Multilevel Least-Squares Finite Element Methods for Hyperbolic Partial Differential Equations

by

Luke Nathan Olson

B.A., Luther College, 1997

M.S., University of Iowa, 1999

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Applied Mathematics

2003

This thesis entitled:
Multilevel Least-Squares Finite Element Methods for Hyperbolic Partial Differential Equations
written by Luke Nathan Olson
has been approved for the Department of Applied Mathematics

Thomas A. Manteuffel

Stephen F. McCormick

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Olson, Luke Nathan (Ph.D., Applied Mathematics)

Multilevel Least-Squares Finite Element Methods for Hyperbolic Partial Differential Equations

Thesis directed by Prof. Thomas A. Manteuffel

Least-Squares Finite Element Methods (LSFEMs) for partial differential equations of hyperbolic type are studied in 2-D. The focus of our studies is on linear convection problems that exhibit discontinuities in the solution and nonlinear conservation laws that develop shocks.

The linear equations are rewritten as a minimization problem associated with an L^2 functional. The space of admissible boundary data is studied, a trace theorem and Poincaré inequality are developed, and well-posedness of the formulation is proved. Both conforming and nonconforming finite element spaces are considered. A discontinuous Least-Squares Finite Element Method (DLSFEM) is proposed. Convergence properties and solution quality for discontinuous solutions are investigated in detail for conforming and nonconforming finite elements of increasing polynomial degree. We present strategies for effectively solving the linear system in a multigrid setting.

Finally, an $H(div)$-conforming least-squares method is introduced that is able to handle nonlinear hyperbolic conservation laws. This formulation is related to a least-squares H^{-1}-minimization. Convergence of the numerical approximation to a weak solution of the conservation law is proved and the theory is supported by numerical tests.

To Kjellrun

Acknowledgements

The work presented in this dissertation is made possible through the guidance and support of my advisor Tom Manteuffel, Steve McCormick and Hans De Sterck. I thank them for their patient teaching and continual direction in my research and professional development and for the many "meetings" we have had at Gold Hill, Copper Mountain, and elsewhere on the trail. I also extend my gratitude to John Ruge for his great software package, FOSPACK, and for his extensive insight into our problem.

I was fortunate enough to share an office with Scott MacLachlan and Chad Westphal. They have been extremely helpful colleagues and very good friends, along with the rest of the Grandview Group: Oliver Röhrle, Marian Brezina, James Brannick, Jeff Heys, Eunjung Lee, and Schorsch Schmidt.

I also wish to thank the Department of Applied Mathematics and NSF VIGRE for funding this research along with the Center for Applied Scientific Computing at Lawrence Livermore National Lab for summer support.

Finally, I thank my wife, Kjellrun, and my family for their encouragement.

Contents

Tables

Table

Figures

Figure

Chapter 1

Introduction

Linear hyperbolic PDEs allow for discontinuous solutions (so-called contact discontinuities) when the boundary data is discontinuous. It is difficult to develop numerical methods that are high-order accurate where the solution is smooth, yet sharply resolve discontinuities without introducing spurious oscillations [5]. For wide classes of elliptic PDEs, optimal multilevel iterative solution algorithms have been developed for the discrete linear algebraic systems that require only $O(n)$ operations, where n is the number of unknowns (see e.g. [79, 20] and references therein). However, for hyperbolic and mixed elliptic-hyperbolic PDEs, such optimal iterative solvers have been elusive, although some promising results have been reported [80].

The general philosophy behind the approach pursued in this dissertation is to combine adaptive least-squares (LS) finite element discretizations on space-time domains with global implicit solves using optimal iterative methods, in particular, algebraic multigrid (AMG). There are important difficulties that have to be overcome. Optimal $O(n)$ solvers are still an active research topic for general hyperbolic and mixed elliptic-hyperbolic PDEs. A strong motivation for our choice of LS discretizations is that optimal solvers are more easily designed for the symmetric positive-definite (SPD) matrices that result from LS discretizations. We remedy the extra smearing at discontinuities that LS introduces, by adaptive refinement based on the natural, sharp error estimator that the LS functional provides. The scope of this dissertation encompasses the theoretical aspects of the LSFEM for continuous and discontinuous elements, as well as a numerical study of AMG performance.

In Chapter 2, we begin with an overview of essential background material. We outline relevant aspects of hyperbolic conservation law theory and highlight important properties of the solution, which is necessary in understanding the underlying difficulty in developing appropriate discretizations. A summary of popular numerical methods in finite difference and finite element settings is presented. Advantages and shortcomings of these methods are discussed in order to properly motivate the methods we subsequently develop. We follow with an overview of least-squares finite element methods in Chapter 3. In particular, the least-squares finite element method for hyperbolic PDEs with smooth solutions is introduced. For hyperbolic problems with smooth solutions, finite element convergence is nearly optimal, as we show both theoretically and numerically. However, the convergence rates are significantly reduced when the solution is continuous, but not pointwise differentiable. Chapter 3 mainly discusses the results of existing least-squares finite element methods applied to continuous flows, while, in Chapter 4, we propose a formal study of least-squares finite element methods for hyperbolic PDEs with discontinuous solutions. We propose a provably convergent finite element method in both a conforming and nonconforming setting that holds for discontinuous flows. A thorough numerical study is presented and we discuss the convergence and solution quality for high-order finite elements.

The performance of a multilevel iterative solver is addressed in Chapter 5. We show that a standard geometric multigrid method is insufficient and motivate the use of an algebraic based scheme. Proper implementation of the AMG algorithm is discussed and some encouraging initial results are presented. With this information, we correct the least-squares functional and theoretically justify its minimization. The resulting discretization is shown to yield optimal/near-optimal multigrid performance for our test cases.

In Chapter 6, we address nonlinear hyperbolic conservation laws. We discuss weak solutions in a new context and reformulate the PDE to emphasize the smoothness of the flux vector. The reformulation introduces the flux vector, or the associated flux potential, explicitly as additional dependent variables. An $H(div)$-conforming least-squares finite element method is proposed and we use a Gauss-Newton nonlinear solution strategy in an adaptive grid contin-

uation framework to study the problem numerically. Exact discrete conservation has previously been considered a strong requirement for nonlinear hyperbolic conservation laws. However, we show that our method converges to a weak solution without this property. A numerical study is presented and a theoretical footing is established that confirms these claims.

