ \forall\ v \in H^1_+.$$

The reverse can be proved similarly.

□

Note that this formulation uses convexity of H^1_+; the optimal regularity of

u is $C^{1,1}_{loc}$:

Theorem 5 (regularity). *Suppose u is a solution to* (4.2.5) *(or* (4.2.6)*), then there exists a positive constant C such that*

$$\|\Delta u\|_\infty \leq C\Big(1+\frac{1}{k}\|u_n\|_\infty + \|\Delta u_n\|_\infty\Big).$$

Moreover, for each compact $K \subset \Omega$ there exists a positive constant $c(K) > 0$ such that

$$\sup_{i,j}\sup_{K}|D_{ij}u(x)| \leq c.$$

The proof follows from [8], where penalty argument is applied together with a non-degeneracy argument, which we refer to Lemma 1.2 from [8].

Remark 12. *This upper bound can be improved by applying energy gradient flow. By competing u with u_n in $E[u] \leq E[u_n]$, we have*

$$\frac{1}{k}\|u-u_n\|_2^2 \leq \|\nabla u_n\|_2^2 + \|u_n\|_1.$$

By applying the penalty argument in [8], we can derive that

$$\|\Delta u\|_2 \leq C(\|u_n\|_{H^2}+1).$$

Time-discrete variational inequality and Rothe's method

In this section, we will show the energy minimization scheme has a limit as the time steps $k \to 0$ that solves the parabolic variational inequality (4.2.3). We consider the energy minimization scheme as in previous sections and by Proposition 4.2.4, it suffices to show the following lemma.

Lemma 7 (Rothe's method)**.** *Suppose for each $j = 1,\cdots M$, where $M = T/k$, u_j is the unique minimizer for $E_j(u)$ as defined in Theorem 4. Then $u = \lim_{k\to 0} u_M(x,t)$ exists and solves the parabolic variational inequality* (4.2.3). *Here u_M is the associated linear interpolation defined by*

$$u_M(x,t) := (1-\theta)\cdot u_j(x) + \theta \cdot u_{j+1}(x)\ ,\ \textit{for}\ t = (j+\theta)k,\ \theta \in [0,1).$$

Proof of Lemma 7. Our proof follows [51, 73]. First, note that u_j is the unique minimizer of E_j and as discussed earlier in Proposition 4.2.4, it satisfies the elliptic variational inequality (4.2.5):

$$\int_0^1 u_j' \cdot (v' - u_j') + \frac{u_j}{k}(v - u_j)\ dx \geq \int_0^1 \left(\frac{u_{j-1}}{k} - 1\right) \cdot (v - u_j) \text{for all } v \in H_+^1. \quad (4.2.7)$$

Taking $v = u_{j-1}$, one can derive

$$\langle u_j', u_{j-1}' - u_j'\rangle + \frac{1}{k}\langle u_j, u_{j-1} - u_j\rangle \geq \frac{1}{k}\langle u_{j-1}, u_{j-1} - u_j\rangle - \langle 1, u_{j-1} - u_j\rangle.$$

Similarly we take $v = u_j$ for $j-1$'s inequality case

$$\langle u_{j-1}', u_j' - u_{j-1}'\rangle + \frac{1}{k}\langle u_{j-1}, u_j - u_{j-1}\rangle \geq \frac{1}{k}\langle u_{j-2}, u_j - u_{j-1}\rangle - \langle 1, u_j - u_{j-1}\rangle.$$

Adding the two inequalities above gives

$$\frac{1}{k}\|u_j - u_{j-1}\|_2^2 + \|u_j' - u_{j-1}'\|_2^2 \leq \frac{1}{k}\langle u_j - u_{j-1}, u_{j-1} - u_{j-2}\rangle.$$

Note that when $j=1$, we choose $v=u_0$ and hence

$$\frac{1}{k}\|u_1-u_0\|_2^2+\|u_1'-u_0'\|_2^2\le\left|\langle u_0',u_1'-u_0'\rangle\right|+|\langle 1,u_0-u_1\rangle|\le\left(\|u_0''\|_2+1\right)\cdot\|u_1-u_0\|_2.$$

Therefore we obtain that

$$\|\frac{u_j-u_{j-1}}{k}\|_2\le C$$

for any $j=1,\cdots,M$ and a positive absolute constant C. Note that $\dot{u_M}(t)=\frac{u_j-u_{j-1}}{k}$, therefore by Arzelà-Ascoli, $u_M(t)$ converges to some function u in $C([0,T],L^2((0,1)))$. Then we can define

$$\widetilde{u_M}(t)=u_j\ ,\text{ for }t\in[jk,(j+1)k),$$

similar to Lemma 8 and Remark 14, $\widetilde{u_M}$ converges to the same u. Indeed, u_M' converges to u' weakly in $L^2((0,T),L^2((0,1)))$. As a result, rewrite (4.2.7): for any $v\in H_+^1$

$$\langle\dot{u_M}(t),v(t)-\widetilde{u_M}(t)\rangle+\langle\widetilde{u_M}',v'-\widetilde{u_M}'\rangle\ge-\langle 1,v-\widetilde{u_M}\rangle,\qquad(4.2.8)$$

which holds for a.e. $t\in(0,T)$. For arbitrary $\tau_1<\tau_2$ in $[0,T]$,

$$\int_{\tau_1}^{\tau_2}\langle\dot{u_M}(t),v(t)-\widetilde{u_M}(t)\rangle+\langle\widetilde{u_M}',v'-\widetilde{u_M}'\rangle\ dt\ge-\int_{\tau_1}^{\tau_2}\langle 1,v-\widetilde{u_M}\rangle\ dt,\qquad(4.2.9)$$

letting $k\to 0$, we have the desired result

$$\int_{\tau_1}^{\tau_2}\langle\dot{u}(t),v(t)-u(t)\rangle+\langle u',v'-u'\rangle\ dt\ge-\int_{\tau_1}^{\tau_2}\langle 1,v-u\rangle\ dt,\qquad(4.2.10)$$

for almost every $\tau_1 < \tau_2$ in $[0,T]$.

□

4.2.5 A regularized formulation

We introduce a formulation using a regularization method with parameter ϵ proposed first in [6]. Here, we will see convergence in regularized solutions $u_\epsilon(x,t)$ as $\epsilon \to 0$ to the other OD formulations.

$$\partial_t u_\epsilon = \partial_{xx} u_\epsilon - f_\epsilon(u_\epsilon), \tag{4.2.11}$$

where

$$f_\epsilon(u_\epsilon) = \begin{cases} 1 & u_\epsilon > \epsilon \\ \dfrac{u_\epsilon}{\epsilon} & u_\epsilon \le \epsilon, \end{cases} \tag{4.2.12}$$

with same initial condition $u_0(x)$. Note that $f_\epsilon(x)$ is a Lipschitz function and as a result u_ϵ exists as a smooth solution for each $\epsilon > 0$ with $u_\epsilon(x,t) > 0$ for all $x > 0$ and $t > 0$.

We consider u_{ϵ_1} and u_{ϵ_2} with $\epsilon_1 < \epsilon_2$. Denote their difference by $w = u_{\epsilon_1} - u_{\epsilon_2}$, then

$$\partial_t w - \partial_{xx} w = -f_{\epsilon_1}(u_{\epsilon_1}) + f_{\epsilon_2}(u_{\epsilon_2}).$$

Note that,

$$-f_{\epsilon_1}(u_{\epsilon_1})+f_{\epsilon_2}(u_{\epsilon_2})=\begin{cases} 0\,, & \text{if } u_{\epsilon_1}>\epsilon_1\,,\ u_{\epsilon_2}>\epsilon_2 \\ -\dfrac{u_{\epsilon_1}}{\epsilon_1}+1\,, & \text{if } u_{\epsilon_1}\le\epsilon_1\,,\ u_{\epsilon_2}>\epsilon_2 \\ -1+\dfrac{u_{\epsilon_2}}{\epsilon_2}\,, & \text{if } u_{\epsilon_1}>\epsilon_1\,,\ u_{\epsilon_2}\le\epsilon_2 \\ -\dfrac{u_{\epsilon_1}}{\epsilon_1}+\dfrac{u_{\epsilon_2}}{\epsilon_2}\,, & \text{if } u_{\epsilon_1}\le\epsilon_1\,,\ u_{\epsilon_2}\le\epsilon_2 \end{cases}. \tag{4.2.13}$$

We observe that

$$\begin{aligned} -1+\frac{u_{\epsilon_2}}{\epsilon_2}\le 0\,, &\qquad \text{if } u_{\epsilon_1}>\epsilon_1\,,\ u_{\epsilon_2}\le\epsilon_2 \\ -\frac{u_{\epsilon_1}}{\epsilon_1}+\frac{u_{\epsilon_2}}{\epsilon_2}=-\frac{w}{\epsilon_1}+u_{\epsilon_2}\cdot\left(\frac{1}{\epsilon_2}-\frac{1}{\epsilon_1}\right)\,, &\qquad \text{if } u_{\epsilon_1}\le\epsilon_1\,,\ u_{\epsilon_2}\le\epsilon_2 \\ u_{\epsilon_1}<u_{\epsilon_2}\,, &\qquad \text{if } u_{\epsilon_1}\le\epsilon_1\,,\ u_{\epsilon_2}>\epsilon_2; \end{aligned}$$

so if we assume the maximal value of w is achieved at (x_0,t_0) with $x_0\in[0,1]$ and $t_0>0$, then $w(x_0,t_0)>0$, $\partial_t w(x_0,t_0)=0$ and $\partial_{xx}w(x_0,t_0)<0$. It contradicts all 4 cases discussed above. It gives a partial result of the following statement:

Theorem 6. *Suppose a sequence of classical functions* $\{u_\epsilon\}$ *solve* (4.2.11)*, then* u_ϵ *is monotonically decreasing as* ϵ *decreases to 0. Moreover, the limiting function*

$$\lim_{\epsilon\to 0}u_\epsilon=u$$

holds pointwisely. This limiting function u *solves the variational inequality* (4.2.3)*.*

The details of the proof can be found in Section 4.7. An alternate conver-

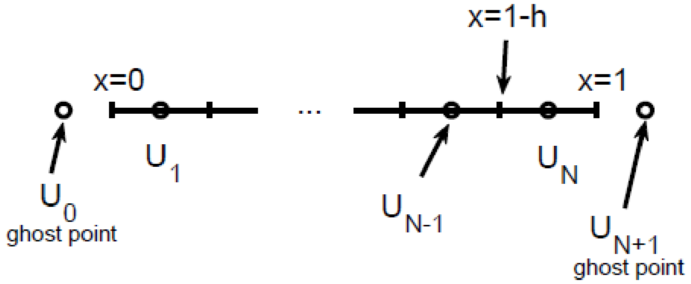

Figure 4.3: Cell centered finite difference spatial approximation of the 1D mapped domain formulation

gence statement and proof is given in Section 4.8. While there is theoretical insight to be gained from this formulation, it is unnatractive for numerical approximation for application purposes as free interface locations are not easily identified from $\epsilon > 0$ results.

4.3 Numerical Approximation

4.3.1 Mapped domain method

We consider the discretization of the mapped domain formulation (4.2.2) in space using cell centred finite differences. We first discretize in space, leaving time continuous (known as a Method of Lines – MoL – discretization) with approximations $u^j(t) \approx u((j-1/2)h,t)$, $j = 1 \dots N$ where h is the uniform grid spacing with N subintervals of $y \in [0,1]$. The interface location $s(t)$ is approximated by $S(t)$.

Boundary conditions are implemented using ghost points [84] $u^0(t) \approx u(-h/2,t)$

and $u^{N+1}(t) \approx u(1+h/2,t)$ depicted in Figure 4.3. Boundary conditions at $y = 1$ are implemented using second order averages and differences:

$$(u^{N+1} + u^N)/2 \quad = \quad 0 \tag{4.3.1}$$

$$(u^{N+1} - u^N)/h \quad = 0 \tag{4.3.2}$$

which implies that $u^N = u^{N+1} = 0$. The no-flux boundary condition at $y = 0$ is approximated similarly. The MoL discretization for the interior equations is

$$D_2 u^j + S\dot{S} y D_1 u^j - S^2 \dot{u}^j - S^2 = 0. \tag{4.3.3}$$

where D_2 and D_1 are the standard centered second order finite difference operators. The system (4.3.1,4.3.2,4.3.3) is a Differential Algebraic Equation (DAE) [3] and has index one. The computation shown in Figure 4.1 (left) is of this system using Implicit (Backward) Euler time stepping with Newton iterations for the resulting nonlinear system at each time step. In a computational study, we observe errors of size $O(h^2)+O(k)$ where k is the time step, as expected for a second order spatial and first order temporal discretization.

Remark 13. *The convergence of the method has not been proved. The missing direct analysis discussed in Remark 10 could give insight.*

4.3.2 Energy minimization method

In this section, we continue the discretization of the gradient flow formulation from Section 4.2.4 and discretize in space with $u_n^i \approx u(ih, nk)$. We consider the discretization in one spatial dimension for ease of presentation but the argument extends to higher dimensions. The energy minimization problem (4.2.5) is approximated by the discrete minimization of

$$E_{n+1}^N = \sum_{i=0}^{N-1} \frac{h}{2}\left(\frac{u^{i+1}-u^i}{h}\right)^2 + h\cdot\sum_{i=0}^{N}\left(\frac{1}{2k}\left(u^i-u_n^i\right)^2+u^i\right), \tag{4.3.4}$$

where N the number of grid points with $N = 1/h$, assuming without loss of generality that all positive values of u are captured in the interval $(0,1)$. We solve this minimization problem subject to all non-negative discrete data $\vec{u}_{n+1}^N := (u_{n+1}^1, u_{n+1}^2, \cdots, u_{n+1}^N)$. This is a convex, quadratic minimization problem with linear, inequality constraints and so has a unique global minimum. We show below that the solution to the discrete optimization problem converges to the OD solutions as $h,k \to 0$. In Section 4.3.3 we discuss the technique we use to solve the optimization problem. This method gives the computational results shown in Figures 4.1 (right) and 4.2.

Denote $M = T/k$, we use $\{\vec{u}_n\}_{n=1}$ to define an approximate solution:

$$u_{N,M}(x,t) := \begin{cases} (1-t)\cdot u_0(x) + t\cdot u_1^N(x) & \text{for } t\in[0,k) \\ \cdots & \\ (1-t)\cdot u_i^N(x) + t\cdot u_{i+1}^N(x) & \text{for } t\in[ik,(i+1)k) \\ \cdots & \\ (1-t)\cdot u_{M-1}^N(x) + t\cdot u_M^N(x) & \text{for } t\in[(M-1)k,T], \end{cases} \tag{4.3.5}$$

where $u_0(x)$ is the initial condition and for $1 \le j \le M$,

$$u_j^N(x) = \sum_{i=1}^{N-1} u_j^i \cdot \Delta(Nx-i), \tag{4.3.6}$$

where $\Delta(x) = (1-|x|)^+$, the linear approximation function, therefore the pointwise limit

$$u_j = \lim_{N\to\infty} u_j^N \tag{4.3.7}$$

exists. For convenience we also define

$$\widetilde{u_j^N}(x) := \sum_{i=1}^{N} u_j^i \cdot \chi_{((i-1)h,ih)}(x) \tag{4.3.8}$$

where $\chi_I(x)$ is the characteristic function on the interval I and

$$u_M(x,t) := (1-t)\cdot u_j(x) + t\cdot u_{j+1}(x)\text{ , for } t \in [jk,(j+1)k). \tag{4.3.9}$$

Remark 14. *We observe that both approximations $u_j^N(x)$ and $\widetilde{u_j^N}(x)$ will converge to the same limit in $L^2(0,1)$ and similarly for $u_{N,M}(x,t)$ and $\widetilde{u_{N,M}}(x,t)$. If the energy is bounded then by Poincaré inequality and Arzelà-Ascoli Thm, uniform convergence can be obtained.*

Theorem 7. *Suppose $\vec{u}_{n+1}^N$ solves the discrete minimization problem* (4.3.4)*, then*

$$u(x,t) = \lim_{M\to\infty}\lim_{N\to\infty} u_{N,M}$$

with $h = 1/N$ and $k = T/M$ exists in $\mathscr{J}$ and u is the solution to the variational inequality (4.2.3) *that is*

$$\int_0^t\int_0^1 u_t\cdot(v-u) + \int_0^t\int_0^1 u'\cdot(v'-u') \geq \int_0^t\int_0^1 u - v;\ \textit{for all } v\in\mathscr{J},\ \textit{a.e. } t\in(0,T).$$

The proof relies on 2 lemmas. To start with, we give definitions of gamma convergence of energy functionals shown in Lemma 8 as given in [27]:

Definition 4.3.1 (Gamma convergence). *We say that the sequence of functionals* $\{\mathscr{E}_n\} : X \to \mathbb{R} \cup \{-\infty, +\infty\}$ *where* X *is a metric space,* Γ*-converges to* $\mathscr{E}$ *if the following conditions satisfied:*

i *whenever* $x_n \to x$*,* $\mathscr{E}(x) \leq \liminf_n \mathscr{E}_n(x_n)$*;*

ii *for any* $x \in X$*, there exists* $x_n \to x$ *in* X *such that* $\limsup_n \mathscr{E}_n(x_n) \leq \mathscr{E}(x)$.

