UNIVERSITY OF CALIFORNIA, SAN DIEGO

A Development of Discretization Techniques for Some Elliptic and Hyperbolic PDE

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics with Specialization in Computational Science

by

Jonathan W. Serencsa

Committee in charge:

Professor Michael Holst, Chair
Professor Steve Shkoller, Co-Chair
Professor Randolph Bank
Professor Thomas Bewley
Professor Julius Kuti

2012

The dissertation of Jonathan W. Serencsa is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2012

DEDICATION

This thesis is dedicated to my grandfather, Joseph J. Zucca.

EPIGRAPH

Chance favors the prepared mind

—Louis Pasteur

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures viii

Acknowledgements ix

Vita and Publications x

Abstract xi

Chapter 1 Error Estimates for Positive Discrete Solutions to the Diffusive Logistic Equation 1
1.1 Introduction 1
1.2 Analysis of the Continuous Problem 5
1.3 Analysis of the Discrete Problem 10
1.4 A Critical Technical Lemma 13
1.5 Galerkin Error Estimates 16
1.5.1 H^1 Estimates 16
1.5.2 L^2 Estimates 18
1.5.3 L^∞ Estimates 20
1.6 Numerical Results 21
1.7 Conclusion 25

Chapter 2 A Space-Time Smooth Artificial Viscosity Method for Nonlinear Conservation Laws 27
2.1 Introduction 27
2.1.1 Smoothing conservation laws 27
2.1.2 Numerical discretization 28
2.1.3 Outline of the paper 32
2.2 The C-method 33
2.2.1 Compressible Euler equations 34
2.2.2 Classical artificial viscosity 36
2.2.3 C-method for compressible Euler 37
2.2.4 Regularization of initial data for use with FEM-C . . 38
2.2.5 A compressive modification of the forcing G in the C-equation 40

2.2.6 Moving to the discrete level 40
2.2.7 The C-method under a Galilean-transformation . . . 41
2.2.8 Regularization through the C-equation 43
2.2.9 Convergence of the C-method in the limit of zero mesh size 45
2.2.10 The C-equation as a gradient flow 49
2.3 Numerical Schemes 50
2.3.1 Notation for discrete solutions 50
2.3.2 FEM-C and FEM-$|u_x|$: A Second-Order Continuous-Galerkin Finite-Element Scheme 51
2.3.3 WENO-C: A Simple WENO scheme using the C-method 52
2.3.4 NT: A Second-order Central-Differencing scheme of Nessayhu-Tadmor 54
2.3.5 WENO-G: WENO with Godunov-based upwinding . 55
2.4 Sod shock-tube problem 55
2.5 Osher-Shu shock-tube problem 57
2.6 Woodward-Colella Blast Wave 59
2.7 Leblanc shock-tube problem 62
2.7.1 Strategy One: A C equation for the energy density . 62
2.7.2 Strategy Two: a new type of viscosity for the energy density 64
2.8 Concluding Remarks 67

Bibliography 69

LIST OF FIGURES

Figure 1.1: The Square domain $\Omega = [0,1] \times [0,1]$ 22
Figure 1.2: L^2, H^1, and L^∞-error plots for FEM solutions to the DLE posed on the square with $b = 22$,$c = 15$ 23
Figure 1.3: The L-shaped domain $\Omega = [0,1] \times [0,1] \setminus [\frac{1}{2},1] \times [\frac{1}{2},1]$ 24
Figure 1.4: L^2, H^1, and L^∞-error plots for FEM solutions to the DLE posed on the L-shaped domain with $b = 22$,$c = 15$ 24

Figure 2.1: A comparison of the artificial viscosity profile produced by the C-method and the classical Richtmyer-type approach for the Sod shock tube at $t = 0.2$. 43
Figure 2.2: Application of FEM-C to a very slowly moving shock 44
Figure 2.3: Comparison of FEM-C and FEM-$|u_x|$, for the Sod shock-tube experiment with $N = 100$, $t = 0.2$. $\beta = 0.5$ for both FEM-C and FEM-$|u_x|$. 55
Figure 2.4: Comparison of FEM-C and FEM-$|u_x|$, for the Sod shock-tube experiment with $N = 100$, $t = 0.2$. $\beta = 0.5$ for FEM-C and $\beta = 3.0$ for FEM-$|u_x|$. 56
Figure 2.5: Comparisons of FEM-C against NT and WENO schemes, for the Sod shock-tube experiment with $N = 100$ and $t = 0.2$. 57
Figure 2.6: Comparisons of FEM-C against NT and WENO-G schemes, for the Osher-Shu shock-tube experiment with $N = 200$ and $t = 0.36$. . 58
Figure 2.7: Comparisons of WENO-C with WENO-G and our WENO scheme with artificial viscosity deactivated, for the Osher-Shu shock-tube experiment with $N = 200$ and $t = 0.36$. 59
Figure 2.8: Comparisons of FEM-C against NT and WENO-G schemes, for the Woodward-Colella blast-tube experiment with $N = 400$ and $T = 0.038$. 60
Figure 2.9: WENO with and without stabilization applied to the Woodward-Colella blast-tube experiment with $N = 400$ and $t = 0.038$. 61
Figure 2.10: Comparison of WENO-C against WENO-G, for the Woodward-Colella blast-tube experiment with $N = 400$ and $t = 0.038$. 61
Figure 2.11: Internal energy plots for WENO-C for the Leblanc shock-tube experiment at $t = 6$. 65
Figure 2.12: Internal energy plots for the Leblanc shock-tube experiment at $t = 6$ using WENO-LF with and without the C-equation. 67

ACKNOWLEDGEMENTS

I would like to acknowledge my two academic mentors, Mike Holst and Steve Shkoller for their continued support and for serving as chairs on my committee. Their friendship and guidance has always shaped my life and career in a positive way.

I would also like to acknowledge my family and friends for keeping me sane through the entire process of my graduate studies.

Chapter 1, in part, is currently being prepared for submission for publication of the material. The dissertation author was the primary investigator and author of this material. I would like to acknowledge the co-author, Michael Holst.

Chapter 2, in full, has been accepted for publication in Journal of Computational Physics. The dissertation author was the primary investigator of this paper. I would like to acknowledge the co-authors, Jon Reisner and Steve Shkoller.

VITA

2006	Bachelor of Science in Mathematics *with honors*, University of California, Davis
2006-2011	Teaching Assistant, Department of Mathematics, University of California, San Diego
2010	Master of Science in Applied Mathematics, University of California, San Diego
2011-2012	Research Assistant, Department of Mathematics, University of California, Davis
2012	Doctor of Philosophy in Mathematics *with Specialization in Computational Science*, University of California, San Diego

ABSTRACT OF THE DISSERTATION

A Development of Discretization Techniques for Some Elliptic and Hyperbolic PDE

by

Jonathan W. Serencsa
Doctor of Philosophy in Mathematics with Specialization in Computational Science

University of California San Diego, 2012

Professor Michael Holst, Chair
Professor Steve Shkoller, Co-Chair

In this thesis, we consider the discretization of two different PDE which govern physical phenomenon. First, we consider the diffusive diffusive logistic equation and develop several new results on weak solutions and on their approximation by Galerkin-type methods. Our goal is to establish a rate of convergence for Galerkin approximations to solutions of this problem, and thus we first consider the continuous model, and briefly review the literature on the known solution theory. We then state and prove a new result on existence and uniqueness of weak solutions. Moreover, we provide numerical results to provide evidence of our theoretical results. Second, we consider the nonlinear systems of conservation laws which propagate shock waves, rarefactions, and contact discontinuities, and introduce what we call the C-method. We shall focus our attention on the compressible Euler equations in one space dimension. The novel feature of our approach involves the coupling of a linear scalar reaction-diffusion equation to our system of conservation laws, whose solution $C(x,t)$ is the coefficient to an additional (and artificial) term added to the flux, which determines the location, localization, and strength of the artificial viscosity. Near shock discontinuities, $C(x,t)$ is large and localized, and transitions smoothly in space-time to zero away from discontinuities. Our approach is a provably convergent, spacetime-regularized variant of the original idea of Richtmeyer and Von Neumann, and is provided at the level of the PDE, thus allowing a

host of numerical discretization schemes to be employed. We demonstrate the effectiveness of the C-method with three different numerical implementations and apply these to a collection of classical problems. All three schemes yield higher-order discretization strategies, which provide sharp shock resolution with minimal overshoot and noise, and compare well with higher-order WENO schemes that employ approximate Riemann solvers, outperforming them for the difficult Leblanc shock tube experiment.

Chapter 1

Error Estimates for Positive Discrete Solutions to the Diffusive Logistic Equation

1.1 Introduction

In this article we consider the Diffusive Logistic Equation (DLE), which is a widely accepted model for modeling the steady-state population distribution of a single species occupying a domain in space. While the nonlinearity is rather simple, and sub-critical in at least up to three space dimensions, it is not monotone, which prevents one from using a number of standard techniques usually available for both the analysis and approximation of nonlinear problems. Another problematic feature of the DLE is the inability to guarantee uniqueness, unless we require the solution to be positive. While this positivity constraint is of practical importance when modeling the distribution of a species (and thus "negative" population makes no sense), it presents considerable difficulty when trying to construct solutions (both existence arguments as well as explicit numerical constructions). As a result of these interesting features, the DLE is often used as a model of more complex nonlinear elliptic equations and elliptic systems. Our goal in this article is to develop several new results on weak solutions to the DLE, and on their approximation by Galerkin-type methods. In particular, we wish to establish a spe-

cific rate of convergence for Galerkin approximations to solutions of this problem, in order to gain insight into the types of techniques that might be useful for establishing convergence rates for more general nonlinear problems.

The classical formulation of the DLE is as follows. Given a bounded domain $\Omega \subset \mathbb{R}^n$ for $n \leqslant 3$ whose boundary is of class C^2, for some $\alpha \in (0, 1]$ and b, c positive constants, we look for $u \in C^{2+\alpha}(\bar{\Omega})$ satisfying

$$-\Delta u + bu^2 = cu \quad \text{in } \Omega, \tag{1.1.1}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{1.1.2}$$

The DLE models the steady-state population density of a single species occupying the region Ω where the boundary $\partial\Omega$ is "hostile" to the species (for more details, see [1] or [2]).

For this classical formulation, we have the following (also classical) result.

Theorem 1.1.1. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with boundary $\partial\Omega$ of class C^2 and (λ, ϕ) be the principle eigenpair of $-\Delta$ on Ω with homogeneous Dirichlet boundary conditions. If $c > \lambda$ then there exists a unique $u \in C^{2+\alpha}(\bar{\Omega})$ satisfying* (1.1.1)*,* (1.1.2) *and $u > 0$ in Ω. Moreover, we have the pointwise bound*

$$\frac{c - \lambda}{b}\phi \leqslant u \leqslant \frac{c}{b} \quad \textit{in } \bar{\Omega}. \tag{1.1.3}$$

Proof. See for example [3]. □

Though the above result asserts the well-posedness of the classical formulation, this requires the domain in question Ω to posses rather strong regularity properties. In practice, when trying to provide a discrete formulation, one must first provide some sort of approximation to the actual domain. Such approximations will generally be some (possibly non-convex) polygon that interpolates at the boundary. In order to be able to show that our discrete solutions converge to the continuous solution, it is desirable that one can solve the continuous problem on the discrete domain. In order to do so, one must weaken their notation of a solution, to a point that the classical formulation no longer makes sense. With this in mind, in the first part of the paper we will generalize Theorem 1.1.1 to *weak solutions* (see Theorem 1.2.2 in §1.2 below). We will then prove a

similar result (see Theorem 1.3.4 in §1.3 below) where we search for discrete Galerkin approximations u_h living in some finite dimensional subspace $V_h \subset H_0^1(\Omega)$. Once we have developed well-posedness results for the continuous and discrete formulations, we will establish a lemma regarding the "linearization" of the nonlinearity about the continuous solution, which will be critical in proving the following rate of convergence result: For h sufficiently small, a Galerkin approximation u_h of the solution u to the DLE satisfies:

$$||u - u_h||_{H^1(\Omega)} = O(h^{\frac{1}{2}+s}),$$
$$||u - u_h||_{L^2(\Omega)} = O(h^{1+2s}),$$
$$||u - u_h||_{L^\infty(\Omega)} = O(h^{\frac{3-n}{2}+s}).$$

where s is determined by the elliptic regularity satisfied by the domain Ω.

In the case Ω is of class $C^{1,1}$ or convex-Lipschitz, for which Ω possesses full H^2-elliptic regularity and $s = \frac{1}{2}$, there is a wide variety of results which yield the above rates of convergence for finite element approximations to a general class of linear and nonlinear elliptic PDE with a variety of boundary conditions (see [4], [5], [6], [7], [8], [9], [10]). Moreover, the results of [8] suggest that our proposed L^∞ bound is rather poor since one can expect a rate of $O(h^2|\log(h)|^{n/4+1})$. However, this requires either convexity or smoothness of the domain and one must exclude the case where $\partial\Omega$ is a general (non-convex) polygon/polyhedra.

In the case where Ω is a non-convex Lipschitz domain and lacks the full H^2-elliptic regularity, $s \in (0, \frac{1}{2})$, the vast number of results we reference above still hold, but with a reduced order of convergence, as expressed in our desired error bounds. Moreover, the techniques used in [8] to achieve optimal L^∞ convergence estimates can not be applied without full H^2-elliptic regularity.

Despite the extensive amount of literature for nonlinear elliptic problems, the convergence theory for the DLE does not fit into the general framework required by these results. Subsequently, we must examine the problem in a far more specific way so as to deal with the positivity constraints and the lack of monotonicity.

We shall define the required assumptions on the approximating spaces V_h more precisely later in the paper but note that the necessary assumptions hold for the sequence

for spaces V_h given by piecewise-linear Lagrange finite elements over meshes defined using triangles or tetrahedra, with the further requirement that Ω can be subdivided exactly using triangles or tetrahedra with strictly acute angles. We further note that for a rate of convergence to be of practical use, the constants (which must not depend on the discretization) must be *a priori* computable quantities. However, keeping track of how large the constants must be would be overly burdening. With that said, we establish the convention that C will be a general positive constant, that only depends on problem parameters (i.e. Ω, n, b, c). Furthermore, we will commonly use the phrase "for h sufficiently small" and such should be interpreted in the sense that "how small h must be," is again computable based solely on problem parameters.

Outline of the paper. The remainder of the article is structured as follows. In §1.2, we first consider the continuous model, and briefly review the literature on the known solution theory. We then state and prove a basic result on existence and uniqueness of weak solutions. The proof will be established through a sequence of lemmas, using a combination of fixed-point arguments, compactness techniques, and maximum principle arguments. In §1.3, we develop a discrete analogue for Galerkin approximations under reasonable assumptions on the approximation spaces. The main proof will involve tracing the proof of the continuous result. Both the continuous and discrete results make it possible to establish *a priori* error estimates for Galerkin approximations. A critical technical result is first given in §1.4, which exploits a subtle relationship to an auxiliary problem. The main *a priori* Galerkin error estimates are then established in §1.5, and are then subsequently used to characterize the rate of convergence of such approximations in a precise way. The error estimates are established by combining maximum principles with a careful analysis of the spectral structure of the linearized problem, and by exploiting the subtle relationship to an auxiliary problem given in §1.4. The final convergence result given in §1.5, is applicable to general finite element approximations to positive solutions in bounded, non-convex polygonal domains in both two and three space dimensions. We show that, under reasonable assumptions on the approximation spaces and on the details of the discretization, the Galerkin method converges at a fixed rate of convergence (Theorem 1.5.8). Finally, in §1.6 we provided numerical data which confirm our results.

1.2 Analysis of the Continuous Problem

Though there are numerous references devoted to the study of classical solutions to some form of the DLE ([2], [3], [11], [12], [13]) it seems that analogous results for weak solutions, critical for our error analysis, is apparently not in the existing literature on the DLE. Thus, we devote the following section to laying down the framework for, and subsequently proving, the following theorem, which can be viewed as a generalization of Theorem 1.1.1 to weak solutions.

Before we state our result, we make the following definition:

Definition 1.2.1. *A Lipschitz domain $\Omega \subset \mathbb{R}^n$ possesses H^s-elliptic regularity if, given any elliptic operator L, there exists a constant $C > 0$ such that*

$$\|w\|_{H^s(\Omega)} \leq C\|Lw\|_{L^2(\Omega)},$$

for all $w \in H^s(\Omega) \cap H_0^1(\Omega)$.

We know that any two-dimensional Lipschitz domain possesses $H^{\frac{3}{2}+s}$-elliptic regularity for $s \in (0, \frac{1}{2}]$ [15]. Moreover, in *any* space dimension, convex Lipschitz domains possess H^2-elliptic regularity . However, despite our best efforts, we do not know whether non-convex Lipschitz domains in three dimensions possess $H^{\frac{3}{2}+s}$-elliptic regularity for $s \in (0, \frac{1}{2}]$.

Theorem 1.2.2. *Let $\Omega \subset \mathbb{R}^n$ ($n \leqslant 3$) possess $H^{\frac{3}{2}+s}$-elliptic regularity for $s \in (0, \frac{1}{2}]$. Furthermore, let (λ, ϕ) be the principle eigenpair of $-\Delta$ on Ω with homogeneous boundary conditions, with the convention that ϕ is normalized w.r.t to L^∞ and $\phi > 0$ in Ω. If $c > \lambda$ then there there exists a unique $u \in H^{\frac{3}{2}+s}(\Omega) \cap H_0^1(\Omega)$ satisfying*

$$\int_\Omega \nabla u \cdot \nabla v + bu^2 v - cuvdx = 0 \quad \textit{for all } v \in H_0^1(\Omega), \tag{1.2.1a}$$

$$u > 0 \quad \textit{in } \Omega. \tag{1.2.1b}$$

Moreover, $u \in L^\infty(\Omega)$ and we have the pointwise bound

$$\frac{c-\lambda}{b}\phi \leqslant u \leqslant \frac{c}{b} \quad a.e. \textit{ in } \Omega.$$

The proof of Theorem 1.2.2 will be established through a sequence of Lemmas. To this end, define $a \in \mathcal{L}(H_0^1(\Omega) \times H_0^1(\Omega), \mathbb{R})$ and $B : H_0^1(\Omega) \to L^2(\Omega)$ via the following

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v dx \quad \text{and} \quad (B(u), v) = \int_\Omega (bu^2 - cu)v dx.$$

Clearly a is bounded, symmetric, and bilinear with

$$|a(u,v)| \leqslant \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \quad \text{for all } u, v \in H_0^1(\Omega),$$

and (via the Poincare inequality) coercive with constant $m > 0$ satisfying

$$m\|u\|_{H^1(\Omega)}^2 \leqslant a(u,u) \quad \text{for all } u \in H_0^1(\Omega).$$

We also have a maximum principle for a i.e. if $u \in H_0^1(\Omega)$ satisfies

$$a(u, v^+) \leqslant (\geqslant)\ 0 \quad \text{for all } v \in H_0^1(\Omega),$$

then $u \leqslant (\geqslant)\ 0$ $a.e.$ in Ω, where we define

$$v^+ := \max\{v, 0\}.$$

For results pertaining to maximum principles for weak solutions to elliptic PDE, see [14].

On the other hand, B satisfies two other useful properties which we assert in the following lemma.

Lemma 1.2.3. *B is a bounded operator (nonlinear) from $H^1(\Omega) \to L^2(\Omega)$ which moreover satisfies the following property: If $u_k, u \in H^1(\Omega) \cap L^\infty(\Omega)$ with $u_k \to u$ $a.e.$ in Ω then*

$$\|B(u_k) - B(u)\|_{L^2(\Omega)} \to 0.$$

Proof. The boundedness of B follows from the estimate

$$\|B(u)\|_{L^2(\Omega)}^2 \leqslant C\left(\|u\|_{L^4(\Omega)}^4 + \|u\|_{L^2(\Omega)}^2\right) \leqslant C\|u\|_{H^1(\Omega)}^2,$$

which follows by Sobolev imbedding since we have restricted ourselves to $n = 2, 3$.

The second claim then follows from the dominated convergence theorem along with the fact that as a map from $\mathbb{R} \to \mathbb{R}$, $f(x) = bx^2 - cx$ is continuous. □

We also establish a few notational conventions.

Definition 1.2.4. *Given $u_1, u_2 \in H^1(\Omega) \cap L^\infty(\Omega)$ with $u_1 \leqslant u_2$ a.e. in Ω, we define the order interval $[u_1, u_2] \subset H^1(\Omega) \cap L^\infty(\Omega)$ via*

$$[u_1, u_2] = \{u \in H^1(\Omega) \cap L^\infty(\Omega) : u_1 \leqslant u \leqslant u_2 \text{ a.e. in } \Omega\},$$

and the homogeneous variant

$$[u_1, u_2]_0 = [u_1, u_2] \cap H_0^1(\Omega).$$

Definition 1.2.5. *We say $G : H_0^1(\Omega) \to L^2(\Omega)$ is* **monotone increasing** *(resp.* **monotone decreasing***) on an order interval $[u_1, u_2]$ if*

$$(G(w_1) - G(w_2), v^+) \leqslant (\geqslant)\ 0,$$

for any $w_1, w_2 \in [u_1, u_2]$ with $w_1 \leqslant w_2$ and $v \in H_0^1(\Omega)$.

Before we prove existence, we have the two following a priori results.

Lemma 1.2.6. *If $u \in H_0^1(\Omega)$ satisfies* (1.2.1a)*, then we have $u \in H^{\frac{3}{2}+s}(\Omega)$ for some $s \in (0, \frac{1}{2}]$ and subsequently, $u \in L^\infty(\Omega)$.*

Proof. Sobolev Embedding and the fact that $u \in H_0^1(\Omega)$ gives us $cu - bu^2 \in L^2(\Omega)$ and elliptic regularity gives us $u \in H^{\frac{3}{2}+s}(\Omega)$ for some $s \in (0, \frac{1}{2}]$ (see [15]). The embedding of $H^{\frac{3}{2}+s}(\Omega)$ into $L^\infty(\Omega)$ for $n \leqslant 3$ finishes the proof. □

We also have the following uniqueness result.

Lemma 1.2.7. *Let $u_1, u_2 \in H_0^1(\Omega)$ be solutions to* (1.2.1a) *and suppose that $u_1, u_2 > 0,\ a.e.$ in Ω. Then $u_1 \equiv u_2$.*

Proof. By (1.2.6) we know that the positivity of u_1, u_2 is well-defined. Taking each solution as test functions for the opposite solution in (1.2.1a) gives us

$$\int_\Omega b(u_1 - u_2)u_1u_2\, dx = 0.$$

The result follows from the fact that $bu_1u_2 > 0$ $a.e.$ □

Proving Theorem 1.2.2 will now proceed as follows. First we establish the existence of positive, ordered sub- and supersolutions $\underline{u}, \overline{u}$. Second, we propose an iterative process that generates pointwise, monotone sequences $\{\underline{u}^k\}$ and $\{\overline{u}^k\}$ which remain in the homogeneous interval $[\underline{u}, \overline{u}]_0$ for $k \geqslant 1$. Finally, we use the various modes of convergence to show that our sequences converge to positive solutions of (1.2.1a) which must be equal by our uniqueness result, Lemma 1.2.7.