The contributions of this dissertation to the fields of hyperbolic conservation laws, finite element methods, and multigrid iterative methods are threefold and interrelated. First, a comprehensive study of least-squares methods for linear hyperbolic PDEs is conducted for solutions with discontinuities. Previously, this has not been considered in depth. We provide a theoretical foundation and numerical evidence to support the least-squares finite element approach for this problem. The LSFEM and DLSFEM that are outlined in this dissertation are closely related to previous formulations, but have been considered only in the context of smooth solutions. A comparative numerical investigation of conforming and nonconforming finite elements with varying polynomial degree is also presented and unique to this thesis. Second, a formal study of the Algebraic Multigrid (AMG) method for the least-squares finite element discretization of hyperbolic PDEs is initiated. Certainly, multigrid methods have been developed and scrutinized for PDEs of hyperbolic type. However, a careful study of the standard Ruge-Stuüben implementation applied to the system resulting from a least-squares finite element discretization of linear hyperbolic PDEs has not surfaced. We present numerical evidence indicating the need for special treatment of the boundary conditions in order to achieve optimal convergence of the multilevel process. Moreover, we propose a novel strategy to amend the difficulties introduced by the least-squares discretization. This is accomplished by supplying additional terms to the PDE. With this, optimal/near-optimal AMG performance is achieved for our test cases. Third, a new least-squares finite element method is introduced for nonlinear hyperbolic conservation laws. The PDEs are reformulated and a least-squares method is proposed that conforms with the nature of the solution. We commence a study of this approach, developing the theoretical justification and validating the results numerically. The contributions noted above are mainly presented in Chapters 4, 5, and 6.

Chapter 2

Introduction to Hyperbolic Conservation Laws

In this chapter, we outline the derivation of hyperbolic conservation laws and highlight various analytical properties of these models. We follow with a summary of concepts and objectives in numerical solution methods for this class of problems. Our discussion in this chapter is important for understanding more general PDEs of hyperbolic type, which we study throughout the dissertation. Moreover, reviewing traditional computational approaches to conservation laws motivates the methods we develop.

We open with a general discussion of hyperbolic conservation laws and their applications in Section 2.1. The basic differential form of these equations is derived and discussed in Section 2.2. We note that this form allows for discontinuous solutions, which are the focus of this dissertation. In Section 2.3, we investigate nonlinear hyperbolic PDEs and give a brief overview of shocks, rarefactions, and discontinuous solutions. To understand the numerical solution of nonlinear hyperbolic equations, we first develop a firm foundation of the linear setting used throughout this dissertation. In Section 2.4, we briefly consider the numerical methods most widely used in solving large classes of hyperbolic conservation laws. Understanding the advantages and shortcomings of conventional numerical techniques is important for motivating the methods we develop.

2.1 Conservation laws

PDEs of hyperbolic type, often in the form of conservation laws, are an important class of equations that arise in many applications of continuum physics. The Euler equations of gas dynamics form a widely used system of hyperbolic conservations laws. These equations are used to model phenomena arising in such fields as astrophysics, aerodynamics, and other areas involving the fluid dynamics of gases. Isotropic and isothermal versions of the Euler equations are also typically studied in the area of hyperbolic conservation laws. Although the Euler equations appear basic even in their general form, numerical analysts often use simpler scalar equations to examine and isolate the behavior of a given method. We also assume this approach by focusing our study on the linear advection equation and the (nonlinear) Burger's equation.

2.1.1 Conservation and Applications

Conservation laws are derived directly from physical principles. These laws allow us to model conservation of energy, mass, momentum, etc., and from here the applications are endless. The dependent or conserved variable we seek to describe in the model represents a density of a conserved quantity, such as mass. We describe the conservation of variables with a set of equations that balances the fluxes across a given volume element in the domain. More precisely, as we discuss in further detail in Section 2.2, suppose that $u(x,t)$ is an unknown density function of a conserved quantity and $F(u)$ is a known flux function. Given a 1-D test volume element, $[x_1,\, x_2]$, we know that a change in the conserved quantity (mass) can only occur by fluxes at the edges of the test element. We thus arrive at

$$\frac{d}{dt}\int_{[x_1,\,x_2]} u(x,t)\,dx = F(u(x_1,t)) - F(u(x_2,t)). \tag{2.1}$$

This relation describes conservation in **integral** form. The corresponding **differential** form

$$u_t + (F(u))_x = 0 \tag{2.2}$$

is easily derived, keeping in mind that solutions are **weak** since (2.2) holds in a distribution sense. A notable feature of this equation is that it is a nonlinear time-dependent PDE. An important

goal in this thesis is the treatment of (2.2) in a space-time domain. We discuss the advantages and implications of such a treatment in Section 2.4.3 and Chapter 4. Equations of type (2.2) are the focus of this dissertation, especially those that admit weak solutions with discontinuities. For the case of nonlinear hyperbolic conservation laws, we study solutions involving shocks and rarefactions, for the linear model, we consider contact discontinuities.

The differential form, (2.2) of (2.1), arises in many applications:

Linear advection or transport equation. This is the standard model to describe transport of a substance along a flow. Here we present an example that may be nonconservative and allows a source (forcing) term:

$$u_t + a(x,t)u_x = f(x,t). \tag{2.3}$$

In this case, $a(x,t)$ denotes the advection speed at a point (x,t) in the domain and $f(x,t)$ is a known source term.

Particle or reactive transport. The source can also depend on the concentration ψ and is often in the form of a reaction term:

$$\frac{1}{\nu}\psi_t + \nabla \cdot F(\psi) = K\psi + f, \tag{2.4}$$

where K represents the absorption/reaction operator [63] and ν the particle speed.

Euler equations. The Euler equations of compressible gas dynamics are a special case of the compressible Navier-Stokes equations for fluids. Conservation of mass, momentum, and energy is described in 1-D by

$$\begin{bmatrix} \rho \\ \rho u \\ \rho e \end{bmatrix}_t + \begin{bmatrix} \rho u \\ \rho u^2 + p \\ (\rho e + p)u \end{bmatrix}_x = \mathbf{0}, \tag{2.5}$$

where ρ, u, p, and e represent mass density, velocity, pressure and specific total energy, respectively [62].

MHD. A magnetohydrodynamical (MHD) representation of a plasma describes the plasma as a conducting fluid. This allows, for example, the modeling of shocks that arise in many astrophysical applications. In 3-D, we have

$$\frac{d}{dt}\begin{bmatrix}\rho\\ \rho\mathbf{u}\\ \mathbf{B}\\ \rho e\end{bmatrix}+\nabla\cdot\begin{bmatrix}\rho\mathbf{u}\\ \rho\mathbf{u}\mathbf{u}+\mathcal{P}\\ \mathbf{u}\mathbf{B}^T-\mathbf{B}\mathbf{u}\\ (\rho e+\mathcal{P})\mathbf{u}\end{bmatrix}=\mathbf{0}, \tag{2.6}$$

where ρ, $\mathbf{u}$, $\mathbf{B}$, and p are the mass density, velocity, magnetic field and pressure, respectively. $\mathcal{P}$ is a pressure tensor involving p and $\mathbf{B}$, while e is the specific total energy, which depends on u, $\mathbf{B}$, ρ, and $\mathcal{P}$ [30].