The following is a relevant property of Γ-convergence:

Proposition 4.3.2. *Given a metric space* X *and suppose a sequence of functionals defined* $\mathscr{E}_n$ *defined in* X Γ*-converges to* $\mathscr{E}$*. Assume that for each* n*,* x_n *is a minimizer of* $\mathscr{E}_n$*, and if* $\overline{x}$ *is a cluster point of* $\{x_n\}$*, then* $\overline{x}$ *is a minimizer of* $\mathscr{E}$*.*

We refer the proof to [27]. In what follows

$$E_{n+1} = \int_0^1 \frac{1}{2}(u')^2 + \frac{1}{2k}(u - u_n)^2 + u$$

with $u(x) = u_{n+1}(x)$ defined in (4.3.7).

Lemma 8 (Gamma convergence of discrete functionals). *For each* n*,* E^N_{n+1} Γ*-converges to* E_{n+1} *as* $N \to \infty$ *or equivalently* $h \to 0$ *in* $L^2((0,1))$.

Proof of Lemma 8. We follow the proof in [27].

To show (i): let $u^N \in L^2((0,1))$ such that $\liminf E^N_{n+1}(u^N) < +\infty$ and therefore there exists a subsequence u^{N_k} such that $\lim E^{N_k}_{n+1}(u^{N_k}) = \liminf E^N_{n+1}(u^N)$. For each k, there exists $\vec{u}^{N_k}_{n+1} \in \mathbb{R}^{N_k+2}$ such that $u^{N_k} = u^{N_k}_{n+1}(x)$ as in the definition and $E^{N_k}_{n+1}(u^{N_k}) = E^{N_k}_{n+1}(u^{N_k}_{n+1})$. By the previous Remark 14, both $u^{N_k}_{n+1}$ and

$\widetilde{u^{N_k}_{n+1}}$ converge to the same limit u in L^2, we have

$$\sum_{i=0}^{N_k-1} \frac{h}{2}\left(\frac{u^{i+1}_{n+1}-u^{i}_{n+1}}{h}\right)^2 = \int_0^1 \left(\frac{du^{N_k}_{n+1}}{dx}\right)^2,$$

and thus

$$\int_0^1 (u')^2 \le \lim_k \int_0^1 \left(\frac{du^{N_k}_{n+1}}{dx}\right)^2 \le \liminf_N \sum_{i=0}^{N-1} \frac{h}{2}\left(\frac{u^{i+1}_{n+1}-u^{i}_{n+1}}{h}\right)^2.$$

On the other hand,

$$h\cdot\sum_{i=0}^{N}\left(\frac{1}{2k}\left(u^{i}-u^{i}_{n}\right)^2+u^{i}\right) = \int_0^1 \frac{1}{2k}(\widetilde{u^{N_k}_{n+1}}-\widetilde{u^{N_k}_{n}})^2+\widetilde{u^{N_k}_{n+1}}.$$

Applying the uniform convergence we obtain that

$$\int_0^1 \frac{1}{2k}(u-u_n)^2+u \le \lim_k \int_0^1 \frac{1}{2k}(\widetilde{u^{N_k}_{n+1}}-\widetilde{u^{N_k}_{n}})^2+\widetilde{u^{N_k}_{n+1}} \le \liminf_N h\cdot\sum_{i=0}^{N}\left(\frac{1}{2k}\left(u^{i}-u^{i}_{n}\right)^2+u^{i}\right).$$

These two estimates lead to $E_{n+1}(u) \le \liminf E^N_{n+1}(u^N)$.

It remains to prove (ii): suppose $u \in L^2((0,1))$ with $E_{n+1}(u) < +\infty$, so $u \in H^1$ and hence continuous. We then let $u^i_{n+1} = u(i/N)$ hence define the vector $\vec{u}^N_{n+1}$ with the piecewise linear approximation $u^N_{n+1}(x)$ and piecewise constant approximation $\widetilde{u^N_{n+1}}(x)$. They converge to u uniformly by remark 14. Then it can be shown similarly that

$$\limsup E^N_{n+1}(u^N_{n+1}) \le E_{n+1}(u).$$

□

As a result of Lemma 8 and Proposition 4.3.2, we obtain the following corollary immediately:

Corollary 8. *Suppose u^N_{n+1} are minimizers of E^N_{n+1} then u^N_{n+1} converges to a function u_{n+1} in $L^2((0,1))$ up to a subsequence as $h \to 0$ and such u_{n+1} is the minimizer of E_{n+1}.*

Now that $u_j := \lim_N u^N_j$ is the minimizer of the continuous functional E_j for $j = 1, \cdots, M$; it remains to show that $u(x,t) = \lim_{M\to\infty} u_M(x,t)$ solves the variational inequality (4.2.3). Recalling the Rothe's Method (Lemma 7) and combining results of Lemma 8 and Lemma 7, we therefore complete the proof of Theorem 7.

4.3.3 Discrete Optimization Scheme

We consider the details of the discrete optimization problem (4.3.4). The corresponding Lagrangian problem is

$$\begin{aligned}
&-D_2u + \frac{u}{k} + \lambda = \frac{u_n}{k} - 1, && \\
&\lambda_j < 0,\ u^j = 0 && \forall j \in J \\
&\lambda_i = 0,\ u^i \geq 0 && \forall i \in I,
\end{aligned}$$

where I and J are a disjoint partition of the grid points. The partitions divide those points J where the values are at the constraint and those points I ("I" for inactive constraint) with positive solution values where the corresponding derivative of E^N must be zero. Note that $\lambda_j < 0$ for $j \in J$ corresponds to $\partial E^N/\partial u^j > 0$, a necessary and sufficient condition for optimality (the KKT conditions [82]). There are many techniques available to solve such quadratic optimization problems with linear inequality constraints. We take advantage of the simple structure of the problem and the fact that there is little change in the index sets from one time step to the next in the following algorithm. It is an iterative algorithm with vectors $u^{(m)}$, $\lambda^{(m)}$ at each iteration. The matrix $A = I/k - D_2$, where I is the identity.

Algorithm

Step 1 Initialize $u^{(0)} \geq 0$ (component-wise), $\lambda^{(0)} = \min\{0, \frac{u_n}{k} - 1 - Au^{(0)}\}$. Set $m = 0$. Repeat steps 2-5 until the convergence criteria in step 3 is reached.

Step 2 Construct the index sets

$$\mathcal{J}^{(m)} = \{j : \lambda^{(m),j} < 0\},$$
$$I^{(m)} = \{j : \lambda^{(m),j} = 0\}.$$

For any $i \in I^{(m)}$ such that $u^{(m),i} < 0$ move i to $\mathcal{J}^{(m)}$.

Step 3 If $\mathcal{J}^{(m)} = \mathcal{J}^{(m-1)}$, the solution $u = u^{(m)}$. Stop.

Step 4 Solve for $u^{(m+1)}$ and λ using

$$Au^{(m+1)} + \lambda = \frac{u_n}{k} - 1,$$
$$\lambda = 0 \text{ on } I^{(m)},$$
$$u^{(m+1)} = 0 \text{ on } \mathcal{J}^{(m)}.$$

This is equivalent to solving sequentially for $(u^{(m+1)}, \lambda)$ that satisfy

$$\begin{aligned}
&A_{II} u_I^{(m+1)} = (\frac{u_n}{k} - 1)_I, \\
&u_J^{(m+1)} = 0, \\
&\lambda = \frac{u_n}{k} - 1 - A u^{(m+1)}.
\end{aligned}$$

Here vector subscripts I and J give the sub-vectors with those components and A_{II} is the block of the matrix A corresponding to the I components.

Step 5 Update $\lambda^{(m+1)} = \min\{0, \lambda\}$. Increment m.

Theorem 9. *Let $0 \leq u^{(0)} \leq u$ (component-wise). The algorithm above converges in finitely many steps.*

Proof. A proof is found following closely the ideas from [53] for a similar approach to the elliptic obstacle problem. Monotone behaviour in the index sets $I^{(m)}$ is shown and since N is finite, the algorithm converges in finite steps. Use is made of the properties that the sub-matrix A_{II}^{-1} has positive entries (A_{II} is monotone) and A_{IJ} has non-positive entries (values zero or $-1/h^2$) for any index sets I and J. □

Remark 15. *While the proof of iteration convergence above is limited to starting conditions $0 \leq u^{(0)} \leq u$, we implement the method with $u^{(0)} = u^n$ and starting index sets from the converged iterations at time step n. This initialization falls out of the scope of the analysis but works well (no failures, few iterations) in practice.*

Remark 16. *Similar index (active set) iteration methods have been used in capturing methods for other implicit boundary value problems. Two of these are discussed in Section 4.5. A general theory for the convergence of these iteration strategies is not known, but they can perform well in practice.*

4.4 Additional Results

4.4.1 Final state decay

Consider the 1D case. Under sufficient regularity assumptions, $u_{xx} = 1$ at the free boundary. If $u_t \leq 0$ ($u_{xx} \leq 1$) for all x at some time t_*, then by the maximum principle for u_{xx} (which formally obeys the heat equation) we have $u_t \leq 0$ for all x and all $t > t_*$. Thus we expect that the solution will decay uniformly after some time. We show a related result for the implicit time step, spatially continuous formulation of Section 4.2.4.

We assume $u_m \leq u_{m-1}$ (pointwise) for any $m = 1, 2, \cdots n$ as an induction hypothesis. Consider

$$\begin{aligned}\mathscr{L}(u_{n+1} - u_n) := -\Delta(u_{n+1} - u_n) + \frac{u_{n+1} - u_n}{k} &= \left(\frac{u_n}{k} - 1\right)\chi_{\{u_{n+1}>0\}} - \left(\frac{u_{n-1}}{k} - 1\right)\chi_{\{u_n>0\}}\\ &= \frac{u_n - u_{n-1}}{k}\cdot\chi_{\{u_{n+1}>0,u_n>0\}} + \left(\frac{u_n}{k} - 1\right)\cdot\chi_{\{u_{n+1}>0,u_n=0\}} - \left(\frac{u_{n-1}}{k} - 1\right)\chi_{\{u_{n+1}=0,u_n>0\}}.\end{aligned}$$

Note that

$$\frac{u_n - u_{n-1}}{k}\cdot\chi_{\{u_{n+1}>0,u_n>0\}} \leq 0$$

by induction and

$$\left(\frac{u_n}{k} - 1\right)\cdot\chi_{\{u_{n+1}>0,u_n=0\}} = -\chi_{\{u_{n+1}>0,u_n=0\}} \leq 0,$$

also note in $\{u_{n+1} = 0, u_n > 0\}$, one automatically has $u_{n+1} \leq u_n$. Rewrite $w = u_{n+1} - u_n$ and we then get

$$-\Delta w + \frac{w}{k} = \frac{u_n - u_{n-1}}{k} \cdot \chi_{\{u_{n+1}>0, u_n>0\}} - \chi_{\{u_{n+1}>0, u_n=0\}} - \left(\frac{u_{n-1}}{k} - 1\right) \chi_{\{u_{n+1}=0, u_n>0\}}$$

Suppose the w achieves maximum value at x_0 in the interior of Ω (Hopf's Lemma guarantees such x_0 doesn't locate at the boundary) such that $w(x_0) > 0$, then $\nabla w(x_0) = 0$ and $\Delta w(x_0) < 0$; this leads to $-\Delta w + \frac{w}{k} > 0$ at x_0, which contradicts to $-\Delta w + \frac{w}{k} < 0$ in $\{w > 0\}$. We can conclude that $w \leq 0$ in Ω, giving $u_{n+1} \leq u_n$.

Remark 17. *We conjecture that there is a generic form of solutions in the end state in the limit as a connected set of* $u > 0$ *disappears. Simulations such as that shown in Figure 4.2 suggest that the end state in 2D tends to a circular shape.*

4.4.2 Stable traveling wave solution

Little is known of the regularity of interfaces in higher dimensions. We conjecture that they will be smooth away from topological changes, with a statement similar to the 1D case (Conjecture 4.2.1). To give some indication of the regularity of a receding front in 2D, we show the linear stability of a planar front. We consider a 1D traveling wave solution in the form of $f(\xi) = u(x - ct)$ for some negative c:

$$\begin{cases} f'' + cf' = 1, \ \xi \in (-\infty, 0) \\ f(0) = f'(0) = 0 \qquad\qquad \text{free boundary condition} \end{cases}$$

By direct computation, we find that

$$f(\xi) = \frac{\xi}{c} - \frac{1}{c^2} + \frac{e^{-c\xi}}{c^2}. \tag{4.4.1}$$

To consider its linear stability, we take the Anstaz $\Phi = f(\xi) + e^{\lambda t + i\mu y} v(\xi)$ in

$$\Phi_t = \Phi_{\xi\xi} + c\Phi_\xi + \Phi_{yy} - 1$$

where $\mu \in \mathbb{R}$ and $\lambda \in \mathbb{C}$. This leads to

$$(\lambda + \mu^2)v = v'' + cv' \tag{4.4.2}$$

with the linearized condition $v(0) = 0$ and $v'(0)$ determining the linear change in interface position. To have appropriate decay in v as $\xi \to -\infty$ we must have two positive roots r of the auxiliary equation

$$r^2 + cr - (\lambda + \mu^2) = 0$$

which can only occur if $Re(\lambda) < -\mu^2$. Thus we have stability of the front to perturbations with parabolic decay.

4.5 Other Implicit Free Boundary Value Problems

4.5.1 A biharmonic problem

The OD problem is the simplest second order implicit free boundary problem. The simplest fourth order problem is the following biharmonic problem

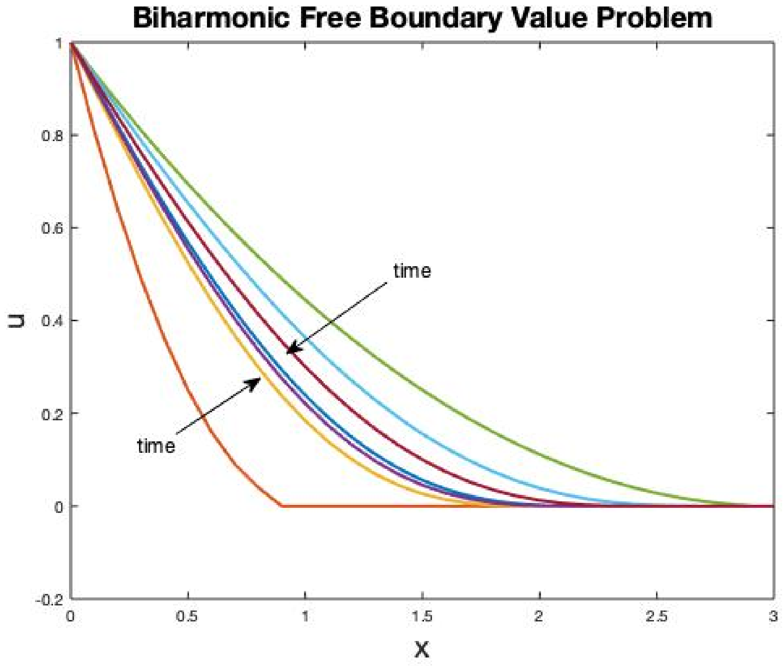

Figure 4.4: Two computations at three times each for the biharmonic free boundary value problem with physical boundary conditions $u(0) = 1$, $u_{xx} = 0$ approaching the analytic steady state solution shown in dark blue.

shown in 1D for $u(x,t)$:

$$u_t = -u_{xxxx} - 1$$

with conditions $u = 0$, $u_x = 0$, and $u_{xxx} = 0$ at the implicitly defined free boundary $x = s(t)$ and $u \equiv 0$ for $x > s(t)$. This can be derived from the scaled, linear, viscoelastic motion of a beam above a flat, rigid surface. Note that another boundary value problem occurs if $u_{xxx} = 0$ is replaced by $u_{xx} = 0$. However, the third order condition is correct for this application [78] and also gives the gradient flow structure described below.

We consider the time discretization of this problem as in Section 4.2.4 and see that it is a discrete L_2 gradient flow on the energy

$$\mathscr{E}^n := \int \frac{1}{2}|\Delta u_n|^2 + u_n$$

with $u_n \in H^2_+$. We form a fully discrete scheme as was done in Section 4.3.2 and compute the discrete optimization at each time step using index iterations as described in Section 4.3.3. The convergence of the method follows the same ideas as presented for the OD problem. Some computational results are shown in Figure 4.4.