Lemma 1.2.8. *There exist sub- and supersolutions $\underline{u}, \overline{u} \in H^1(\Omega) \cap L^\infty(\Omega)$ satisfying*

$$a(\underline{u}, v^+) + (B(\underline{u}), v^+) \leqslant 0 \quad \textit{and} \quad a(\overline{u}, v^+) + (B(\overline{u}), v^+) \geqslant 0 \tag{1.2.2}$$

for all $v \in H_0^1(\Omega)$ and

$$0 < \underline{u} \leqslant \overline{u},$$

almost everywhere in Ω.

Proof. Since we follow the convention that the principle eigenfunction is positive and normalized such that $\|\phi\|_{L^\infty(\Omega)} = 1$ taking $\underline{u} = \frac{c-\lambda}{b}\phi$, given $v \in H_0^1(\Omega)$ we also have $v^+ \in H_0^1(\Omega)$ and

$$a(\underline{u}, v^+) + (B(\underline{u}), v^+) = \frac{(c-\lambda)^2}{b} \int_\Omega (\phi - 1)\phi v^+ \, dx.$$

Since $b > 0$ and $(1-\phi)\phi\varphi \leqslant 0$ for any $\varphi \in C_0^\infty(\Omega)$ with $\varphi \geqslant 0$, $\underline{u}$ being a subsolution follows from a density argument. Moreover, the condition $c > \lambda$ ensures $\underline{u} > 0$ in Ω.

We can then take $\overline{u} = \frac{c}{b}$ as a supersolution and the ordering $\underline{u} \leqslant \overline{u}$ follows because ϕ is normalized in $L^\infty(\Omega)$. □

It is rather easy to see that B is **not** monotone on $[\underline{u}, \overline{u}]$ and thus we define a shifted variant

$$(B_c(u), v) = (B(u) - cu, v),$$

which is monotone decreasing on $[\underline{u}, \overline{u}]$ and still possesses the boundedness and continuity properties. Moreover, we also have that the shifted bilinear form defined by

$$a_c(u, v) = a(u, v) + c(u, v),$$

maintains a maximum principle, its boundedness, and coercivity.

Now, due to the monotone property of B_c and the maximum principle for a_c, we provide an iteration along with the following result.

Lemma 1.2.9. *The sequences* $\{\underline{u}^k\}$ *and* $\{\overline{u}^k\}$ *defined by*

$$a_c(\underline{u}^{k+1}, v) + (B_c(\underline{u}^k), v) = 0 \quad \textit{and} \quad a_c(\overline{u}^{k+1}, v) + (B_c(\overline{u}^k), v) = 0,$$

for all $v \in H_0^1(\Omega)$*, with* $\underline{u}^0 = \underline{u}$ *and* $\overline{u}^0 = \overline{u}$ *remain in the order interval* $[\underline{u}, \overline{u}]$ *and are ordered in the following sense*

$$\underline{u}^k \leqslant \underline{u}^{k+1} \leqslant \overline{u}^{k+1} \leqslant \overline{u}^k, \quad \textit{for } k = 0, 1, \ldots. \tag{1.2.3}$$

Proof. The iterations are defined by a sequence of linear elliptic PDE, and thus standard existence arguments and the maximum principle allows us to conclude the sequences are well-defined and remain in $H_0^1(\Omega) \cap L^\infty(\Omega)$ for $k \geqslant 1$. Since $\underline{u}$ and $\overline{u}$ are ordered sub- and supersolutions, we have

$$a_c(\underline{u} - \underline{u}^1, v^+) \leqslant 0 \quad \text{and} \quad a_c(\overline{u}^1 - \overline{u}, v^+) \leqslant 0,$$

for all $v \in H_0^1(\Omega)$. Furthermore, since B_c is monotone decreasing on $[\underline{u}, \overline{u}]$ and $\underline{u} \leqslant \overline{u}$ we also have

$$a_c(\underline{u}^1 - \overline{u}^1, v^+) \leqslant 0,$$

for all $v \in H_0^1(\Omega)$. Thus we have the desired ordering for $k = 1$. Proceeding inductively we achieve (1.2.3). □

We are now able to prove Theorem 1.2.2.

Proof. (Theorem 1.2.2) To be more concise, we first consider only the sequence $\underline{u}^k$. By Lemma 1.2.9 the sequence $\underline{u}^k$ is uniformly bounded, pointwise monotonic in L^∞ and thus there exists $\underline{v} \in L^\infty(\Omega)$ such that **any** subsequence of $\underline{u}^k$ converges pointwise a.e to $\underline{v}$.

This pointwise convergence allows us to conclude that

$$\|B_c(\underline{u}^k)\|_{L^2(\Omega)},$$

is uniformly bounded which in turn implies that $\underline{u}^k$ is uniformly bounded in $H_0^1(\Omega)$. Any subsequence $\underline{u}^{k_j}$ must also be uniformly bounded in $H_0^1(\Omega)$ and thus there exists $\underline{w} \in H_0^1(\Omega)$ and a further subsequence $\underline{u}^{k_{j_l}}$ with

$$\underline{u}^{k_{j_l}} \rightharpoonup \underline{w} \quad \text{in } H_0^1(\Omega),$$

but we also have

$$\underline{u}^{k_{j_l}} \to \underline{v} \quad \text{pointwise a.e.}$$

Invoking the uniform pointwise boundedness of $\underline{u}^{k_{j_l}}$ along with the necessary *strong* L^2 convergence of $\underline{u}^{k_{j_l}}$ to $\underline{w}$ we have $\underline{v} = \underline{w}$ a.e. in Ω. Since $\underline{u}^{k_j}$ was an arbitrary subsequence, we have

$$\begin{aligned}
\underline{u}^k &\rightharpoonup \underline{v} \quad \text{in } H_0^1(\Omega), \\
\underline{u}^k &\to \underline{v} \quad \text{in } L^2(\Omega), \\
\underline{u}^k &\to \underline{v} \quad \text{pointwise a.e. in} \Omega.
\end{aligned}$$

Similarly, we have the same result for the respective convergences of $\overline{u}^k$ to $\overline{v}$.

Finally, the weak-H_0^1 and pointwise convergence allows us to conclude that $\underline{v}$ and $\overline{v}$ satisfy (1.2.1a). Moreover, since $\underline{u} > 0$ in Ω and $\underline{v}, \overline{v} \in [\underline{u}, \overline{u}]_0$, Lemmas 1.2.6 and 1.2.7 allows us to conclude that $\underline{v} \equiv \overline{v}$. We then write $u \equiv \underline{v} \equiv \overline{v}$ as the unique positive solution to (1.2.1a). □

1.3 Analysis of the Discrete Problem

We now consider ones ability to generate approximate solutions. We first let $X_h \subset H^1(\Omega) \cap L^\infty(\Omega)$ denote a finite-dimensional subspace and then define the zero trace analog

$$V_h = X_h \cap H_0^1(\Omega),$$

and we wish to find conditions such that the following Galerkin formulation is well-posed.

Find $u_h \in V_h$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h + b u_h^2 v_h - c u_h v_h \, dx = 0 \quad \text{for all } v_h \in V_h, \tag{1.3.1a}$$

$$u_h > 0 \quad \text{in } \Omega. \tag{1.3.1b}$$

In general, the above formulation will not be well-posed for general V_h, however, under some reasonable assumptions we can achieve well-posedness. We now list the following assumptions;

Assumption 1.3.1. *V_h must possess a basis $\{\varphi_{h,k}\}$ for which $\varphi_{h,k} \geqslant 0$ in Ω. Thus we define*

$$V_h^+ = \{v_h \in V_h : v_h = \sum_{k=1}^{N_h} \xi_{h,k}\varphi_{h,k} \ \text{ with } \ \xi_{h,k} \geqslant 0\},$$

which is a convex subset of V_h with the property that $v_h \in V_h^+$ has $v_h(x) = 0$ for some $x \in \Omega$ if and only if some coefficient with respect to the non-negative basis is zero. Moreover, we require the basis to be extendable to one for X_h with the same properties and we define X_h^+ in the same way.

Assumption 1.3.2. *The shifted bilinear form a_c must satisfy a discrete maximum principle; i.e. if $z_h \in X_h$ satisfies*

$$a_c(z_h, v_h) \geqslant 0 \quad \textit{for all } v_h \in V_h^+, \tag{1.3.2}$$

then $z_h \in X_h^+$ and thus we necessarily have the opposite direction;

$$a_c(z_h, v_h) \leqslant 0 \quad \textit{for all } v_h \in X_h^+,$$

then $-z_h \in X_h^+$ which we just write as $z_h \in X_h^-$.

Assumption 1.3.3. *The principle eigenvalue of the discrete Laplacian, denoted λ_h must satisfy*

$$\lambda_h < c,$$

and we must be able to take the principle eigenfunction ϕ_h to satisfy

$$\phi_h > 0 \quad \textit{in } \Omega \quad \textit{and} \quad ||\phi_h||_{L^\infty(\Omega)} = 1.$$

The above assumptions enable us to use similar techniques to those used for the continuos problem to prove well-posedness of the discrete problem. We further note that a large variety of standard approximation spaces will not in general satisfy the above assumptions. For now, and in the following presentation, we postpone discussing the existence of a space V_h which satisfies the above assumptions until §1.6.

We then can show the following theorem;

Theorem 1.3.4. *If $V_h \subset H_0^1(\Omega) \cap L^\infty(\Omega)$ satisfies Assumptions 1.3.1, 1.3.2, and 1.3.3 then there exists a unique $u_h \in V_h$ satisfying* (1.3.1a) *and* (1.3.1b).

Moreover,

$$u_h \in [\frac{c-\lambda_h}{b}\phi_h, \frac{c}{b}].$$

Proof. The proof of Theorem 1.3.4 follows a similar procedure as that of Theorem 1.2.2. Indeed, Assumption 1.3.3 allows us to construct approximate, positive sub- and super-solutions

$$\underline{u}_h = \frac{c-\lambda_h}{b}\phi_h \quad \text{and} \quad \overline{u}_h = \frac{c}{b},$$

where we remark that $\overline{u}_h \notin V_h$ is not an issue, analogous to how $\overline{u}$ does not have zero trace in the continuous case. One can easily verify that $\underline{u}_h$ and $\overline{u}_h$ satisfy

$$a(\underline{u}_h, v_h) + (B(\underline{u}_h), v_h) \leqslant 0 \quad \text{and} \quad a(\overline{u}_h, v_h) + (B(\overline{u}_h), v_h) \geqslant 0,$$

for all $v_h \in V_h^+$.

With the initial iterates $\underline{u}_h^0 = \underline{u}_h$ and $\overline{u}_h^0 = \overline{u}_h$, we define $\underline{u}_h^{k+1}, \overline{u}_h^{k+1} \in V_h$ satisfying

$$a_c(\underline{u}_h^{k+1}, v_h) + (B_c(\underline{u}_h^k), v_h) = 0 \quad \text{and} \quad a_c(\overline{u}_h^{k+1}, v_h) + (B_c(\overline{u}_h^k), v_h) = 0, \tag{1.3.3}$$

for all $v_h \in V_h$. Choosing a basis adhering to Assumption 1.3.1, one solves the above problems for $\underline{u}_h^{k+1}$ and $\overline{u}_h^{k+1}$ which reduce to solving a finite dimensional linear system. Existence and uniqueness follows since $V_h \subset H_0^1(\Omega)$ and a_c is coercive on $H_0^1(\Omega)$. Moreover, we have the estimates

$$\|\underline{u}_h^{k+1}\|_{H^1(\Omega)} \leqslant C\|B_c(\underline{u}_h^k)\|_{L^2(\Omega)} \quad \text{and} \quad \|\overline{u}_h^{k+1}\|_{H^1(\Omega)} \leqslant C\|B_c(\overline{u}_h^k)\|_{L^2(\Omega)}. \tag{1.3.4}$$

The proof of the required ordering

$$\underline{u}_h \leqslant \underline{u}_h^1 \leqslant \cdots \leqslant \underline{u}_h^k \leqslant \cdots \leqslant \overline{u}_h^k \leqslant \cdots \leqslant \overline{u}_h^1 \leqslant \overline{u}_h, \tag{1.3.5}$$

follows along the same line of reasoning as that for (1.2.3) due to Assumption 1.3.2; we omit the details.

We write an equivalent statement of (1.3.5),

$$\underline{\xi}_{h,j}^k \leqslant \underline{\xi}_{h,j}^{k+1} \leqslant \overline{\xi}_{h,j}^{k+1} \leqslant \overline{\xi}_{h,j}^k, \tag{1.3.6}$$

for $k = 1, 2, \ldots$ and $j = 1, 2, \ldots, N_h$ where $\underline{\xi}_{h,j}^k$ and $\overline{\xi}_{h,j}^k$, denote the coefficients of $\underline{u}_h^k$ and $\overline{u}_h^k$ with respect to the bases ensured by Assumption 1.3.1.

Using (1.3.6) we conclude $\underline{\xi}_{h,j}^k$ and $\overline{\xi}_{h,j}^k$ are bounded monotonic sequences of real numbers, and thus each converge to some $\underline{\zeta}_{h,j}$ and $\overline{\zeta}_{h,j}$. Since the initial discrete sub- and supersolutions were strictly positive in Ω, we necessarily have that each $\underline{\zeta}_{h,j}$ and $\overline{\zeta}_{h,j}$ are positive for all j. Defining $\underline{v}_h, \overline{v}_h \in V_h$ via

$$\underline{v}_h = \sum_{j=1}^{N_h} \underline{\zeta}_{h,j}\varphi_{h,j} \quad \text{and} \quad \overline{v}_h = \sum_{j=1}^{N_h} \overline{\zeta}_{h,j}\varphi_{h,j}.$$

Moreover, we have

$$\underline{v}_h, \overline{v}_h \in [\frac{c-\lambda_h}{b}\phi_h, \frac{c}{b}]$$

satisfying the positivity requirement (1.3.1b). Since all norms are equivalent in finite dimensions, the convergence of the coefficients guarantee that

$$\underline{u}_h^k \to \underline{v}_h \quad \text{and} \quad \overline{u}_h^k \to \overline{v}_h \quad \text{in } H_0^1(\Omega)$$

Passing to the limits in (1.3.3) we have $\underline{v}_h$ and $\overline{v}_h$ satisfy (1.3.1a) and (1.3.1b).

The uniqueness of positive solutions follows the same reasoning as in the proof of Lemma 1.2.7 with $H_0^1(\Omega) \cap L^\infty(\Omega)$ replaced with V_h. Thus, we write $u_h = \underline{v}_h = \overline{v}_h$ as the unique, discrete positive solution.

□

1.4 A Critical Technical Lemma

Our ability to achieve rate of convergence results in Theorem 1.5.8 depends heavily on a subtle realization. If we first define the following modified bilinear form given $g \in L^\infty(\Omega)$ via

$$a(g; w, v) = \int_\Omega \nabla w \cdot \nabla v + gwv \, dx \quad \text{for } w, v \in H_0^1(\Omega),$$

then our unique solution satisfies

$$a(bu - c; u, v) = 0 \quad \text{for all } v \in H_0^1(\Omega), \tag{1.4.1}$$

and thus the bilinear form $a(bu - c; \cdot, \cdot)$ is not coercive. However, if we are to shift by some positive function $g \in L^\infty(\Omega)$, we can show that the resulting bilinear form

$$a(bu + g - c, \cdot, \cdot),$$

is coercive, which is the statement of the following lemma:

Lemma 1.4.1. *For any $g \in L^\infty(\Omega)$ with $g > 0$ in Ω the symmetric bilinear form*

$$a((bu+g)-c,\cdot,\cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R},$$

where u is the unique positive solution to (1.2.1a)*, then there exists a positive constant $\alpha > 0$ such that*

$$a((bu+g)-c,w,w) \geqslant \alpha ||w||^2_{H^1(\Omega)} \quad \textit{for all } w \in H_0^1(\Omega),$$

and subsequently satisfies a maximum principle, i.e if

$$a((bu+g)-c,w,v^+) \leqslant (\geqslant) 0 \quad \textit{for any } v \in H_0^1(\Omega),$$

than $w \leqslant (\geqslant) 0$ $a.e.$ in Ω.

The proof of Lemma 1.4.1 will depend on the following technical lemma, which shows that we have sufficient "wiggle room" for the spectrum of A_u.

Lemma 1.4.2. *Let $\eta \in L^\infty(\Omega)$ and $a(\eta;\cdot,\cdot)$ satisfy*

$$a(\eta;w,w) \geqslant K\|w\|^2_{L^2(\Omega)} \quad \textit{for all } w \in H_0^1(\Omega), \tag{1.4.2}$$

for some $K > 0$. Then there exists $\epsilon \in (0,1)$ sufficiently small depending on η and K such that

$$(1-\epsilon)\|\nabla w\|_{L^2(\Omega)} + (\eta w, w) \geqslant \frac{K}{2}\|w\|^2_{L^2(\Omega)} \quad \textit{for all } w \in H_0^1(\Omega). \tag{1.4.3}$$

Proof. For any $w \in H_0^1(\Omega)$ and $\epsilon \in (0,1)$ we have

$$\begin{aligned}(1-\epsilon)\|\nabla w\|^2_{L^2(\Omega)} + (\eta w, w) &= (1-\epsilon)\left\{\|\nabla w\|^2_{L^2(\Omega)} + (\eta w, w)\right\} + \epsilon(\eta w, w) \\ &\geqslant (1-\epsilon)K\|w\|^2_{L^2(\Omega)} - \epsilon\|\eta\|_{L^\infty(\Omega)}\|w\|^2_{L^2(\Omega)}\end{aligned}$$

and thus taking

$$\epsilon \leqslant \frac{K}{2(K+\|\eta\|_{L^\infty(\Omega)})},$$

we have the desired result.

□

Proof. (**Lemma 1.4.1**) If we interpret (1.4.1) as saying that (u, c) is an eigenpair to the following eigenvalue problem:

Find $w \in H_0^1(\Omega)$ and $\eta \in \mathbb{R}$ satisfying

$$a(bu; w, v) = \eta(w, v) \quad \text{for all } v \in H_0^1(\Omega), \tag{1.4.4}$$

we then make the following observations. First, since u is a positive $L^\infty(\Omega)$ function, the bilinear form $a(bu; \cdot, \cdot)$ is a symmetric, coercive bilinear form over $H_0^1(\Omega) \times H_0^1(\Omega)$ and thus (1.4.4) only possesses a non-trivial solution for η in a countable set we denote $\Lambda(a_{bu})$. Moreover, $\Lambda(a_{bu})$ has a minimal positive value. By (1.4.1), we know that $c \in \Lambda(a_{bu})$ and since u is thus a strictly positive eigenfunction we necessarily have

$$c = \min \Lambda(a_{bu}).$$

Since we have assumed that g is a strictly positive $L^\infty(\Omega)$ function, we then have

$$\min \Lambda(a_{bu}) < \min \Lambda(a_{bu+g}),$$

which implies the existence of some $\delta > 0$ such that

$$\min \Lambda(a_{bu+g}) = c + \delta.$$

This in turn implies that

$$\int_\Omega |\nabla w|^2 + (bu + g)w^2 \, dx \geqslant (c + \delta)\|w\|^2_{L^2(\Omega)} \quad \text{for all } w \in H_0^1(\Omega),$$

or

$$a((bu + g) - c; w, w) \geqslant \delta\|w\|^2_{L^2(\Omega)} \quad \text{for all } w \in H_0^1(\Omega).$$

To achieve the coercivity bound, we choose $0 < \epsilon < 1$ sufficiently small to ensure the result of Lemma 1.4.2, and thus we have

$$\begin{aligned} a((bu + g) - c; w, w) &= \int_\Omega |\nabla w|^2 + [(bu + g) - c]w^2 \, dx \\ &= \epsilon\|\nabla w\|^2_{L^2(\Omega)} + (1 - \epsilon)\|\nabla w\|^2_{L^2(\Omega)} + ([(bu + g) - c]w, w) \\ &\geqslant \epsilon\|\nabla w\|^2_{L^2(\Omega)} + \frac{\delta}{2}\|w\|^2_{L^2(\Omega)} \\ &\geqslant \min\{\epsilon, \frac{\delta}{2}\}\|w\|^2_{H^1(\Omega)}, \end{aligned}$$

and taking $\alpha = \min\{\epsilon, \frac{\delta}{2}\}$ we have the desired coercivity bound.

To see that $a(bu + g - c; \cdot, \cdot)$ satisfies a maximum principle, let $v = w$ yielding

$$0 \geqslant a((bu + g) - c, w, w^+) \geqslant \alpha \|w\|^2_{H^1(\Omega^{(0)})},$$

where $\Omega^{(0)} := \{x \in \Omega \ : \ u(x) \geqslant 0\}$ and thus $w \leqslant 0$ in Ω. Similarly, taking $v = -w$ achieves the reverse inequality.

□

1.5 Galerkin Error Estimates

We now make assumptions on the approximating spaces V_h regarding how well they approximate the space $H^1_0(\Omega)$.

Assumption 1.5.1. *For our given sequence of finite dimensional spaces $V_h \subset H^1_0(\Omega)$ indexed by parameter h, for $s \in (0, \frac{1}{2}]$ there exists some constant $C > 0$ which does not depend on h such that*

$$\inf_{v_h \in V_h} \|w - v_h\|_{H^1(\Omega)} \leqslant C h^{\frac{1}{2}+s} \|w\|_{H^{\frac{3}{2}+s}(\Omega)} \quad \textit{for all } w \in H^{\frac{3}{2}+s}(\Omega). \tag{1.5.1}$$

Assumption 1.5.2. *Our spaces $V_h \subset H^1_0(\Omega)$ also satisfy*

$$\inf_{v_h \in V_h} \|w - v_h\|_{L^\infty(\Omega)} \leqslant C h^{\frac{3-n}{2}+s} \|w\|_{H^{\frac{3}{2}+s}(\Omega)} \quad \textit{for all } w \in H^{\frac{3}{2}+s}(\Omega). \tag{1.5.2}$$

The above assumptions are in the form of standard approximation estimates for classical finite element spaces. We defer a thorough explanation to §1.6.

1.5.1 H^1 Estimates

We examine the error between u and u_h measured in the H^1. Indeed, we have the error, $u - u_h$ satisfying:

Lemma 1.5.3. *The approximate solutions u_h converge strongly to u w.r.t the H^1_0 topology and we have the rate*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant C h^{\frac{1}{2}+s} \|u\|_{H^{\frac{3}{2}+s}(\Omega)}, \tag{1.5.3}$$

where $C > 0$ is independent of h.