Maxwell's equations. The hyperbolic form of Maxwell's equations arises when the current is assumed to be zero. The equations are given by

$$\mathbf{E}_t-\frac{1}{\epsilon\mu}\nabla\times\mathbf{B}=0 \tag{2.7}$$

$$\mathbf{B}_t+\nabla\times\mathbf{E}=0, \tag{2.8}$$

which can be written in standard hyperbolic form as the system

$$\mathbf{u}_t+\sum_k\mathbf{A}_k(\mathbf{x})\mathbf{u}_{\mathbf{x}_k}+\mathbf{B}\mathbf{u}=\mathbf{0}, \tag{2.9}$$

where $u=\begin{bmatrix}\mathbf{E}\\ \mathbf{B}\end{bmatrix}$ and $\mathbf{A}_k$, and $\mathbf{B}$ depend on the material's permittivity, ϵ, and magnetic permeability, μ [62].

Elastic waves in solid mechanics. These equations model acoustic waves in a solid. Shear waves and surface waves in the material are coupled and a hyperbolic system is obtained in the equations of linear elasticity. The resulting system is closely related to the standard acoustic wave equation [62].

2.1.2 Classification of the First-Order Linear Hyperbolic PDE

Just as analytical techniques for solving partial differential equations are developed for certain classes of equations, so are numerical solution methods. The problems discussed throughout this dissertation are PDEs of **hyperbolic** type, and we must properly define this class of equations. Classification of a first-order system of PDEs as hyperbolic is rooted in the traditional method for characterizing second-order PDEs (cf. [53], [41], [75], [36]). Classification not only formalizes the equations we study, but also offers a deeper look at the nature of the PDE in relation to other classes of equations.

The quadratic curve

$$ax^2 + 2bxy + cy^2 = 0 \tag{2.10}$$

is **nondegenerate** when $\begin{vmatrix} a & b \\ b & c \end{vmatrix} \neq 0$. These (real) nondegenerate curves arise geometrically in slices of a cone and appear as three distinct patterns. For $ac - b^2 > 0$, $ac - b^2 < 0$, and $ac - b^2 = 0$, these curves define an ellipse, parabola, and a hyperbola, respectively. The geometric identification carries directly over to the classification of second-order PDEs. Let

$$a(x,y)u_{xx} + 2b(x,y)u_{xy} + c(x,y)u_{yy} = d(x,y,u,u_x,u_y), \tag{2.11}$$

be a linear second-order PDE, where d is linear in u, u_x, and u_y. Let Γ be a curve in $\Omega \in \mathbf{R}^2$ that is described parametrically by

$$x = X(s), \qquad y = Y(s). \tag{2.12}$$

Following the motivation presented in [53], notice that

$$\frac{du_x}{ds} = u_{xx}X'(s) + u_{xy}Y'(s) \tag{2.13}$$

and

$$\frac{du_y}{ds} = u_{xy}X'(s) + u_{yy}Y'(s). \tag{2.14}$$

Equations (2.11), (2.13), and (2.14) define the values u_{xx}, u_{yy}, and u_{xy} uniquely if

$$\Delta = \det \begin{bmatrix} a & 2b & c \\ X' & Y' & 0 \\ 0 & X' & Y' \end{bmatrix} = aY'^2 - 2bX'Y' + cX'^2 \neq 0. \tag{2.15}$$

This means that the solution away from the curve can be obtained through continuation, given the boundary values for $\frac{du_x}{ds}$ and $\frac{du_y}{ds}$. If $\Delta = 0$, then the curve Γ is defined to be **characteristic**, a term that appears often in partial differential equations of hyperbolic type. Boundary conditions cannot be specified on a characteristic curve. In the characteristic case, we can write (2.15) as

$$\frac{Y'(s)}{X'(s)} = \frac{dy}{dx} = \frac{b \pm \sqrt{b^2 - ac}}{a}. \tag{2.16}$$

This results in different cases depending on the sign of $ac - b^2$. With $ac - b^2 > 0$, the right side of (2.16) is not real and thus does not define a meaningful characteristic. We follow the geometric equivalent (2.10) and label this type as **elliptic**. If $ac - b^2 = 0$, then (2.16) defines one characteristic curve and corresponds to a **parabolic** PDE. With $ac - b^2 < 0$, (2.16) defines two characteristic curves, and we define this case to be hyperbolic.

Intuitively, this classification scheme is consistent. Elliptic equations typically involve a diffusion process and result in smooth solutions (characteristics in a diffusive solution are not meaningful). A parabolic equation, on the other hand, is usually time dependent, resulting in a preferred direction (time). Parabolic equations provide a transition to hyperbolic problems, like the the wave equation,

$$u_{tt} = C^2 u_{xx}, \tag{2.17}$$

where C is a constant and t is time. The solution of (2.17) propagates from an initial profile along two characteristics. This is also consistent with the classification discussed above.

Remark 2.1. *Although we have only classified PDEs of hyperbolic type in 2-D, the terminology extends to second-order equations in higher dimensions. A linear second-order PDE in* $\mathbf{R}^d$ *has*

the form

$$\sum_{i,j=1}^{d} a_{ij} u_{x_i x_j} + \sum_{k=1}^{d} b_k u_{x_k} + cu = d, \tag{2.18}$$

where a_{ij}, b_k, c, and d are only functions of $\mathbf{x}$ for each i,j, and k. As shown in [75], there exists a transformation such that the high-order terms in (2.18) can be written as

$$\sum_{i=1}^{d} d_i u_{x_i x_i}, \tag{2.19}$$

where d_i are the eigenvalues of $A = \{a_{ij}\}$ and can be scaled so that $d_i = 0, \pm 1$ for $i = 1 \ldots d$. Then (2.18) is elliptic if $d_i < 0$ for all i or if $d_i > 0$ for all i. Similarly, the PDE is parabolic if $d_i = 0$ for some $i = j$ and the d_i have the same sign for all other $i \neq j$. Finally, (2.18) is hyperbolic if there is exactly one $d_i \neq 0$ with a different sign than the remaining nonzero eigenvalues.

We now consider first-order systems of partial differential equations. Recall the wave equation (2.17) and write this second-order hyperbolic PDE as the first-order system

$$\begin{bmatrix} v \\ w \end{bmatrix}_t + \begin{bmatrix} 0 & -C^2 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix}_x = \mathbf{0}. \tag{2.20}$$

Writing this system as

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = \mathbf{0}, \tag{2.21}$$

notice that the eigenvalues of $\mathbf{A}$ are $\pm C$. We have real and distinct eigenvalues for this matrix, which leads to linearly independent eigenvectors. Let $\mathbf{R}$ be the matrix of (right) eigenvectors of $\mathbf{A}$. Since the eigenvectors are linear independent, $\mathbf{R}$ is invertible and $\mathbf{A}$ is diagonalizable by

$$\mathbf{\Lambda} = \mathbf{R}^{-1}\mathbf{A}\mathbf{R}, \tag{2.22}$$

where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues.