Remark 18. *There has been considerable mathematical interest in the elliptic obstacle problem as discussed in the introduction. This is the steady state of the OD problem with nonzero physical boundary conditions. The steady state of the biharmonic problem (in higher dimensions) described in this section would also be mathematically interesting. Its analysis would be complicated by the lack of a maximum principle.*

4.5.2 Vector problems

The free boundary in complex fluids with yield stress is of implicit type and is well studied [42]. Numerical approaches include regularization (increased viscosity in the unyielded region) and an Augmented Lagrangian approach to the non-smooth optimization problem that comes from a discretization of a variational inequality formulation. The literature on this problem is focussed on capturing the unyielded region rather than considering the free boundary directly.

Implicit free boundaries in porous media flow can occur when phase change is present. Boundaries between dry and two-phase (where there is liquid and vapour present) regions were studied in [9, 45]. The work in [9] had important implications to simulations of water management in fuel cells. However, many theoretical questions were left unanswered and this became the motivation of the corresponding author to attempt the current work.

We present below a class of implicit free boundary value problems that generalizes the OD problem. The problems are presented in 1D with a single free boundary at $x = s(t)$ with $\mathbf{u}^l(x,t)$ having n components for $x < s(t)$ and $\mathbf{u}^r(x,t)$ having m components for $x > s(t)$. Near the interface we take

$$\mathbf{u}_t^* = D^*\mathbf{u}_{xx}^* + \mathbf{a}^*$$

for $* \in \{l,r\}$, D^* positive diagonal matrices, and $\mathbf{a}^*$ constant vectors. At the boundary, we take

$$B\begin{bmatrix} \mathbf{u}^l \\ \mathbf{u}_x^l \\ \mathbf{u}^r \\ \mathbf{u}_x^r \end{bmatrix} = \mathbf{0}$$

where B is an $(m+n+1)\times(2m+2n)$ matrix of full rank. This class can be reached from a wider class by taking affine combinations of solution components and x, and as an approximation of some nonlinear problems. A problem statement can be made by adding far field conditions, n on the left and m on the right. With these far field conditions we label the class as n+m implicit free boundary value problems. The OD problem is the only well defined example of the 1+0 class. The model in [9] is of class 2+2, although one of the components has degenerate diffusion at the free boundary.

There are several open questions related to problems of this type motivated by the current work on the OD problem. Which lead to well defined problems? (this could depend on the sign of entries of **a** as discussed in Remark 10). Which have gradient flow or variational inequality structure? Which allow a capturing formulation with index iteration similar to that described in Section 4.3.3? (true of the model in [9]).

4.6 Proof of Theorem 4.2.3

With the help of the minimality of u, we consider a competing function $u+\varepsilon\phi$ where ϕ is an arbitrary smooth function that is compactly supported inside Ω. By the definition of $\tilde{E}[u]$, it follows that

$$\tilde{E}[u+\varepsilon\phi] \geq \tilde{E}[u],$$

that is

$$\varepsilon \int \nabla u \cdot \nabla \phi + \frac{\varepsilon^2}{2}\int |\nabla \phi|^2 + \frac{\varepsilon}{k}\int u\phi + \frac{\varepsilon^2}{2k}\int \phi^2 \geq -\int (1-\frac{u_n}{k})[(u+\varepsilon\phi)^+ - u]. \tag{4.6.1}$$

Note that

$$\int (1-\frac{u_n}{k})[(u+\varepsilon\phi)^+ - u] = \varepsilon\int_{\{u+\varepsilon\phi\geq 0\}} (1-\frac{u_n}{k})\phi - \int_{\{u+\varepsilon\phi<0\}} (1-\frac{u_n}{k})u,$$

ignoring the $O(\varepsilon^2)$ terms in (4.6.1), we have

$$\begin{aligned}\varepsilon\int\nabla u\cdot\nabla\phi+\frac{\varepsilon}{k}\int u\phi-\varepsilon\int_{\{u+\varepsilon\phi\geq 0\}}(1-\frac{u_n}{k})^-\phi+\int_{\{u+\varepsilon\phi<0\}}(1-\frac{u_n}{k})^-u\\ \geq-\varepsilon\int_{\{u+\varepsilon\phi\geq 0\}}(1-\frac{u_n}{k})^+\phi+\int_{\{u+\varepsilon\phi<0\}}(1-\frac{u_n}{k})^+u.\end{aligned}\tag{4.6.2}$$

In fact we have

$$0\leq\int_{\{u+\varepsilon\phi<0\}}(1-\frac{u_n}{k})^{\pm}u<-\varepsilon\int_{\{u+\varepsilon\phi<0\}}(1-\frac{u_n}{k})^{\pm}\phi,$$

hence (4.6.2) turns out to be

$$\int\nabla u\cdot\nabla\phi+\frac{1}{k}\int u\phi-\int_{\{u+\varepsilon\phi\geq 0\}}(1-\frac{u_n}{k})^-\phi-\int_{\{u+\varepsilon\phi<0\}}(1-\frac{u_n}{k})^-\phi\geq-\int_{\{u+\varepsilon\phi\geq 0\}}(1-\frac{u_n}{k})^+\phi.$$

Moreover, we also recall that $u\geq 0$, then in L^1 sense as $\varepsilon\to 0$,

$$\begin{cases}\chi_{\{u+\varepsilon\phi\geq 0\}}\to\chi_{A_\phi\cup\{u>0\}}\\ \chi_{\{u+\varepsilon\phi<0\}}\to\chi_{\{u=0\}\cap\{\phi<0\}},\end{cases}$$

where $A_\phi:=\{u=0\}\cap\{\phi\geq 0\}$. Clearly, A_ϕ and $\{u>0\}$ are disjoint. This leads to

$$\begin{aligned}\int\nabla u\cdot\nabla\phi+\frac{1}{k}\int u\phi-\int\chi_{A_\phi\cup\{u>0\}}(1-\frac{u_n}{k})^-\phi-\int\chi_{\{u=0\}\cap\{\phi<0\}}(1-\frac{u_n}{k})^-\phi\\ \geq-\int\chi_{A_\phi\cup\{u>0\}}(1-\frac{u_n}{k})^+\phi,\end{aligned}$$

or equivalently,

$$\int \nabla u \cdot \nabla \phi + \frac{1}{k}\int u\phi + \int \chi_{A_\phi \cup \{u>0\}}(1-\frac{u_n}{k})\phi - \int \chi_{\{u=0\}\cap\{\phi<0\}}(1-\frac{u_n}{k})^{-}\phi \geq 0. \tag{4.6.3}$$

Define a distribution

$$T(\phi) := \int \nabla u \cdot \nabla \phi + \frac{1}{k}\int u\phi + \int \chi_{\{u>0\}}(1-\frac{u_n}{k})\phi,$$

then by (4.6.3),

$$T(\phi) \geq -\int_{A_\phi}(1-\frac{u_n}{k})\phi + \int_{\{u=0\}\cap\{\phi<0\}}(1-\frac{u_n}{k})^{-}\phi.$$

Since ϕ is arbitrary, we may replace it with $-\phi$ and as a result,

$$\begin{cases} T(\phi) \geq -\int_{A_\phi}(1-\frac{u_n}{k})\phi + \int_{\{u=0\}\cap\{\phi<0\}}(1-\frac{u_n}{k})^{-}\phi \\ T(\phi) \leq -\int_{\{u=0\}\cap\{\phi\leq 0\}}(1-\frac{u_n}{k})\phi + \int_{\{u=0\}\cap\{\phi>0\}}(1-\frac{u_n}{k})^{-}\phi. \end{cases} \tag{4.6.4}$$

Therefore, $|T(\phi)| \leq C\|\phi\|_\infty$ for some positive constant C, thus by a density argument we derive that T is a radon measure, i.e. there exists a density function $\rho(x)$ such that

$$T(\phi) = \int_\Omega \rho\phi \, dx.$$

However, by (4.6.4), we get $\rho = 0$ a.e. in $\{u > 0\}$; moreover, by definition of T we get $\rho = 0$ a.e. in $\{u = 0\}$. This shows that $T(\phi) = 0$, or

$$-\Delta u + \frac{1}{k}u + \chi_{\{u>0\}}(1-\frac{u_n}{k}) = 0$$

in the weak sense. Equivalently,

$$\frac{u - u_n \cdot \chi_{\{u>0\}}}{k} = \Delta u - \chi_{\{u>0\}}.$$

4.7 Proof of Theorem 6

As the discussion in Section 4.2.5 above showed, $u = \lim_{\epsilon \to 0} u_\epsilon$ exists pointwisely by monotonicity. It remains to show u is the solution to (4.2.3), that is

$$\int_0^t \int_0^1 \partial_t u \cdot (v-u) + \int_0^t \int_0^1 u' \cdot (v' - u') \geq \int_0^t \int_0^1 u - v; \text{ for all } v \in \mathscr{J}, \text{ a.e. } t \in (0,T).$$

We show the result in one spatial dimension, but the proof applies to 2D and 3D with minor modifications. Intuitively, suppose that f is a smooth approximation, then by maximum principle $|\partial_x u_\epsilon| \leq \max |u_0'(x)|$ for any $x \in [0,1]$ and $\epsilon > 0$. Thus $|\partial_x u| \leq \max |u_0'(x)|$, therefore by Dini's Theorem, such convergence is uniform and as a result, $u \in \mathscr{J}$ because u also satisfies the boundary condition and initial condition. Once we have such uniform boundedness of $\partial_x u_\epsilon$, $\partial_x u_\epsilon$ converges to $\partial_x u$ weakly and as a result,

$$\lim_\epsilon \int_0^t \int_0^1 u_\epsilon' \cdot (v' - u') = \int_0^t \int_0^1 u' \cdot (v' - u');$$

and

$$\lim_\epsilon \int_0^t \int_0^1 -f_\epsilon(u_\epsilon) \cdot (v-u) = -\int_0^t \int_0^1 \chi_{\{u>0\}} \cdot (v-u) = -\int_0^t \int_0^1 v - u + \int_0^t \int_0^1 \chi_{\{u=0\}} \cdot v.$$

Indeed we have weak convergence of $\partial_t u_\epsilon$ thanks to the equation:

$$\lim_\epsilon \int_0^t \int_0^1 \partial_t u_\epsilon \cdot (v-u) = \lim_\epsilon \int_0^t \int_0^1 -f(u_\epsilon)\cdot(v-u) - u_\epsilon' \cdot (v'-u').$$

Since u_ϵ converges to u pointwisely and strongly in $L^2((0,T);L^2((0,1)))$, then up to a subsequence

$$\lim_\epsilon \int_0^t \int_0^1 \partial_t u_\epsilon \cdot (v-u) = \int_0^t \int_0^1 \partial_t u \cdot (v-u).$$

Note that $v \geq 0$,

$$-\int_0^t \int_0^1 \chi_{\{u>0\}} \cdot (v-u) \geq -\int_0^t \int_0^1 v-u.$$

therefore

$$\int_0^t \int_0^1 \partial_t u \cdot (v-u) + \int_0^t \int_0^1 u' \cdot (v'-u') \geq \int_0^t \int_0^1 u-v; \text{ for all } v \in \mathscr{J}, \text{ a.e. } t \in (0,T).$$

Indeed, we only require the H^1 uniform boundedness of u_ϵ. To see this without using smooth $f(u_\epsilon)$ we write down u_ϵ in the mild form:

$$u_\epsilon(t) = e^{t\Delta}u_0 + \int_0^t e^{(t-s)\Delta}(f_\epsilon(u_\epsilon))\, ds\ ,$$

where $e^{t\Delta}$ represents convolution with heat kernel. As a result, for any first order differential operator D we have

$$Du_\epsilon = De^{t\Delta}u_0 + \int_0^t De^{(t-s)\Delta}(f_\epsilon(u_\epsilon))\, ds$$

and hence

$$\|Du_\epsilon\|_2 \leq \|De^{t\Delta}u_0\|_2 + \int_0^t \|De^{(t-s)\Delta}f(u_\epsilon)\|_2 \, ds.$$

Note that $e^{t\Delta}u_0$ solves the standard heat equation with initial data u_0, we have

$$\|De^{t\Delta}u_0\|_2 = \|e^{t\Delta}Du_0\|_2 \lesssim \|Du_0\|_2 \lesssim 1,$$

for any $t \in (0,T)$. On the other hand,

$$\|De^{(t-s)\Delta}f(u_\epsilon)\|_2 \lesssim \|De^{(t-s)\Delta}f(u_\epsilon)\|_\infty = |K * f(u_\epsilon)| \ ,$$

where K is the kernel corresponding to $De^{(t-s)\Delta}$. Since $|f| \leq 1$,

$$\begin{aligned} |K * f(u_\epsilon)| &\leq \|K\|_2 \cdot \|f(u_\epsilon)\|_2 \\ &\lesssim \|K\|_2. \end{aligned}$$

We see that from Fourier side

$$\begin{aligned} \|K\|_2^2 &\lesssim \sum_{k\in\mathbb{Z}} |k|^2 e^{-2(t-s)|k|^2} \\ &= \sum_{|k|\geq 1} |k|^2 e^{-2(t-s)|k|^2} \\ &\lesssim \int_1^\infty e^{-2(t-s)r^2} r^2 \, dr \ . \end{aligned}$$

Observe that

$$\begin{aligned} \int_1^\infty e^{-2(t-s)r^2} r^2 \, dr &= \frac{\sqrt{2\pi}[1-\operatorname{erf}(\sqrt{2(t-s)})] + 4\sqrt{t-s}e^{-2(t-s)}}{16(t-s)^{3/2}} \\ &\lesssim \frac{1-\operatorname{erf}(\sqrt{2(t-s)})}{(t-s)^{3/2}} + \frac{e^{-2(t-s)}}{t-s}, \end{aligned}$$

where $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$, the Gauss error function. Therefore,

$$\|De^{\gamma\Delta} f(u_\epsilon)\|_2 \lesssim \frac{\left(1-\operatorname{erf}(\sqrt{2(t-s)})\right)^{1/2}}{(t-s)^{3/4}} + \frac{e^{-(t-s)}}{(t-s)^{1/2}}.$$

Now we would assume $t \geq 1$, as the other case $t < 1$ is easier. Let $\gamma = t-s$, we split the following integral into 2 parts:

$$\int_0^t \|De^{\gamma\Delta} f(u_\epsilon)\|_2 \, d\gamma = \int_0^1 \|De^{\gamma\Delta} f(u_\epsilon)\|_2 \, d\gamma + \int_1^t \|De^{\gamma\Delta} f(u_\epsilon)\|_2 \, d\gamma.$$

(i) $\gamma > 1$**:** Then we have

$$\frac{\left(1-\operatorname{erf}(\sqrt{2\gamma})\right)^{1/2}}{\gamma^{3/4}} \lesssim \frac{e^{-\gamma}}{\gamma^{5/4}},$$

thus

$$\begin{aligned}\int_1^t \|De^{\gamma\Delta} f(u_\epsilon)\|_2 \, d\gamma &\lesssim \int_1^t \frac{e^{-\gamma}}{\gamma^{3/4}} + \frac{e^{-\gamma}}{\gamma^{1/2}} \, d\gamma \\ &\lesssim \int_1^t \frac{e^{-\gamma}}{\gamma^{1/2}} \, d\gamma \\ &\lesssim \int_1^\infty \frac{e^{-\gamma}}{\gamma^{1/2}} \, d\gamma \\ &\lesssim 1\ .\end{aligned}$$

(ii) $\gamma \leq 1$**:** We use another estimate for $\|K * f(u_\epsilon)\|_2$. We compute from the

Fourier side:

$$\begin{aligned}
\|K * f(u_\epsilon)\|_2^2 &= \sum_{|k|\geq 1} |k|^2 e^{-2\gamma|k|^2} |\widehat{f(u_\epsilon)}(k)|^2 \\
&\leq \max_{|k|\geq 1}\left\{|k|^2 e^{-2\gamma|k|^2}\right\} \cdot \sum_{|k|\geq 1} |\widehat{f(u_\epsilon)}(k)|^2 \\
&\lesssim \max_{|k|\geq 1}\left\{|k|^2 e^{-2\gamma|k|^2}\right\} \cdot \|f(u_\epsilon)\|_2^2 \\
&\lesssim \max_{|k|\geq 1}\left\{|k|^2 e^{-2\gamma|k|^2}\right\}.
\end{aligned}$$

Define $g(x) = x^2 e^{-2\gamma x^2}$, where $x \geq 0$. Then,

$$g'(x) = x e^{-2\gamma x^2}\left(1 - 2\gamma x^2\right) \ ,$$

this shows the maximum achieves at $x = \frac{1}{\sqrt{2\gamma}}$ and hence

$$g(x) \leq g(\frac{1}{\sqrt{2\gamma}}) \leq \frac{1}{\gamma}$$

thus

$$\|De^{\gamma\Delta} f(u_\epsilon)\|_2 \lesssim \frac{1}{\sqrt{\gamma}} \ ,$$

As a result,

$$\int_0^1 \|De^{\gamma\Delta} f(u_\epsilon)\|_2 \, d\gamma \lesssim \int_0^1 \frac{1}{\sqrt{\gamma}} \, d\gamma \cdot \|f(u_\epsilon)\|_2 \lesssim 1 \ .$$

As a result,

$$\|Du_\epsilon\|_2 \lesssim 1,$$

for any $t \in (0,T)$ and the bound is independent of ϵ.