Proof. The error $u - u_h$ satisfies

$$a(u - u_h, v_h) + (B(u) - B(u_h), v_h) = 0 \quad \text{for any } v_h \in V_h, \tag{1.5.4}$$

and thus we have, for any $v_h \in V_h$,

$$\begin{aligned} a(u - u_h, u - u_h) &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h), \\ &= a(u - u_h, u - v_h) - (B(u) - B(u_h), v_h - u_h), \\ &= a(u - u_h, u - v_h) + (B(u) - B(u_h), u - v_h) - (B(u) - B(u_h), u - u_h). \end{aligned}$$

When then have the estimates

$$\begin{aligned} &\int_\Omega |\nabla(u - u_h)|^2 + [b(u + u_h) - c]\,|u - u_h|^2\, dx \\ &\qquad \leqslant M(\|u - u_h\|_{H^1(\Omega)} + \|B(u) - B(u_h)\|_{L^2(\Omega)})\|u - v_h\|_{H^1(\Omega)} \\ &\qquad \leqslant (M + L)\|u - u_h\|_{H^1(\Omega}\|u - v_h\|_{H^1(\Omega)}, \end{aligned}$$

and since u_h is strictly positive in Ω we have a coercivity constant $\alpha_h > 0$ such that

$$\alpha_h\|u - u_h\|^2_{H^1(\Omega)} \leqslant (M + L)\|u - u_h\|_{H^1(\Omega)}\|u - v_h\|_{H^1(\Omega)}.$$

After dividing through by $\alpha_h\|u - u_h\|_{H^1(\Omega)}$ and taking the infimum over all v_h in V_h we have the quasi-optimal error estimate

$$\|u - u_h\|_{H^1(\Omega)} \leqslant \frac{M + K}{\alpha_h} \inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}. \tag{1.5.5}$$

By Assumption 1.5.1, and the fact that $M + K$ is independent of h, all that is left for us to show is a uniform lower bound on α_h. Indeed, we know that u_h satisfies the pointwise lower bound

$$u_h \geqslant \frac{c - \lambda_h}{b}\phi_h,$$

where ϕ_h are the approximate principle eigenfunctions. Note that we have constants $\tilde{C}, \bar{C} > 0$ such that

$$\|\phi - \phi_h\|_{L^\infty(\Omega)} \leqslant \tilde{C}h^s\|\phi\|_{H^{\frac{3}{2}+s}(\Omega)} \quad \text{and} \quad |\lambda - \lambda_h| \leqslant \bar{C}h^{1+2s}, \tag{1.5.6}$$

(see, for example [15]) and by Assumption 1.3.3

$$\phi_h > 0 \quad \text{in } \Omega,$$

Thus, if we take h sufficiently small, we have

$$u_h \geqslant \frac{c-\lambda}{2b}\phi,$$

and thus there exists some constant $\tilde{\alpha} > 0$, which no longer depends on h, such that we have the uniform coercivity bound

$$\tilde{\alpha}\|w\|^2_{H^1(\Omega)} \leqslant \int_\Omega |\nabla w|^2 + [b(u+u_h) - c]\, w^2 dx \quad \text{for all } w \in H^1_0(\Omega),$$

for h sufficiently small. This allows us to replace α_h with $\tilde{\alpha}$ in (1.5.5) and conclude the proof. □

1.5.2 L^2 Estimates

Using the standard Aubin-Nitsche trick and the H^1-error estimates we prove above, we prove that L^2-error converges with a rate which is faster than that of the H^1-error. First, we state and prove a crucial result.

Corollary 1.5.4. *For any $w \in L^\infty$ such that $w \geqslant u$, $a(\cdot,\cdot) + (B'(w)\cdot,\cdot)$ is a coercive, bilinear form over H^1_0 and subsequently possesses a maximum principle.*

Proof. Using the identity

$$a(\cdot,\cdot) + (B'(w)\cdot,\cdot) = a((bu+g) - c;\cdot,\cdot),$$

where $g = bu + 2b(w-v) > 0$, we invoke Lemma 1.4.1 to achieve the desired result. □

Corollary 1.5.4, which asserts that linearizations about $w \geqslant u$ are invertible, allows us to prove the following result.

Lemma 1.5.5. *The error $u - u_h$ satisfies the following rate-of-convergence estimate:*

$$\|u - u_h\|_{L^2(\Omega)} \leqslant Ch^{1+2s}\|u\|_{H^{\frac{3}{2}+s}(\Omega)}, \tag{1.5.7}$$

where $C > 0$ is independent of h.

Proof. We first note that we have

$$(B(u) - B(u_h), w) = (B'(u)(u-u_h), w) - b\int_\Omega (u-u_h)^2 w dx \tag{1.5.8}$$

for any $w \in H_0^1(\Omega)$. Now, we let $w \in H_0^1(\Omega)$ be the unique solution to the following adjoint problem,

$$a(v,w) + (B'(u)v, w) = (u - u_h, v) \quad \text{for all } v \in H_0^1(\Omega), \tag{1.5.9}$$

where existence and uniqueness is a result of Lemma 1.5.4. Since u is sufficiently regular, we have an elliptic regularity result

$$\|w\|_{H^{\frac{3}{2}+s}(\Omega)} \leqslant C\|u - u_h\|_{L^2(\Omega)}.$$

Setting $v = u - u_h$ in (1.5.9) we have

$$\begin{aligned}
\|u - u_h\|_{L^2(\Omega)} &= a(u - u_h, w) + (B'(u)(u - u_h), w) \\
&= a(u - u_h, w) + (B(u) - B(u_h), w) + b\int_\Omega (u - u_h)^2 w \, dx \\
&= a(u - u_h, w - w_h) + (B(u) - B(u_h), w - w_h) + b\int_\Omega (u - u_h)^2 w \, dx \\
&\leqslant (M + K)\|u - u_h\|_{H^1(\Omega)}\|w - w_h\|_{H^1(\Omega)} + b\int_\Omega |u - u_h|^2|w|dx,
\end{aligned}$$

for any $w_h \in V_h$.

Since $n \leqslant 3$, we have the embedding

$$H^{\frac{3}{2}+s}(\Omega) \subset L^\infty(\Omega),$$

for $s > 0$, and thus we have

$$\|w\|_{L^\infty(\Omega)} \leqslant C\|u - u_h\|_{H^1(\Omega)} \leqslant Ch^{\frac{1}{2}+s}\|u\|_{H^{\frac{3}{2}+s}(\Omega)}.$$

Using this result, along with choosing w_h appropriately as in (1.5.1)

$$\|u - u_h\|^2_{L^2(\Omega)} \leqslant Ch^{\frac{1}{2}+s}\|u - u_h\|_{H^1(\Omega)}\|u - u_h\|_{L^2(\Omega)} + \tilde{C}h^{\frac{1}{2}+s}\|u\|_{H^{\frac{3}{2}+s}(\Omega)}\|u - u_h\|^2_{L^2(\Omega)}.$$

Then, by taking h sufficiently small so that

$$\tilde{C}h^{\frac{1}{2}+s}||u||_{H^{\frac{3}{2}+s}(\Omega)} < \frac{1}{2},$$

and dividing by $||u - u_h||_{L^2(\Omega)}$ we arrive at

$$||u - u_h||_{L^2(\Omega)} \leqslant 2Ch^{\frac{1}{2}+s}||u - u_h||_{H^1(\Omega)}.$$

Using the previously shown H^1 error estimate, we then arrive at

$$||u - u_h||_{L^2(\Omega)} \leqslant Ch^{1+2s}||u||_{H^{\frac{3}{2}+s}(\Omega)}.$$

□

1.5.3 L^∞ Estimates

Using the above error estimates, we prove an L^∞-error estimate using a simple inverse estimate. We expect (and we shall see in our experiments) that our rate of convergence is not optimal for domains which are convex. A number of results have been shown for linear and nonlinear problems where one can achieve $O(h^2|\log(h)|^{1+\frac{n}{4}})$ convergence but these proofs are rather technical. Moreover, these proofs rely on full H^2-elliptic regularity which is not generally the case for Lipschitz non-convex domains. With this said, we provide evidence that our estimate is sharp when the domain is non-convex. The technique of our simple proof is based upon a proof given in [16].

The following assumption is known to hold for most standard finite element spaces.

Assumption 1.5.6. *Let V_h be such that the following estimate holds:*

$$||v_h||_{L^\infty(\Omega)} \leqslant Ch^{1-\frac{n}{2}}||v_h||_{H^1(\Omega)} \quad \textit{for all } v_h \in V_h. \tag{1.5.10}$$

Lemma 1.5.7. *Under assumptions 1.5.1, 1.5.2 and 1.5.6 the error $u - u_h$ satisfies the following L^∞ rate-of convergence estimate:*

$$\|u - u_h\|_{L^\infty(\Omega)} \leqslant Ch^{\frac{3-n}{2}+s}\|u\|_{H^2(\Omega)}.$$

where $C > 0$ is independent of h.

Proof. To examine the L^∞ error, take $v_h \in V_h$ and we have

$$\begin{aligned}
\|u - u_h\|_{L^\infty(\Omega)} &\leqslant \|u - v_h\|_{L^\infty(\Omega)} + \|v_h - u_h\|_{L^\infty(\Omega)} \\
&\leqslant \|u - v_h\|_{L^\infty(\Omega)} + Ch^{1-\frac{n}{2}}\|v_h - u_h\|_{H^1(\Omega)} \\
&\leqslant \|u - v_h\|_{L^\infty(\Omega)} + Ch^{1-\frac{n}{2}}\left\{\|u - u_h\|_{H^1(\Omega)} + \|u - v_h\|_{H^1(\Omega)}\right\}.
\end{aligned}$$

Then, choosing v_h appropriately, we the desired result. □

We can summarize the last few results in the following theorem, the proof of which can be assembled from the results above.

Theorem 1.5.8. *Let $V_h \subset H_0^1(\Omega)$ be a sequence of subspaces satisfying Assumptions 1.3.1, 1.3.2, 1.3.3, 1.5.1, and 1.5.2 and let u_h be the unique, positive, discrete solution satisfying*

$$\int_\Omega \nabla u_h \cdot \nabla v_h + (bu_h^2 - cu_h)v_h \, dx = 0 \quad \textit{for all } v_h \in V_h. \tag{1.5.11}$$

Then, for h sufficiently small, we have the following orders of convergence

$$\|u - u_h\|_{H^1(\Omega)} = O(h^{\frac{1}{2}+s}),$$
$$\|u - u_h\|_{L^2(\Omega)} = O(h^{1+2s}),$$
$$\|u - u_h\|_{L^\infty(\Omega)} = O(h^{\frac{3-n}{2}+s}).$$

1.6 Numerical Results

We provide some numerical experiments to confirm our theoretical results. At this point, we only assume the existence of the finite dimensional approximating space V_h, which satisfy certain assumptions. Now, we explicitly define the spaces V_h which satisfy these assumptions. Indeed, given a domain Ω, let $\mathcal{T}_h$ denote a quasi-uniform triangularization (see, for example [17]). Further, let

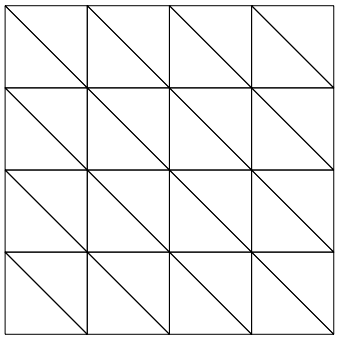
$$V_h^k = \{v \in H_0^1(\Omega) : v|_T \in \mathcal{P}_k(T), \forall\, T \in \mathcal{T}_h\}$$

be the standard finite element space defined on $\mathcal{T}_h$, where, for $k \geq 1$, $\mathcal{P}_k(T)$ denotes the space of polynomials of degree k on the triangle T. Requiring that the maximum angle of each triangle $T \in \mathcal{T}_h$ be less than $\frac{\pi}{2}$, we assert that the space V_h^1 satisfies our required assumptions. This is documented in the following result:

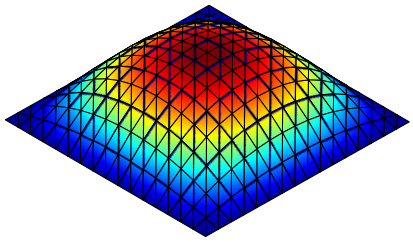
Lemma 1.6.1. *Let $\mathcal{T}_h$ be a quasi-uniform triangularization of Ω and suppose there exists some $\gamma > 0$ such that the maximum angle of each $T \in \mathcal{T}_h$ is less than or equal to $\frac{\pi}{2} - \gamma$. Then, the space V_h^1 satisfies Assumptions 1.3.1, 1.3.2, 1.3.3, 1.5.1, 1.5.2, and 1.5.6.*

Proof. It is easy to see that the standard choice of piecewise-linear Lagrange shape functions satisfies the requirements of Assumption 1.3.1. The fact that V_h^1 satisfies Assumption 1.3.2 depends on the requirement that the maximum angle of each triangle is less than $\frac{\pi}{2} - \gamma$ and is proven in [18]. Assumption 1.3.3 follows from (1.5.6), which is proven in [15], and the satisfaction of Assumption 1.3.2. Finally, the fact that V_h^1 satisfies Assumptions 1.5.1, 1.5.2 and 1.5.6 is a classical result that can be found in [17].

□



(a) The initial triangulation of the square domain. $h = \frac{\sqrt{2}}{4}$



(b) A computed FEM solution for $h = \frac{\sqrt{2}}{16}$

Figure 1.1: The Square domain $\Omega = [0,1] \times [0,1]$

To provide evidence for our theoretical results, we perform convergence studies on two domains in $\mathbb{R}^2$. Though our results establish the existence of the unique positive solution u_h by virtue of a monotone iteration, we find the convergence of the iteration to be quite slow in practice. Thus, we use a nonlinear Newton iteration which begins with the initial iterate $u^0 = \overline{u}^0 = \frac{c}{b}$. We find this to greatly accelerate the original monotone iteration, but we are unable to rigorously prove any results pertaining to this procedure.

The first domain we consider is the unit square $\Omega = [0,1] \times [0,1]$. Since this domain is a convex polygon, and possesses full H^2-elliptic regularity, our above results should hold with $s = \frac{1}{2}$. To show this, we start with an initial triangulation of the domain which is shown in Figure 1.1(a). Note that the maximum angle on this triangulation is equal to $\frac{\pi}{2}$, suggesting a loss in the maximum principle, but experiments show that there is no issue with angles of $\frac{\pi}{2}$. This suggests that the maximum angle condition of [18] is sufficient but may not be necessary.

We solve (1.3.1) with $b = 22, c = 15$ on a sequence of uniformly refined meshes. Since $15 = c > \lambda_1 = \pi^2$ (the principle eigenvalue of the Laplace operator on the region $[0,1] \times [0,1]$), we know that (1.2.1) has a unique positive solution u. However, a closed form solution is not known and we must compare with an approximate solution on an extremely refined mesh. Specifically, for the initial mesh we have $h_0 = \frac{\sqrt{2}}{4}$, and we solve for u_h for $h \in \{h_0, \frac{h_0}{2}, \frac{h_0}{4}, \frac{h_0}{8}, \frac{h_0}{16}, \frac{h_0}{32}, \frac{h_0}{64}\}$ and compare with an 'Exact' solution

using $h = \frac{h_0}{512}$.

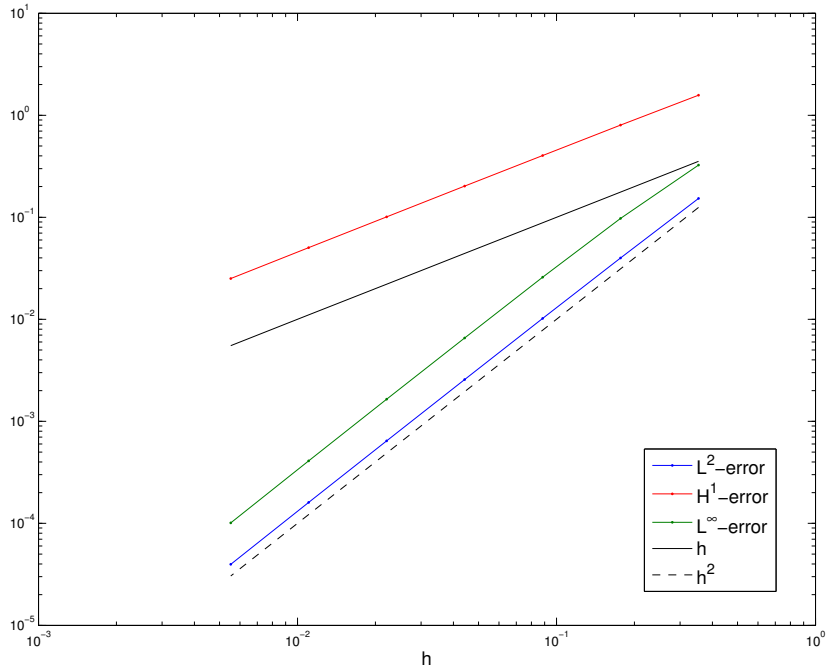
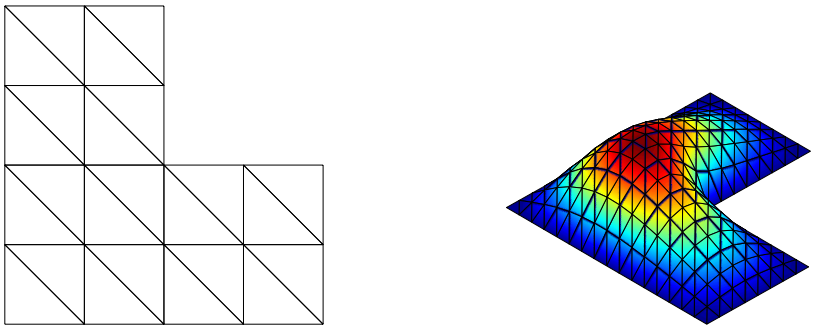


Figure 1.2: L^2, H^1, and L^∞-error plots for FEM solutions to the DLE posed on the square with $b = 22$,$c = 15$

In Figure 1.6 we plot the resulting errors, measured in L^2, H^1 and L^∞. We also include plots of the quantities h and h^2 for a basis for comparison. We see that our L^2 and H^1 rate-of-convergence results are demonstrated by these plots, but as we have discussed, the L^∞ rate is more like $O(h^2)$ than the rate $O(h)$ that our results would suggest.

The second domain we consider is the L-shaped domain $\Omega = [0,1] \times [0,1] \setminus [\frac{1}{2},1] \times [\frac{1}{2},1]$ for which we show an initial triangulation in Figure 1.3(a). As this domain is non-convex with a reentrant corner of $\theta = \frac{3\pi}{2}$, this domain only possesses $H^{\frac{7}{4}}$ regularity and thus our results suggest our estimates with $s = \frac{1}{4}$. For this domain, we perform the same experiments as for the square domain, with the same choices of parameters $b = 22, c = 15$. The principle eigenvalue of the L-shaped domain is quite close to that of the square, and thus $c = 15$ still admits a positive solution.

In Figure 1.6 we plot the resulting errors where we see that our experiments match up



(a) The initial triangulation of the L-shaped domain. $h = \frac{\sqrt{2}}{4}$

(b) A computed FEM solution for $h = \frac{\sqrt{2}}{16}$

Figure 1.3: The L-shaped domain $\Omega = [0,1] \times [0,1] \setminus [\frac{1}{2},1] \times [\frac{1}{2},1]$

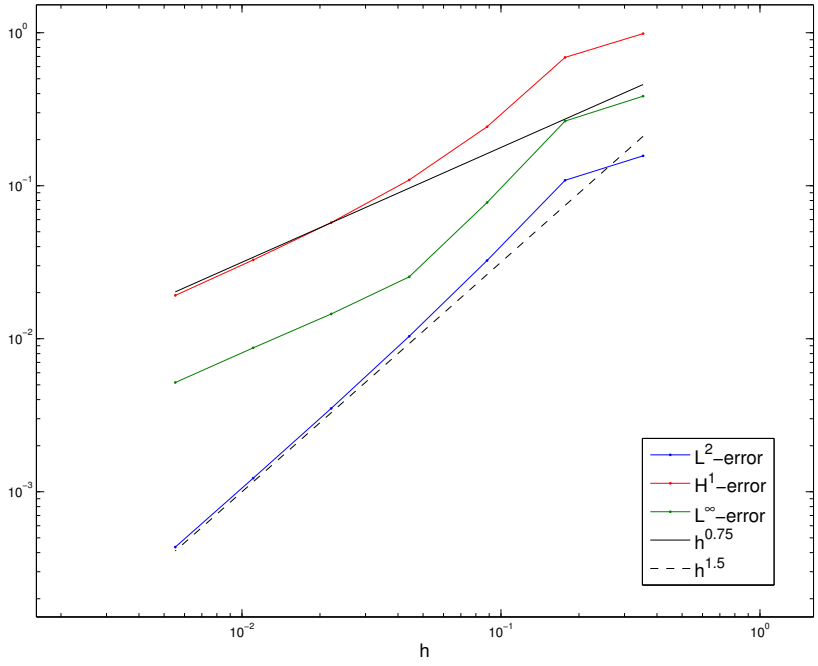


Figure 1.4: L^2, H^1, and L^∞-error plots for FEM solutions to the DLE posed on the L-shaped domain with $b = 22$,$c = 15$

quite well with our theoretical results. We further note that, unlike was the case for the square where our L^∞ convergence results were not sharp, for the L-shaped domain, our result does seem to be sharp. This suggests that the techniques used to prove improved rates of convergence do not work for the non-convex case.

1.7 Conclusion

In this article we considered the diffusive logistic equation as a model of more complex semilinear elliptic systems, and developed a number of new results on weak solutions and on their approximation by Galerkin-type methods. Our motivation was to establish a rate of convergence for Galerkin approximations to solutions of this problem, as a step towards developing techniques for proving rates of convergence for more general nonlinear problems. To this end, we first considered the continuous model, and briefly reviewed the known solution theory. We then proved a new basic result on existence and uniqueness of weak solutions. The proof involved several lemmas, combining fixed-point arguments, compactness techniques, and maximum principles. Under reasonable assumptions on the approximation spaces, using similar arguments we developd a discrete analogue for Galerkin approximations. Both the continuous and discrete results were critical building blocks for developing *a priori* error estimates for Galerkin approximations, and then subsequently using the estimates to characterizing the rate of convergence of such approximations in a precise way. The necessary error estimates were established by (again) exploiting (discrete) maximum principles, by analyzing the spectral structure of the linearized problem in some detail, and then by exploiting the subtle relationship to an auxiliary problem. Our final convergence result holds for finite element approximations to positive solutions of the diffusive logistic equation in bounded, non-convex polygonal domains in both two and three space dimensions. We showed that, under reasonable assumptions on the approximation spaces and on the details of the discretization, the Galerkin method converges at a fixed (optimal) rate of convergence. Our numerical experiments confirm the theoretical predictions.

Although our focus was on the diffusive logistics equation in this article, the techniques developed in the paper are quite general. In particular, the more general elliptic

problem

$$-\nabla \cdot (a\nabla u) + b(u) = f \quad \text{in } \Omega, \tag{1.7.1}$$

$$-n \cdot (a\nabla u) + c(u) = g_1 \quad \text{on } \partial_1\Omega, \tag{1.7.2}$$

$$u = g_0 \quad \text{on } \partial_0\Omega, \tag{1.7.3}$$

can be analyzed using similar techniques under reasonable assumptions on the functions a, b, c, f, g_0, and g_1. Moreover, certain classes of coupled elliptic systems can be handled using extensions of the arguments here; this will be pursued by the authors in a second article.