Equation (2.21) can then be written as

$$\mathbf{u}_t + \mathbf{R}\mathbf{\Lambda}\mathbf{R}^{-1}\mathbf{u}_x = \mathbf{0}. \tag{2.23}$$

Multiplying by $\mathbf{R}^{-1}$ and letting $\tilde{\mathbf{u}} = \mathbf{R}^{-1}\mathbf{u}$, we arrive at

$$\tilde{\mathbf{u}}_t + \mathbf{\Lambda}\tilde{\mathbf{u}}_x = \mathbf{0}. \tag{2.24}$$

This is a decoupled system of first-order equations.

The resulting first-order form for the (second-order) wave equation is indicative of the form for general first-order systems of hyperbolic PDEs. The following definition, used in [59],[62],[53] et al., characterizes the first-order system of linear PDEs.

Definition 2.2. *The linear first-order system of equations*

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = \mathbf{0} \tag{2.25}$$

are said to be ***hyperbolic*** *if* $\mathbf{A}$ *is diagonalizable with real eigenvalues.*

We can extend this definition to matrices $\mathbf{A}(u,t,x)$. The problem is hyperbolic at a given point (u,t,x) if $\mathbf{A}(u,t,x)$ is diagonalizable with real eigenvalues. The eigenvalues λ_i determine the slopes of the characteristic curves. Furthermore, nonlinear PDEs in conservation form (2.2) are called hyperbolic if the Jacobian matrix $F'(u)$, where $F(u)$ is the flux function, satisfies these conditions for any admissible u.

Remark 2.3. *Extending Definition 2.2 to d spatial dimensions follows similarly. A general system*

$$\mathbf{u}_t + \sum_{k=1}^{d} \mathbf{A}_k \mathbf{u}_{\mathbf{x}_k} = \mathbf{0} \tag{2.26}$$

is called hyperbolic if the matrix

$$\sum_{k=1}^{2} n^k \mathbf{A}_k \tag{2.27}$$

is diagonalizable with real eigenvalues for any direction $\mathbf{n} = (n^1, \ldots, n^d)$ *(cf. [62]).*

2.2 Multidimensional Nonlinear Systems of Hyperbolic Conservation Laws

In this section, we explore multidimensional nonlinear systems of hyperbolic conservation laws in more detail. The majority of this thesis concerns two-dimensional problems (two spatial

dimensions or one dimension in both space and time), although the methodology we develop extends in a fairly straightforward way to higher dimensions. We consider both scalar and systems of hyperbolic PDEs in our formulations, thus motivating a closer look at the general form of conservation laws. A differential form of a general multidimensional nonlinear hyperbolic conservation law is first justified in Section 2.2.1. In Section 2.2.2, weak solutions are discussed. Throughout this preliminary section, we assume that the PDE satisfies the basic properties of a first-order hyperbolic PDE outlined in Section 2.1.2. Flux functions are assumed to be smooth and convex, and we impose further constraints as needed.

2.2.1 Derivation of Conservation Laws

Let Ω be a domain in $\mathbb{R}^d$. Let $\mathbf{u} \in \mathbb{R}^m$ represent a vector of real-valued variables,

$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} u_1(\mathbf{x}) \\ u_2(\mathbf{x}) \\ \vdots \\ u_m(\mathbf{x}) \end{bmatrix}, \tag{2.28}$$

where $\mathbf{x} \in \mathbb{R}^d$. Assume that there is a vector-valued function $\mathbf{F} = (\mathbf{F}_1(\mathbf{u}), \mathbf{F}_2(\mathbf{u}), \cdots, \mathbf{F}_d(\mathbf{u}))$ that defines the flux of $\mathbf{u}$ in Ω. $\mathbf{F}$ is a vector of vector-valued functions $\mathbf{F}_i(\mathbf{u}) \in \mathbb{R}^m$, for $i = 1, \ldots, d$. Each $\mathbf{F}_i(\mathbf{u})$ defines the flux in the x_i-direction. Assume that F_{ij} is a convex (or concave) function of u_j for $i = 1, \ldots, d$ and $j = 1, \ldots, m$. This assumption simplifies the analytical discussion and more general assumptions are regarded as "non-classical" in field of hyperbolic conservation laws.

Let s parameterize the $d-1$ manifold $\partial\Omega$ and let $\mathbf{n}(s) = (n_1(s), n_2(s), \ldots, n_d(s))$ be the unit outward normal to $\partial\Omega$ at $\mathbf{x}(s) = (x_1(s), x_2(s), \ldots, x_d(s))$. We define

$$\mathbf{n} \cdot \mathbf{F} = n_1\mathbf{F}_1(\mathbf{u}) + \cdots + n_d\mathbf{F}_d(\mathbf{u}) \tag{2.29}$$

and

$$\nabla \cdot \mathbf{F} = \partial_{x_1}\mathbf{F}_1(\mathbf{u}) + \cdots + \partial_{x_d}\mathbf{F}_d(\mathbf{u}). \tag{2.30}$$

Conservation of $\mathbf{u}$ means that a temporal change in the volume integral of $\mathbf{u}$ can only be due to fluxes across the boundary of the test domain. We then have

$$\frac{d}{dt}\int_{\Omega'} \mathbf{u}(\mathbf{x},t)\,d\mathbf{x} = -\int_{\partial\Omega'} \mathbf{n}\cdot\mathbf{F}(\mathbf{u})\,ds, \tag{2.31}$$

where Ω' is an arbitrarily small, non-moving volume in Ω. The right side of equation (2.31) is the net flux across the boundary of Ω'. Applying the Gauss Divergence Theorem component-wise, we arrive at

$$\frac{d}{dt}\int_{\Omega'} \mathbf{u}(\mathbf{x},t)\,d\mathbf{x} = -\int_{\Omega'} \nabla\cdot\mathbf{F}(\mathbf{u})\,d\mathbf{x}, \tag{2.32}$$

and

$$\int_{\Omega'} \mathbf{u}_t(\mathbf{x},t) + \nabla\cdot\mathbf{F}(\mathbf{u})\,d\mathbf{x} = \mathbf{0}. \tag{2.33}$$

Since (2.33) holds for any test volume, we can conclude that the integrand is zero on Ω, which defines the **differential** form of the conservation law as

$$\mathbf{u}_t + \nabla\cdot\mathbf{F}(\mathbf{u}) = \mathbf{0}. \tag{2.34}$$

As we turn to the numerical solution of hyperbolic conservation laws in the remaining chapters of this dissertation, we use the differential form as a basis for development. However, problems of the form (2.34), with solutions that possess discontinuities, are the major focus in our study. We consider discontinuous solutions that arise from jumps in the prescribed boundary data and also from the formation of shocks as time progresses. Such solutions do not satisfy (2.34) pointwise, but we implicitly assume that the solution actually satisfies the integral form, (2.31). The notion of a weak solution allows us to interpret discontinuous $\mathbf{u}$ as a valid solutions to the conservation law.