4.8 Another Proof of the Regularization Result

We recall the variational inequality setting (4.2.3), that is to solve $u \in H^1_+$

$$\int_0^t \langle \partial_t u - \Delta u + 1, v - u \rangle \geq 0, \ \forall \ v \in \mathscr{J}.$$

As in [49], it then has an equivalent formulation, that is to solve $u(t)$ and $\lambda^*(t)$:

$$\begin{cases} \partial_t u - \Delta u + 1 = -\lambda^*(t) \geq 0 \\ u \geq 0, \ \langle u(t), \lambda^*(t) \rangle = 0, \ \forall \ t > 0. \end{cases} \tag{4.8.1}$$

To approach this, we introduce a regularized approximation family: we aim to find u_c for any $c > 0$ such that the following holds weakly:

$$\partial_t u_c - \Delta u^c + 1 + \min\left(0, -1 + cu^c\right) = 0.$$

By defining $\lambda^c = \min(0, -1 + cu^c)$, we can rewrite the above scheme as

$$\partial_t u^c - \Delta u^c + 1 + \lambda^c = 0.$$

It is typical to write the regularization term in this way in some literature, but the approach is the same as the regularization in Section 4.2.5 with $c = 1/\epsilon$. We then discretize it in time: for any $\phi \in H^1$, the following holds

$$\left\langle \frac{u^c_{n+1} - u^c_n}{k}, \phi \right\rangle + \left\langle \nabla u^c_{n+1}, \nabla \phi \right\rangle + \langle 1, \phi \rangle + \left\langle \min(0, -1 + cu^c_{n+1}), \phi \right\rangle = 0, \tag{4.8.2}$$

where u^c_0 is chosen to be u_0. We write u_n instead of u^c_n for simplicity. Note that the operator $A(u) := \frac{u}{k} - \Delta u + \min(0, -1 + cu)$ is coercive and monotone. As a result, there exists a unique solution $u_{n+1} \in H^1$ for sufficiently small $k > 0$ independent of $c > 0$. To show $u_{n+1} \in H^1_+$, we prove by induction. Assuming $u_n \in H^1_+$, we test the (4.8.2) with $(u_{n+1})^-$. Therefore we derive that

$$\begin{aligned}&\frac{1}{k}\langle u_{n+1},(u_{n+1})^-\rangle + \langle \nabla u_{n+1},\nabla (u_{n+1})^-\rangle + \langle 1,(u_{n+1})^-\rangle + \langle \min(0,-1+cu_{n+1}),(u_{n+1})^-\rangle\\ =&\frac{1}{k}\langle u_n,(u_{n+1})^-\rangle \le 0.\end{aligned}$$

We observe that $\langle \nabla u_{n+1},\nabla(u_{n+1})^-\rangle = \langle \nabla(u_{n+1})^-,\nabla(u_{n+1})^-\rangle \ge 0$. Moreover, $\langle 1,(u_{n+1})^-\rangle +$ $\langle \min(0,-1+cu_{n+1}),(u_{n+1})^-\rangle = c\langle (u_{n+1})^-,(u_{n+1})^-\rangle \ge 0$. We thus obtain that $\langle u_{n+1},(u_{n+1})^-\rangle \le 0$ and hence $u_{n+1} \in H^1_+$. We then define

$$u^c_M(x,t) = u_n + \frac{t-nk}{k}(u_{n+1}-u_n), \text{ for } t \in [nk,(n+1)k),$$

where $M = T/k$. By the same argument in Lemma 7, we have u^c_M converges to function u^c in $L^2(0,T;H^1)$ as $M \to \infty$ up to a subsequence. In fact, it is easy to see that u_c is the solution to (4.8.2). On the other hand, we show that u^c converges to u^* as $c \to \infty$.

Theorem 10 (Monotonicity). *Let u^c_{n+1} and u^c be defined as above. If $0 < c \le b$, then $u^c_{n+1} \ge u^b_{n+1}$ for all $n = 0,1,2,\cdots$. Therefore $u^c(t) \ge u^b(t)$ as a direct application.*

Proof. The proof is given by induction. Suppose $u^c_n \ge u^b_n$ and for each n define λ^c_n by

$$\lambda^c_{n+1} = \min(0,-1+cu^c_{n+1}).$$

Then the proof is similar to the one showing $u^c_{n+1} \geq 0$, we have that

$$\begin{aligned}&\frac{1}{k}\langle u^c_{n+1} - u^b_{n+1}, (u^c_{n+1} - u^b_{n+1})^-\rangle + \left\langle \nabla(u^c_{n+1} - u^b_{n+1}), \nabla(u^c_{n+1} - u^b_{n+1})^- \right\rangle \\ &+ \left\langle \lambda^c_{n+1} - \lambda^b_{n+1}, (u^c_{n+1} - u^b_{n+1})^- \right\rangle = \frac{1}{k}\langle u^n_c - u^b_n, (u^c_{n+1} - u^b_{n+1})^-\rangle \leq 0.\end{aligned}$$

Note that $cu^c_{n+1} - bu^b_{n+1} \leq cu^c_{n+1} - cu^b_{n+1}$ for $c \leq b$ and hence $\langle \lambda^c_{n+1} - \lambda^b_{n+1}, (u^c_{n+1} - u^b_{n+1})^-\rangle \geq 0$. We thus obtain that $u^c_{n+1} \geq u_b$.

□

As a corollary of the monotonicity, we obtain the existence of $u(t)$ and it solves (4.8.1). Uniqueness can be proved similarly as in [49].

Chapter 5

Non-uniqueness of the stationary surface quasi-geostrophic equation

In this chapter, we show the existence of nontrivial stationary weak solutions to the surface quasi-geostrophic equations on the two dimensional periodic torus.

5.1 Introduction

Consider the two dimensional dissipative surface quasi-geostrophic (SQG) equations for $\theta = \theta(x,t) : \mathbb{T}^2 \times [0,\infty) \to \mathbb{R}$:

$$\begin{cases} \partial_t \theta + u \cdot \nabla \theta = -\nu \Lambda^\gamma \theta, \qquad \text{in } \mathbb{T}^2 \times (0,\infty); \\ u = \nabla^\perp \Lambda^{-1} \theta = (-\partial_2 \Lambda^{-1} \theta, \partial_1 \Lambda^{-1} \theta) = (-\mathscr{R}_2 \theta, \mathscr{R}_1 \theta); \\ \theta|_{t=0} = \theta_0, \end{cases} \tag{SQG}$$

where $\nu \geq 0$ is the viscosity, $0 < \gamma \leq 2$ and $\mathbb{T}^2 = [-\pi,\pi]^2$ is the periodic torus. Here the unknown scalar function θ denotes the potential temperature in the context of geophysical fluid dynamics [44, 68]. This transport equation models the evolution of the temperature in a fast rotating stratified fluid and can be derived from a more complete 3D system via Boussinesq approximation [68]. In equation (SQG), $\mathscr{R} = (\mathscr{R}_1, \mathscr{R}_2)$ is the pair of Riesz transforms and $\nabla^\perp = (-\partial_2, \partial_1)$. For $s \geq 0$ the fractional Laplacian $\Lambda^s = (-\Delta)^{\frac{s}{2}}$ is defined by (under suitable assumptions on θ) $\widehat{\Lambda^s \theta}(k) = |k|^s \hat{\theta}(k)$ for $k \in \mathbb{Z}^2$. For negative s the formula is restricted to nonzero wave numbers. We consider solutions with zero mean, i.e. $\int_{\mathbb{T}^2} \theta(x,t)\, dx = 0$, which is invariant under the dynamics thanks to incompressibility. The purpose of this work is to construct stationary weak solutions to (SQG). By using integration by parts, one way to define stationary weak solutions to (SQG) is to drop the $\partial_t \theta$ term and require

$$-\int_{\mathbb{T}^2} \theta u \cdot \nabla \phi \, dx = -\nu \int_{\mathbb{T}^2} \theta \Lambda^\gamma \phi \, dx, \quad \forall \phi \in C^\infty(\mathbb{T}^2). \tag{5.1.1}$$

However, this definition requires the strong assumption $\theta \in L^2$ which did not take into account of the incompressibility condition. On the other hand, it is possible to define stationary weak solutions using the mere $\dot{H}^{-\frac{1}{2}}$-regularity. The starting point is to note that the operators $\mathscr{R}_j$, $j = 1, 2$ are skew-symmetric, i.e. $\langle \mathscr{R}_j f, g \rangle = -\langle f, \mathscr{R}_j g \rangle$ where $\langle , \rangle$ denotes the usual L^2 (real) inner product. Using this one can derive for $\theta \in L^2$ (below $[A,B] = AB - BA$ is the usual commutator):

$$\langle \theta \mathscr{R}_j \theta, \phi \rangle = -\frac{1}{2} \langle \theta, [\mathscr{R}_j, \phi] \theta \rangle, \qquad \forall \phi \in C^\infty(\mathbb{T}^2).$$

Since $\|[R_j,\phi]\theta\|_{\dot{H}^{\frac{1}{2}}} \lesssim \|\phi\|_{H^3}\|\theta\|_{\dot{H}^{-\frac{1}{2}}}$ (see Proposition 5.5.1), it is then not difficult to see that $\dot{H}^{-\frac{1}{2}}$-regularity suffices for defining a stationary weak solution.

Definition 5.1.1. *We say $\theta \in \dot{H}^{-\frac{1}{2}}(\mathbb{T}^2)$ with zero mean is a stationary weak solution to* (SQG) *if*

$$\frac{1}{2}\int_{\mathbb{T}^2}(\Lambda^{-\frac{1}{2}}\theta)\cdot\Lambda^{\frac{1}{2}}([\mathscr{R}^{\perp},\nabla\psi]\theta)dx = -\nu\int_{\mathbb{T}^2}(\Lambda^{-\frac{1}{2}}\theta)\Lambda^{\gamma+\frac{1}{2}}\psi dx, \quad \forall\,\psi\in C^{\infty}(\mathbb{T}^2),$$

where $[\mathscr{R}^{\perp},\nabla\psi]\theta = -[\mathscr{R}_2,\partial_1\psi]\theta + [\mathscr{R}_1,\partial_2\psi]\theta$.

In the non-steady case, weak solutions in $L^2_{t,\mathrm{loc}}\dot{H}_x^{-\frac{1}{2}}$ can be defined similarly by employing time-dependent test functions. Resnick [70] proved the global existence of a weak solution to (SQG) for $\nu \ge 0$ and $0<\gamma\le 2$ in $L^\infty_t L^2_x$ for any initial data $\theta_0\in L^2_x(\mathbb{T}^2)$. Marchand [60] obtained a global weak solution in $L^\infty_t H_x^{-\frac{1}{2}}$ for $\theta_0\in \dot{H}_x^{-\frac{1}{2}}(\mathbb{R}^2)$ or $L^\infty_t L^p_x$ for $\theta_0\in L^p_x(\mathbb{R}^2)$, $p\ge\frac{4}{3}$, when $\nu>0$ and $0<\gamma\le 2$. Note that in Marchand's result, the inviscid case $\nu=0$ requires $p>4/3$ since the embedding $L^{\frac{4}{3}}\hookrightarrow \dot{H}^{-\frac{1}{2}}$ is not compact, whereas for the diffusive case one has extra $L^2_t\dot{H}^{\frac{\gamma}{2}-\frac{1}{2}}$ conservation by construction.

For non-stationary smooth solutions with zero mean, one has conservation ($\nu=0$) or dissipation ($\nu>0$) of $\dot{H}^{-\frac{1}{2}}$-Hamiltonian. Indeed for $\nu=0$ by using the identity (below $P_{<J}$ is a smooth frequency projection to $\{|k|\le \text{constant}\cdot 2^J\}$)

$$\frac{1}{2}\frac{d}{dt}\|\Lambda^{-\frac{1}{2}}P_{<J}\theta\|_2^2 = -\int P_{<J}(\theta\mathscr{R}^{\perp}\theta)\cdot P_{<J}\mathscr{R}\theta dx,$$

one can prove the conservation of $\|\Lambda^{-\frac{1}{2}}\theta\|_2^2$ under the assumption $\theta\in L^3_{t,x}$ (see also [48]). We also mention that for the non-dissipative case in the positive direction uniqueness of SQG patches with moving boundary satisfying the arc-chord condition was obtained in recent [24].

In this chapter, we prove the non-uniqueness of stationary weak solutions to (SQG).

Theorem 5.1.2. *For any $\nu \geq 0$, $\gamma \in (0, \frac{3}{2})$, and $\frac{1}{2} \leq \alpha < \frac{1}{2} + \min(\frac{1}{6}, \frac{3}{2} - \gamma)$, there exist infinitely many stationary weak solutions θ to* (SQG) *with zero mean satisfying $\Lambda^{-1}\theta \in C^{\alpha}(\mathbb{T}^2)$.*

Remark 5.1.3. *The restriction $\gamma < \frac{3}{2}$ in Theorem 5.1.2 can be seen by a crude heuristic using the plane wave ansatz localized around frequency λ. The domination of nonlinearity versus dissipation yields $\|\Lambda^{-1}\theta\|_{\infty} \gg \lambda^{\gamma-2}$. The Hölder regularity of $\Lambda^{-1}\theta$ yields $\|\Lambda^{-1}\theta\|_{\infty} \lesssim \lambda^{-\alpha}$ where $\alpha > \frac{1}{2}$. Thus $\gamma \leq 2 - \alpha < \frac{3}{2}$.*

One can apply the convex integration scheme [29, 30] to general active scalar models such as $\partial_t \theta + \nabla \cdot (\theta u) = 0$ where $\widehat{u} = m(k)\widehat{\theta}(k)$ and $m(k)$ is a general Fourier multiplier. By using a plane wave ansatz $\theta = a_k e^{i\lambda k \cdot x} + a_k^* e^{-i\lambda k \cdot x}$ with $|k| = 1$ and $\lambda \gg 1$, one can extract the non-oscillatory part of $\nabla \cdot (\theta u)$ as $\nabla \cdot (|a_k|^2 (m(-\lambda k) + m(\lambda k)))$ which vanishes if m is odd. This is known as the odd multiplier obstruction [28, 48, 77]. Previously the non-uniqueness results were established only for active scalar equations with non-odd multipliers [48, 77]. In [10] this issue was resolved for the time-dependent SQG, by using the momentum equation[1] for $v = \Lambda^{-1}u$ and rewriting the nonlinearity $u \cdot \nabla v - (\nabla v)^T \cdot u$ as the sum of a divergence of a 2-tensor, and a gradient of a scalar function. In particular, weak solutions $\Lambda^{-1}\theta \in C_t^{\sigma} C_x^{\beta}$, $\frac{1}{2} <$

[1] This approach originates from an exposition in [83], which dates back to Resnick's thesis [70].

$\beta < \frac{4}{5}, \sigma < \frac{\beta}{2-\beta}$, with any prescribed energy $\|\Lambda^{-\frac{1}{2}}\theta(t)\|_2 = e(t) \in C_c^\infty$ were constructed when $\nu \geq 0, 0 < \gamma < 2-\beta$. Note that the restriction $\beta - 1 < 1-\gamma$ accords with the critical $\|\theta\|_{L_t^\infty \dot{C}^{1-\gamma}}$ norm. Recently Isett and Ma [47] give another direct approach at the level of θ. For some more recent application of convex integration to other fluid models, see [20, 23, 58] and the references therein.