Chapter 1, in part, is currently being prepared for submission for publication of the material. The dissertation author was the primary investigator and author of this material. I would like to acknowledge the co-author, Michael Holst.

Chapter 2

A Space-Time Smooth Artificial Viscosity Method for Nonlinear Conservation Laws

2.1 Introduction

2.1.1 Smoothing conservation laws

The initial-value problem for a general nonlinear system of conservation laws can be written as an evolution equation,

$$\partial_t U(x,t) + \operatorname{div} F(U(x,t)) = 0 \text{ with } U|_{t=0} = U_0\,, \tag{2.1.1}$$

for an m-vector U defined on $(D+1)$-dimensional space-time. Such partial differential equations (PDE) are both ubiquitous and fundamental in science and engineering, and include the compressible Euler equations of gas dynamics, the magneto-hydrodynamic (MHD) equations modeling ionized plasma, the elasticity equations of solid mechanics, and numerous related physical systems which possess complicated nonlinear wave interactions.

It is well known that solutions of (2.1.1) can develop finite-time shocks, even when the initial data is smooth, in which case, discontinuities of U are propagated according to the so-called Rankine-Hugoniot conditions (see §2.2.1 below). It is important

to develop stable and robust numerical algorithms which can approximate shock-wave solutions. Even in one-space dimension, nonlinear wave interaction such as two shock waves colliding, is a difficult problem when considering accuracy, stability and monotonicity. The challenge is maintaining higher-order accuracy away from the shock while approximating the discontinuity in an order-Δx smooth transition region where Δx denotes the spatial grid size.

As we describe below, a variety of clever discretization schemes have been developed and employed, particularly in one-space dimension, to approximate discontinuous solution profiles in an essentially non-oscillatory (ENO) fashion. These include, but are not limited to, total variation diminishing (TVD) schemes, flux-corrected transport (FCT) schemes, weighted essentially non-oscillatory (WENO) schemes, discontinuous Galerkin methods, artificial diffusion methods, exact and approximate Riemann solvers, and a host of variants and combinations of these techniques.

We develop a robust parabolic-type regularization of (2.1.1), which we refer to as the C-method, which couples a modified set of m equations for U with an additional linear scalar reaction-diffusion equation for a new scalar field $C(x,t)$. Thus, instead of (2.1.1), we consider a system of $m+1$ equations, which use the solution $C(x,t)$ as a coefficient in a carefully chosen modification of the flux. As we describe in detail below, the solution $C(x,t)$ is highly localized in regions of discontinuity, and transitions smoothly (in both x and t) to zero in regions wherein the solution is smooth. Further, as $\Delta x \to 0$, we recover the original hyperbolic nonlinear system of conservation laws (2.1.1).

2.1.2 Numerical discretization

In the case of 1-D gas dynamics, the construction of non-oscillatory, higher-order, numerical algorithms such as ENO by Harten, Engquist, Osher & Chakravarthy [19] and Shu & Osher [20], [21]; WENO by Liu, Osher, & Chan [22] and Jiang & Shu [23]; MUSCL by Van Leer [24], Colella [25], and Huynh [26]; or PPM by Colella & Woodward [27] requires carefully chosen *reconstruction* and *numerical flux*.

Such numerical methods evolve cell-averaged quantities; to calculate an accurate approximation of the flux at cell-interfaces, these schemes reconstruct kth-order ($k \geq 2$)

polynomial approximations of the solution (and hence the flux) from the computed cell-averages, and thus provide kth-order accuracy away from discontinuities. See, for example, the convergence plots of Greenough & Rider [28] and Liska & Wendroff [29]. Given a polynomial representation of the solution, a strategy is chosen to compute the most accurate cell-interface flux, and this is achieved by a variety of algorithms. Centered numerical fluxes, such as Lax-Friedrichs, add dissipation as a mechanism to preserve stability and monotonicity. On the other hand, *characteristic-type* upwinding based upon exact (Godunov) or approximate (Roe, Osher, HLL, HLLC) Riemann solvers, which preserve monotonicity without adding too much dissipation, tend to be rather complex and PDE-specific; moreover, for strong shocks, other techniques may be required to dampen post-shock oscillations or to yield entropy-satisfying approximations (see Quirk [30]). Again, we refer the reader to the papers [28], [29] or Colella & Woodward [31] for a thorough overview, as well as a comparison of the effectiveness of a variety of competitive schemes.

Majda & Osher [32] have shown that *any* numerical scheme is *at best*, first-order accurate in the presence of shocks or discontinuities. The use of higher-order numerical schemes is, nevertheless, imperative for the elimination of error-terms in the Taylor expansion (in mesh-size) and the subsequent limiting of truncation error. Moreover, higher-order schemes tend to be less dissipative than there lower-order counterparts, as discussed by Greenough & Rider [28]; therein, a comparison between a 2nd-order PLMDE scheme and a 5th-order WENO scheme demonstrates the improved resolution of intricate fine structure afforded by 5th-order WENO, while simultaneously providing far less clipping of local extrema than PLMDE.

In multi-D, similar tools are required to obtain non-oscillatory numerical schemes, but the multi-dimensional analogues to those described above are generally limited by mesh considerations. For structured grids (such as products of uniform 1-D grids), dimensional splitting is commonly used, decomposing the problem into a sequence of 1-D problems. This technique is quite successful, but stringent mesh requirements prohibits its use on complex domains. Moreover, applications to PDE outside of variants of the Euler equations may be somewhat limited. For further discussion of the limitations of dimensional splitting, we refer the reader to Crandall & Majda [33], and Jiang &

Tadmor [34]. For unstructured grids, dimensional splitting is not available and alternative approaches must be employed, necessitated by the lack of multi-D Riemann solvers. WENO schemes on unstructured triangular grids have been developed in Hu & Shu [35], but using simplified methods, which employ reduced characteristic decompositions, can lead to a loss of monotonicity and stability.

Algorithms that explicitly introduce diffusion provide a simple way to stabilize higher-order numerical schemes and subsequently remove non-physical oscillations near shocks. In the mathematical analysis of conservation laws (and in the truncation error of certain discretization schemes), the simplest parabolic-regularization is by the addition of a uniform linear viscosity. Choosing a constant $\beta > 0$, which depends upon mesh-size Δx and sometimes velocity or wave-speed, and adding

$$\beta(\Delta x)\partial_x^2 U(x,t) \tag{2.1.2}$$

to the right hand side of (2.1.1) provides a uniformly parabolic regularization of the hyperbolic conservation laws, and its discrete implementation smears sharp discontinuities across $O(\Delta x)$-regions and thus adds stabilization, but unfortunately, at the cost of accuracy. With the addition of uniform linear viscosity, shocks and discontinuities are captured in a non-oscillatory fashion, but the transition region from left to right state, which approximates the discontinuity, tends to grow over time. Moreover, since viscosity is applied uniformly over the *entire* domain $\mathcal{I}$, the benefits of a higher-order scheme (away from the discontinuity) may be lost, and the accuracy may reduce to merely first-order (at best). For practical implementation in a numerical scheme, the use of viscosity should be localized in regions of shock (so as to stabilize the scheme), limited at contact discontinuities (to avoid over-smearing the sharp transition), and very small in smooth regions away from discontinuities. Achieving these requirements allows higher-order approximation of smooth flow and sharp, non-oscillatory, resolution of shocks and discontinuities. Naturally, this necessitates that the amount of added viscosity be a function of the solution.

The pioneering papers of Richtmyer [36], Von Neumann & Richtmyer [37], Lax & Wendroff [38], and Lapidus [39] suggest the introduction of nonlinear artificial viscosity to equations (2.1.1) in a form similar to the following expression:

$$\beta(\Delta x)^2 \partial_x \left(|\partial_x u(x,t)| \, \partial_x U(x,t) \right). \tag{2.1.3}$$

We refer the reader to the classical papers of Gentry, Martin, & Daly [40] and Harlow & Amsden [41] for an interesting discussion on artificial viscosity. Specifically, Gentry, Martin, & Daly [40] define the nonlinear viscosity of the type (2.1.3) to be artificial viscosity, and show that the linear viscosity (2.1.2), scaled by the magnitude of local velocity, arises as truncation error (in finite-difference approximations). The latter is responsible for stabilizing the *transport* of sound waves, while (2.1.3) stabilizes the *steepening* of sound waves.[1]

We are primarily concerned with the steepening of sound waves, and shall term artificial viscosity of the type (2.1.3) as *classical artificial viscosity*. Formally, the use of (2.1.3) produces the required amount of viscosity near shocks but allows for second-order accuracy in smooth regions. On the other hand, the diffusion coefficient $|\partial_x u(x,t)|$ is precisely the quantity which loses regularity (or smoothness) near shock discontinuities. Also, the constant β must be larger than one to control numerical oscillations behind the shock wave, which in turn overly diffuses the waves and produces incorrect wave speeds.

Alternative procedures have been proposed. For streamline upwind Petrov-Galerkin schemes (SUPG), Hughes & Mallet [42] and Shakib, Hughes, & Johan [43] use residual-based artificial viscosity. Guermond & Pasquetti [44] present a similar, entropy-residual-based scheme for use in spectral methods. Persson & Peraire [45] develop a method based upon decay of local interpolating polynomials for discontinuous Galerkin schemes. Later, Barter & Darmofal [46] use a reaction-diffusion equation to provide a regularized variant of this approach.

Our approach is similar to [46] in that it uses a reaction-diffusion equation to calculate a smooth distribution of artificial viscosity. Instead of regularizing a DG-based noise-indicator that allows for the growth of viscosity near shocks, we regularize the classical artificial viscosity of [39], using a gradient-based approach for this source term. This approach yields both a discretization-independent and PDE-independent methodology which can be generalized to multiple dimensions by regularizing a similar viscosity to that in Löhner, Morgan, & Peraire [47].

In 1-D, our approach proves to be a simple way of circumventing the need for char-

[1] We are indebted to the anonymous referee for clarifying this point for us.

acteristic or other *a priori* information of the exact solution to remove oscillations in higher-order schemes. Due to the simple and discretization-independent nature of our method, we expect our methodology to be useful for a wide range of applications.

2.1.3 Outline of the paper

In §2.2, we introduce the C-method for the compressible Euler equations in one space dimension. We show that the C-method is Galilean invariant and that solutions of the C-method converge to the entropy solutions of the Euler equations in the limit of zero mesh size. We also show the relative smoothness of our new viscosity coefficient with respect to the classical artificial viscosity of Richtmyer and Von Neumann, and we demonstrate the ability of the C-method to remove downstream oscillation in slowly moving shocks.

In §2.3, we give a brief outline of the numerical schemes whose solutions are used in this paper. First, we outline a second-order, continuous Galerkin finite-element method. Second, we outline a simple WENO-based finite-volume scheme which performs upwinding using only the sign of the velocity (no Riemann-solvers or characteristic decompositions in primitive variables). The resulting schemes applied to the C-method are referred to as FEM-C and WENO-C, respectively. Third, we outline the central-finite-difference scheme of Nessyahu and Tadmor (NT), a simple scheme, easily generalizable to multi-D [48]. Like our FEM-C scheme, the NT-scheme is at best, second-order, and does not require specialized techniques for upwinding. Fourth, we outline a Godunov-type characteristic decomposition-based WENO scheme (WENO-G) developed by Rider, Greenough & Kamm [49] which utilizes a variant of a Godunov/Riemann-solver as upwinding, providing a very competitive scheme for modeling the collision of very strong shocks.

In §2.4, we consider the classical shock-tube problem of Sod. With the Sod shock problem, we apply our FEM-C scheme and compare with the classical viscosity approach. We then compare the FEM-C scheme with the two standalone methods, NT and WENO-G.

In §2.5, we consider the moderately difficult problem of Osher-Shu, modeling the interaction of a mild shock with an entropy wave. We compare FEM-C to NT and

WENO-G in which the differences are more significant than in the Sod-shock comparisons. We show that WENO-C compares well with WENO-G; on the other hand, the simple WENO scheme without the C-method and without the Gudonov-based characteristic solver also does well in modeling the Osher-Shu test case.

In §2.6, we consider the numerically challenging Woodward-Colella blast wave simulation, which models the collision of two strong interacting shock fronts. Though the FEM-C scheme performs better than NT, both second-order schemes are somewhat outperformed by the higher-order WENO-G method (with characteristic solver). On the other hand, WENO-C compares well with WENO-G, having slightly less damped amplitudes with the same shock resolution.

Finally, in §2.7, we consider the Leblanc shock-tube, an extremely difficult test case consisting of a very strong shock. For this problem, we devise two strategies to demonstrate the use of the C-method. In the first strategy, we use our simplified WENO-C scheme with a right-hand side term for the energy equation that relies on a second C-equation which smooths gradients of E/ρ. We obtain an excellent approximation of the notoriously difficult contact discontinuity for internal energy, while maintaining an accurate shock speed; simultaneously, we avoid generating large overshoots at the contact discontinuity, which would indeed occur without the use of the C-method. For our second strategy, we show that WENO with the Lax-Friedrichs flux can be significantly improved with the addition of the C-method. We call this algorithm WENO-LF-C, and show that by using just one C-equation (as we have for all of the other test cases), we can sharply resolve the contact discontinuity for the internal energy, with accurate wave speed, and without overshoots.

2.2 The C-method

We begin with a description of the 1-D compressible Euler equations, written as a 3x3 system of conservation laws. We then explain our parabolic regularization, which we call the C-method.

2.2.1 Compressible Euler equations

The compressible Euler equations set on a 1-D space domain $\mathcal{I} \subset \mathbb{R}$, and a time interval $[0, T]$ are written in vector-form as the following coupled system of nonlinear conservation laws:

$$\partial_t \mathbf{u}(x,t) + \partial_x \mathbf{F}(\mathbf{u}(x,t)) = 0, \qquad x \in \mathcal{I}\,, t > 0, \tag{2.2.1a}$$

$$\mathbf{u}(x,0) = \mathbf{u}_0(x), \quad x \in \mathcal{I}\,, t = 0, \tag{2.2.1b}$$

where the 3-vector $\mathbf{u}(x,t)$ and *flux function* $\mathbf{F}(\mathbf{u}(x,t))$ are defined, respectively, as

$$\mathbf{u} = \begin{pmatrix} \rho \\ m \\ E \end{pmatrix} \quad \text{and} \quad \mathbf{F}(\mathbf{u}) = \begin{pmatrix} m \\ \frac{m^2}{\rho} + p \\ \frac{m}{\rho}\,(E+p) \end{pmatrix},$$

and

$$\mathbf{u}_0(x) = \begin{pmatrix} \rho_0(x) \\ m_0(x) \\ E_0(x) \end{pmatrix}$$

denotes the initial data for the problem. The variables $\rho,\ m$, and E denote the *density*, *momentum*, and *energy density* of a compressible gas, while $p = \mathcal{P}(\rho, m, E)$ denotes the *pressure* function. It is necessary to choose an equation-of-state $\mathcal{P}(\rho, m, E)$, and we use the ideal gas law, for which

$$p = (\gamma - 1)\left(E - \frac{m^2}{2\rho}\right), \tag{2.2.2}$$

where γ denotes the adiabatic constant. The equations (2.2.1) are indeed conservation laws, as they represent the conservation of mass, momentum, and energy in the evolution of a compressible gas. The velocity field $u(x,t)$ is obtained from momentum and density via the identity

$$u = \frac{m}{\rho}\,.$$

Inverting the relation (2.2.2), we see that the energy density E is a sum of kinetic and potential energy density functions:

$$E = \underbrace{\frac{\rho\, u^2}{2}}_{\text{kinetic}} + \underbrace{\frac{p}{\gamma - 1}}_{\text{potential}}\,.$$

The gradient (or Jacobian) of the flux vector $\mathbf{F}(\mathbf{u})$ is given by

$$D\mathbf{F}(\mathbf{u}) = \begin{bmatrix} 0 & 1 & 0 \\ \frac{(\gamma-3)m^2}{2\rho^2} & \frac{(3-\gamma)m}{\rho} & \gamma-1 \\ -\gamma\frac{Em}{\rho^2} + (\gamma-1)\frac{m^3}{\rho^3} & \frac{\gamma E}{\rho} + (1-\gamma)\frac{3m^2}{2\rho^2} & \frac{\gamma m}{\rho} \end{bmatrix}$$

with eigenvalues

$$\lambda_1 = u + c\,,\ \lambda_2 = u\,,\ \lambda_3 = u - c\,, \tag{2.2.3a}$$

where c denotes the sound speed (see, for example, Toro [50]). These eigenvalues determine the wave speeds.

The behavior of the various wave patterns is greatly influenced by the speed of propagation; as such, we define the maximum *wave speed* to be

$$[S(\mathbf{u})](t) = \max_{i=1,2,3} \max_{x\in\mathcal{I}} \{|\lambda_i(x,t)|\}\ . \tag{2.2.3b}$$

We are interested in solutions with shock waves and contact discontinuities. The Rankine-Hugoniot (R-H) conditions determine the speed s of the moving shock discontinuity, as well as the speed of a contact discontinuity. For a shock wave discontinuity, the R-H condition can be stated

$$F(\mathbf{u}_l) - F(\mathbf{u}_r) = s(\mathbf{u}_l - \mathbf{u}_r)$$

where the subscript l denotes the state to the left of the discontinuity, and the subscript r denotes the state to the right of the discontinuity. In general, the following three jump conditions must hold:

$$\begin{aligned} m_l - m_r &= s(\rho_l - \rho_r) \\ \left(\frac{(3-\gamma)m_l^2}{2\rho_l^2} + (\gamma-1)E_l\right) - \left(\frac{(3-\gamma)m_r^2}{2\rho_r^2} + (\gamma-1)E_r\right) &= s(m_l - m_r) \\ \left(\gamma\frac{E_l m_l}{\rho_l} - \frac{\gamma-1}{2}\frac{m_l^3}{\rho_l^2}\right) - \left(\gamma\frac{E_r m_r}{\rho_r} - \frac{\gamma-1}{2}\frac{m_r^3}{\rho_r^2}\right) &= s(E_l - E_r)\,. \end{aligned}$$

There can be non-uniqueness for weak solutions that have jump discontinuities, unless entropy conditions are satisfied (see the discussion in §2.2.9). So-called viscosity solutions $\mathbf{u}_{vis}$ are known to satisfy the entropy condition (and are hence unique) and are

defined as the limit as $\epsilon \to 0$ of a sequence of solutions $\mathbf{u}^\epsilon$ to the following parabolic equation:

$$\partial_t \mathbf{u}^\epsilon + \partial_x \mathbf{F}(\mathbf{u}^\epsilon) = \epsilon \partial_{xx} \mathbf{u}^\epsilon, \quad t > 0, \tag{2.2.4a}$$

$$\mathbf{u}^\epsilon = \mathbf{u}_0, \qquad t = 0\,. \tag{2.2.4b}$$

In the isentropic setting, for bounded initial data $\mathbf{u}_0$ with bounded variation, solutions $\mathbf{u}^\epsilon$ converge to the entropy solution $\mathbf{u}_{vis}$ of (2.2.1) as $\epsilon \to 0$ (see DiPerna [51] and Lions, Perthame, & Souganidis [52]). For non-isentropic dynamics, the same result holds if the initial data has small total variation (see Bianchini & Bressan [53]). Moreover, if the initial data $\mathbf{u}_0$ is regularized, then solutions to (2.2.4) are smooth in both space and time, and the discontinuity is approximated by a smooth function, transitioning from the left-state to the right-state over an interval whose length is $O(\epsilon)$.

Some of the classical finite-differencing schemes, such as the Lax-Friedrichs discretization, is dissipative to second-order and effectively behaves as a discrete version of (2.2.4). The uniform nature of such diffusion does not distinguish between discontinuous and smooth flow regimes, and thus adds unnecessary dissipation in regions of the wave profile which do not require any numerical stabilization. Such uniform dissipation contributes to a non-physical damping of entropy waves as well as over-diffusion and smearing of contact discontinuities, and may lead to errors in wave speeds.

2.2.2 Classical artificial viscosity

The idea of adding localized artificial viscosity to capture discontinuous solution profiles in numerical simulations dates back to Richtmyer [36], Von Neumann & Richtmyer [37], Lax & Wendroff [38], Lapidus [39] and a host of other reseachers. The idea behind *classical artificial viscosity* is to refine the uniform viscosity on the right-hand side of equation (2.2.4a) with

$$\partial_t \mathbf{u}^\epsilon + \partial_x \mathbf{F}(\mathbf{u}^\epsilon) = \beta \epsilon^2 \partial_x (|\partial_x u^\epsilon| \partial_x \mathbf{u}^\epsilon), \quad t > 0\,, \tag{2.2.5}$$

for a suitably chosen constant $\beta > 0$, which may depend upon the numerical discretization scheme.

When the velocity u exhibits a jump discontinuity (i.e., at a shock), the quantity $|\partial_x u^\epsilon|$ is $O(\frac{1}{\epsilon})$; however, away from shocks, where the velocity is smooth, $|\partial_x u^\epsilon|$ remains uniformly bounded in ϵ, and in such smooth regions, (2.2.5) adds significantly less viscosity than (2.2.4a). On the other hand, as we shall demonstrate in Figure 2.1, the use of $|\partial_x u^\epsilon|$ as a coefficient in the smoothing operator, can lead to spurious oscillations in the solution, caused by the lack of regularity in the quantity $|\partial_x u^\epsilon|$.

Formally, the use of the localizing coefficient $|\partial_x u^\epsilon|$ corrects for the over-dissipation of the uniform viscosity in (2.2.4), and a variety of numerical methods have employed some variant of this idea, achieving methods that are nominally non-oscillatory near shocks while maintaining second-order accuracy away from shocks. However, as we have already noted, the quantity $|\partial_x u^\epsilon|$ may become highly irregular near shock discontinuities, and may thus require setting the constant $\beta \gg 1$ in order to stabilize incipient numerical oscillations (see §2.4 for evidence to this observation). While this increase in β does not effect the asymptotic accuracy of the scheme, it is clearly beneficial to take β as small as possible to preserve the correct wave amplitude and wave speed.

The loss of regularity of the coefficient $|\partial_x u^\epsilon|$ suggests that a smoothed version of $|\partial_x u^\epsilon|$ would greatly benefit the dynamics. Smoothing $|\partial_x u^\epsilon|$ in space is not sufficient, as we must ensure smoothness in time as well. Hence, we propose our C-method, which indeed provides a regularized version of (2.2.5) and allows for the use of much smaller values of β (less localized artificial dissipation), higher accuracy, and practical viability.

2.2.3 C-method for compressible Euler

Analogous to (2.2.5), we control the amount of viscosity in (2.2.4a) by the use of a function $C^\epsilon(x,t)$ of space and time, and parameterized by $\epsilon := \Delta x > 0$. This function $C^\epsilon(x,t)$ is the solution to a reaction-diffusion equation which is forced by normalized modulus of the gradient of $\mathbf{u}^\epsilon$; the diffusion mechanism smooths the rough diffusion coefficient, while the reaction mechanism tries to minimize the support of spatial support of C^ϵ.