2.2.2 Weak Solutions and Uniqueness

The integral definition, (2.31), is a **weak** formulation in the sense that shocks, jumps, and, in general, profiles with low continuity are valid solutions. Solutions that satisfy the integral form, or the differential form in a distribution sense, are called weak solutions. Although the

weak solution satisfies the integral form of conservation laws, we need to be concerned about the uniqueness of this solution. In fact, in many cases, for a given choice of prescribed boundary and initial data, there are more than one weak solution that satisfy the integral form of the conservation law. This difficulty frequently arises in nonlinear applications where shocks are forming after a finite time [62].

There may be many valid solutions to the weak form of the PDE, but only one solution is stable and thus physically acceptable. As a result, many solution methods are designed to ensure that the PDE allows only the physical solution or that a numerical solution approximates the physical solution. In many applications, a hyperbolic conservation law is used to model a particular phenomenon and viscosity is assumed to be zero, as we outlined in Section 2.1. The absence of diffusion is central to hyperbolic conservation laws, by definition. Many physical problems contain a tiny amount of diffusion that is ignored by the hyperbolic PDE, yet this small amount of viscosity has physical significance. A popular idea is to view the physically meaningful solution to the conservation law as the limit case of a problem with a small amount of "artificial" diffusion. As the diffusive term diminishes, we expect to arrive at the solution to the purely hyperbolic form. Introducing this small amount of dissipation into the conservation laws is also referred to as "vanishing viscosity" and forces the PDE to be well-posed in the sense of a unique solution. This technique is valuable in analyzing the problem mathematically. However, the form and classification of the problem changes and often increases the computational complexity. Convergence analysis for problems with added numerical viscosity requires special attention. If the rate of convergence of the numerical solution is too slow, then the converged approximation may not be the physical solution that we expect in the limit case. That is, the amount of diffusion added may not be well balanced with the computational mesh size. However, ensuring convergence to the physical solution by requiring a minimum rate of convergence may be unattainable and reveals that the method of vanishing viscosity is not necessarily a viable approach.

Another strategy is to impose additional physical constraints that ensure the selection of

the desired solution. The basic idea is to identify admissible solutions by verifying that certain physical conditions hold. This approach originated in gas dynamics, where it is known that the **entropy** across shocks should increase. Other applications maintain similar properties, and these types of conditions are known as entropy conditions. Similar to the case of artificial diffusion, correct entropy conditions are often difficult to define in general applications. We do not investigate the question of convergence to the entropy solution in this thesis. However, it is possible that our approach implicitly imposes such entropy conditions [5, 56, 60, 62], because the entropy solution seems to be what we obtain in our numerical tests.

2.2.3 Source Terms

Although we have developed hyperbolic PDEs in the form of conservation laws, resulting in homogeneous right sides, we can describe these equations more generally. Specifically, consider the **balance law** [62] equation

$$u_t + f(u)_x = \delta(x, t, u). \tag{2.35}$$

A source term, δ, arises in many of the applications described in Section 2.1. For example, particle transport models include source terms to represent neutron generators, scattering terms, and contributions from secondary particles [63]. Sources are also due to chemical reactions in fluids and external forces on the medium, such as gravity and forces due to motion [62]. Finite difference and finite volume methods tend to treat balance laws differently than conservation laws through splittings and fractional step methods. The least-squares formulation we develop in Chapter 4 naturally allows non-zero right sides, while the $H(div)$-conforming solution methods we derive in Chapter 6 need special consideration. For these methods, the source term can be absorbed in the equation by using a modified lifting argument, and we address these issues specifically in the respective chapters. In general, we develop our methods in a least-squares setting, where the effects of a source term on the numerical approximation have been addressed by Bochev and Choi [11].

2.3 Discontinuities in Hyperbolic PDEs

As described to in the previous sections, solutions to hyperbolic conservation laws may contain discontinuities. Discontinuities may be propagated across the domain due to jumps in the prescribed boundary data, or can emerge in the form of a shock in the case of nonlinear equations. The notion of characteristics is essential in understanding this latter type of solution. Hyperbolic problems are purely anisotropic, so a perturbation must follow a particular direction that is defined by characteristic curves. We give an overview of characteristics in Section 2.3.1 to better understand the behavior of the solution and to clarify and motivate several proofs presented in the derivation of least-squares finite element methods in Chapter 4. We present a brief description of shocks and discontinuities in Section 2.3.2.

2.3.1 Method of Characteristics

Characteristics were mentioned in Section 2.1.2 as attributes that are unique to PDEs classified as hyperbolic. Solutions to hyperbolic PDEs are entirely defined by these characteristics, and we articulate this idea by considering simple forms of hyperbolic conservation laws. Although analytical solutions to these basic problems are found quite easily, exact solutions to more complicated problems are not always feasible. However, studying the behavior of basic model problems does build intuition for more complex systems of conservation laws, and this approach what we we use throughout the dissertation.

Consider the scalar form of the conservation law (2.34) in 1-D and defined on $\mathbb{R} \times [0,\, T]$:

$$u_t + f(u)_x = 0, \qquad x \in \mathbb{R},\, t \in [0,\, T], \tag{2.36a}$$

$$u(x, 0) = u_0(x), \quad x \in \mathbb{R},\, t \in [0,\, T], \tag{2.36b}$$

where $u_0(x) \in L^\infty(\mathbb{R})$ is a given initial condition and $f \in C^\infty$ is a know flux function. Problem (2.36) is called the Cauchy problem and is commonly used to discuss various solution qualities. The total derivative in the direction of transport along the curve $x(t)$, is zero by virtue of the

conservation law:

$$\frac{du(X(t),t)}{dt} = \frac{\partial u(X(t),t)}{\partial t} + \frac{dX(t)}{dt} \cdot \frac{\partial u(X(t),t)}{\partial x} = 0. \tag{2.37}$$

Writing (2.36a) as

$$u_t + f'(u)u_x = 0, \tag{2.38}$$

the curves, $X(t) : t \mapsto (t, X(t))$, now define **characteristics** of the solution according to he following

$$\frac{dX(t)}{dt} = f'(u(X(t),t)). \tag{2.39}$$

As an example, consider the linear case when $f'(u) = a$. The characteristics are given by

$$\frac{dX(t)}{dt} = a. \tag{2.40}$$

Letting $x_0 = X(0)$, we have $X(t) = at + x_0$. and recalling that $u(X(t),t) =$ constant, we have

$$u(at + x_0, t) = u(x_0, 0) = u_0(x_0) \tag{2.41}$$

and

$$u(x,t) = u_0(x - at). \tag{2.42}$$

Here we see that the entire solution is defined by the values prescribed on the inflow boundary. This is the nature of hyperbolic PDEs. Problems with variable coefficients and nonlinear equations can be solved in a similar manner, although solutions are not typically constant along characteristics, and characteristic curves may intersect.