The modest goal of this chapter is to introduce another approach to overcome the odd multiplier obstruction by working directly with the scalar function $f = \Lambda^{-1}\theta$ and developing a concise framework tailor-made for similar problems. From our analysis it appears that the indirect momentum formulation emphasized in [10] is not needed and one can settle the problem directly using the special structure of SQG. Returning to the plane wave ansatz, a decisive step for the SQG nonlinearity is to identify the nontrivial non-oscillatory part after removing the $\nabla^\perp$-direction. More precisely, consider $f = \sum_l a_l(x)\cos(\lambda l \cdot x)$ where $|l| = 1$ and $\lambda \gg 1$, then (see Lemma 5.2.1)

$$\Lambda f = \sum_l \Big(\lambda f + (l\cdot\nabla)a_l \sin(\lambda l \cdot x) + (T_{\lambda l}^{(1)} a_l)\cos(\lambda l \cdot x) + (T_{\lambda l}^{(2)} a_l)\sin(\lambda l \cdot x)\Big).$$

By a short computation we arrive at

$$\Lambda f \nabla^\perp f \overset{\circ}{\approx} -\frac{1}{4}\lambda \sum_l (l\cdot\nabla)(a_l^2) l^\perp + \text{error terms},$$

where the notation $\overset{\circ}{\approx}$ is defined in (5.1.2). We then use a novel algebraic lemma (Lemma 5.2.2) to obtain nontrivial projection in the gradient direction. One should note that in the above computation, the leading $O(\lambda^2)$ term vanishes which completely accords with the odd multiplier obstruction problem mentioned earlier. What is remarkable is that in the next $O(\lambda)$ term there is nontrivial non-oscillatory contribution coming from the commutator piece $[\Lambda, a_l]\cos\lambda x$. This seems to be the crucial technical difference between SQG and Euler.

Our next result is about the weak rigidity of solutions in the time-dependent case. It improves Theorem 1.3 of [48] all the way from $L_t^p L_x^2$, $p>2$ to $L_t^2 \dot{H}^{-\frac{1}{2}+}$. The proof can be found in Section 5.5.

Theorem 5.1.4 (Weak rigidity). *Let $\nu \geq 0$ and $0<\gamma\leq 2$. Suppose $f = \lim_n \theta_n$ is a weak limit of solutions* (SQG) *in $L_t^2 \dot{H}^s$ for $s > -\frac{1}{2}$. Then f must also be a weak solution.*

Notation in this chapter

For a real number X, we use X^+ for $X+\epsilon$ when $\epsilon>0$ is sufficiently small. For any two vector functions v and w, we denote

$$\boxed{v \overset{\circ}{\approx} w, \quad \text{if} \quad v = w + \nabla^\perp p} \tag{5.1.2}$$

holds for some smooth scalar function p. The mean of f on $\mathbb{T}^2$ is denoted by $\overline{f} = \frac{1}{(2\pi)^2}\int_{\mathbb{T}^2} f(x)dx$. We define the function space $C_0^\infty(\mathbb{T}^2)$ as

$$C_0^\infty(\mathbb{T}^2) = \left\{ f \in C^\infty(\mathbb{T}^2) : \overline{f} = 0 \right\}. \tag{5.1.3}$$

For any $1 \leq p \leq \infty$, we denote $\|f\|_p = \|f\|_{L^p(\mathbb{T}^2)}$ as the usual Lebesgue norm. For f on $\mathbb{T}^2$, we follow the Fourier transform convention $\hat{f}(k) = \frac{1}{(2\pi)^2}\int_{\mathbb{T}^2} f(x)e^{-ix\cdot k}dx$ and $f(x) = \sum_{k\in\mathbb{Z}^2}\hat{f}(k)e^{ik\cdot x}$. The convolution operation $*$ is defined by $(f *$

$g)(x) = \frac{1}{(2\pi)^2}\int_{\mathbb{T}^2} f(x-y)g(y)dy$, which implies $\widehat{f*g}(k) = \hat{f}(k)\hat{g}(k)$ and $\widehat{fg}(k) = \sum_{l\in\mathbb{Z}^2}\hat{f}(l)\hat{g}(k-l)$.

For $s \in \mathbb{R}$, the homogeneous $\dot{H}^s$-Sobolev norm is defined by $\|f\|_{\dot{H}^s(\mathbb{T}^2)} = \left(\sum_{0\neq k\in\mathbb{Z}^2}|k|^{2s}|\hat{f}(k)|^2\right)^{\frac{1}{2}}$.

Parameters

Throughout this chapter, we fix parameters as follows. $\nu \geq 0$, $0 < \gamma < \frac{3}{2}$, $0 < \beta < \min\{\frac{1}{3}, 3-2\gamma\}$,

$$\lambda_n = \left\lceil \lambda_0^{b^n} \right\rceil, \quad r_n = \lambda_n^{-\beta}, \quad \mu_{n+1} = (\lambda_{n+1}\lambda_n)^{\frac{1}{2}}, \qquad n \in \mathbb{N}\cup\{0\}, \tag{5.1.4}$$

where $\lceil\cdot\rceil$ denotes the ceiling function. Here $\lambda_0 \in \mathbb{N}$, $b = 1^+$, will be chosen in Proposition 5.3.1. The Hölder exponent in Theorem 5.1.2 is $\alpha = \frac{1}{2} + \frac{\beta}{2b} - \epsilon_0 > \frac{1}{2}$ by taking first $b-1$ sufficiently small and then ϵ_0 sufficiently small. See also Section 5.6 for more explicit dependence of constants.

5.2 Construction of the perturbation

For $f = \Lambda^{-1}\theta$ the steady-state SQG equation is $\nabla\cdot\left(\Lambda f \nabla^\perp f\right) = -\nu\Lambda^{\gamma+1}f$ which follows from $\Lambda f\nabla^\perp f \overset{\circ}{\approx} \nu\Lambda^{\gamma-1}\nabla f$. The idea is to find approximate solutions $(f_{\leq n}, q_n) \in C_0^\infty(\mathbb{T}^2)\times C_0^\infty(\mathbb{T}^2)$ solving the relaxed equation

$$\Lambda f_{\leq n}\nabla^\perp f_{\leq n} \overset{\circ}{\approx} \nu\Lambda^{\gamma-1}\nabla f_{\leq n} + \nabla q_n, \tag{5.2.1}$$

such that $q_n \to 0$ in the limit. This will be done inductively.

Writing $f_{\le n+1} = f_{\le n} + f_{n+1}$, we first show that for given q_n one can solve

$$\Lambda f_{n+1} \nabla^\perp f_{n+1} + \nabla q_n \overset{\circ}{\approx} \text{small error}, \tag{5.2.2}$$

where the left hand side is the main piece in

$$(\Lambda f_{n+1} \nabla^\perp f_{n+1} + \nabla q_n) + \Lambda f_{\le n} \nabla^\perp f_{n+1} + \Lambda f_{n+1} \nabla^\perp f_{\le n} \overset{\circ}{\approx} \nabla q_{n+1} + \nu \Lambda^{\gamma-1} \nabla f_{n+1}. \tag{5.2.3}$$

5.2.1 Derivation of the leading order part

Consider the ansatz ($f = f_{n+1}$)

$$f(x) = \sum_l a_l(x) \cos(\lambda l \cdot x), \tag{5.2.4}$$

where the frequency of a_l is much smaller than λ and the summation over l is finite.

Lemma 5.2.1 (Leibniz). *Let $|l| = 1$, $\lambda l \in \mathbb{Z}^2$, and $g(x) = a(x)\cos(\lambda l \cdot x)$. Then,*

$$\Lambda g = \lambda g + (l \cdot \nabla a) \sin(\lambda l \cdot x) + (T^{(1)}_{\lambda l} a) \cos(\lambda l \cdot x) + (T^{(2)}_{\lambda l} a) \sin(\lambda l \cdot x),$$

where

$$\widehat{T^{(1)}_{\lambda l} a}(k) = \left(\frac{|\lambda l + k| + |\lambda l - k|}{2} - \lambda \right) \widehat{a}(k), \quad \widehat{T^{(2)}_{\lambda l} a}(k) = i \left(\frac{|\lambda l + k| - |\lambda l - k|}{2} - l \cdot k \right) \widehat{a}(k). \tag{5.2.5}$$

Proof. We begin with the following simple fact: if $\widehat{T_m g}(k) = m(k)\widehat{g}(k)$, then

$\forall n \in \mathbb{Z}^2$, $T_m(g(x)e^{in\cdot x}) = (T_{m_1}g)e^{in\cdot x}$, where $m_1(k) = m(k+n)$. Noting that $\widehat{\Lambda g}(k) = |k|\widehat{g}(k)$, we have

$$\Lambda(a(x)\cos(\lambda l\cdot x)) = \frac{1}{2}\Lambda(a(x)e^{i\lambda l\cdot x}) + \frac{1}{2}\Lambda(a(x)e^{-i\lambda l\cdot x}) = \frac{1}{2}\Lambda_{m_1}(a)e^{i\lambda l\cdot x} + \frac{1}{2}\Lambda_{m_2}(a)e^{-i\lambda l\cdot x},$$

where $\widehat{\Lambda_{m_1}a}(k) = |k+\lambda l|$ and $\widehat{\Lambda_{m_2}a}(k) = |k-\lambda l|$. The desired identity then follows by rearranging terms. □

By using Lemma 5.2.1, we have

$$\Lambda f \nabla^{\perp} f \overset{\circ}{\approx} \boxed{\text{main}} + \boxed{\text{non-oscillatory error}} + \boxed{\text{oscillatory error}}, \tag{5.2.6}$$

where (below $l^\perp = (-l_2, l_1)^\intercal$ for $l = (l_1, l_2)^\intercal$)

$$\boxed{\text{main}} = -\frac{1}{4}\lambda\sum_l (l\cdot\nabla)(a_l^2)l^\perp,$$

$$\boxed{\text{non-oscillatory error}} = -\frac{1}{2}\lambda\sum_l (T^{(2)}_{\lambda l}a_l)a_l l^\perp + \frac{1}{2}\sum_l (T^{(1)}_{\lambda l}a_l)\nabla^\perp a_l,$$

$$\boxed{\text{oscillatory error}} = \frac{1}{2}\sum_l (l\cdot\nabla a_l + T^{(2)}_{\lambda l}a_l)(\lambda a_l l^\perp \cos(2\lambda l\cdot x) + \nabla^\perp a_l \sin(2\lambda l\cdot x)) \tag{osc1}$$

$$-\frac{1}{2}\sum_l (T^{(1)}_{\lambda l}a_l)(\lambda a_l l^\perp \sin(2\lambda l\cdot x) - \nabla^\perp a_l \cos(2\lambda l\cdot x)) \tag{osc2}$$

$$-\lambda\sum_{l\neq l'}(l\cdot\nabla a_l + T^{(2)}_{\lambda l}a_l)a_{l'}(l')^\perp \sin(\lambda l\cdot x)\sin(\lambda l'\cdot x) \tag{osc3}$$

$$+\sum_{l\neq l'}(l\cdot\nabla a_l + T^{(2)}_{\lambda l}a_l)\nabla^\perp a_{l'}\sin(\lambda l\cdot x)\cos(\lambda l'\cdot x) \tag{osc4}$$

$$-\lambda\sum_{l\neq l'}(T^{(1)}_{\lambda l}a_l)a_{l'}(l')^\perp\cos(\lambda l\cdot x)\sin(\lambda l'\cdot x) \tag{osc5}$$

$$+\sum_{l\neq l'}(T^{(1)}_{\lambda l}a_l)\nabla^\perp a_{l'}\cos(\lambda l\cdot x)\cos(\lambda l'\cdot x). \tag{osc6}$$

Note that the leading-order term $\lambda f\nabla^\perp f$ in $\Lambda f\nabla^\perp f$ vanishes since $\nabla^\perp\left(\frac{\lambda}{2}f^2\right) \overset{\circ}{\approx} 0$.

5.2.2 Matching

We begin with a simple yet powerful lemma.

Lemma 5.2.2 (Algebraic Lemma)**.** *For a given $Q \in C_0^\infty(\mathbb{T}^2)$, we have the de-*

composition identity

$$\sum_{j=1}^{2} l_j^{\perp}(l_j \cdot \nabla)(\mathscr{R}_j^o Q) \overset{\circ}{\approx} \nabla Q,$$

where $l_1 = (\frac{3}{5}, \frac{4}{5})^{\intercal}$, $l_2 = (1,0)^{\intercal}$, *and the Riesz-type transforms* $\mathscr{R}_j^o$, $j = 1,2$ *are defined by*

$$\widehat{\mathscr{R}_1^o}(k_1,k_2) = \frac{25(k_2^2 - k_1^2)}{12|k|^2}, \quad \widehat{\mathscr{R}_2^o}(k_1,k_2) = \frac{7(k_2^2 - k_1^2)}{12|k|^2} + \frac{4k_1k_2}{|k|^2}. \tag{5.2.7}$$

Proof. This follows from the identity $\sum_{j=1}^{2}(l_j^{\perp} \cdot \nabla)(l_j \cdot \nabla)(\mathscr{R}_j^o Q) = \Delta Q$. □

Proposition 5.2.3. *Set* l_j *and* $\mathscr{R}_j^o$, $j = 1,2$ *as in Lemma 5.2.2. For given* $q_n \in C_0^{\infty}(\mathbb{T}^2)$, *choose* $C_0 \geq 2$ *to be a fixed constant and*

$$a_{j,n+1}^{\text{perfect}} = 2\sqrt{\frac{r_n}{5\lambda_{n+1}}}\sqrt{C_0 + \mathscr{R}_j^o \frac{q_n}{r_n}}, \tag{5.2.8}$$

where (λ_{n+1}, r_n) *are taken as in* (5.1.4). *Then*

$$-\frac{1}{4} \cdot (5\lambda_{n+1}) \cdot \Big(\sum_{j=1}^{2} l_j^{\perp}(l_j \cdot \nabla)(a_{j,n+1}^{\text{perfect}})^2\Big) + \nabla q_n \overset{\circ}{\approx} 0. \tag{5.2.9}$$

Proof. The proof follows from applying Lemma 5.2.2 to $Q = q_n$. □

We now choose

$$f_{n+1}(x) = \sum_{j=1}^{2} a_{j,n+1}(x)\cos(5\lambda_{n+1} l_j \cdot x), \qquad a_{j,n+1} = P_{\leq \mu_{n+1}} a_{j,n+1}^{\text{perfect}}, \tag{5.2.10}$$

where $\widehat{P_{\leq \mu_{n+1}} g}(k) = \psi(\frac{k}{\mu_{n+1}})\widehat{g}(k)$, and $\psi \in C_c^{\infty}(\mathbb{R}^2)$ satisfies $\psi(k) = 0$ for $|k| \geq 1$, and $\psi(k) = 1$ for $|k| \leq \frac{1}{2}$. We have $\Lambda f_{n+1} \nabla^{\perp} f_{n+1} + \nabla q_n \overset{\circ}{\approx}$ small error. In the next section we estimate the errors.

5.3 Error estimates

In this section we prove the following proposition which is the key in the whole iteration procedure.

Proposition 5.3.1. *Given* $\nu \ge 0$, $0 < \gamma < \frac{3}{2}$, $0 < \beta < \min\left(\frac{1}{3}, 3 - 2\gamma\right)$, *there exists* $b_0 = b_0(\nu, \gamma, \beta)$ *such that for any* $0 < b - 1 < b_0$ *we can find* $\Lambda_0 = \Lambda_0(\nu, \gamma, \beta, b)$ *for which the following holds. If* $\lambda_0 \ge \Lambda_0$ *and* $(f_{\le n}, q_n)$ *satisfies*

- *the frequencies of* $f_{\le n}$ *and* q_n *are localized to* $\le 6\lambda_n$ *and* $\le 12\lambda_n$, *respectively,*

- $\|f_{\le n}\|_{C^\alpha(\mathbb{T}^2)} \le 100$ *and* $\|q_n\|_X \le r_n$ *where*

$$\|q\|_X := \|q\|_\infty + \sum_{j=1}^{2} \|\mathscr{R}_j^o q\|_\infty, \tag{5.3.1}$$

and $\mathscr{R}_j^o$ *is defined in* (5.2.7). *Then there exists* $q_{n+1} \in C_0^\infty(\mathbb{T}^2)$ *solving* (5.2.3) *with frequency localized to* $\le 12\lambda_{n+1}$, f_{n+1} *defined by* (5.2.10) *satisfying*

$$\|q_{n+1}\|_X \le r_{n+1}. \tag{5.3.2}$$

We now explain the motivation for choosing the X-norm in (5.3.1). First of all, $q = q_n$ represents the residual error at step n and in the Hölderian context an ideal choice is to use $\|q\|_\infty$ only. However, there are Riesz-type operators $\mathscr{R}_j^o$, $j = 1, 2$ which appear somewhat inevitably in the "matching" process (see for example Proposition 5.2.3 and especially (5.2.8)). For this reason it is necessary to include $\|\mathscr{R}_j^o q\|_\infty$ in the working X-norm. To prove Proposition 5.3.1, we need several technical lemmas.