For fixed $\mathbf{u}_0$ we choose $\beta > 0$ to be $O(1)$. Then, for each $\epsilon > 0$, we let

$$\mathbf{u}^\epsilon(x,t) = \begin{pmatrix} \rho^\epsilon(x,t) \\ m^\epsilon(x,t) \\ E^\epsilon(x,t) \end{pmatrix} \quad \text{and} \quad C^\epsilon(x,t)$$

denote the solution of the following parabolic system of (viscous) conservation laws:

$$\partial_t \mathbf{u}^\epsilon + \partial_x \mathbf{F}(\mathbf{u}^\epsilon) = \partial_x \left(\tilde{\beta} \epsilon^2 C^{\epsilon,\delta} \partial_x \mathbf{u}^\epsilon \right), \quad t > 0, \tag{2.2.6a}$$

$$\partial_t C^\epsilon - \epsilon\, S(\mathbf{u}^\epsilon) \partial_x^2 C^\epsilon + \frac{S(\mathbf{u}^\epsilon)}{\epsilon} C^\epsilon = S(\mathbf{u}^\epsilon) G(\partial_x u^\epsilon), \qquad t > 0, \tag{2.2.6b}$$

$$(\mathbf{u}^\epsilon, C^\epsilon) = (\mathbf{u}_0^\epsilon, G(\partial_x u_0^\epsilon)), \qquad t = 0, \tag{2.2.6c}$$

where $C^{\epsilon,\delta} = C^\epsilon + \delta$ for a *fixed* positive constant $0 < \delta < \Delta x$, and $\tilde{\beta} = \beta \frac{\max\limits_{\mathcal{I}} |\partial_x u^\epsilon|}{\max\limits_{\mathcal{I}} C^\epsilon}$. The forcing to equation (2.2.6b) is defined as

$$G(\partial_x u^\epsilon) = \frac{|\partial_x u^\epsilon|}{\max\limits_{\mathcal{I}} |\partial_x u^\epsilon|}, \tag{2.2.7}$$

$S(\mathbf{u}^\epsilon)$ is defined by (2.2.3), and $\mathbf{u}_0^\epsilon$ denotes a regularization of the initial data which we discuss below. We also note that the scaling factor in $\tilde{\beta}$, given by $\frac{\max\limits_{\mathcal{I}} |\partial_x u^\epsilon|}{\max\limits_{\mathcal{I}} C^\epsilon}$, is included only to make comparisons with the classical artificial viscosity approach, but is in no way necessary.

2.2.4 Regularization of initial data for use with FEM-C

Unlike numerical algorithms which advance cell-averaged quantities, the finite-element method relies upon polynomial interpolation of nodal values, and requires solutions to be continuous across element boundaries in order for the interpolation to converge. As such, the use of discontinuous initial data produces Gibbs-type oscillations, at least on very short time intervals. To avoid this spurious behavior, it is advantageous to smooth discontinuous initial profiles when using a finite-elements.

More specifically, we provide a hyperbolic-tangent smoothing for initial data $\mathbf{u}_0^\epsilon$ for our FEM-C scheme. Since pointwise evaluation is well-defined for smooth functions, the finite-element discretization scheme can interpolate the regularized data and generate appropriate initial states.

For an interval $[a, b]$, we denote the indicator function

$$\mathbf{1}_{[a,b]}(x) = \begin{cases} 1, & x \in [a, b], \\ 0, & x \notin [a, b], \end{cases} \tag{2.2.8}$$

and consider initial conditions with components of the form

$$(\mathbf{u}_0(x))^i = \sum_{j=1}^{L_i} \mathbf{1}_{[a_j^i, b_j^i]}(x) f_j^i(x),$$

where the collection $\left\{[a_j^i, b_j^i]\right\}_{j=1}^{L_i}$ is pairwise disjoint,

$$\bigcup_{j=1}^{L_i} [a_j^i, b_j^i] = [a, b], \quad \text{for all } i = 1, 2, \ldots m,$$

and each of the f_j^i's are smooth. The i-th component of $\mathbf{u}_0$ is subsequently smooth over each of the L_i intervals, but may contain jump discontinuities at the boundaries of the regions $[a_j^i, b_j^i]$.

We then define the regularized initial condition

$$(\mathbf{u}_0^\epsilon(x))^i = \sum_{j=1}^{L_i} \mathbf{1}^\epsilon_{[a_j^i, b_j^i]}(x) f_j^i(x),$$

where

$$\mathbf{1}^\epsilon_{I_j^i}(x) = \frac{1}{2}\left[\tanh\left(\frac{x - a_j^i}{\epsilon}\right) - \tanh\left(\frac{x - b_j^i}{\epsilon}\right)\right].$$

This regularization procedure achieves approximations of exponential-order away from discontinuities; near discontinuities, it is a first-order approximation, when measured in the L^1-norm. Specifically, if $(\mathbf{u}_0)^i$ is smooth in $\omega \subset \mathcal{I}$, then the $L^1(\omega)$-norm of the error

$$\| (\mathbf{u}_0)^i - (\mathbf{u}_0^\epsilon)^i \|_{L^1(\omega)} = \int_\omega \left| (\mathbf{u}_0(x))^i - (\mathbf{u}_0^\epsilon(x))^i \right| \, dx = O(\epsilon^p) \tag{2.2.9}$$

for any positive integer p. Alternatively, if $\mathbf{u}_0^i$ is discontinuous somewhere in $\Omega \subset \mathcal{I}$, the $L^1(\Omega)$-norm of the error

$$\| \left(\mathbf{u}_0^i\right) - (\mathbf{u}_0^\epsilon)^i \|_{L^1(\Omega)} = O(\epsilon). \tag{2.2.10}$$

These observations assert that our regularization of the initial data allows for higher-order approximation of the initial data and is analogous to the averaging procedure required by Majda & Osher [32].

2.2.5 A compressive modification of the forcing G in the C-equation

The function G in (2.2.7) is chosen in such a manner so that C^ϵ is large where there are sharp transitions in the velocity field $u^\epsilon(x,t)$. In compressive regions (i.e., where $\partial_x u^\epsilon < 0$), sharp transitions over lengths of $O(\epsilon)$ correspond to shocks and artificial viscosity is required so that $\mathbf{u}^\epsilon$ remains smooth. In expansive regions, corresponding to rarefactions, artificial viscosity is not generally necessary.

These observations motivate the following alternative forcing function:

$$G_{comp}(\partial_x u^\epsilon) = \frac{|\partial_x u^\epsilon|}{\max\limits_{\mathcal{I}} |\partial_x u^\epsilon|} \mathbf{1}_{(-\infty,0)}\left(\partial_x u^\epsilon\right) \tag{2.2.11}$$

where the indicator function $\mathbf{1}_{(-\infty,0)}$ introduces viscosity only in regions of compression.

The ability to use such a switch is heavily dependent on the use of a space-time smoothing. Since the velocity in many numerical schemes may become oscillatory near shocks, such a switch can become discontinuous between adjacent cells/elements. However, the space-time nature of the C-equation resolves this issue, providing a smooth artificial viscosity profile.

This modified function G_{comp} typically increases accuracy in Euler simulations, but can lead to a loss of stabilization. For our FEM-C approach, where the stabilizing effects of artificial viscosity are necessary to dampen noise, the use of G_{comp} is restricted to the problems of Sod and Osher-Shu, which contain only moderately strong shocks.

2.2.6 Moving to the discrete level

The use of the C-equation yields smooth solutions $\mathbf{u}^\epsilon$ and thus we expect that a variety of higher-order discretization techniques, with sufficiently small Δt and Δx, could provide accurate, non-oscillatory approximations. In our implementation, artificial viscosity spreads discontinuities over regions of size $O(\beta\epsilon)$. Thus, given a particular initial condition, final time, discretization scheme, etc., we choose $\beta > 0$ such that the scaling $\epsilon = \Delta x$ produces non-oscillatory profiles.

We also note that the initial condition for C^ϵ, given in (2.2.6c) is chosen so to guarantee the coefficients of diffusion in (2.2.6a) are smooth up to $t = 0$. Moreover, choosing

such initial conditions for C^ϵ allows one to recover the classical artificial viscosity as $\epsilon \to 0$. As stated, these initial conditions may require a smaller time-step (by a factor of 10) for the first few time-steps. In practice, taking $C^\epsilon \equiv 0$ is an effective simplification to eliminate the need for smaller initial time-steps. Alternatively, we can solve an elliptic PDE for C^ϵ at the initial time and similarly eliminate that concern.

2.2.7 The C-method under a Galilean-transformation

We begin our discussion for the case of constant entropy. The Galilean invariance of the isentropic Euler equations results from the advective nature of the PDE. Since we solve a modified equation (coupled with the additional C-equation) it is of interest to know to what extent Galilean invariance is preserved. For simplicity, we assume that

$$p(x,t) = \rho(x,t)^2 .$$

(The choice $\gamma = 2$ corresponds to the shallow water equations, but any other choice of $\gamma > 1$ works in the same fashion.)

Given a fixed $v \in \mathbb{R}$ we define the change in independent variables

$$\tilde{x} = x - vt, \quad \tilde{t} = t,$$

denoting $\phi(\tilde{x},\tilde{t}) = (x,t)$ and the analogous change in the dependent variables

$$\tilde{\rho}(\tilde{x},\tilde{t}) = \rho(\tilde{x} + v\tilde{t},\tilde{t}), \quad \tilde{u}(\tilde{x},\tilde{t}) = u(\tilde{x} + v\tilde{t},\tilde{t}) - v. \tag{2.2.12}$$

A simple calculation yields

$$\partial_{\tilde{t}}\tilde{\rho} + \partial_{\tilde{x}}(\tilde{\rho}\tilde{u}) = [\partial_t\rho + \partial_x(\rho u)] \circ \phi, \tag{2.2.13a}$$

$$\partial_{\tilde{t}}(\tilde{\rho}\tilde{u}) + \partial_{\tilde{x}}(\tilde{\rho}\tilde{u}^2 + \tilde{p}) = \left[\partial_t(\rho u) + \partial_x(\rho u^2)\right] \circ \phi + \partial_{\tilde{x}}\tilde{p} - v\,[\partial_t\rho + \partial_x(\rho u)] \circ \phi. \tag{2.2.13b}$$

We further have that

$$\tilde{p}(\tilde{x},\tilde{t}) = p(\tilde{x} + v\tilde{t},\tilde{t}), \tag{2.2.14}$$

so that the mass and momentum equations are, in fact, Galilean invariant in the absence of artificial viscosity.

With the C-method employed, (2.2.13) transforms to

$$\partial_{\tilde{t}}\tilde{\rho} + \partial_{\tilde{x}}(\tilde{\rho}\tilde{u}) = [\partial_x(\mathcal{C}\partial_x\rho)] \circ \phi, \tag{2.2.15a}$$

$$\partial_{\tilde{t}}(\tilde{\rho}\tilde{u}) + \partial_{\tilde{x}}(\tilde{\rho}\tilde{u}^2 + \tilde{p}) = (\partial_x\{\mathcal{C}\partial_x[\rho(u-v)]\}) \circ \phi, \tag{2.2.15b}$$

where we let $\mathcal{C} = \epsilon^2\tilde{\beta}C$, and drop the ϵ superscript for notational convenience.

Examining (2.2.6b), we see that the equation for C is not Galilean invariant, but this is not a physical quantity, but can rather be viewed as a parameter to the modified system of conservation laws. As such we *define* the behavior of C under Galilean transformations as follows:

$$\tilde{\mathcal{C}}(\tilde{x}, \tilde{t}) = \mathcal{C}(\tilde{x} + v\tilde{t}, \tilde{t}).$$

With this definition of $\tilde{C}$, we find that

$$\partial_{\tilde{t}}\tilde{\rho} + \partial_{\tilde{x}}(\tilde{\rho}\tilde{u}) = \left[\partial_{\tilde{x}}(\tilde{\mathcal{C}}\partial_{\tilde{x}}\tilde{\rho})\right], \tag{2.2.16a}$$

$$\partial_{\tilde{t}}(\tilde{\rho}\tilde{u}) + \partial_{\tilde{x}}(\tilde{\rho}\tilde{u}^2 + \tilde{p}) = \left\{\partial_{\tilde{x}}[\tilde{\mathcal{C}}\partial_{\tilde{x}}(\tilde{\rho}\tilde{u})]\right\}, \tag{2.2.16b}$$

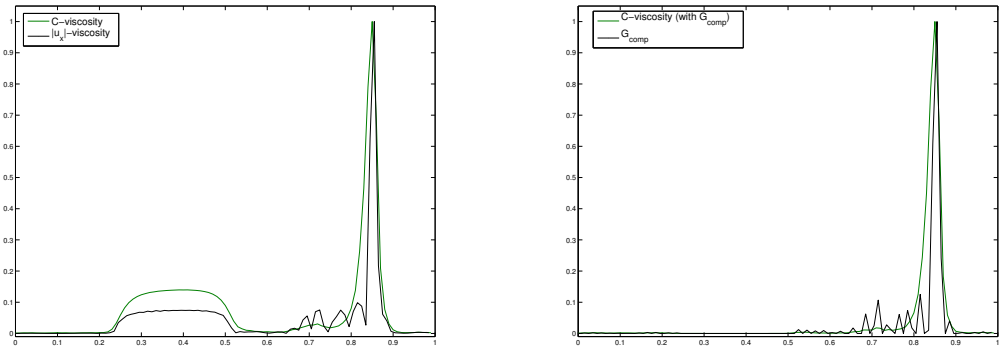
and hence the C-method for isentropic Euler retains the Galilean invariance.

We remark that in the absence of artificial viscosity on the right-hand side of the mass equation, the artificial flux term in the momentum equation is modified according to (2.3.7) below. This modification ensures Galilean invariance when the mass equation is left unchanged, which is the strategy employed for our WENO-C scheme.

Next, since the Galilean symmetry is for smooth solutions (for which classical derivatives are well-defined), and since smooth velocity fields simply transport the entropy function, it is thus a consequence of the transport of entropy, that Galilean invariance holds for the non-isentropic case as well. The importance of a numerical approximation to capture the Galilean invariant solution is fundamental to the initiation of the Kelvin-Helmholtz instability and other basic instabilities present in the Euler equation; see Robertson, Kravtsov, Gnedin, and Rudd [54] for a thorough discussion. In this connection, we next examine long wavelength instabilities which can arise for very slowly moving shock waves.

2.2.8 Regularization through the C-equation

It is of interest to examine the relative smoothness of C (we drop the superscript ϵ) to its rough counterpart $|u_x|$, and to determine the effect of this smoothing relative to the classical artificial viscosity approach. In Figure 2.1 we provide two plots demonstrating the effect of the C-method. In Figure 2.1(a) we see that the C-equation provides a smoothened viscosity profile compared to the classical approach. Alternatively, in Figure 2.1(b) we plot C using the compression-switch modification G_{comp} versus using purely the quantity G_{comp} (not smoothed by the C-equation) as a viscosity. In both cases we see how the C-method provides a far smoother profile with roughly the same magnitude as the non-smoothened approach.



(a) The C-equation with $G(u_x)$ (b) The C-equation with $G_{comp}(u_x)$

Figure 2.1: A comparison of the artificial viscosity profile produced by the C-method and the classical Richtmyer-type approach for the Sod shock tube at $t = 0.2$.

A useful feature of the C-method is the ability to tune parameters in the C-equation to generate non-oscillatory behavior. Though we are quite explicit on the form of the C-equation in (2.2.6b), a simple modification allows for the diffusion coefficient to be problem dependent, i.e. allowing for a choice of positive constant $\kappa > 0$ and replacing the diffusion term with

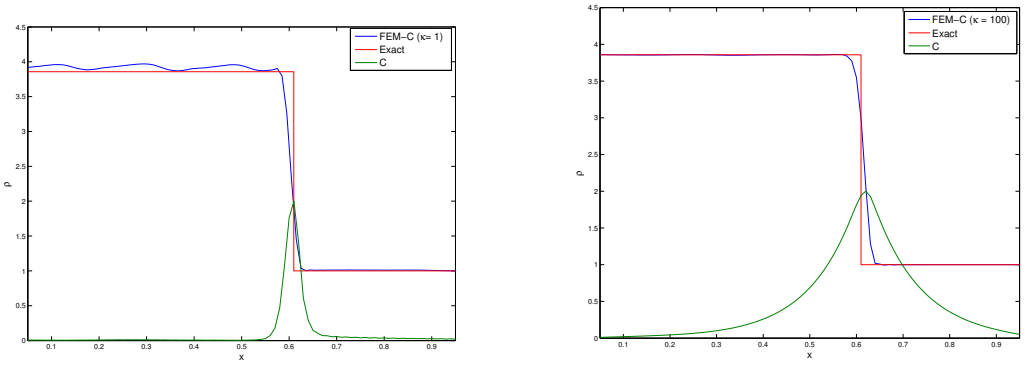
$$-\kappa\epsilon\, S(\mathbf{u}^\epsilon)\partial_x^2 C^\epsilon\,.$$

In most of the forthcoming experiments, we fix $\kappa = 1$, but we note that choosing larger κ can yield smoother solution profiles as the profile of C will be less localized. The parameter κ is a time-relaxation parameter, and can be viewed in an analogous fashion to

the time-relaxation parameter present in Cahn-Hilliard and Ginzburg-Landau theories. For very slow moving shocks, the time-relaxation can be adjusted to scale with the shock speed.[2]

We find this to be an effective approach for the flattening procedure discussed in [27] for removing oscillations that form to the left of a slowly right-moving shock. Moreover, Roberts [55] concludes that a differentiable form of the numerical flux construction appears necessary to remove downstream long-wavelength oscillations caused by slow shock motion. We use the C-method to analyze this.

Using the slow-shock initial conditions outlined in Quirk [30], in Figure 2.2 we show the success of the FEM-C (outlined below in §2.3.2) in removing these oscillations when choosing $\kappa = 1$ (Fig. 2.2(a)) and $\kappa = 100$ (Fig. 2.2(b)).



(a) $\kappa = 1$ (b) $\kappa = 100$

Figure 2.2: Application of FEM-C to a very slowly moving shock

[2]We note that κ is inversely proportional to the Mach number and its precise functional relation shall be examined in future work.

2.2.9 Convergence of the C-method in the limit of zero mesh size

The isentropic case

We sketch the proof for the isentropic Euler equations given by

$$(\rho u)_t + (\rho u^2 + p)_x = 0, \tag{2.2.17a}$$

$$\rho_t + (\rho u)_x = 0\,, \tag{2.2.17b}$$

$$p(\rho) = \rho^\gamma\,, \tag{2.2.17c}$$

where $\gamma > 1$.

To simplify the notation, we set $\epsilon^2\tilde{\beta} = \epsilon$, and set the momentum $m^\epsilon = \rho^\epsilon u^\epsilon$. Following (2.2.6), we write the C-method version of (2.2.17) as

$$m^\epsilon_t + [(m^\epsilon)^2/\rho^\epsilon + p^\epsilon]_x = \epsilon \left(C^{\epsilon,\delta} m^\epsilon_x\right)_x\,, \tag{2.2.18a}$$

$$\rho^\epsilon_t + m^\epsilon_x = \epsilon \left(C^{\epsilon,\delta} \rho^\epsilon_x\right)_x\,, \tag{2.2.18b}$$

$$p^\epsilon(\rho^\epsilon) = (\rho^\epsilon)^\gamma\,, \tag{2.2.18c}$$

$$C^\epsilon_t - \epsilon\, S(\mathbf{u}^\epsilon) C^\epsilon_{xx} + \frac{S(\mathbf{u}^\epsilon)}{\epsilon} C^\epsilon = S(\mathbf{u}^\epsilon) G(u^\epsilon_x)\,, \tag{2.2.18d}$$

or (2.2.18a,b) can be equivalently written in terms of the vector $\mathbf{u}^\epsilon = (m^\epsilon, \rho^\epsilon)$ and flux $\mathbf{f}(\mathbf{u}^\epsilon) = ((m^\epsilon)^2/\rho^\epsilon + (\rho^\epsilon)^\gamma\ ,\ m^\epsilon)$ as

$$\mathbf{u^\epsilon}_t + \mathbf{f}(\mathbf{u}^\epsilon)_x = \epsilon \left[\mathcal{C}\mathbf{u^\epsilon}_x\right]_x\,, \tag{2.2.18'}$$

where $\mathcal{C}$ denotes a diagonal 2x2 matrix with entries $C^{\epsilon,\delta}$ which is strictly positive-definite. Recall that $G(u^\epsilon) = |u^\epsilon_x|/\max|u^\epsilon_x|$, satisfies $G \geq 0$, and that $S(\mathbf{u}^\epsilon) = \max(|u^\epsilon + c|, |u^\epsilon - c|)$, with c denoting the sound speed. On any time interval $[0, T]$, the maximum wave speed $S(\mathbf{u}^\epsilon)$ is uniformly strictly positive; thus, as the initial data for $C^\epsilon_{t=0} \geq 0$, the maximum principle shows that $C^\epsilon(x,t)$ must be non-negative. We remark that while the use of $C^{\epsilon,\delta} = C^\epsilon + \delta$ as the coefficient is not required for the numerics, as δ is taken much smaller than the mesh size Δx, strict positivity of $\mathcal{C}$ simplifies the proof of regularity of solutions to (2.2.18) as well as the convergence argument.

To avoid issues with spatial boundaries, we shall assume periodic boundary conditions for our spatial domain. Note that in this case, the fundamental theorem of calculus shows that $\frac{d}{dt}\int \rho(x,t)\,dx = 0$ and that mass is conserved.

The basic energy law

In order to prove that solutions to (2.2.18) converge to solutions of (2.2.17), we must establish ϵ-independent estimates for solutions of (2.2.18). To do so, we multiply equation (2.2.18a) by u^ϵ, integrate over our spatial domain, and make use of the equation (2.2.18b) to find that any weak solution to (2.2.18) must verify the basic energy law

$$\frac{d}{dt}\left[\int \frac{1}{2}\rho^\epsilon (u^\epsilon)^2\,dx + \frac{1}{\gamma-1}\int p^\epsilon\,dx\right] \leq \\ -\epsilon\int C^{\epsilon,\delta}\rho^\epsilon\,(u_x^\epsilon)^2\,dx - \epsilon\gamma\int C^{\epsilon,\delta}(\rho^\epsilon)^{\gamma-2}(\rho_x^\epsilon)^2\,dx\,. \quad (2.2.19)$$

(The inequality in (2.2.19) is due to the lower semi-continuity of weak convergence and is replaced with equality for solutions which are sufficiently regular.) Thus, the total energy of isentropic gas dynamics is dissipated according to the right-hand side of (2.2.19), and for each $\epsilon > 0$, we see that u_x^ϵ and ρ_x^ϵ are square-integrable (in L^2) for almost every instant of time, if the density $\rho^\epsilon \geq \lambda > 0$, that is, if ρ^ϵ avoids vacuum. We shall explain below that this is indeed the case.