2.3.2 Shocks and Rarefactions

In this section, we review various solution types that are considered throughout the remainder of this dissertation. A large part of this thesis focuses on discontinuous solutions, and we present them here in the form of discontinuous boundary data transported across the domain and in the form of shocks formed by nonlinearities. The analytic solutions presented satisfy the integral or weak form of the equation.

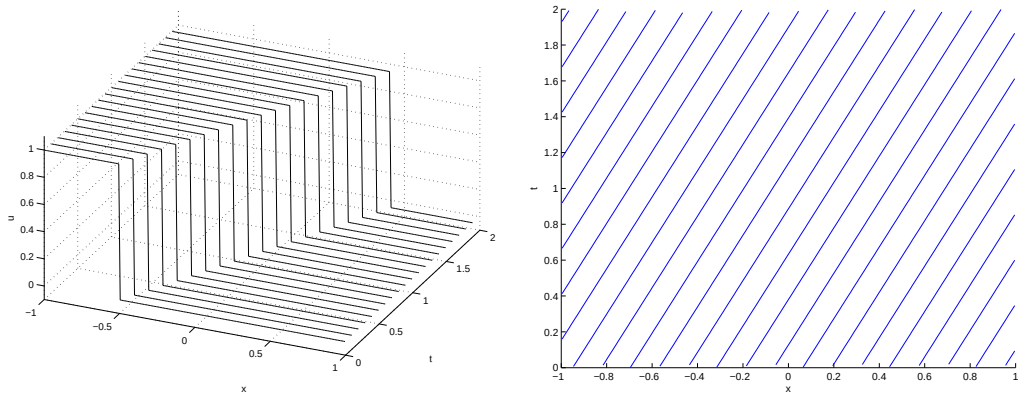
Consider again the conservation law

$$u_t + f(u)_x = 0, \qquad x \in [-1,\, 1]\, t \in [0,\, T], \tag{2.43a}$$

$$u(0,t) = g(t), \qquad t \in [0,\, T], \tag{2.43b}$$

$$u(x,0) = u_0(x), \quad x \in [-1,\, 1]. \tag{2.43c}$$

Analytic solutions to this problem are easily derived using the method of characteristics in certain cases. We present several examples to describe different types of solution profiles, which are important in Chapters 4 and 6.



(a) Solution profile (b) Characteristics

Figure 2.1: Linear solution with advection speed equal to $\frac{1}{2}$.

Linear advection. Consider the linear form of the flux function, $f(u) = au$, where $a = \frac{1}{2}$, for example. Let the initial data be defined as

$$u(x,0) = u_0(x) = \begin{cases} 1, & \text{if } x < 0.5; \\ 0, & \text{if } x \geq 0.5. \end{cases} \tag{2.44}$$

and let $f = 0$. We arrive at

$$u(x,t) = u(x - at, 0) = u_0(x - at). \tag{2.45}$$

Figure 2.1 illustrates the solution characteristics for this example.

Variable advection. Here we consider the balance equation

$$u_t + c(x)u_x = f(x,t), \tag{2.46}$$

in place of (2.43a). Let $c(x) = x$, $f = 0$, and define the initial data as

$$u(x,0) = u_0(x) = \begin{cases} 1, & \text{if } x < 0.25; \\ 0, & \text{if } x \geq 0.25. \end{cases} \tag{2.47}$$

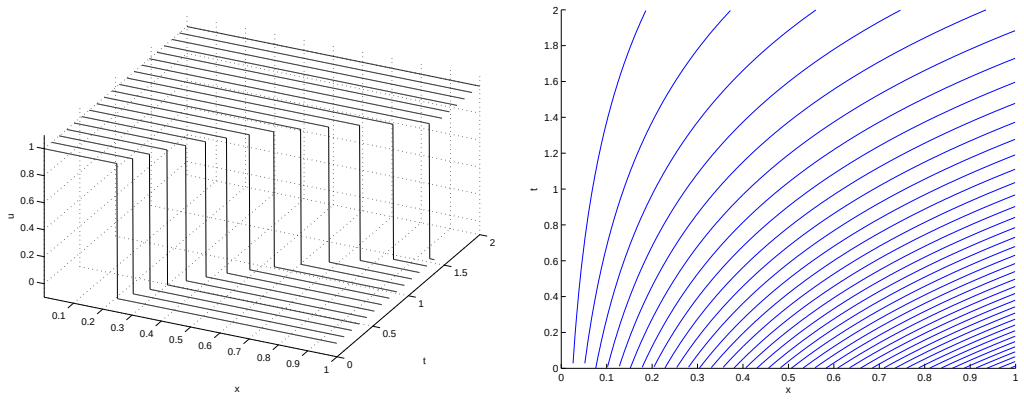
We arrive at

$$u(x,t) = u_0(xe^{-t}), \tag{2.48}$$

since the characteristics are defined by $x(t) = x_0 e^t$. Figure 2.2 illustrates the solution characteristics for this example. In case of variable and constant advection, the solution is constant along the characteristics curves. This behavior is apparent in both of the respective plots.

Shock formation. The flux function $f(u) = \frac{u^2}{2}$ corresponds to the so called Burger's equation, a nonlinear hyperbolic conservation law. We prescribe continuous initial data

$$u(x,0) = u_0(x) = \begin{cases} 0.75, & \text{if } x < -0.6; \\ -x + 0.15, & \text{if } x < -0.6; \\ 0.25, & \text{if } x \geq -0.1. \end{cases} \tag{2.49}$$



(a) Solution profile

(b) Characteristics

Figure 2.2: Solution to a problem with spatially dependent velocity.