Lemma 5.3.2. *Suppose $a : \mathbb{T}^2 \to \mathbb{R}$, $a \in L^\infty(\mathbb{T}^2)$ with $\operatorname{supp}(\widehat{a}) \subset \{|k| \le \mu\}$ and $\mu \ge 10$. Let $m \in C^\infty(\mathbb{R}^2 \setminus \{0\})$ be a homogeneous function of degree 0 and $\mathscr{R}$ is the Fourier multiplier defined by $\widehat{\mathscr{R}f}(k) = m(k)\widehat{f}(k)$,then we have $\|\mathscr{R}a\|_\infty \lesssim \|a\|_\infty \log\mu$. Here the implied constant depends on m.*

Proof. With no loss we can assume $\overline{a} = 0$. Using the Littlewood-Paley decomposition [80], splitting into low and high frequencies and choosing integer $J \sim 2\log\mu$, we obtain

$$\begin{aligned}\|\mathscr{R}a\|_\infty &\lesssim (J+3)\|a\|_\infty + 2^{-J}\|\nabla a\|_\infty \\ &\lesssim (J+3+2^{-J}\mu)\|a\|_\infty \lesssim \|a\|_\infty \log\mu.\end{aligned}$$

□

We now state two useful facts. Assume $f \in C^\infty(\mathbb{T}^2)$ and $K \in L^1(\mathbb{R}^2)$ with $m(\xi) = \int_{\mathbb{R}^2} K(z)e^{-i\xi\cdot z}dz$. Then[2]

$$(T_m f)(x) := \sum_k m(k)\hat{f}(k)e^{ik\cdot x} = \int_{\mathbb{R}^2} K(z)f(x-z)dz, \tag{5.3.3}$$

$$\|T_m f\|_{L_x^p(\mathbb{T}^2)} \le \|K\|_{L_x^1(\mathbb{R}^2)}\|f\|_{L_x^p(\mathbb{T}^2)},\ \forall\, 1 \le p \le \infty. \tag{5.3.4}$$

[2] Here and below we still denote by f its periodic extension to all of $\mathbb{R}^2$.

Assume $f,g \in C^{\infty}(\mathbb{T}^2)$ and $K \in L^1(\mathbb{R}^2 \times \mathbb{R}^2)$ with

$$m(\xi,\eta) = \int_{\mathbb{R}^2 \times \mathbb{R}^2} K(z_1,z_2) e^{-i\xi\cdot z_1 - i\eta\cdot z_2} dz_1 dz_2.$$

Then

$$T_m(f,g)(x) := \sum_k \Big(\sum_{k' \in \mathbb{Z}^2} m(k',k-k')\hat{f}(k')\hat{g}(k-k') \Big) e^{ik\cdot x} \tag{5.3.5}$$

$$= \int_{\mathbb{R}^2 \times \mathbb{R}^2} K(z_1,z_2) f(x-z_1) g(x-z_2) dz_1 dz_2, \tag{5.3.6}$$

and consequently $\|T_m(f,g)\|_{L^r_x(\mathbb{T}^2)} \le \|K\|_{L^1_x(\mathbb{R}^2\times\mathbb{R}^2)} \|f\|_{L^p_x(\mathbb{T}^2)} \|g\|_{L^q_x(\mathbb{T}^2)}$ for any $1 \le r,p,q \le \infty$ with $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$.

Lemma 5.3.3. *Assume* $b_0 : \mathbb{T}^2 \to \mathbb{R}$ *with* $\operatorname{supp}(\widehat{b_0}) \subset \{|k| \le \mu\}$ *and* $10 \le \mu \le \frac{1}{2}\lambda$. *Then (see* (5.2.5)*)*

$$\|T^{(1)}_{\lambda l} b_0\|_\infty \lesssim \lambda^{-1}\mu^2 \|b_0\|_\infty,$$

$$\|T^{(2)}_{\lambda l} b_0\|_\infty \lesssim \lambda^{-2}\mu^3 \|b_0\|_\infty,$$

$$\|\Delta^{-1}\nabla T^{(2)}_{\lambda l} b_0\|_X \lesssim \|b_0\|_\infty \lambda^{-2}\mu^2 \log\mu.$$

Proof. We show only the first one as the rest are similar. Choose $\phi_1 \in C^\infty_c(\mathbb{R}^2)$ such that $\phi_1(\xi) \equiv 1$ for $|\xi| \le 1$ and $\phi_1(\xi) \equiv 0$ for $|\xi| \ge 1.1$. Denote $\phi_2(z) = |l + z| + |l - z| - 2$ and note that for $|z| \le \frac{2}{3}$ we have $\phi_2(z) = \sum_{i,j=1}^2 h_{ij}(z) z_i z_j$ for some $h_{ij} \in C^\infty$. By (5.3.4) it suffices to show $\|F\|_{L^1_x(\mathbb{R}^2)} \lesssim \lambda^{-2}\mu^2$ for $F(x) = \int_{\mathbb{R}^2} \phi_2(\lambda^{-1}\xi)\phi_1(\mu^{-1}\xi) e^{i\xi\cdot x} d\xi$. This follows from a change of variable $\mu^{-1}\xi \to \xi$ and integration by parts. For the third estimate one can extract an extra gradient from the symbol and then use Lemma 5.3.2. □

Lemma 5.3.4. *Let* $\operatorname{supp}(\widehat{b_0}) \subset \{|k| \le \mu\}$, $\mu \le \frac{1}{2}\lambda$. *Then for*[3] *some* $K_i = \mathscr{F}^{-1}(m_i)$ *with* $\|K_i\|_{L^1(\mathbb{R}^4)} \lesssim 1$, *we have*

$$b_0 T^{(2)}_{\lambda l} b_0 = \frac{\mu^2}{\lambda^2} \sum_{i=1}^{4} \partial_{x_i} T_{m_i}(b_0, b_0),$$
$$(T^{(1)}_{\lambda l} b_0)\partial_{x_1} b_0 = \frac{\mu^2}{\lambda} \sum_{i=1}^{4} \partial_{x_i} T_{m_i}(b_0, b_0),$$
$$(T^{(1)}_{\lambda l} b_0)\partial_{x_2} b_0 = \frac{\mu^2}{\lambda} \sum_{i=1}^{4} \partial_{x_i} T_{m_i}(b_0, b_0).$$

Proof. Observe that for $|z| \le \frac{2}{3}$, $\phi(z) = |l+z| - |l-z| - 2l\cdot z = \sum_{i,j,k=1}^{2} h_{ijk}(z) z_i z_j z_k$ for some $h_{ijk} \in C^\infty$. Choose $\phi_1 \in C_c^\infty(\mathbb{R}^2)$ such that $\phi_1(\xi) \equiv 1$ for $|\xi| \le 1$ and $\phi_1(\xi) \equiv 0$ for $|\xi| \ge 1.1$. By using parity of ϕ, we have

$$\widehat{b_0 T^{(2)}_{\lambda l} b_0}(k) = \frac{i}{4}\lambda \sum_{k' \in \mathbb{Z}^2} (\phi(\lambda^{-1}k') - \phi(\lambda^{-1}(k'-k)))\widehat{b_0}(k')\widehat{b_0}(k-k')$$
$$= -\frac{i}{4} \sum_{k' \in \mathbb{Z}^2} \int_0^1 k \cdot (\nabla\phi)(\lambda^{-1}(k' - \theta k)) d\theta \phi_1(\mu^{-1}k')\phi_1(\mu^{-1}(k-k'))\widehat{b_0}(k')\widehat{b_0}(k-k').$$

Note that

$$(\nabla\phi)(\frac{k'-\theta k}{\lambda})\phi_1(\frac{k'}{\mu})\phi_1(\frac{k-k'}{\mu}) = \lambda^{-2} \sum_{1\le i,j\le 2} \tilde{h}_{ij}(\frac{k'-\theta k}{\lambda})(k'-\theta k)_i (k'-\theta k)_j \phi_1(\frac{k'}{\mu})\phi_1(\frac{k-k'}{\mu}),$$

where $\tilde{h}_{ij} \in C_c^\infty(\mathbb{R}^2)$. The result then follows from (5.3.6) by checking the L^1 bound of the kernel. The case for $T^{(1)}_{\lambda l}$ is similar. □

[3] Here $\mathscr{F}^{-1}$ denotes Fourier inverse transform on $\mathbb{R}^2 \times \mathbb{R}^2$. See (5.3.6).

Proof of Proposition 5.3.1. Rewrite (5.2.3) as

$$\begin{aligned}\nabla q_{n+1} &\overset{\circ}{\approx} \underbrace{\Lambda f_{n+1}\nabla^{\perp} f_{n+1} + \nabla q_n}_{\text{Mismatch error}} + \underbrace{\Lambda f_{n+1}\nabla^{\perp} f_{\le n} + \Lambda f_{\le n}\nabla^{\perp} f_{n+1}}_{\text{Transport error}} \underbrace{-\nu\nabla\Lambda^{\gamma-1} f_{n+1}}_{\text{Dissipation error}} \\ &=: \nabla q_M + \nabla q_T + \nabla q_D.\end{aligned}$$

Frequency localization of q_{n+1} can be easily deduced from q_M, q_T, and q_D which are defined below. For convenience, we shall write $a_{j,n+1}$ as a_j in the computation below.

`Mismatch error.` By (5.2.6), we can further decompose the mismatch error as

$$\begin{aligned}\nabla q_M &\overset{\circ}{\approx} (\boxed{\text{main}} + \nabla q_n) + \boxed{\text{non-oscillatory error}} + \boxed{\text{oscillatory error}} \\ &\overset{\circ}{\approx} \nabla q_{M1} + \nabla q_{M2} + \nabla q_{M3}.\end{aligned}$$

We first estimate q_{M1}. To ease the notation we write $a_j^{\text{per}} = 2\sqrt{\frac{r_n}{\lambda_{n+1}}}\sqrt{C_0 + \mathscr{R}_j^o \frac{q_n}{r_n}}$ and $a_j = P_{\le \mu_{n+1}} a_j^{\text{per}}$. By using a fattened frequency projection $\tilde{P}_{\le \mu_{n+1}}$ which is frequency localized to $\{|k| \le 4\mu_{n+1}\}$, we have

$$\begin{aligned}&-\frac{1}{4}\cdot(5\lambda_{n+1})\cdot\sum_{j=1}^{2} l_j^{\perp}(l_j\cdot\nabla)a_j^2 + \nabla q_n - \nabla q_{M1} \\ &= -\frac{5}{4}\lambda_{n+1}\sum_{j=1}^{2} l_j^{\perp}(l_j\cdot\nabla)\tilde{P}_{\le \mu_{n+1}}((P_{\le \mu_{n+1}} a_j^{\text{per}})^2) + \nabla q_n - \nabla q_{M1} \\ &= -\frac{5}{4}\lambda_{n+1}\sum_{j=1}^{2} l_j^{\perp}(l_j\cdot\nabla)\tilde{P}_{\le \mu_{n+1}}\Big(-2a_j^{\text{per}} P_{>\mu_{n+1}} a_j^{\text{per}} + (P_{>\mu_{n+1}} a_j^{\text{per}})^2\Big) - \nabla q_{M1} \overset{\circ}{\approx} 0.\end{aligned}$$

Thus we can solve $q_{M1} \in C_0^\infty(\mathbb{T}^2)$ as

$$q_{M1} = -\frac{5}{4}\lambda_{n+1}\sum_{j=1}^{2}\Delta^{-1}\nabla\cdot\left(l_j^\perp(l_j\cdot\nabla)\tilde{P}_{\leq\mu_{n+1}}\left(-2a_j^{\text{per}}P_{>\mu_{n+1}}a_j^{\text{per}}+(P_{>\mu_{n+1}}a_j^{\text{per}})^2\right)\right). \tag{5.3.7}$$

Note that q_{M1} is frequency localized to $\{|k| \leq 4\mu_{n+1}\}$. By Lemma 5.3.2, we obtain

$$\|q_{M1}\|_X \lesssim \log\mu_{n+1}\cdot\lambda_{n+1}\sum_{j=1}^{2}\|a_j^{\text{per}}\|_\infty\|P_{>\mu_{n+1}}a_j^{\text{per}}\|_\infty \lesssim \log\mu_{n+1}\cdot(\mu_{n+1}^{-1}\lambda_n)^2 r_n. \tag{5.3.8}$$

Note that both $\boxed{\text{non-oscillatory error}}$ and $\boxed{\text{oscillatory error}}$ have zero means, so we define

$$q_{M2} = \Delta^{-1}\nabla\cdot\boxed{\text{non-oscillatory error}}, \quad q_{M3} = \Delta^{-1}\nabla\cdot\boxed{\text{oscillatory error}}$$

in $C_0^\infty(\mathbb{T}^2)$. To estimate q_{M2}, we claim that

$$\|\Delta^{-1}\nabla\cdot((T_{n+1,j}^{(1)}a_j)\nabla^\perp a_j)\|_X + \|\Delta^{-1}\nabla\cdot(5\lambda_{n+1}(T_{n+1,j}^{(2)}a_j)a_jl_j^\perp)\|_X \lesssim r_n\lambda_{n+1}^{-2}\mu_{n+1}^2\log\mu_{n+1}.$$

This is because by Lemma 5.3.4 and 5.3.2 we have

$$\|\Delta^{-1}\nabla\cdot(5\lambda_{n+1}(T_{n+1,j}^{(2)}a_j)a_jl_j^\perp)\|_X \lesssim (\log\mu_{n+1})\lambda_{n+1}(\frac{\mu_{n+1}}{\lambda_{n+1}})^2\frac{r_n}{\lambda_{n+1}} \lesssim r_n(\frac{\mu_{n+1}}{\lambda_{n+1}})^2\log\mu_{n+1}.$$

The other term can be estimated similarly. Then, it leads to

$$\|q_{M2}\|_X \lesssim r_n \lambda_{n+1}^{-2} \mu_{n+1}^2 \log \mu_{n+1}. \tag{5.3.9}$$

Next we estimate q_{M3}. Denote $T^{(i)}_{n+1,j} = T^{(i)}_{5\lambda_{n+1}l_j}$ for $i, j = 1, 2$. By Lemma 5.3.3, we have

$$\|T^{(1)}_{n+1,j} a_j\|_\infty \lesssim \lambda_{n+1}^{-1} \mu_{n+1}^2 \sqrt{\frac{r_n}{\lambda_{n+1}}}, \quad \|T^{(2)}_{n+1,j} a_j\|_\infty \lesssim \lambda_{n+1}^{-2} \mu_{n+1}^3 \sqrt{\frac{r_n}{\lambda_{n+1}}}. \tag{5.3.10}$$

Since all terms in (oscillatory error) have the frequency localized to $\sim \lambda_{n+1}$ provided that $48\lambda_n \leq \lambda_{n+1}$, the estimate for q_{M3} easily follows from (5.3.10):

$$\begin{aligned}
&\|\Delta^{-1}\nabla\cdot(\mathrm{osc1})\|_X \\
&\lesssim \sum_{j=1}^{2} \|\Delta^{-1}\nabla\cdot(l_j\cdot\nabla a_j + T^{(2)}_{n+1,j}a_j)(\lambda_{n+1}a_j l_j^\perp \cos(2\lambda_{n+1}l_j\cdot x) + \nabla^\perp a_j \sin(2\lambda_{n+1}l_j\cdot x))\|_X \\
&\lesssim \sum_{j=1}^{2} \lambda_{n+1}\|\Delta^{-1}\nabla\cdot\left(a_j l_j\cdot\nabla a_j l_j^\perp \cos(2\lambda_{n+1}l_j\cdot x)\right)\|_X \\
&\quad + \|\Delta^{-1}\nabla\cdot\left(\nabla a_j\cdot l_j \nabla^\perp a_j \sin(2\lambda_{n+1}l_j\cdot x)\right)\|_X \\
&\quad + \lambda_{n+1}\|\Delta^{-1}\nabla\cdot\left(a_j T^{(2)}_{n+1,j}a_j l_j^\perp \cos(2\lambda_{n+1}l_j\cdot x)\right)\|_X \\
&\quad + \|\Delta^{-1}\nabla\cdot\left(T^{(2)}_{n+1,j}a_j \nabla^\perp a_j \sin(2\lambda_{n+1}l_j\cdot x)\right)\|_X \\
&\lesssim \sum_{j=1}^{2} \|\nabla a_j\|_X\|a_j\|_X + \lambda_{n+1}^{-1}\|\nabla a_j\|_X\|\nabla^\perp a_j\|_X \\
&\quad + \|T^{(2)}_{n+1,j}a_j\|_X\|a_j\|_X + \lambda_{n+1}^{-1}\|T^{(2)}_{n+1,j}a_j\|_X\|\nabla^\perp a_j\|_X \\
&\lesssim \left(\frac{\lambda_n}{\lambda_{n+1}}\right) r_n.
\end{aligned}$$

Similarly,

$$\|\Delta^{-1}\nabla\cdot(\text{osc2})\|_X \lesssim \sum_{j=1}^{2}\|T^{(1)}_{n+1,j}a_j\|_\infty(\|a_j\|_\infty+\lambda_{n+1}^{-1}\|\nabla^\perp a_j\|_\infty)\lesssim\left(\frac{\lambda_n}{\lambda_{n+1}}\right)r_n.$$

The estimates for (osc3)-(osc6) are similar (using $2/\sqrt{5}\le|l_1\pm l_2|\le 4/\sqrt{5}$) and therefore

$$\|q_{M3}\|_X\lesssim\left(\frac{\lambda_n}{\lambda_{n+1}}\right)r_n. \tag{5.3.11}$$

Combining (5.3.8), (5.3.9), and (5.3.11) and using $b>1$, $\beta<1$, we can find $\Lambda_M=\Lambda_M(\beta,b)$ such that for any $\lambda_0\ge\Lambda_M$, we get $q_M=q_{M1}+q_{M2}+q_{M3}\in C_0^\infty(\mathbb{T}^2)$ satisfying (see also Section 5.6)

$$\|q_M\|_X\le\frac{1}{3}r_{n+1}.$$

`Transport error`. Define

$$q_T=\Delta^{-1}\nabla\cdot(\Lambda f_{n+1}\nabla^\perp f_{\le n}+\Lambda f_{\le n}\nabla^\perp f_{n+1})\in C_0^\infty(\mathbb{T}^2).$$

Since $\Lambda f_{n+1}\nabla^\perp f_{\le n}+\Lambda f_{\le n}\nabla^\perp f_{n+1}$ is frequency-localized to $\sim\lambda_{n+1}$, using $\|f_{\le n}\|_{C^\alpha}\le 100$, we get

$$\|q_T\|_X\lesssim\|f_{n+1}\|_\infty(\|\nabla^\perp f_{\le n}\|_\infty+\|\Lambda f_{\le n}\|_\infty)\le C_\alpha\lambda_n^{1-\alpha}\sqrt{\frac{r_n}{\lambda_{n+1}}}\le\frac{1}{3}r_{n+1}$$

for some constant $C_\alpha>0$. We can find $\Lambda_T=\Lambda_T(\beta,b)$ such that for any $\lambda_0\ge\Lambda_T$ the last inequality holds since $b>1$ and $\beta<\frac{1}{5}$.