Regularity of solutions $\mathbf{u}^\epsilon$

Suppose that for each instant of time, $u^\epsilon(t)$ and its derivatives $u_x^\epsilon(t)$ and $u_{xx}^\epsilon(t)$ are all square-integrable in space. The reaction-diffusion equation (2.2.18d) is a uniformly parabolic equation. By our assumption, and as a consequence of Sobolev's theorem, $u_x^\epsilon(t)$ is a bounded function; furthermore, the right-hand side of (2.2.18d) is square-integrable in space, for every instant of time. It is standard, from the regularity theory of uniformly parabolic equations, that for each time t, $C^\epsilon(t)$ then has two spatial (weak) derivatives which are square-integrable. This, in turn, shows that for $\epsilon > 0$, solutions $\mathbf{u}^\epsilon$ possess three spatial (weak) derivative which are square-integrable for almost every instant of time, and we have verified our assumption. This implies that solutions $\mathbf{u}^\epsilon$ are classically differentiable in both space and time.

Furthermore, by using the symmetrizing matrix $\begin{bmatrix} \rho^\epsilon & 0 \\ 0 & \gamma(\rho^\epsilon)^{\gamma-2} \end{bmatrix}$ we can show that $(u^\epsilon(\cdot,t), \rho^\epsilon(\cdot,t))$ are, independently of ϵ and t, uniformly bounded in the Sobolev space H^2 (consisting of measurable functions with two weak derivatives in L^2), and thus we

may take a pointwise limit of this sequence as $\epsilon \to 0$, in the event that the time-interval is sufficiently small as to ensure that a shock has not yet formed. Of course, we are interested, in convergence to discontinuous profiles, so we address this next.

Convergence to the entropy solution

We shall now provide a sketch of the limit as $\epsilon \to 0$. A function $\eta : \mathbb{R}^2 \to \mathbb{R}$ is called an *entropy* for (2.2.17) with *entropy flux* $q : \mathbb{R}^2 \to \mathbb{R}$ if smooth solutions $\mathbf{u}$ satisfy the additional conservation law

$$\eta(\mathbf{u})_t + q(\mathbf{u})_x = 0\,. \tag{2.2.20}$$

In non-conservative form, (2.2.17) and (2.2.20) are written as

$$\mathbf{u}_t + \nabla f(\mathbf{u})\mathbf{u}_x = 0\,, \;\; \nabla\eta(\mathbf{u})\mathbf{u}_t + \nabla q(\mathbf{u})\mathbf{u}_x = 0\,,$$

from which we obtain the compatibility condition between η and q,

$$\nabla\eta(\mathbf{u})\,\nabla f(\mathbf{u}) = \nabla q(\mathbf{u})\,. \tag{2.2.21}$$

The pair (η, q) satisfy (2.2.20) if and only if condition (2.2.21) holds. Moreover, a weak solution to (2.2.17) is the unique entropy solution if

$$\eta(\mathbf{u})_t + q(\mathbf{u})_x \leq 0\,. \tag{2.2.22}$$

For isentropic gas dynamics we can set

$$\eta(m,\rho) = \frac{m^2}{2\rho} + \frac{\rho^\gamma}{\gamma - 1}$$

which is the total energy, with corresponding entropy flux

$$q(m,\rho) = \left[\frac{m^2}{2\rho} + \frac{\gamma}{\gamma-1}\rho^\gamma\right]\frac{m}{\rho}\,.$$

We observe that $\nabla^2\eta(m,\rho)$ is strongly convex as long as $\rho > 0$.

For the sequence of solution $\mathbf{u}^\epsilon$ of (2.2.18), suppose that as $\epsilon \to 0$, $\mathbf{u}^\epsilon$ converges boundedly (almost everywhere) to a weak solution $\mathbf{u}$ of (2.2.17). We claim that if (η, q)

satisfy (2.2.20), then (2.2.22) holds in the distributional sense. To see that this is the case, we take the inner-product of $\nabla\eta(\mathbf{u}^\epsilon)$ with equation (2.2.18'), and find that

$$\begin{aligned}\eta(\mathbf{u}^\epsilon)_t + q(\mathbf{u}^\epsilon)_x &= \epsilon\nabla\eta(\mathbf{u}^\epsilon)\left[\mathcal{C}\mathbf{u}^\epsilon_x\right]_x \\ &= \epsilon\left[\mathcal{C}\eta(\mathbf{u}^\epsilon)_x\right]_x - \epsilon[\mathbf{u}^\epsilon_x]^T\,\mathcal{C}\,\nabla^2\eta(\mathbf{u}^\epsilon)\mathbf{u}^\epsilon_x\,.\end{aligned}$$

Integrating over the spatial domain and then over the time interval $[0,T]$ yields

$$\int \eta(\mathbf{u}^\epsilon(x,T))dx - \int \eta(\mathbf{u}^\epsilon(x,0))dx = -\epsilon\int_0^T\int [\mathbf{u}^\epsilon_x]^T\,\mathcal{C}\,\nabla^2\eta(\mathbf{u}^\epsilon)\mathbf{u}^\epsilon_x\,dx\,dt\,,$$

from which it follows that

$$\int_0^T\int |\sqrt{\epsilon}\mathbf{u}^\epsilon_x|^2\,dx\,dt \le \bar{c} \tag{2.2.23}$$

where the constant $\bar{c}$ depends upon δ, the minimum value of density, and the entropy in the initial data. For a smooth, non-negative test function ψ with compact support in the strip $\mathcal{I}\times(0,T)$,

$$\begin{aligned}\iint \eta(\mathbf{u}^\epsilon)\phi_t + q(\mathbf{u}^\epsilon)\phi_x\,dxdt = \sqrt{\epsilon}\iint \mathcal{C}(\sqrt{\epsilon}\mathbf{u}^\epsilon)_x\phi_x\,dxdt \\ + \iint \epsilon[\mathbf{u}^\epsilon_x]^T\,\mathcal{C}\,\nabla^2\eta(\mathbf{u}^\epsilon)\mathbf{u}^\epsilon_x\phi\,dxdt\,.\end{aligned}$$

Thanks to (2.2.23), the first term on the right-hand side goes to zero like ϵ, while the second term is non-negative, since $\nabla^2\eta(\mathbf{u}^\epsilon)$ is positive-definite (since η is strongly convex) as is $\mathcal{C}$. Thus, as $\epsilon\to 0$, we recover the entropy inequality (2.2.22).

It remains to discuss the assumptions concerning the bounded convergence of $\mathbf{u}^\epsilon$ to $\mathbf{u}$, as well as the uniform bound from below $\rho^\epsilon \ge \nu > 0$. The argument relies on finding a priori bounds on the amplitudes of solutions to (2.2.18). If it is the case that uniformly in $\epsilon > 0$,

$$|\mathbf{u}^\epsilon| \le M \text{ and } 0 < \nu \le \rho^\epsilon\,,$$

then the compensated-compactness approach for isentropic Euler pioneered by DiPerna [51] and made much more general by Lions, Perthame, & Souganidis [52] provides a subsequence of $\mathbf{u}^\epsilon$ converging pointwise (almost everywhere) to a solution $\mathbf{u}$ of (2.2.17).

For isentropic gas dynamics, our approximation (2.2.18) preserves the invariant quadrants of the inviscid dynamics (just as in the case of uniform artificial viscosity) and

provides the bound $|\mathbf{u}^\epsilon| \leq M$ as long as $0 < \nu \leq \rho^\epsilon$ for some ν. In particular, the Riemann invariants $w = u + \frac{2\gamma}{\gamma-1}\rho^{\sqrt{\gamma-1}}$ and $z = u - \frac{2\gamma}{\gamma-1}\rho^{\sqrt{\gamma-1}}$ satisfy $w(x,t) \leq \sup w|_{t=0}$ and $-z(x,t) \leq \sup(-z_{t=0})$ and the intersection of these half-planes provides the invariant quadrant (see Chueh, Conley, & Smoller [56]), and hence the desired bound $|\mathbf{u}^\epsilon| \leq M$ as long as vacuum is avoided.

Finally, the fact that we have the lower-bound $0 < \nu \leq \rho^\epsilon$ is an immediate consequence of the strong maximum principle.

2.2.10 The C-equation as a gradient flow

Notice that equilibrium solutions to the C-equation are minimizers of the following functional (we drop the superscript ϵ):

$$\mathcal{E}_G(C) = \int \left(\frac{\epsilon}{2} C_x^2 - G(u_x)C + \frac{1}{2\epsilon} C^2 \right) dx \,.$$

In the absence of a forcing function $G(u_x)$, this reduces to

$$\mathcal{E}_0(C) = \frac{1}{2} \int \left(\epsilon\, C_x^2 + \frac{1}{\epsilon} C^2 \right) dx \,. \tag{2.2.24}$$

The first term is commonly referred to as the Dirichlet energy and its minimizers are harmonic functions. The second term can be viewed as a *penalization of the Dirichlet energy*. In particular, because the *energy* functional is bounded by a constant independent of $\epsilon > 0$, the penalization term constrains C to be $O(\sqrt{\epsilon})$. Thus, minimizers are trying to be harmonic while minimizing their support.

The C-equation can be written as a classical gradient flow equation

$$\frac{dC}{dt} = -S(\mathbf{u}) \nabla \mathcal{E}_G(C) \,,$$

where the gradient is computed relative to the L^2 inner-product. Thus the heat operator in the C-equation, $\partial_t - \epsilon \partial_x^2$, smooths the forcing in space-time, while the reaction term $\frac{S(u)}{\epsilon} C$ minimizes the support of the smoothed profile. This is very much related to the theories of Cahn-Hilliard and Ginzburg-Landau gradient flows, and we intend to examine this connection in subsequent papers.

2.3 Numerical Schemes

We describe two very different numerical algorithms in the context of our C-method. First, we outline a classical continuous finite-element discretization, yielding FEM-C and FEM-$|u_x|$ (based on classical artificial viscosity). Second, we discuss a simple WENO-based scheme for compressible Euler that upwinds solely based on the sign of the velocity u. To this scheme, we apply a slightly modified C-method resulting in our WENO-C algorithm.

For the purpose of comparison, we also implement two additional numerical methods. The first is a second-order central-differencing scheme of Nessayhu-Tadmor (NT), a nice and simple method which serves as a base-line for our FEM-C comparisons. The second scheme is a very competitive WENO scheme that utilizes a Godunov-based upwinding based upon characteristic decompositions (WENO-G). This will serve as a benchmark for our WENO-C scheme.

2.3.1 Notation for discrete solutions

To compute approximations to (2.2.1), we subdivide space-time into a collection of spatial nodes $\{x_i\}$ and temporal nodes $\{t_n\}$. We denote the computed approximate solution by

$$\mathbf{u}_i^n \approx \mathbf{u}(x_i, t_n),$$

noting that for fixed i and n, $\mathbf{u}_i^n$ is a 3-vector of solution components, i.e.,

$$\mathbf{u}_i^n = \begin{bmatrix} \rho_i^n \\ m_i^n \\ E_i^n \end{bmatrix}.$$

It is important to note that we use the notation $\mathbf{u}_i^n$ for both pointwise approximations to $\mathbf{u}$, (acquired via FEM-C) and approximations to the cell-average values of $\mathbf{u}$ (acquired via WENO-C).

A subscripted quantity w_i denotes the vector itself *and* the individual components of the vector. We overload this notation so to not cause any confusion between functions defined over a continuum versus those defined only at a finite number of points.

In FEM-C and WENO-C, we discretize (2.2.6) (or some slight modification) with $\epsilon = \Delta x$, and use the above notation for the computed solution. We also denote the approximation to C by C_i^n.

2.3.2 FEM-C and FEM-$|u_x|$: A Second-Order Continuous-Galerkin Finite-Element Scheme

We choose a second-order continuous-Galerkin finite-element scheme to provide a discretization of (2.2.6), subsequently defining our FEM-C scheme.

We subdivide $\mathcal{I}$ with $N+1$ (for N even)-uniformly spaced nodes $\{x_i\}$ separated by a distance Δx. In the FEM community, spatial discretization size is more commonly referred to by *element-width*; to maintain consistency with the literature, we refer to the inter-nodal regions as *cells*. Since we use a continuous FEM, the degrees-of-freedom are defined at the cell-edges (as opposed to cell-centers)[3].

For use in our FEM implementation, it is useful to consider the variational form of (2.2.6). At the continuum level, $(\mathbf{u}^\epsilon, C^\epsilon)$ satisfy

$$\int_{\mathcal{I}} \left[\partial_t \mathbf{u}^\epsilon \cdot \Phi - \mathbf{F}(\mathbf{u}^\epsilon) \cdot \partial_x \Phi + \beta \epsilon^2 \frac{\max\limits_{\mathcal{I}} |\partial_x u^\epsilon|}{\max\limits_{\mathcal{I}} C^\epsilon} C^\epsilon \partial_x \mathbf{u}^\epsilon \cdot \partial_x \Phi \right] dx = 0\,, \tag{2.3.1a}$$

$$\int_{\mathcal{I}} \left[\partial_t C^\epsilon \phi + S(\mathbf{u}^\epsilon) \left(\epsilon \partial_x C^\epsilon \partial_x \phi + \frac{1}{\epsilon} C^\epsilon \phi \right) \right] dx = \int_{\mathcal{I}} S(\mathbf{u}^\epsilon) G(\partial_x u^\epsilon) \phi \, dx \tag{2.3.1b}$$

for almost every t, for all vector-valued test functions Φ, and all scalar-valued test functions ϕ.

Using the finite-element spatial discretization based on piecewise second-order Lagrange polynomials, we construct operators $\mathcal{A}_{\text{FEM}}$ and $\mathcal{B}_{\text{FEM}}$, corresponding to the non-time-differentiated terms in (2.3.1a) and (2.3.1b), respectively. Using these discrete operators, we write the semi-discrete form of (2.3.1a) and (2.3.1b) as

$$\partial_t \begin{bmatrix} \mathbf{u}_i \\ C_i \end{bmatrix} + \begin{bmatrix} \mathcal{A}_{\text{FEM}}(\mathbf{u}_i, C_i) \\ \mathcal{B}_{\text{FEM}}(\mathbf{u}_i, C_i) \end{bmatrix} = 0 \tag{2.3.2}$$

[3] When we compare our FEM-C scheme with other, *cell-averaged* schemes, we perform an averaging procedure based upon averages between nodes.

where $\mathbf{u}_i$ and C_i represent the nodal values of an approximation to $\mathbf{u}^\epsilon$ and C^ϵ for which $\epsilon = \Delta x$ (see §2.2.6). For a standard reference on the details of this procedure, see Larsson & Thomée [57].

The time-differentiation in (2.3.2) is approximated by a diagonally-implicit second-order time-stepping procedure; first we predict $\mathbf{u}_i^{n+1}$ to and solve the implicit set of equations for C_i^{n+1} and follow by implicitly solving for $\mathbf{u}_i^{n+1}$ using C_i^{n+1}. Our fully discrete scheme is given by

$$\tilde{\mathbf{u}}_i^{n+1} = \mathbf{u}_i^n + \mathcal{A}_{FEM}(\mathbf{u}_i^n, C_i^n), \tag{2.3.3a}$$

$$C_i^{n+1} = C_i^n + \frac{t_{n+1} - t_n}{2}\left[\mathcal{B}_{FEM}(\tilde{\mathbf{u}}_i^{n+1}, C_i^{n+1}) + \mathcal{B}_{FEM}(\mathbf{u}_i^n, C_i^n)\right], \tag{2.3.3b}$$

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n + \frac{t_{n+1} - t_n}{2}\left[\mathcal{A}_{FEM}(\mathbf{u}_i^{n+1}, C_i^{n+1}) + \mathcal{A}_{FEM}(\mathbf{u}_i^n, C_i^n)\right]. \tag{2.3.3c}$$

For smooth solutions, where artificial viscosity is not necessary, our FEM-C scheme is second-order accurate in both space and time when the error is measured in the L^1-norm. Moreover, the addition the artificial viscosity obtained through the C-method is formally a second-order perturbation (in Δx) and we have verified this accuracy when $\beta > 0$ (again, for smooth $\mathbf{u}_0$). For $\mathbf{u}_0$ containing jump discontinuities, the given scheme is no longer second-order accurate on all of $\mathcal{I}$ but preserves second-order accuracy in the smooth regions away from discontinuities.

For the classical artificial viscosity schemes (2.2.5), the C-equation is no longer used but we require a similar step to predict the velocity for use in the diffusion coefficient. This analogous scheme, is referred to as the FEM-$|u_x|$ scheme.

2.3.3 WENO-C: A Simple WENO scheme using the C-method

Our WENO-based scheme is motivated by Leonard's finite volume schemes ([58], pg. 65). Upwinding is performed solely based on the sign of the velocity at cell-edges, and the WENO reconstruction procedure is formally fifth-order.

We divide the interval $\mathcal{I}$ into N equally sized cells of width Δx, identifying the N degrees-of-freedom as cell-averages over cells centered at the x_i. The cell edges are denoted using the fraction index, i.e.

$$x_{i+1/2} = \frac{x_i + x_{i+1}}{2}$$

Subsequently, we denote a cell-averaged quantity by w_i and its values at the left and right endpoints by $w_{i-1/2}$ and $w_{i+1/2}$, respectively.

Given a vector w_i, corresponding to cell-average values, and vectors $z_{i-1/2}$, $z_{i+1/2}$ corresponding to left and right cell-edge values, we define the jth component of vector

$$\big[\text{WENO}(w_i, z_{i\pm1/2})\big]_j = \frac{1}{\Delta x}\left(\tilde{w}_{j+1/2} z_{j+1/2} - \tilde{w}_{j-1/2} z_{j-1/2}\right)$$

where the cell-edge values of $\tilde{w}_{j+1/2}$ are calculated using a fifth-order WENO reconstruction, upwinding based upon the sign of $z_{j+1/2}$.

For the flux in the energy equation, we use

$$\big[\text{WENO}_E(E_i, u_{i\pm1/2})\big]_j = \frac{1}{\Delta x}\Big(\tilde{E}_{j+1/2} u_{j+1/2} \frac{(1+\frac{p_j}{E_j}) + (1+\frac{p_{j+1}}{E_{j+1}})}{2} - \tilde{E}_{j-1/2} u_{j-1/2} \frac{(1+\frac{p_{j-1}}{E_{j-1}}) + (1+\frac{p_j}{E_j})}{2}\Big). \quad (2.3.4)$$

Using this simplified WENO-based reconstruction, we construct the operators $\mathcal{A}_{\text{WENO}}$ and $\mathcal{B}_{\text{WENO}}$ where

$$\left[\mathcal{A}_{\text{WENO}}\left(\begin{bmatrix}\rho_i\\ m_i\\ E_i\end{bmatrix}, C_i\right)\right] = \begin{bmatrix}\text{WENO}(\rho_i, u_{i\pm1/2})\\ \text{WENO}(m_i, u_{i\pm1/2}) + \tilde{\partial} p_i - \frac{\tilde{\partial}_C u_{i+1/2} - \tilde{\partial}_C u_{i-1/2}}{\Delta x}\\ \text{WENO}_E(E_i, u_{i\pm1/2})\end{bmatrix} \quad (2.3.5a)$$

$$\mathcal{B}_{\text{WENO}}\left(\begin{bmatrix}\rho_i\\ m_i\\ E_i\end{bmatrix}, C_i\right) = S(\mathbf{u}_i)\left[\frac{C_i}{\Delta x} - G(\tilde{\partial} u_i)\right] - \frac{\tilde{\partial}_S C_{i+1/2} - \tilde{\partial}_S C_{i-1/2}}{\Delta x}. \quad (2.3.5b)$$

where for a general quantity w_i, defined at the cell-centers, we denote

$$w_{i+1/2} = \frac{w_{i+1} + w_i}{2}, \quad \tilde{\partial} w_i := \frac{w_{i+1} - w_{i-1}}{2\Delta x}, \quad \tilde{\partial} w_{i+1/2} = \frac{w_{i+1} - w_i}{\Delta x}.$$

We also use the shorthand notation

$$\tilde{\partial}_C u_{i+1/2} = \beta\, \Delta x^2\, \max_i \left|\tilde{\partial} u_{i+1/2}\right|\, \frac{C_{i+1/2}}{\max_i C_i}\, \rho_{i+1/2}\, \tilde{\partial} u_{i+1/2},$$

and

$$\tilde{\partial}_S C = \Delta x\, S(\mathbf{u}_i)\, \tilde{\partial} C_{i+1/2}.$$

Using the above definitions, we define the semi-discrete form

$$\partial_t \begin{bmatrix} \mathbf{u}_i \\ C_i \end{bmatrix} + \frac{1}{\Delta x} \begin{bmatrix} \mathcal{A}_{\text{WENO}}(\mathbf{u}_i, C_i) \\ \mathcal{B}_{\text{WENO}}(\mathbf{u}_i, C_i) \end{bmatrix} = 0 \tag{2.3.6}$$

and we generate the sequence of iterates $\mathbf{u}_i^n$ and C_i^n with a standard fourth-order Runge-Kutta time-stepper.

The resulting discretization outlined above is a slight variation on that outlined in (2.2.6). While the amount of artificial viscosity $C(x,t)$ is controlled by only the velocity, we only add artificial viscosity to the momentum equation. This change is based upon the fact that WENO already minimizes the production of numerical oscillations and the addition of artificial viscosity is primarily intended on stabilizing the solution near strong shocks, whereas standalone WENO may lose stability. Without dissipation on the right-hand side of the mass equation, it is necessary to modify the artificial viscosity on the momentum equation as follows:

$$\epsilon^2 \tilde{\beta} \partial_x (C \partial_x (\rho u)) \rightarrow \epsilon^2 \tilde{\beta} \partial_x (C \rho \partial_x u) \,. \tag{2.3.7}$$

This modification allows the C-method to maintain a basic energy law (in fact, it is the energy law (2.2.19) with the last term on the right-hand side), and simultaneously permits higher accuracy for our WENO-based scheme.

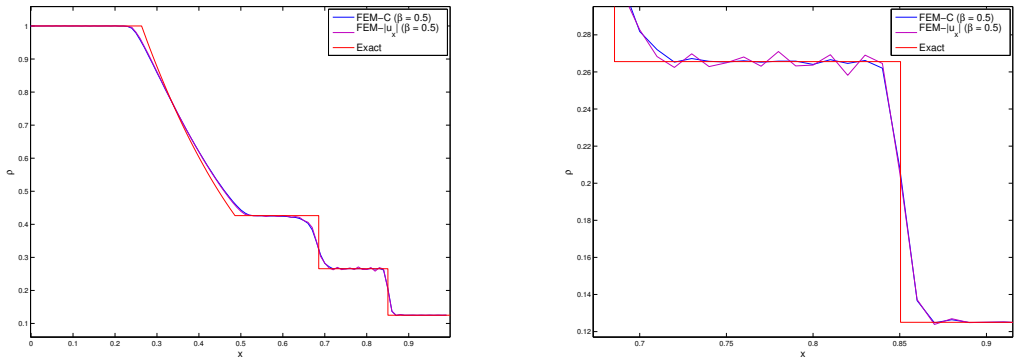
2.3.4 NT: A Second-order Central-Differencing scheme of Nessayhu-Tadmor

The central-differencing scheme of Nessyahu and Tadmor is an extension of the first-order Lax-Fredrichs finite difference scheme in which linear, MUSCL-based reconstructions are used to yield a second-order accurate scheme. The resulting scheme is extremely easy to implement (a FORTRAN code for 2-D problems is given in the Appendix of [34]) and does not require the use of Riemann solvers or characteristic directions for the purpose of upwinding. The NT scheme allows for various choices of limiters to enforce TVD or ENO but the UNO-limiter (see Harten & Osher [59]) is the most successful for our range of experiments.