The solution is displayed in Figure 2.3. Notice that the continuous solution evolves into a discontinuity as time progresses. The characteristics eventually intersect and follow a new direction defined by the shock speed. To determine the shock speed, s, of the physically admissible solution, the classical approach is to consider the integral form given by

$$\frac{d}{dt}\int_a^b u(x,t)\,dx = f(u(a,t)) - f(u(b,t)). \tag{2.50}$$

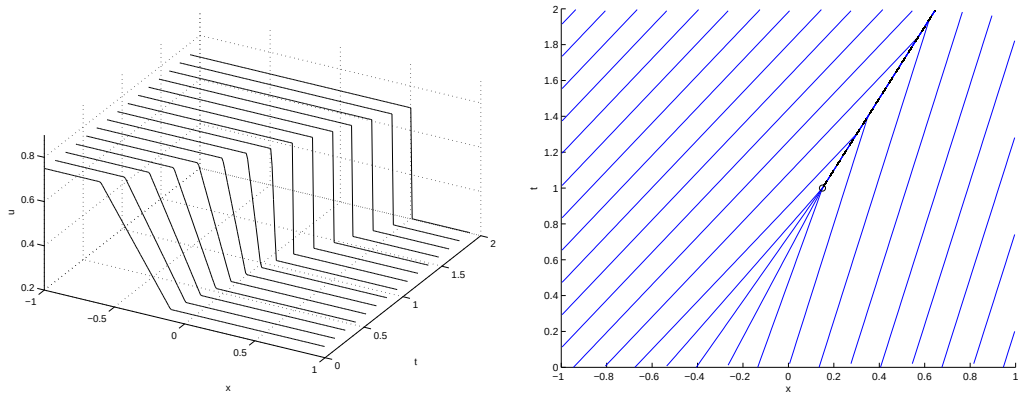
Let u_l and u_r be the speed on the left and right of the shock respectively. In our case, $u_l = .75$ and $u_r = .25$. If $u_l < u_r$, then rarefaction is observed. For a time t after a shock has emerged, we have

$$\frac{d}{dt}\int_a^b u(x,t)\,dx = \frac{d}{dt}\int_a^{st} u_l\,dx + \int_{st}^b u_r\,dx = s(u_l - u_r). \tag{2.51}$$

Thus,

$$s = \frac{f(u_l) - f(u_r)}{u_l - u_r} = \frac{u_l + u_r}{2}. \tag{2.52}$$

Alternatively, we can derive the shock speed with respect to functions in $H(div)$. We leave this analysis to Chapter 6. Obtaining the correct shock speed in a numerical scheme validates the approximation to some extent. However, this is only one indicator that the physically valid solution is obtained.



(a) Solution profile

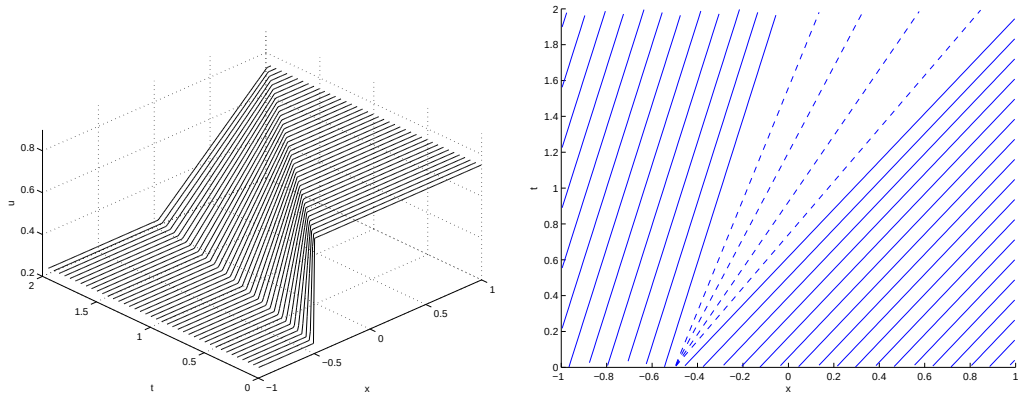
(b) Characteristics

Figure 2.3: Shock formation in the Burger's equation ($\circ$ marks the start of the emerging shock).

Rarefaction. Here we consider a phenomenon called rarefaction. In this case, a discontinuity spreads out as time evolves. Again, consider the flux function $f(u) = \frac{u^2}{2}$ for the Burger's equation. We prescribe discontinuous initial data:

$$u(x,0) = u_0(x) = \begin{cases} 0.25, & \text{if } x < -0.5; \\ 0.75, & \text{if } x \geq -0.5. \end{cases} \tag{2.53}$$

The solution is displayed in Figure 2.4. Notice that the shock **rarefies** in time. The dashed lines indicate characteristics emanating from the same point.



(a) Solution profile

(b) Characteristics

Figure 2.4: Rarefaction formation in the Burger's equation.

2.4 Numerical Overview

Numerical methods for PDEs of hyperbolic type have been a large area of computational research. For certain hyperbolic problems, specific numerical methods have been successfully developed, but generality and computational cost continue to pose problems in the field.

Discontinuous solutions are a significant challenge in scientific computation. Solutions of low continuity often emerge when modeling with hyperbolic conservation laws. Methods with low-order discrete spatial and temporal discretizations (e.g., First-Order Upwind) often smear the solution near the discontinuity, while higher-order methods (e.g., Lax-Wendroff type) dis-

play poor solution quality in the form of dispersion. This dispersion is revealed by oscillations in the solution as well as overshoots or undershoots near a discontinuity. This is a difficulty common in the development of adequate numerical methods for PDEs of hyperbolic type. Since shortcomings of a discretization are often most apparent in a region surrounding a discontinuity in the analytic solution, much effort has been devoted to locating this discontinuity. The so-called shock tracking approach uses shock indicators or estimators to determine the position of the discontinuity and an appropriate weighting or adjusted discretization to overcome the disadvantages found in the standard approach [62]. Complexity issues arise, however, particularly in time-dependent problems and when the spatial dimension is higher than 1. The goal is then to successfully "capture" the shock. We can obtain good convergence properties in regions of smooth solutions and we seek sharp resolution of the discontinuities in other regions, while avoiding smearing and other undesirable solution properties. This is the motivation of much of this dissertation.

In this section, we highlight standard numerical techniques for hyperbolic problems. We point out particular advantages and certain shortcomings of these methods. Classical numerical schemes have focused on the same concerns that we have addressed in Section 2.2.2, and we look further into the computational issues outlined above. Typical concerns in scientific computing, such as parallelizability, iterative solution methods for the linear systems arising in implicit schemes, and adaptivity, are at the forefront of research in computational hyperbolic conservation laws. Although we do not detail all of these issues in our brief overview in this section, we point out the status of various methods in terms of these computational aspects. In Section 2.4.1, we outline finite difference and finite volume methods for hyperbolic PDEs. Section 2.4.2 addresses finite element approaches. The least-squares method that we develop later is based on the ideas in this section. Finally, in Section 2.4.3, we discuss aspects of time-dependent PDEs, including explicit and implicit time-stepping methodologies as well as full space-time based approaches.

2.4.1 Finite Difference/Volume Methods

Finite difference techniques often expose basic difficulties in solving a problem numerically. This is particularly the case for hyperbolic PDEs, where the shortcomings are easily identified by considering discontinuous solutions. In many cases, finite difference methods work quite well for continuous solution profiles. However, problems arise when the solution has reduced smoothness [62].