`Dissipation error.` We define $q_D = -\nu\Lambda^{\gamma-1}f_{n+1} \in C_0^\infty(\mathbb{T}^2)$ which satisfies

$$\|q_D\|_X \le C_2\lambda_{n+1}^{\gamma-1}\|f_{n+1}\|_\infty \le 5C_2\lambda_{n+1}^{\gamma-1}\sqrt{\frac{r_n}{\lambda_{n+1}}} \le \frac{1}{3}r_{n+1},$$

for some $C_2 = C_2(\nu,\gamma) > 0$. Since $\beta < 3-2\gamma$, we can find sufficiently small $b_0 = b_0(\nu,\gamma,\beta)$ such that for any $1 < b < b_0+1$ there exists $\Lambda_D = \Lambda_D(\nu,\gamma,\beta,b)$ which leads the last inequality for any $\lambda_0 \ge \Lambda_D$.

Collecting the estimates, we obtain $\|q_{n+1}\|_X \le r_{n+1}$ if $\lambda_0 > \Lambda_0$ where $\Lambda_0 = \max(\Lambda_M,\Lambda_T,\Lambda_D)$. □

5.4 Proof of Theorem 5.1.2

Proof of Theorem 5.1.2. WLOG we take $C_0 = 2$ in Proposition 5.2.3. Fix $\nu \ge 0$, $0 < \gamma < \frac{3}{2}$ and choose parameters as in (5.1.4). Choose b and λ_0 as in Proposition 5.3.1. If necessary, we choose larger λ_0 to have $\sum_{m=0}^\infty \lambda_m^{\alpha-\frac{1}{2}-\frac{\beta}{2b}} \le 1$. Take the base step $(f_{\le 0},q_0) = (0,0)$. At n^{th}-step, assume that $(f_{\le n},q_n) \in C_0^\infty(\mathbb{T}^2)\times C_0^\infty(\mathbb{T}^2)$ satisfies

- $(f_{\le n},q_n)$ solves (5.2.1).
- $\text{supp}(\widehat{f_{\le n}}) \subset \{|k| \le 6\lambda_n\}$, $\text{supp}(\widehat{q_n}) \subset \{|k| \le 12\lambda_n\}$ and $\|q_n\|_X \le r_n$,

$$\|f_{\le n}\|_{C^\alpha(\mathbb{T}^2)} \le 50\sum_{m=1}^{n}\lambda_m^\alpha\sqrt{\frac{r_{m-1}}{\lambda_m}} \le 100\sum_{m=0}^{n-1}\lambda_{m+1}^{\alpha-\frac{1}{2}-\frac{\beta}{2b}} \le 100.$$

Then by Proposition 5.3.1 and (5.2.10), at $(n+1)^{\text{th}}$step, we find f_{n+1} and $q_{n+1} \in C_0^\infty(\mathbb{T}^2)$ satisfying

- (f_{n+1}, q_{n+1}) solves (5.2.3).

- $\operatorname{supp}(\widehat{f_{\le n+1}}) \subset \{|k| \le 6\lambda_{n+1}\}$, $\|f_{n+1}\|_{C^\alpha(\mathbb{T}^2)} \le 50\lambda_{n+1}^{\alpha}\sqrt{\frac{r_n}{\lambda_{n+1}}}$, $\operatorname{supp}(\widehat{q_{n+1}}) \subset \{|k| \le 12\lambda_{n+1}\}$, and $\|q_{n+1}\|_X \le r_{n+1}$.

Thus the induction step can be closed and it remains to show that $f_{\le n}$ converges to the desired weak solution. We first check its regularity. Clearly

$$\|f_{\le n'} - f_{\le n}\|_{C^\alpha} \lesssim \sum_{m=n}^{n'-1} \lambda_{m+1}^{\alpha-\frac{1}{2}-\frac{\beta}{2b}}, \quad \forall\, n' \ge n.$$

Thus $f_{\le n} \to f \in C^\alpha(\mathbb{T}^2)$. Now denote $\theta_n = \Lambda f_{\le n}$ and $\theta = \Lambda f$. Clearly

$$\langle \theta_n \Lambda^{-1}\nabla^\perp\theta_n - \nu\Lambda^{\gamma-2}\nabla\theta_{n+1} - \nabla q_{n+1}, \nabla\psi\rangle = 0, \quad \forall\, \psi \in C^\infty(\mathbb{T}^2).$$

We then rewrite the above as

$$\frac{1}{2}\langle \Lambda^{-\frac{1}{2}}\theta_n, \Lambda^{\frac{1}{2}}[\mathscr{R}^\perp, \nabla\psi]\theta_n\rangle + \nu\langle \Lambda^{-\frac{1}{2}}\theta_n, \Lambda^{\gamma+\frac{1}{2}}\psi\rangle + \langle q_n, \Delta\psi\rangle = 0, \quad \forall\, \psi \in C^\infty(\mathbb{T}^2).$$

Since $\Lambda^{-\frac{1}{2}}\theta_n \to \Lambda^{-\frac{1}{2}}\theta$ strongly in L^∞, Proposition 5.5.1 implies that θ solves (SQG). □

Finally we remark that our solution $\theta = \Lambda f$ has an almost explicit form. By using (5.2.10), we have

$$f = \sum_{n=0}^{\infty}\sum_{j=1}^{2} 2\sqrt{\frac{r_n}{5\lambda_{n+1}}}\left(P_{\le\mu_{n+1}}\sqrt{C_0 + R_j^o\frac{q_n}{r_n}}\right)\cos(5\lambda_{n+1} l_j \cdot x).$$

The leading term is an almost explicit Fourier series (one can take C_0 large) and thus our solution is nontrivial.

5.5 Proof of Theorem 5.1.4

In this section, we prove Theorem 5.1.4 based on the following proposition.

Proposition 5.5.1. *Let $\mathscr{R} = \mathscr{R}_j$, $j = 1,2$. Assume $\phi \in H^3$ and $\theta \in \dot{H}^{-\frac{1}{2}}$ ($\overline{\theta} = 0$). Then we have*

$$\|[\mathscr{R},\phi]\theta\|_{\dot{H}^{\frac{1}{2}}} \lesssim \|\phi\|_{\dot{H}^3}\|\theta\|_{\dot{H}^{-\frac{1}{2}}}.$$

Proof. Denote $m(k) = \frac{k_1}{|k|}$. It suffices to show that

$$\|\sum_{k'\neq 0,k} |k|^{\frac{1}{2}}(m(k)-m(k'))\widehat{\phi}(k-k')\widehat{\theta}(k')\|_{l^2_k} \lesssim \||k|^3\widehat{\phi}(k)\|_{l^2_k}\||k|^{-\frac{1}{2}}\widehat{\theta}(k)\|_{l^2_k}. \quad (5.5.1)$$

If $|k'| \lesssim |k-k'|$, then $|k| \lesssim |k-k'|$, and

$$\text{LHS of } (5.5.1) \lesssim \|\sum_{k'\neq 0,k} |k-k'||\widehat{\phi}(k-k')|\cdot|k'|^{-\frac{1}{2}}|\widehat{\theta}(k')|\|_{l^2_k} \lesssim \text{RHS of } (5.5.1).$$

If $|k-k'| \ll |k|$, then $|k| \sim |k'|$, and it suffices to use $|m(k)-m(k')| \lesssim |k-k'|(|k'|+|k|)^{-1}$. □

Proof of Theorem 5.1.4. The point is to use the weak formulation (below $\langle , \rangle$ denotes L^2-inner product in (t,x), and ψ is a time-dependent test function)

$$\langle \partial_t\theta_n, \psi\rangle + \frac{1}{2}\langle \Lambda^{-\frac{1}{2}}\theta_n, \Lambda^{\frac{1}{2}}[\mathscr{R}^{\perp},\nabla\psi]\theta_n\rangle + \nu\langle \Lambda^{-\frac{1}{2}}\theta_n, \Lambda^{\gamma+\frac{1}{2}}\psi\rangle = 0.$$

By using the above together with Proposition 5.5.1, we have[4] $\|\partial_t\theta_n\|_{L^1_t\dot{H}^{-8}} \lesssim 1$.

[4] Here t belongs to an arbitrary compact interval.

Fix any $0 \neq k \in \mathbb{Z}^2$. We have $\|\partial_t \widehat{\theta_n}(k,t)\|_{L_t^1} \lesssim |k|^8$ and $\|\widehat{\theta_n}(k,t)\|_{L_t^2} \lesssim |k|^{-s}$. By further using a diagonal argument, we obtain along a subsequence

$$\|\widehat{\theta_{n_l}}(k,t) - \widehat{f}(k,t)\|_{L_t^2} \to 0 \quad \text{for any fixed } k. \tag{5.5.2}$$

Using $\sup_l \|\theta_{n_l}\|_{L_t^2 \dot{H}^s} \lesssim 1$ (note that $s > -\frac{1}{2}$), we have for any integer J (below $P_{>J}$ denotes frequency projection to the regime $|k| \geq 2^J$)

$$\begin{aligned} \|P_{>J}(\theta_{n_l} - f)\|_{L_t^2 \dot{H}^{-\frac{1}{2}}} &\lesssim 2^{-J(s+\frac{1}{2})} \|\theta_{n_l} - f\|_{L_t^2 \dot{H}^s} && (5.5.3) \\ &\lesssim 2^{-J(s+\frac{1}{2})}. && (5.5.4) \end{aligned}$$

By (5.5.2) and (5.5.4), one obtains the strong convergence $\theta_{n_l} \to f$ in $L_t^2 \dot{H}^{-\frac{1}{2}}$. Since $\|\Lambda^{\frac{1}{2}}[\mathscr{R}^\perp, \nabla\psi](\theta_n - f)\|_2 \lesssim \|\theta_n - f\|_{\dot{H}^{-\frac{1}{2}}}$, it follows that f is the desired weak solution. □

5.6 Bookkeeping of various parameters

In this section we sketch how the choice of various parameters in (5.1.4) take effect on various error terms and the regularity of the weak solution. Recall that (observe from below $\log \mu_{n+1} \sim \log \lambda_n$)

$$\lambda_n = \left\lceil \lambda_0^{b^n} \right\rceil, \quad r_n = \lambda_n^{-\beta}, \quad \mu_{n+1} = (\lambda_n \lambda_{n+1})^{\frac{1}{2}}, \quad \alpha = \frac{1}{2} + \frac{\beta}{2b} - \epsilon_0 > \frac{1}{2}.$$

Mismatch error $\quad r_n \frac{\lambda_n}{\lambda_{n+1}} \log \lambda_n \ll r_{n+1} \iff \lambda_n^{(b-1)(\beta-1)} \log \lambda_n \ll 1.$

Transport error $\quad \lambda_n^{1-\alpha} \sqrt{\frac{r_n}{\lambda_{n+1}}} \ll r_{n+1} \iff \lambda_n^{1-\alpha-\frac{1}{2}\beta-\frac{1}{2}b+b\beta} \ll 1.$

Dissipation error $\quad \lambda_{n+1}^{\gamma-1} \sqrt{\frac{r_n}{\lambda_{n+1}}} \ll r_{n+1} \iff \lambda_{n+1}^{\gamma-\frac{3}{2}+\beta-\frac{\beta}{2b}} \ll 1.$

C^α`-regularity` $\lambda_{n+1}^{\alpha}\sqrt{\frac{r_n}{\lambda_{n+1}}} \ll 1 \iff \lambda_{n+1}^{\alpha-\frac{1}{2}-\frac{1}{2b}\beta} \ll 1.$

Now one can take $\alpha = \frac{1}{2} + \frac{\beta}{2b}$ to do a limiting computation. From the transport error we obtain (the limiting condition)

$$1-\alpha-\frac{1}{2}\beta-\frac{1}{2}b+b\beta = \frac{1-b}{2b}(b-\beta(2b+1)) \Rightarrow \beta < \frac{1}{3}.$$

From the dissipation error we obtain $\frac{\beta}{2} < \frac{3}{2} - \gamma$.

Chapter 6

Conclusion & Future Work

Throughout this thesis, we discussed three different topics.

`1. The AC & CH dynamics.`

In Chapter 3, we have identified the time step scaling for several first and second order schemes for AC and CH under the restriction of fixed local truncation error, σ. In particular, we derive the asymptotic behavior of time-step number with σ and asymptotic parameter ϵ during meta-stable dynamics. These predictions are made under the assumption that the time steps preserve the asymptotic structure of the diffuse interface, a concept we refer to as *profile fidelity*. The predictions are verified in numerical experiments. We see that methods whose dominant local truncation error can be expressed as a pure time derivative have optimal asymptotic performance in this particular limit. BE, TR, and BDF2 all have this desirable property. We believe these methods will also have superior performance for other problems with metastable dynamics. Our numerical results show that BE performs better than expected and we have shown an explanation of this behaviour with formal asymptotics.

The optimal fully implicit methods asymptotically computationally outperform all linearly implicit methods in the limit we consider. We present precise criteria on the computational cost of nonlinear solvers for this comparison. The provably energy stable first and second order SAV schemes had higher computational cost than standard IMEX methods for similar results. As a final result, we present a rigorous proof that large time steps with fully implicit BE can be taken with locally unique solutions that are energy stable. This is done for the 2D radial AC equation in meta-stable dynamics. Eyre-type iteration is also considered in this analytic framework, and it is shown that in general this approach loses profile stability unless very small time steps are taken.

Extending the analysis to the non-radial case and to CH is an interesting question. We observed that the question of global accuracy is not trivial in Section 3.5 and should be considered for other schemes. Accurate local error estimation for these problems is another interesting question to pursue.

2. `The OD problem.`

In Chapter 4 we summarize the ways the Oxygen Depletion problem has been considered in the literature: with interfaces to be tracked, captured, or found as a limit of regularized problems. We fill in a gap in the list of formulations, showing that the OD problem can be considered as a gradient flow with constraint. A new numerical capturing method based on the gradient flow formulation is proposed and a convergence proof given. The equivalence of all formulations is shown. A biharmonic implicit free boundary value problem and a class of vector problems are introduced.

More questions can be asked. Firstly, the regularity of boundary point positions in 1D (Conjecture 4.2.1) and higher dimensions can be studied with help of geometric analysis tools. Secondly analysis of the mapped domain formulation discussed in Remark 10 would possibly extend to a convergence proof of its numerical approximation (Section 4.3.1). We are also interested in an understanding of the general class of vector problems in Section 4.5.2 and an understanding of the limiting behaviour of solutions to the OD problem as discussed in Remark 17. The numerical approach in [63] may be useful to gain insight into the 1D case. Studies of the biharmonic obstacle problem discussed in Remark 18 and the general class of vector problems introduced in Section 4.5.2 can be pursued.

3. The SQG system.

In Chapter 5, we developed a new framework of the convex integration scheme and applied it to the stationary SQG system on the two dimensional periodic torus. We constructed nontrivial stationary weak solutions and therefore showed the uniqueness.