Though NT is easy to implement and is easy generalized to multi-D (yielding the JT-scheme [34]), it merely serves as a base-line comparison for our FEM-C. Both FEM-C and NT are second-order, but FEM-C turns out to be far less diffusive by comparison.

2.3.5 WENO-G: WENO with Godunov-based upwinding

In [49] the authors study a fifth-order, WENO-based discretization, upwinding by virtue of a high-accuracy Godunov-scheme. Their scheme has the usual trait of WENO, offering minimal diffusion near extrema, and has the added stabilization and accuracy of higher-order Godunov solvers. For a more in-depth description, see [49].



(a) Complete density profile

(b) Zoom-in at shock

Figure 2.3: Comparison of FEM-C and FEM-$|u_x|$, for the Sod shock-tube experiment with $N = 100$, $t = 0.2$. $\beta = 0.5$ for both FEM-C and FEM-$|u_x|$.

2.4 Sod shock-tube problem

For the classic Sod shock-tube problem, we consider the domain $\mathcal{I} = [0, 1]$ along with the initial conditions

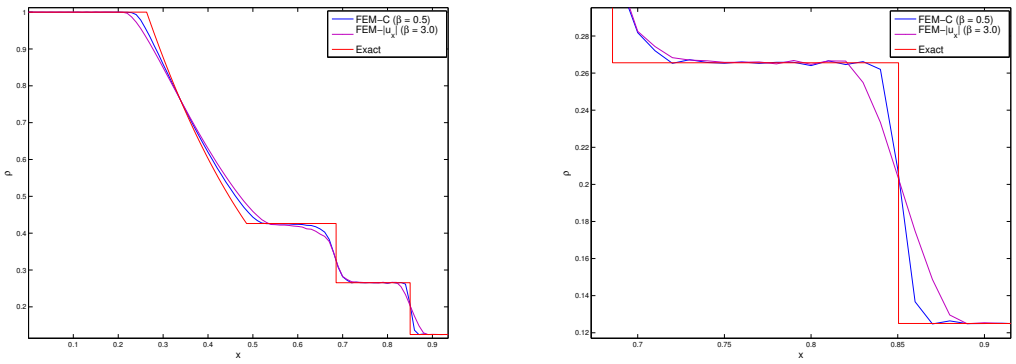
$$\begin{pmatrix} \rho_0(x) \\ m_0(x) \\ E_0(x) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2.5 \end{pmatrix} \mathbf{1}_{[0,\frac{1}{2})}(x) + \begin{pmatrix} 0.125 \\ 0 \\ 0.25 \end{pmatrix} \mathbf{1}_{[\frac{1}{2},1]}(x), \tag{2.4.1}$$

imposing natural boundary conditions at $x = 0$ and $x = 1$. This standard test problem, first considered in Sod [60], is a preliminary test for the viability of numerical schemes.

An exact solution is known for this problem and consists of two nonlinear waves (one shock and one rarefaction) along with a contact discontinuity.

In Figure 2.3(b) we compare the results of FEM-C and FEM-$|u_x|$ at $t = 0.2$ using $N = 100$ cells. We note that this comparison uses the standard choice of G in (2.2.6) since we are merely concerned with the C-equation performing as a smooth version of classical artificial viscosity schemes. Unlike comparisons with the schemes based on cell-averages, we compare the *nodal* values of FEM-C and FEM-$|u_x|$. In this comparison, we choose $\beta = 0.5$ for both schemes and see that the accuracy of both FEM-C and FEM-$|u_x|$ are quite comparable and each scheme resolves the shock in 3 cells. However, we notice noise in FEM-$|u_x|$ near the shock. In Figure 2.3(b) this observation is exemplified and we see that FEM-C is relatively non-oscillatory by comparison.

To limit these oscillations generated by FEM-$|u_x|$, we increase β by a factor of 6 and compare the resulting density in Figure 2.4. In Figure 2.4(b) we can see a significant loss in accuracy when increasing to $\beta = 3$. Furthermore, in Figure 2.4(a) we see FEM-$|u_x|$ requires 6 cells to capture the shock.



(a) Complete density profile

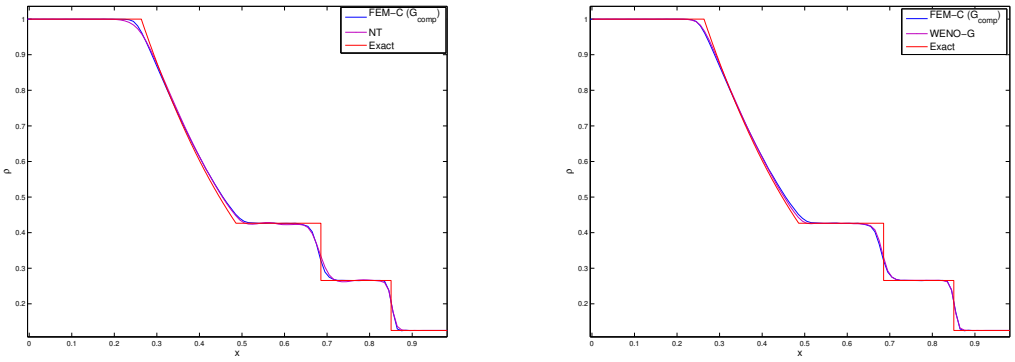
(b) Zoom-in at shock

Figure 2.4: Comparison of FEM-C and FEM-$|u_x|$, for the Sod shock-tube experiment with $N = 100$, $t = 0.2$. $\beta = 0.5$ for FEM-C and $\beta = 3.0$ for FEM-$|u_x|$.

In Figure 2.5 we compare the results of the FEM-C scheme versus NT and WENO-G. Each simulation is performed with $N = 100$ and for the FEM-C scheme we choose $\beta = 0.4$ and now use G_{comp} (see §2.2.5).

Each scheme produces similar resolution of the shock and contact discontinuity,

capturing the shock in 3 cells and the contact discontinuity in 6 cells. The NT-scheme produces small, smooth, non-physical oscillations as the density transitions from the rarefaction to the lower states, and performs the worst at the rarefaction. Both FEM-C and WENO-G are essentially non-oscillatory and despite WENO-C performing slightly better at the rarefaction, the results are virtually indistinguishable at the shock and contact discontinuity.



(a) FEM-C vs. NT (b) FEM-C vs. WENO-G

Figure 2.5: Comparisons of FEM-C against NT and WENO schemes, for the Sod shock-tube experiment with $N = 100$ and $t = 0.2$.

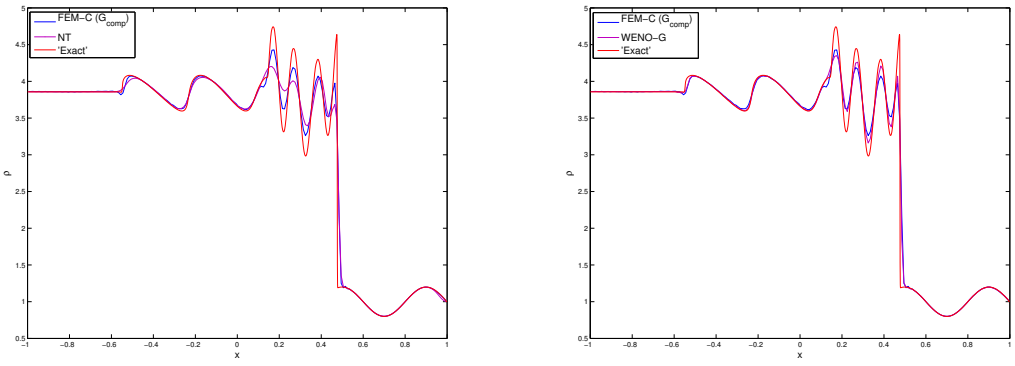
2.5 Osher-Shu shock-tube problem

For the problem of Osher-Shu, we consider the domain $\mathcal{I} = [-1, 1]$ along with initial conditions

$$\begin{pmatrix} \rho_0(x) \\ m_0(x) \\ E_0(x) \end{pmatrix} = \begin{pmatrix} 3.857143 \\ 10.14185 \\ 39.1666 \end{pmatrix} \mathbf{1}_{[-1,-0.8)}(x) + \begin{pmatrix} 1 + 0.2\sin(5\pi x) \\ 0 \\ 2.5 \end{pmatrix} \mathbf{1}_{[-0.8,1]}(x), \tag{2.5.1}$$

imposing natural boundary conditions at $x = -1$ and $x = 1$

This moderately difficult test problem, first considered in [21], proves to be more difficult for numerical schemes due to the evolution a shock-wave which interacts with an entropy-wave; care is required to accurately capture the amplitudes of the post-shock



(a) FEM-C vs. NT

(b) FEM-C vs. WENO-G

Figure 2.6: Comparisons of FEM-C against NT and WENO-G schemes, for the Osher-Shu shock-tube experiment with $N = 200$ and $t = 0.36$.

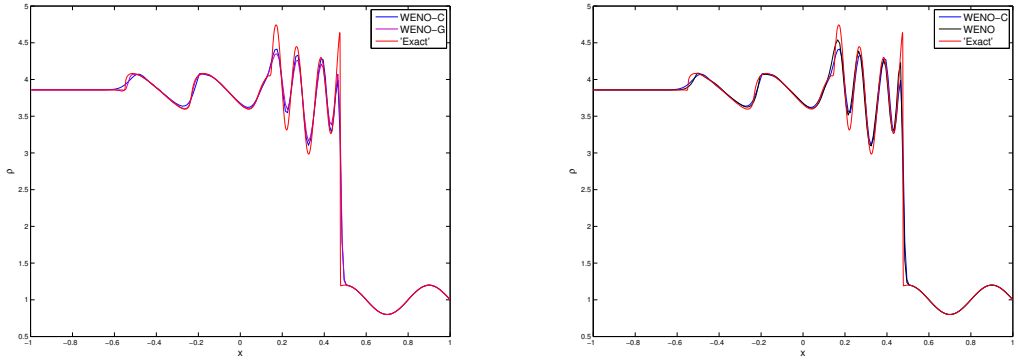
entropy waves. Since the density is not monotone, standard flux limiters may unnecessarily apply too much dissipation at local-extrema, significantly reducing accuracy. An exact solution for this problem is not available and our 'Exact' solution in our plots is generated using the DG-solver furnished in Hesthaven & Warburton [61] with 3200 cells.

In Figure 2.6 we compare the results of FEM-C (we choose $\beta = 0.5$ and use G_{comp}), versus NT and WENO-G at $t = 0.36$. In Figure 2.6(a) we see that NT diffuses the post-shock amplitudes and FEM-C provides improved results. On the other hand, in Figure 2.6(b) we see that all but one of the post-shock amplitudes are slightly better for the WENO-G scheme. This insufficiency of the FEM-C scheme is not completely surprising as the FEM-C is only formally second-order versus the fifth-order accuracy of the WENO-G scheme.

Noting this insufficiency of the FEM-C scheme, we compare the WENO-G scheme with WENO-C in Figure 2.7(a) and see the WENO-C scheme is more accurate in resolving the post-shock amplitudes. This comes at a price however, as we see WENO-G is more accurate in the N-wave region $[-0.6, 0]$.

Furthermore, it is interesting to note that in Figure 2.7(b) where we choose $\beta = 0$ in our simplified WENO-scheme, we see that the C-equation is not necessary for Osher-Shu. As we see in §2.6 this ceases to be the case as the collision of strong shock waves

require stabilization.



(a) WENO-C vs. WENO-G (b) WENO-C vs. WENO

Figure 2.7: Comparisons of WENO-C with WENO-G and our WENO scheme with artificial viscosity deactivated, for the Osher-Shu shock-tube experiment with $N = 200$ and $t = 0.36$.

2.6 Woodward-Colella Blast Wave

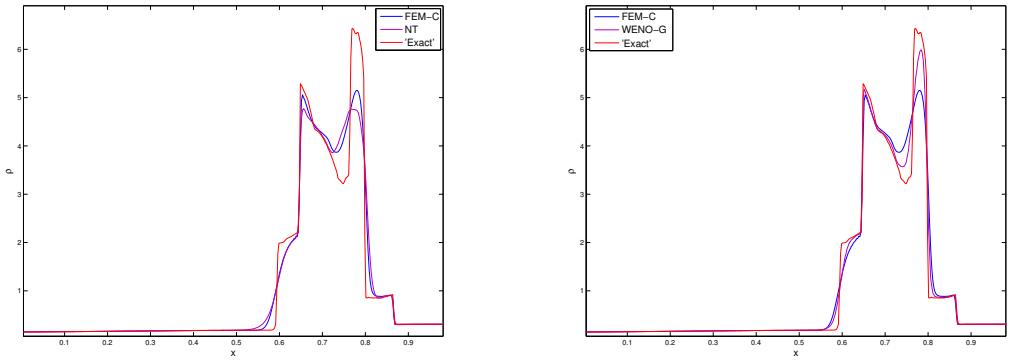
The colliding blast wave problem of Woodward-Collella is posed on the domain $\mathcal{I} = [0, 1]$ with initial conditions

$$\rho_0(x) = 1,$$
$$m_0(x) = 0,$$
$$E_0(x) = 250 \cdot \mathbf{1}_{[0.9,1]} + 0.25 \cdot \mathbf{1}_{[0.1,0.9)} + 2500 \cdot \mathbf{1}_{[0,0.1)},$$

and reflective boundary conditions at $x = 0$ and $x = 1$. This challenging blast wave problem, considered in [31] tests the ability of a numerical scheme to handle collisions between strong shock waves. Any viable scheme generally requires stabilization at these collisions. For the results of a wide range of schemes applied to this problem, see [27]. An exact solution for this problem is not available and the 'Exact' solution in our plots is generated with a 400-cell PPM solver.

As is standard in our sequence of experiments, we provide a comparison of FEM-C ($\beta = 0.5$) with NT and WENO-G in Figure 2.8 at $t = 0.038$. It is interesting to note,

the use of G_{comp} is far too oscillatory in this difficult test problem; we revert to the standard choice of G. We again see that while FEM-C is superior to NT in capturing the amplitude of the two peaks in the density, FEM-C is far too diffusive in comparison to WENO-G.



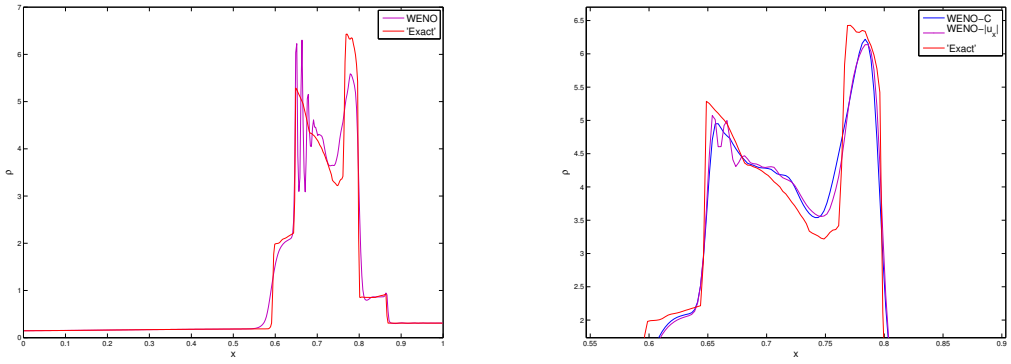
(a) FEM-C vs. NT (b) FEM-C vs. WENO-G

Figure 2.8: Comparisons of FEM-C against NT and WENO-G schemes, for the Woodward-Colella blast-tube experiment with $N = 400$ and $T = 0.038$.

Despite the relative inefficiency of FEM-C compared to WENO-G, it is interesting to note that our FEM-C results (with $N = 1200$) are better than the artificial viscosity schemes use in Colella & Woodward [27]. Our scheme is slightly sharper at the shocks and contact discontinuities and is just as accurate in the height of the two peaks.

Before moving to a comparison of WENO-G and WENO-C, in Figure 2.9(a) we see that our simplified WENO scheme is highly oscillatory due to the strong shock collision, necessitating the use of stabilization. This requirement contrasts to the observations made in §2.5. However, in Figure 2.9(b), we see that the use of a classical artificial viscosity significantly dampens the instability but moderate oscillations occur and the C-method provides similar dampening in a smooth way.

Finally, in Figure 2.10 we demonstrate the relative success of WENO-C versus WENO-G. At the left peak, WENO-G is more accurate, but at the right peak the reverse situation occurs. Each scheme provides very good results, and it is clear that WENO-C is a simple alternative to WENO-G which produces similar results for complicated shock interaction.



(a) WENO (without stabilization) (b) WENO (with stabilization)

Figure 2.9: WENO with and without stabilization applied to the Woodward-Colella blast-tube experiment with $N = 400$ and $t = 0.038$.

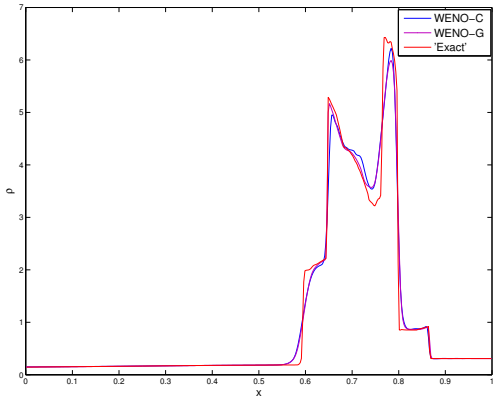


Figure 2.10: Comparison of WENO-C against WENO-G, for the Woodward-Colella blast-tube experiment with $N = 400$ and $t = 0.038$.

2.7 Leblanc shock-tube problem

We conclude our experiments with the Leblanc shock-tube, posed on the domain $\mathcal{I} = [0, 9]$, with initial conditions

$$\begin{pmatrix} \rho_0(x) \\ m_0(x) \\ E_0(x) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 10^{-1} \end{pmatrix} \mathbf{1}_{[0,3)}(x) + \begin{pmatrix} 10^{-3} \\ 0 \\ 10^{-9} \end{pmatrix} \mathbf{1}_{[3,9]}(x)\,, \tag{2.7.1}$$

with natural boundary conditions at $x = 0$ and $x = 9$, and with the adiabatic constant $\gamma = \frac{5}{3}$.

Because the initial energy E_0 jumps eight orders of magnitude, a very strong shock wave is produced, making the Leblanc problem an extraordinarily difficult numerical experiment. First , numerical methods tend to over-estimate the correct shock speed whenever the shock wave in the pressure field is not sharply resolved. Second, numerical approximations tend to produce large overshoots in the internal energy

$$e = \frac{p}{(\gamma - 1)\rho}$$

at the contact discontinuity. We refer the reader to Liu, Cheng, & Shu [62] and Loubére & Shashkov [63] for a discussion of the difficulties in the numerical simulation of the Leblanc problem for a variety of numerical schemes. The second-order finite-element basis that we use for our FEM-C algorithm is not sufficiently high-order to accurately capture wave speeds in Leblanc, but our fifth-order WENO-C scheme is ideally suited for this difficult test case. We shall present two differing strategies for WENO-C, which both capture the correct shock speed and remove overshoots of the internal energy.

2.7.1 Strategy One: A C equation for the energy density

As we introduced the C-method in equation (2.2.6), artificial viscosity is present on the right-hand side of all three conservation laws for momentum, mass, and energy. For the WENO-C algorithm, only viscosity in the momentum equation has been used for the Sod, Osher-Shu, and Woodward-Colella test cases. Due to the strength of the shock in Leblanc, we now return to using artificial viscosity for the energy equation. In our first

strategy for this problem, we solve for one additional linear reaction-difffusion equation for a new C-coefficient to use on the right-hand side of the energy conservation law.

Specifically, to combat the large overshoot in the internal energy e, we solve a second C-equation for the coefficient which we label C_E; the forcing term for the C_E equation uses $|\partial_x(E/\rho)|/\max|\partial_x(E/\rho)|$, replacing $|\partial_x u|/\max|\partial_x u|$ which forces the C-equation for the coefficient C_u, used for the right-hand side of the momentum equation.[4]

In particular, since C_u is found using the G_{comp} forcing, activated only in compressive regions when $u_x < 0$, for the C_E equation, we activate the right-hand side only in expansive regions when $u_x \geq 0$. To be precise, this modified WENO-C scheme replaces the semi-discrete form (2.3.6) with

$$\partial_t \begin{bmatrix} \mathbf{u}_i \\ \mathbf{C}_i \end{bmatrix} + \frac{1}{\Delta x} \begin{bmatrix} \tilde{\mathcal{A}}_{\text{WENO}}(\mathbf{u}_i, \mathbf{C}_i) \\ \tilde{\mathcal{B}}_{\text{WENO}}(\mathbf{u}_i, \mathbf{C}_i) \end{bmatrix} = 0. \tag{2.7.2}$$

The resulting fully-discrete scheme solves for $\mathbf{u}_i^n$ and

$$\mathbf{C}_i^n = \begin{pmatrix} C_{u_i}^n \\ C_{E_i}^n \end{pmatrix}$$

where the modified fluxes $\tilde{\mathcal{A}}_{\text{WENO}}$ and $\tilde{\mathcal{B}}_{\text{WENO}}$ are given by:

$$\begin{bmatrix} \tilde{\mathcal{A}}_{\text{WENO}} \left(\begin{bmatrix} \rho_i \\ m_i \\ E_i \end{bmatrix}, \begin{bmatrix} C_{u_i} \\ C_{E_i} \end{bmatrix} \right) \end{bmatrix} = \begin{bmatrix} \text{WENO}(\rho_i, u_{i\pm 1/2}) \\ \text{WENO}(m_i, u_{i\pm 1/2}) + \tilde{\partial} p_i - \frac{\tilde{\partial}_{C_u} u_{i+1/2} - \tilde{\partial}_{C_u} u_{i-1/2}}{\Delta x} \\ \text{WENO}_E(E_i, u_{i\pm 1/2}) - \frac{\tilde{\partial}_{C_E} E_{i+1/2} - \tilde{\partial}_{C_E} E_{i-1/2}}{\Delta x} \end{bmatrix} \tag{2.7.3a}$$

[4]Gradients of the function E/ρ are similar to gradients of the internal energy e for regions near the contact discontinuity where large overshoots may occur.

$$\left[\tilde{\mathcal{B}}_{\text{WENO}}\left(\begin{bmatrix}\rho_i\\ m_i\\ E_i\end{bmatrix},\begin{bmatrix}C_{u_i}\\ C_{E_i}\end{bmatrix}\right)\right]=$$

$$\begin{bmatrix} S(\mathbf{u}_i)\left[\frac{C_{u_i}}{\Delta x}-G_{comp}(\tilde{\partial}u_i)\right]-\frac{\tilde{\partial}_S C_{u_{i+1/2}}-\tilde{\partial}_S C_{u_{i-1/2}}}{\Delta x}\\ S(\mathbf{u}_i)\left[\frac{C_{E_i}}{\Delta x}-G_{expand}(\tilde{\partial}(E/\rho)_i,\tilde{\partial}u_i)\right]-\frac{\tilde{\partial}_S C_{E_{i+1/2}}-\tilde{\partial}_S C_{E_{i-1/2}}}{\Delta x}. \end{bmatrix} \quad (2.7.3b)$$

The expansive-region forcing for C_E is given by

$$G_{expand}\left(\tilde{\partial}E_i,\tilde{\partial}u_i\right)=\frac{|\tilde{\partial}(E/\rho)_i|}{\max\limits_i|\tilde{\partial}(E/\rho)_i|}\mathbf{1}_{[0,\infty)}(\tilde{\partial}u_i) \quad (2.7.4)$$

and we use the shorthand

$$\tilde{\partial}_{C_u}u_{i+1/2}=\beta_u\,\Delta x^2\,\max_i\left|\tilde{\partial}u_{i+1/2}\right|\,\frac{C_{u_{i+1/2}}}{\max\limits_i C_{u_i}}\,\rho_{i+1/2}\,\tilde{\partial}u_{i+1/2},$$

and

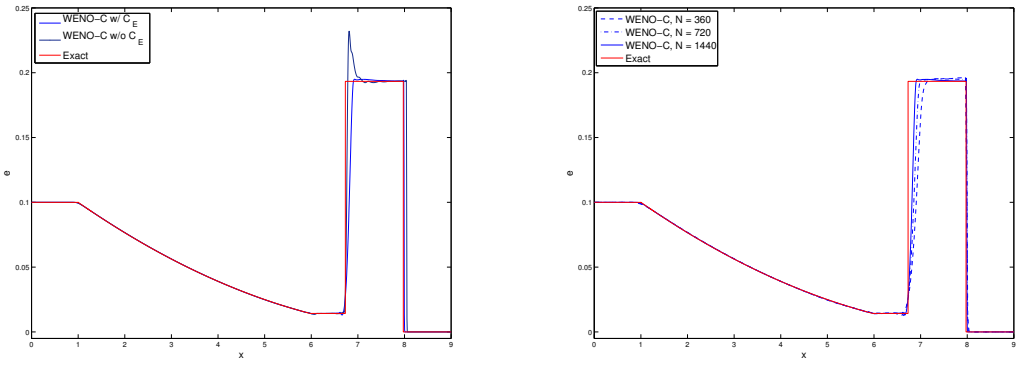
$$\tilde{\partial}_{C_E}E_{i+1/2}=\beta_E\,\Delta x^2\,\max_i\left|\tilde{\partial}u_{i+1/2}\right|\,\frac{C_{E_{i+1/2}}}{\max\limits_i C_{E_i}}\rho_{i+1/2}\,\tilde{\partial}(E/\rho)_{i+1/2}.$$

In Figure 2.11(a) we plot the difference between WENO-C with and without the use of this new equation for C_E. For WENO-C with C_E activated, we choose $\beta_u = 1.0$ and $\beta_E = 0.15$; with the C_E-equation deactivated, we use $\beta_u = 1.0$ and $\beta_E = 0$. Observe that activating the C_E-equation removes the large overshoot at the contact discontinuity. Furthermore, examining the location of the shock, we see that the use of the C_E-equation produces more accurate approximations of the shock speed.

In Figure 2.11(b) we show the results of WENO-C at $N = 360, 720, 1440$. In this plot, we see very little overshoot at each level of refinement and this small overshoot does not grow with refinement.

2.7.2 Strategy Two: a new type of viscosity for the energy density

Our second strategy for the Leblanc problem may be viewed as being motivated by the energy dissipation rate of real fluids, and adheres to our framework of only solving one C-equation, forced by the normalized modulus of the gradient of velocity. The idea



(a) With and without C_E, $N = 1440$ (b) Successive refinements, $N = 360, 720, 1440$

Figure 2.11: Internal energy plots for WENO-C for the Leblanc shock-tube experiment at $t = 6$.

is easy to explain, and we begin by writing the equations for momentum and mass (we drop the superscript ϵ):

$$(\rho u)_t + (\rho u^2 + p)_x = \epsilon^2 \tilde{\beta} (C \rho u_x)_x \,, \tag{2.7.5a}$$

$$\rho_t + (\rho u)_x = 0 \tag{2.7.5b}$$

$$p = (\gamma - 1)\rho\, e \,, \tag{2.7.5c}$$

$$C_t - \epsilon\, S(u) C_{xx} + \frac{S(u)}{\epsilon} C = S(u) G(u_x) \tag{2.7.5d}$$

where

$$\tilde{\beta} = \beta \frac{\max\limits_{\mathcal{I}} \partial_x u}{\max\limits_{\mathcal{I}} C} \,.$$

By multiplying the momentum equation by the velocity u, integrating over the spatial domain, and using the conservation of mass equation, we find the basic energy law:

$$\frac{d}{dt}\left[\int \frac{1}{2}\rho u^2\, dx + \frac{1}{\gamma - 1}\int p\, dx\right] = -\epsilon^2 \tilde{\beta} \int C \rho\, u_x^2\, dx \,. \tag{2.7.6}$$

Note, that when $\epsilon = 0$, the variable E is exactly the energy density; that is, when $\epsilon = 0$, $E = \frac{1}{2}\rho u^2 + \frac{p}{\gamma - 1}$. Thus, we formulate a right-hand side term for the energy equation to ensure the E continues to represent the energy density for $\epsilon > 0$. To do, we choose a right-hand side which will provide the same energy law as (2.7.6). We introduce the following equation:

$$E_t + (uE + up)_x = -\epsilon^2 \tilde{\beta} C \rho\, u_x^2 \,. \tag{2.7.7}$$

The fundamental theorem of calculus shows that integration of (2.7.7) provides the same basic energy law as (2.7.6). Hence, our second strategy employs the equation (2.7.5) together with (2.7.7). The interesting feature of the new right-hand side of the energy equation is its nonlinear structure, quadratic in velocity gradients. This energy loss compensates for entropy production, and can become anti-diffusive near contact discontinuities. As such, we shall discretize this set of equations using the very stable Lax-Friedrichs flux. We remark that the term $\epsilon^2 \tilde{\beta} C \rho \, u_x^2$ is analogous to the viscous dissipation term of the Navier-Stokes-Fourier system and can be found as a truncation error in [40].

As we noted above, to the best of our knowledge, the most commonly used numerical schemes applied to Leblanc tend to exhibit a significant overshoot in the internal energy e at the contact discontinuity. Furthermore, on coarse meshes (< 2000 cells), the speed of the shock tends to be inaccurate. Indeed, this is the case for arguably the most widely used WENO implementation, designated WENO-LF-5-RK-4 by Jiang & Shu [23]. This scheme, which we call WENO-LF, uses a Lax-Friedrichs flux-splitting with a 5th-order WENO reconstruction in space and 4th-order Runge-Kutta in time.

If we examine the contact discontinuity at $x \approx 6.8$ in Figure 2.12(a), at resolutions $N = 360, 720, 1440$ we see that WENO-LF exhibits relative overshoots of 12.8%, 11.8% and 11.4% respectively. This slow decay of the overshoot suggests that WENO-LF suffers from the Gibbs-phenomenon, despite it's attempt to quell oscillatory behavior. Examining the shock at $x \approx 8$ we see that the computed shock speeds are inaccurate.

To address the loss of accuracy exhibited by WENO-LF, we propose the use of the C-equation along with a nonlinear viscosity on the energy equation. Since WENO-LF has an intrinsic artificial viscosity (by virtue of the Lax-Friedrichs splitting) on the right-hand side of the momentum equation, we find that we do not need to explicitly use our artificial viscosity for the momentum (even though this mathematically motivated our nonlinear viscosity for the energy equation). As such, we require a single C-equation which is forced by $G_{comp}(u_x)$.

Keeping consistent with the semi-discrete formulation, we write the WENO-LF-C

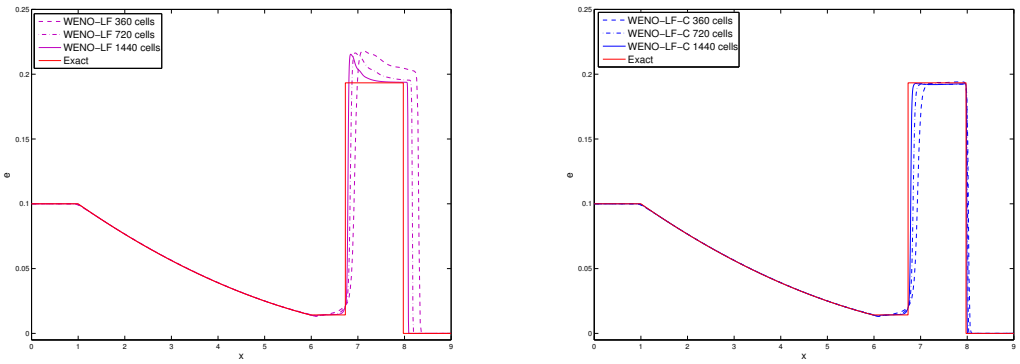
scheme

$$\partial_t \begin{bmatrix} \mathbf{u}_i \\ C_i \end{bmatrix} + \frac{1}{\Delta x} \begin{bmatrix} \mathcal{A}_{\text{WENO-LF}}(\mathbf{u}_i) + \mathcal{H}(\mathbf{u}_i, C_i) \\ \mathcal{B}_{\text{WENO}}(\mathbf{u}_i, C_i) \end{bmatrix} = 0 \tag{2.7.8}$$

where $\mathcal{B}_{\text{WENO}}$ is given by (2.3.5b) and $\mathcal{A}_{\text{WENO-LF}}$ corresponds to the choice of the WENO flux described in [23] (i.e. if $\mathcal{H} \equiv 0$ then (2.7.8) is the same as WENO-LF). The term $\mathcal{H}$ is a discrete approximation of $\tilde{\beta}\epsilon^2 C^{\epsilon,\delta}\rho^\epsilon |\partial_x u^\epsilon|^2$. The operator $\mathcal{H}$ is defined as

$$\mathcal{H}(\mathbf{u}_i, C_i) = \begin{bmatrix} 0 \\ 0 \\ \beta \Delta x^2 \max\limits_i \tilde{\partial} u_{i+1/2} \frac{C_i}{\max\limits_i C_i} \rho_i |\tilde{\partial} u_i|^2 \end{bmatrix}.$$

In Figure 2.12(b) we demonstrate the benefit of WENO-LF-C with $\beta = 5.0$, again at successive refinements of $N = 360, 720, 1440$. The overshoot at the contact discontinuity is relatively non-existent while the shock speeds are far more accurate and appear to converge to the correct speed at a faster rate.



(a) WENO-LF (b) WENO-LF-C

Figure 2.12: Internal energy plots for the Leblanc shock-tube experiment at $t = 6$ using WENO-LF with and without the C-equation.

2.8 Concluding Remarks

We have presented a localized space-time smooth artificial viscosity algorithm, the C-method, and have demonstrated its efficacy on a variety of classical one-dimensional

shock-tube problems. As compared to more established procedures, the C-method has been shown to be highly competitive with regards to accuracy and stability, while being relatively easy to implement. Because of its simplicity, the C-method can readily be extended to multiple space-dimensions and/or utilized in reactive-flow simulations. Of value to reactive flows is the localized *smooth* diffusion provided by the C-method; specifically, the function C can be used to actively influence various mixing-rate-limited reactions occurring near sharp boundaries.

In the future, the gradient-based source term used in the current implementation of the C-method may be combined with a noise-indicator that turns off the current gradient-based source term when it is not needed. Such noise-indicators require a very high-order scheme compatible with DG or 11th-order WENO to name just two examples. By projecting the solution onto a suitable basis, the noise-indicator would activate when small-scale coefficients of this basis do not have sufficient decay; in turn, an indicator function, localized about the region of noise, would activate and force the C equation. This approach is taken in [46], but without any gradient-based forcing functions like our function G or G_{comp}.

For example, with our first strategy, after the rapid initial growth of the internal energy field in the Leblanc shock-tube problem, this field is essentially representative of the advection of a square-wave. Thus, after initial growth, the gradient-based source term in the C equation for energy could be deactivated leading to less diffusion in the downstream contact discontinuity; simultaneously, the noise-indicator would activate if small-scale instabilities were to set in. (This, of course, motivated our second strategy, where the diffusion coefficient $\tilde{\beta}$ used $\max u_x$ rather $\max |u_x|$.)

But, for more general problems, the impact of the activation/deactivation of the source term in the C-method on numerical accuracy is not entirely obvious and is left for future research.

Chapter 2, in full, has been accepted for publication in Journal of Computational Physics. The dissertation author was the primary investigator of this paper. I would like to acknowledge the co-authors, Jon Reisner and Steve Shkoller.

Bibliography

[1] J. Murray, Mathematical Biology I: An Introduction, Springer-Verlag, 2002.

[2] S. Oruganti, J. Shi, R. Shivaji, Diffusive logistic euqation with constant yield harvesting, I: steady states, Transactions of the American Mathematical Society 354 (2002) 3601–3619.

[3] C. Pao, Nonlinear Parabolic and Elliptic Equations, Plenum Press, New York, 1992.

[4] I. Babuska, Error-bounds for finite element method, Numerische Mathematik 16 (1971) 322–333.

[5] M. H. Schultz, L^2 error bounds for the Rayleigh-Ritz-Galerkin method, SIAM Journal on Numerical Analysis 8 (1971) 737–748.

[6] J. Nitsche, Linear Spline-Funktionen und die Methoden von Ritz fur elliptische Randwertprobleme, Archive for Rational Mechanics and Analysis 15 (1970) 348–355.

[7] M. A. Noor, J. R. Whiteman, Error bounds for finite element solutions of mildly nonlinear elliptic boundary value problems, Numerische Mathematik 26 (1976) 107–116.

[8] J. Frehse, R. Rannacher, Asymptotic L^∞-error estimates for linear finite element approximations of quasilinear boundary value problems, SIAM Journal on Numerical Analysis 15 (1978) 418–431.

[9] M. Feistauer, A. Ženíšek, Finite element solution of nonlinear elliptic problems, Numerische Mathematik 50 (1984) 451–475.

[10] S.-S. Chow, Finite element error estimates for non-linear elliptic equations of monotone type, Numerische Mathematik 84 (1989) 373–393.

[11] G. A. Afrouzi, K. J. Brown, On a diffusive logistic equation, Journal of Mathematical Analysis and Applications 225 (1998) 326–339.

[12] R. S. Cantrell, C. Cosner, S. Martínez, Steady state solutions of a logistic equation with nonlinear boundary conditions, Rocky Mountain Journal of Mathematics 41 (2011) 445–455.

[13] Y. Wang, Y. Wang, J. Shi, Exact multiplicity of solutions to a diffusive logistic equation with harvesting, Applied Mathematics and Computation 216 (2010) 1531–1537.

[14] Y. Chen, L. Wu, Second Order Elliptic Equations and Elliptic Systems, Vol. 174 of Translations of Mathematical Monographs, American Mathematical Society, Rhode Island, 1998.

[15] W. Hackbusch, Elliptic Differential Equations: Theory and Numerical Treatment, Springer-Verlag, New York, 1992.

[16] D. Braess, Finite Elements: Theory, fast solvers, and applications in solid mechanics, Cambridge University Press, New York, 2007.

[17] P. Ciarlet, The Finite Element Method for Elliptic Problems, Vol. 4 of Studies in Mathematics and its Applications, North-Holland, 1978.

[18] P. Ciarlet, P.-A. Raviart, Maximum principle and uniform convergence for the finite element method, Computer Methods in Applied Mechanics and Engineering 2 (1973) 17–31.

[19] A. Harten, B. Engquist, S. Osher, S. R. Chakravarthy, Uniformly high-order accurate essentially non-oscillatory schemes, III, Journal of Computational Physics 71 (1987) 231 – 303.

[20] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, Journal of Computational Physics 77 (1988) 439 – 471.

[21] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, II, Journal of Computational Physics 83 (1989) 32 – 78.

[22] X.-D. Liu, S. Osher, T. Chan, Weighted essentially non-oscillatory schemes, Journal of Computational Physics 115 (1994) 200–212.

[23] G.-S. Jiang, C.-W. Shu, Efficient implementation of weighted ENO schemes, Journal of Computational Physics 126 (1996) 202–228.

[24] B. V. Leer, Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method, Journal of Computational Physics 32 (1979) 101 – 136.

[25] P. Colella, A direct eulerian MUSCL scheme for gas dynamics, SIAM Journal on Scientific and Statistical Computing 6 (1985) 104–117.

[26] H. T. Huynh, Accurate upwind methods for the Euler equations, SIAM Journal on Numerical Analysis 32 (1995) 1565–1619.

[27] P. Colella, P. R. Woodward, The piecewise parabolic method (PPM) for gas-dynamical simulations, Journal of Computational Physics 54 (1984) 174 – 201.

[28] J. A. Greenough, W. J. Rider, A quantitative comparison of numerical methods for the compressible Euler equations: fifth-order WENO and piecewise-linear Godunov, Journal of Computational Physics 196 (2004) 259 – 281.

[29] R. Liska, B. Wendroff, Comparison of several difference schemes on 1D and 2D test problems for the Euler equations, SIAM J. Sci. Comput 25 (2003) 995–1017.

[30] J. J. Quirk, A contribution to the great Riemann solver debate, Int. J. Num. Methods Fluids 18 (1994) 555–574.

[31] P. Colella, P. R. Woodward, The numerical simulation of two-dimensional fluid flow with strong shocks, Journal of Computational Physics 54 (1984) 115 – 173.

[32] A. Majda, S. Osher, Propagation of error into regions of smoothness for accuracte difference approximations to hyperbolic equations, Communications on Pure and Applied Mathematics 30 (1977) 671–705.

[33] M. Crandall, A. Majda, The method of fractional steps for conservation laws, Numerische Mathematik 34 (1980) 285–314.

[34] G.-S. Jiang, E. Tadmor, Nonoscillatory central schemes for multidimensional hyperbolic conservation laws, SIAM J. Sci. Comput 19 (1998) 1892–1917.

[35] C. Hu, C.-W. Shu, Weighted essentially non-oscillatory schemes on triangular meshes, Journal of Computational Physics 150 (1999) 97–127.

[36] R. D. Richtmyer, Proposed numerical method for calculation of shocks , LANL Report, LA- 671 (1948) 1–18.

[37] J. von Neumann, R. D. Richtmyer, A method for the numerical calculation of hydrodynamic shocks, Journal of Applied Physics 21 (1950) 232 –237.

[38] P. Lax, B. Wendroff, Systems of conservation laws, Comm. Pure Appl. Math. 13 (1960) 217–237.

[39] A. Lapidus, A detached shock calculation by second-order finite differences, Journal of Computational Physics 2 (1967) 154 – 177.

[40] R. A. Gentry, R. E. Martin, B. J. Daly, An Eulerian differencing method for unsteady compressible flow problems, J. Computational Physics 1 (1966) 87–118.

[41] F. H. Harlow, A. A. Amsden, A numerical fluid dynamics calculation method for all flow speeds, J. Computational Physics 8 (1971) 197–213.

[42] T. J. R. Hughes, M. Mallet, A new finite element formulation for computational fluid dynamics: IV. A discontinuity-capturing operator for multidimensional advective-diffusive systems, Computer Methods in Applied Mechanics and Engineering 58 (1986) 329 – 336.

[43] F. Shakib, T. J. R. Hughes, Z. Johan, A new finite element formulation for computational fluid dynamics: X. The compressible Euler and Navier-Stokes equations, Computer Methods in Applied Mechanics and Engineering 89 (1991) 141 – 219.

[44] J.-L. Guermond, R. Pasquetti, Entropy-based nonlinear viscosity for Fourier approximations of conservation laws, C.R. Acad. Sci. Paris 346 (2008) 801–806.

[45] P.-O. Persson, J. Peraire, Sub-cell shock capturing for discontinuous Galerkin Methods, Tech. rep., AIAA (2006).

[46] G. E. Barter, D. L. Darmofal, Shock capturing with PDE-based artificial viscosity for DGFEM: Part I. Formulation, Journal of Computational Physics 229 (2010) 1810 – 1827.

[47] R. Löhner, K. Morgan, J. Peraire, A simple extension to multidimensional problems of the artificial viscosity due to Lapidus, Communications in Applied Numerical Methods 1 (1985) 141–147.

[48] H. Nessyahu, E. Tadmor, Nonoscillatory central schemes for hyperbolic conservation laws, Journal of Computational Physics 87 (1990) 408–463.

[49] W. J. Rider, J. A. Greenough, J. R. Kamm, Accurate monotonicity-and extrema-preserving methods through adaptive nonlinear hybridizations, Journal of Computational Physics 225 (2007) 1827–1848.

[50] E. F. Toro, Riemann solvers and numerical methods for fluid dynamics, Springer-Verlay Berling Heidelberg, 2009.

[51] R. J. DiPerna, Convergence of approximate solutions to conservation laws, Archive for Rational Mechanics and Analysis 82 (1983) 27–70.

[52] P.-L. Lions, B. Perthame, P. E. Souganidis, Existence and stability of entropy solutions for the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian coordinates, Comm. Pure Appl. Math. 49 (6) (1996) 599–638.

[53] S. Bianchini, A. Bressan, Vanishing viscosity solutions of nonlinear hyperbolic systems, Annals of Mathematics. Second Series 161 (2005) 223–342.

[54] T. A. B.E. Robertson A.V. Kravtsov N.Y. Gnedin, D. Rudd, Computational Eulerian hydrodynamics and Galilean invariance, Mon. Not. R. Astron. Soc. 401 (4) (2010) 2463–2476.

[55] T. W. Roberts, The behavior of flux difference splitting schemes near slowly moving shock waves, Journal of Computational Physics 90 (1990) 141–160.

[56] K. N. Chueh, C. C. Conley, J. A. Smoller, Positively invariant regions for systems of nonlinear diffusion equations, Indiana University Mathematics Journal 26 (1977) 373–392.

[57] S. Larsson, V. Thomée, Partial differential equations with numerical methods, Springer-Verlay Berling Heidelberg, 2003.

[58] B. P. Leonard, The ULTIMATE conservative difference scheme applied to unsteady one-dimensional advection, Comput. Methods Appl. Mech. Engrg. 88 (1991) 17–74.

[59] A. Harten, S. Osher, Uniformly high-order accurate nonoscillatory schemes. I, SIAM Journal on Numerical Analysis 24 (1987) pp. 279–309.

[60] G. Sod, A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws, Journal of Computational Physics 27 (1978) 1 – 31.

[61] J. S. Hesthaven, T. Warburton, Nodal discontinuous Galerkin methods, algorithms, analysis and applications, Vol. 54 of Texts in Applied Mathematics, Springer, 2008.

[62] W. Liu, J. Cheng, C.-W. Shu, High-order conservative Lagrangian schemes with Lax-Wendroff type time discretization for the compressible Euler equations, Journal of Computational Physics 228 (2009) 8872–8891.

[63] R. Loubère, M. Shashkov, A subcell remapping method on staggered polygonal grids for arbitrary-Lagrangian-Eulerian methods, Journal of Computational Physics 209 (2005) 105–138.