Extending our analysis to the time-dependent case and other fluid models such as the Navier–Stokes equations and the Euler equations is interesting and our plane wave ansatz can be applicable.

Bibliography

[1] N. Alikakos, P. Bates, and X. Chen. Convergence of the cahn-hilliard equation to the hele-shaw model. *Archive for Rational Mechanics and Analysis*, 128(2): 165–205, 6 1994.

[2] S. M. Allen and J. W. Cahn. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Metallurgica*, 27(6): 1085 – 1095, 1979.

[3] U. M. Ascher and L.R. Petzold, *Computer methods for ordinary differential equations and differential-algebraic equations.* Society for Industrial and Applied Mathematics, USA. 1st edition, 1998.

[4] U. M. Ascher, S. J. Ruuth, and B. T. R. Wetton. Implicit-explicit methods for time-dependent partial differential equations. *SIAM Journal on Numerical Analysis*, 32(3): 797–823, 1995.

[5] J. W. Barrett, J. F. Blowey and H. Garcke. Finite element approximation of the Cahn-Hilliard equation with degenerate mobility, *SIAM Journal on Numerical Analysis.*, 37(1): 286–318, 2000.

[6] A.E. Berger, M. Ciment, and J.C.W. Rogers. Numerical solution of a

diffusion consumption problem with a free boundary. *SIAM Journal on Numerical Analysis.*, 12(4): 646–672, 1975.

[7] J. F. Blowey, M. I. M. Copetti and C. M. Elliott. Numerical analysis of a model for phase separation of a multicomponent alloy, *IMA Journal of Numerical Analysis*. 16(1): 111–139, 1996.

[8] H. Brezis and D. Kinderlehrer. The smoothness of solutions to nonlinear variational inequalities. *Indiana University Mathematics Journal.*, 23(9): 831–844, 1974.

[9] L. J. Bridge and B. Wetton. A mixture formulation for numerical capturing of a two-phase/vapour interface in a porous medium. *Journal of Computational Physics*, 225: 2043–2068, 2007.

[10] T. Buckmaster, S. Shkoller and V. Vicol. Nonuniqueness of weak solutions to the SQG equation. *Comm. Pure Appl. Math.*, 72(9): 1809–1874, 2019.

[11] L. A. Caffarelli. The obstacle problem revisited. *Journal of Fourier Analysis and Applications.*, 4(4): 383–402, 1998.

[12] J. W. Cahn and J. E. Hilliard. Free energy of a nonuniform system. i. interfacial free energy. *The Journal of Chemical Physics*, 28(2): 258–267, 1958.

[13] L. Carleson. On convergence and growth of partial sums of Fourier series., *Acta Math.*, 116(1): 135–157, 1966.

[14] L. Chen, X. Hu and S. Wise. Convergence analysis of the Fast Subspace Descent method for convex optimization problems, *Mathematics of Computation.*, 89: 2249–2282, 2020.

[15] W. Chen, Y. Liu, C. Wang and S. Wise. Convergence analysis of a fully discrete finite difference scheme for the Cahn-Hilliard-Hele-Shaw equation, *Mathematics of Computation.*, 85: 2231–2257, 2016.

[16] X. Cheng. *On the stability of a semi-implicit scheme of Cahn-Hilliard type equations*. Master thesis, University of British Columbia, 2017.

[17] X. Cheng, Z. Fu and B. Wetton. Equivalent formulations of the oxygen depletion problem, other implicit free boundary value problems, and implications for numerical approximation *arXiv:2105.03538*, 2021.

[18] X. Cheng, H. Kwon and D. Li. Non-uniqueness of steady-state weak solutions to the surface quasi-geostrophic equations. *arXiv:2007.09591*, 2020.

[19] X. Cheng, D. Li, K. Promsilow and B. Wetton. Asymptotic Behaviour of Time Stepping Methods for Phase Field Models. Journal of Scientific Computing, 86(3): 1–34, 2021.

[20] A. Choffrut and L. Székelyhidi. Weak Solutions to the Stationary Incompressible Euler Equations, *SIAM J. Math. Anal.*, 46(6): 4060–4074, 2014.

[21] A. Christlieb, J. Jones, K. Promislow, B. Wetton, and M. Willoughby.

High accuracy solutions to energy gradient flows from material science models. *Journal of Computational Physics*, 257: 193 – 215, 2014.

[22] A. Christlieb, K. Promislow and Z. Xu. On the unconditionally gradient stable scheme for the Cahn-Hilliard equation and its implementation with Fourier method, *Communications in Mathematical Sciences.*, 11(2): 345–360, 2013.

[23] M. Colombo, L. De Rosa and M. Sorella. Typicality results for weak solutions of the incompressible Navier-Stokes equations. *arXiv:2102.03244.*

[24] A. Córdoba, D. Córdoba and F. Gancedo. Uniqueness for SQG patch solutions, *Trans. Amer. Math. Soc. Ser.*, B, 5, 1–31, 2018.

[25] J. Crank, *Free and moving boundary problems.*, Oxford University Press, Walton Street, Oxford, 1984.

[26] D'Alembert. Recherches sur la courbe que forme une corde tenduë mise en vibration (Researches on the curve that a tense cord forms [when] set into vibration),*Histoire de l'académie royale des sciences et belles lettres de Berlin*, 3: 214—219, 1747.

[27] G. Dal Maso. *An introduction to* Γ*-convergence*. Progress in Nonlinear Differential Equations and their Applications, 8. Birkhäuser Boston, Inc., Boston, MA, 1993.

[28] C. De Lellis and L. Székelyhidi Jr. The h-principle and the equations of fluid dynamics. *Bull. Amer. Math. Soc.*, 49(3): 347–375, 2012.

[29] C. De Lellis and L. Székelyhidi Jr. Dissipative continuous Euler flows. *Invent. math.*, 193(2): 377–407, 2013.

[30] C. De Lellis and L. Székelyhidi Jr. On turbulence and geometry: from Nash to Onsager. *Notices AMS*, 66(5): 677–685, 2019.

[31] A.E. Diegel, C. Wang and S. Wise. Stability and convergence of a second-order mixed finite element method for the Cahn–Hilliard equation, *IMA Journal of Numerical Analysis.*, 36(4): 1867–1897, 2015.

[32] A. Einstein, *Relativity: the special and general theory.* New York: Crown, 1961.

[33] C. M. Elliott and A. M. Stuart. The global dynamics of discrete semilinear parabolic equations. *SIAM Journal on Numerical Analysis*, 30(6): 1622–1663, 1993.

[34] L. C. Evans. *Partial differential equations: second edition* ., American Mathematical Society Providence, Rhode Island, 2010.

[35] D. J. Eyre. Unconditionally gradient stable time marching the cahn-hilliard equation. *MRS Proceedings*, 529:39, 1998.

[36] A. Fasano and M. Primicerio. New results on some classical parabolic free-boundary problems. *Quarterly of Applied Mathematics.*, 38(4): 439–460, 1981.

[37] A. Figalli, X. Ros-Oton, and J. Serra. The singular set in the stefan problem. *arXiv*, 2103.13379, 2021.

[38] W. Feng, A.J. Salgado, C. Wang and S. Wise. Preconditioned steepest descent methods for some nonlinear elliptic equations involving p-Laplacian terms. *Journal of Computational Physics.*, 334: 45–67, 2017.

[39] X. Feng and A. Prohl. Numerical analysis of the Cahn-Hilliard equation and approximation for the Hele-Shaw problem, *Interfaces and Free Boundaries.*, 7: 1–28, 2005.

[40] M. Focardi, M.S. Gelli, and E. Spadaro. Monotonicity formulas for obstacle problems with Lipschitz coefficients. *Calculus of Variations and Partial Differential Equations.*, 54(2): 1547–1573, 2015.

[41] U. Frisch and A.N. Kolmogorov. *Turbulence: the legacy of AN Kolmogorov*. Cambridge university press, 1995.

[42] R. Glowinski and A. Wachs. On the numerical simulation of viscoplastic fluid flow. In R. Glowinski and J. Xu, editors, *Numerical Methods for Non-Newtonian Fluids*, volume 16 of *Handbook of Numerical Analysis*, pages 483–717. Elsevier, 2011.

[43] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*. Springer, Berlin, 1996.

[44] I. M. Held, R. T. Pierrehumbert, S. T. Garner and K. L. Swanson. Surface quasi-geostrophic dynamics. *J. Fluid Mech.*, 282:1-20, 1995.

[45] H. Huang, P. Lin and W. Zhou. Moisture transport and diffusive instability during bread baking.,*SIAM Journal on Applied Mathematics.* 68(1): 222-238, 2007.

[46] T. Ilmanen. Convergence of the Allen-Cahn equation to Brakke's motion by mean curvature., *Journal of Differential Geometry.*, 38: 417–461, 1993.

[47] P. Isett and A. Ma. A direct approach to nonuniqueness and failure of compactness for the SQG equation. *Nonlinearity.*, 34(5), 3122–3162, 2021.

[48] P. Isett and V. Vicol. Hölder continuous solutions of active scalar equations. *Ann. PDE.*, 1(1): 1-77, 2015.

[49] K. Ito and K. Kunisch, Parabolic variational inequalities: The Lagrange multiplier approach., *Journal de Mathématiques Pures et Appliquées.*, 85(3): 415–449, 2006.

[50] K. Ito and K. Kunisch, An augmented Lagrangian technique for variational inequalities., *Applied Mathematics and Optimization.*, 21: 223–241, 1990.

[51] J. Kačur, *Method of Rothe in evolution equations.*, Teubner Verlagsgesellschaft, Leipzig, 1985.

[52] T. Kapitula and K. Promislow. *Spectral and dynamical stability of nonlinear waves*, volume 185 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2013.

[53] T. Kärkkäinen, K. Kunisch and P. Tarvainen, Augmented lagrangian active set methods for obstacle problems., *Journal of Optimization Theory and Applications.* 119(3): 499–533, 2003.

[54] D. Li and Z. Qiao. On second order semi-implicit fourier spectral methods for 2d cahn–hilliard equations. *Journal of Scientific Computing*, 70(1): 301–341, Jan 2017.

[55] D. Li and Z. Qiao. On the stabilization size of semi-implicit fourier-spectral methods for 3d cahn hilliard equations. *Communications in Mathematical Sciences*, 15: 1489–1506, 2017.

[56] E. Lindgren and R. Monneau. Pointwise regularity of the free boundary for the parabolic obstacle problem. *Calculus of Variations and Partial Differential Equations.*, 54(1): 299–347, 2015.

[57] J.-L. Lions and G. Stampacchia. Variational inequalities. *Comm. Pure Appl. Math.*, 20: 439–519, 1967.

[58] X. Luo, Stationary solutions and nonuniqueness of weak Solutions for the Navier–Stokes equations in high dimensions. *Arch Rational Mech Anal* 233, 701–747, 2019.

[59] E. Magenes. Topics in parabolic equations: some typical free boundary problems. *Boundary Value Problems for Linear Evolution Partial Differential Equations.*, 29: 239–312, 1977.

[60] F. Marchand. Existence and regularity of weak solutions to the quasi-geostrophic equations in the spaces L^p or $\dot{H}^{-\frac{1}{2}}$. *Comm. Math. Phys.*, 277(1): 45–67, 2008.

[61] J. C. Maxwell. *A Treatise on electricity and magnetism*. Dover, 1954.

[62] F. Miranda, J.F. Rodrigues, and L. Santos. Evolutionary quasi-variational and variational inequalities with constraints on the derivatives. *Advances in Nonlinear Analysis.*, 9(1): 250–277, 2020.

[63] S. L. Mitchell and M. Vynnycky. The oxygen diffusion problem: Analysis and numerical solution. *Applied Mathematical Modelling.*, 39(9): 2763–2776, 2015.

[64] L. Modica and S. Mortola. Un esempio di γ-convergenza. *Boll. Un. Mat. Ital.*, 14(5): 285–299, 1977.

[65] R. Monneau. On the number of singularities for the obstacle problem in two dimensions. *The Journal of Geometric Analysis.*, 13(2): 359–389, 2003.

[66] W. W. Mullins and R. F. Sekerka. Morphological stability of a particle growing by diffusion or heat flow. *Journal of Applied Physics*, 34(2): 323–329, 1963.

[67] L. Onsager. Statistical hydrodynamics. *Nuovo Cim* 6: 279–287, 1949.

[68] J. Pedlosky. *Geophysical fluid dynamics*. Springer, New York, 1982.

[69] R. L. Pego. Front migration in the nonlinear cahn-hilliard equation. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 422(1863): 261–278, 1989.

[70] S.G. Resnick. *Dynamical problems in non-linear advective partial differential equations.* PhD thesis, U of Chicago, 1995.

[71] A. D. Pierce. *Acoustics: an introduction to its physical principles and applications.* Springer, Cham, 2019.

[72] L. I. Rubinstein. *The Stefan problem.* Translations of. Math. Monog. Vol. 27, American Math. Society, Providence R.I., U.S.A., 1971.

[73] M. Rudd and K. Schmitt. Variational inequalities of elliptic and parabolic Type. *Taiwanese Journal of Mathematics.*, 6(3): 287–322, 2002.

[74] O. Savin. Phase transitions, minimal surfaces and a conjecture of de giorgi. *Current Developments in Mathematics 2009*, 101(3): 59 – 113, 2010.

[75] D G. Schaeffer. A new proof of the infinite differentiability of the free boundary in the Stefan problem . *Journal of Differential Equations*, 20: 266–269, 1976.

[76] J. Shen, J. Xu, and J. Yang. The scalar auxiliary variable (sav) approach for gradient flows. *Journal of Computational Physics*, 353: 407 – 416, 2018.

[77] R. Shvydkoy. Convex integration for a class of active scalar equations. *J. Amer. Math. Soc.*, 24(4): 1159–1174, 2011

[78] H. Singh and J.A. Hanna. Pick-up and impact of flexible bodies. *Journal of the Mechanics and Physics of Solids*, 106:46–59, 2017.

[79] J. Stefan. Ueber die theorie der eisbildung, insbesondere über die eisbildung im polarmeere. *Annalen der Physik.*, 278(2): 269–286, 1890.

[80] E. M. Stein. *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals,* volume 43 of Princeton Mathematical Series. Princeton University Press, Princeton, NJ, 1993.

[81] M. Struwe. *Variational methods*, volume 34 of *A Series of Modern Surveys in Mathematics*. Springer-Verlag Berlin Heidelberg, 2008.

[82] R. K. Sundaram. *A first course in optimization theory*. Cambridge University Press, 1996.

[83] T. Tao. Conserved quantities for the surface quasi-geostrophic equation (Wordpress Blog), 2014. https://terrytao.wordpress.com/2014/03/06/conserved-quantities-for-the-surface-quasigeostrophic-equation/

[84] J. W Thomas. *Numerical partial differential equations: finite difference methods*, volume 22 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 1995.

[85] B. P. Vollmayr-Lee and A. D. Rutenberg. Fast and accurate coarsening simulation with an unconditionally stable time step. *Phys. Rev. E*, 68:066703, Dec 2003.

[86] C. Vuik. Some historical notes about the stefan problem. *Nieuw Archief voor Wiskunde 4e serie*, 11: 157–167, 1993.

[87] G. S. Weiss. A homogeneity improvement approach to the obstacle problem. *Inventiones Mathematicae.*, 138(1): 23–50, 1999.

[88] B. Wetton. 2D Allen Cahn Simulation (YouTube Video). https://youtu.be/W7oNaJQ4_kc, December 2018.

[89] B. Wetton. 2D periodic Cahn Hilliard Simulation (YouTube Video). https://youtu.be/gI-S7MfWN5I, March 2018.

[90] S. Wise. Unconditionally stable finite difference, nonlinear multigrid simulation of the Cahn-Hilliard-Hele-Shaw system of equations. *J. Sci. Comput.*, 44: 38–68, 2010.

[91] J. Xu, Y. Li, S. Wu, and A. Bousquet. On the stability and accuracy of partially and fully implicit schemes for phase field modeling. *Computer Methods in Applied Mechanics and Engineering*, 345: 826–853, 2019.

[92] Y. Yan, W. Chen, C. Wang and S. Wise. A second-order energy stable BDF numerical scheme for the Cahn-Hilliard equation, *Communications in Computational Physics.*, 23: 572–602, 2018.

[93] P. Yue, J. J. Feng, C. Liu, and J. Shen. A diffuse-interface method for simulating two-phase flows of complex fluids. *Journal of Fluid Mechanics*, 515: 293-–317, 2004.

[94] [ogbash]. *Simulation of 2D waves using implicit scheme of Finite Difference Method.*, [video File] (2009)., Retrieved June 19th, 2017, from https://www.youtube.com/watch?v=QKSUaRbgpSM.