The finite difference method is elementary, but is still heavily used in Computational Fluid Dynamics (CFD) and other domains where convection processes are considered. Beginning with the scalar conservation law

$$\frac{du}{dt} + \frac{\partial f(u)}{\partial x} = 0, \tag{2.54}$$

consider the linear advection case with $f(u) = au$:

$$\frac{du}{dt} + a\frac{\partial u}{\partial x} = 0, \tag{2.55}$$

for some advection speed $a > 0$. This generalization allows us to assume that the fluid elements are transported in the positive x-direction, an important distinction that some finite difference schemes rely on. Information is received from the negative (**upwind**) x-direction. Let $\{x_i\}_{i=0}^{n}$ be a uniform partition of the interval $[x_a, x_b]$, with $x_0 = x_a$ and $x_n = x_b$. Let $\{t_j\}_{j=0}^{m}$ be a uniform partition of the temporal interval $[0, T]$, with $t_0 = 0$ and $t_m = T$. Use h and k to denote the spatial and temporal mesh size, respectively. We write $u_i^j = u(x_i, t_j)$ to simplify the notation.

If we assume the technique of first deriving a semi-discretization (SD) followed by a standard ODE solver, then we can easily arrive at a number of full discretizations (FD). The most obvious starting point is to use central differencing in space. This is a second-order accurate method and we have

$$u_i'(t) + a\frac{u_{i+1}(t) - u_{i-1}(t)}{2h} = 0. \tag{2.56}$$

Now, using either Backward or Forward Euler's method, both first-order methods in time, we

arrive at

$$\frac{u_i^{j+1} - u_i^j}{k} + a\frac{u_{i+1}^j - u_{i-1}^j}{2h} = 0, \tag{2.57}$$

$$\frac{u_i^{j+1} - u_i^j}{k} + a\frac{u_{i+1}^{j+1} - u_{i-1}^{j+1}}{2h} = 0, \tag{2.58}$$

respectively. A von Neumann type Fourier stability analysis [49] reveals that the explicit discretization is numerically unstable. The scheme in (2.57) is unconditionally unstable, while the implicitly defined discretization (2.58) is unconditionally stable. Implicit finite difference methods are rarely used for time-dependent hyperbolic equations largely due to the linear solve required at each time step [61]. Efficiently solving these discrete systems is an active area of research. Implicit schemes and space-time formulations are gaining popularity in finite element approaches and we motivate the methods developed throughout the dissertation along this direction.

We continue with explicit solution methods since this technique is more representative of the current work being conducted in finite difference and finite volume methods. For the linear advection equation (2.55), we can tailor our spatial discretization to accommodate the upwind anisotropy. The FD becomes

$$u_i^{j+1} = u_i^j - \frac{ak}{h}(u_i^j - u_{i-1}^j). \tag{2.59}$$

Now, the stability of the numerical method is dependent on the Courant-Friedrichs-Levy condition (CFL), which requires

$$k < \frac{h}{a}. \tag{2.60}$$

For time-marching methods, the CFL condition limits the size of the time step. In practice, this limitation often forces m to be much larger than n and increases the computational cost severely.

As we can see from Figure 2.5, using the First-Order Upwind (FOU) scheme in (2.59) results in an approximation that is smeared around the discontinuity. Higher-order methods are meant to sharpen the resolution of the method both around discontinuities and in regions of

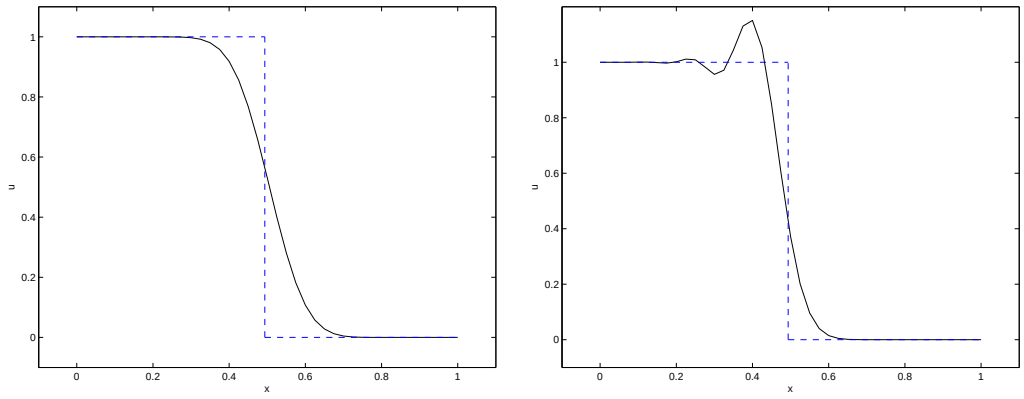
smooth solutions. A classic choice is the Lax-Wendroff method that results from a Taylor series expansion:

$$u_i^{j+1} = u_i^j - \frac{ak}{2h}(u_{i+1}^j - u_{i-1}^j) + \frac{k^2}{2h^2}a^2(u_{i-1}^j - 2u_i^j + u_{i+1}^j). \tag{2.61}$$

Notice that the last term is simply a finite difference stencil for εu_{xx}, where $\varepsilon = \frac{k^2a^2}{2}$. The Lax-Wendroff method can then be thought of as a method with artificial diffusion added. The diffusion forces a smoother solution to the PDE at the location of the discontinuity, resulting in higher accuracy of the numerical approximation. Figure 2.5 shows a significant improvement in resolving the discontinuity, but spurious oscillations are present and large overshoots exist.

The FOU method for the linear case (2.59) can also be extended to nonlinear equations. Methods such as FOU are called conservative methods, since the conservation law holds discretely, and are the basis for so-called finite volume methods. On the discrete level, we want to ensure that changes in the discrete sum are due only to fluxes at the endpoints, so we introduce **cells** or volumes to help force this property. Figure 2.6 illustrates this case. We consider the conservation law in the space-time cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [t_j, t_{j+1}]$ to arrive at

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x, t_{n+1})\,dx = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x, t_n)\,dx - \int_{t_n}^{t_{n+1}} (f(u(x_{i+\frac{1}{2}})) - f(u(x_{i-\frac{1}{2}})))\,dt. \tag{2.62}$$



(a) First-Order Upwind finite difference approximation of a discontinuity.

(b) Lax-Wendroff second-order finite difference approximation of a discontinuity.

Figure 2.5: Explicit time-stepping schemes.

Let $\bar{u}_i^j$ be an approximation to u_i^j averaged over the cell defined above and let $\bar{f}^j_{i\pm\frac{1}{2}}$ denote an approximation to the time average of the flux on the spatial boundary of the cell. The function $\bar{f}$ is called the **numerical** flux function. We can then write (2.62) as

$$\bar{u}_i^{j+1} = \bar{u}_i^j - \frac{k}{h}(\bar{f}^j_{i+\frac{1}{2}} - \bar{f}^j_{i-\frac{1}{2}}) \quad [cf.(2.59)]. \tag{2.63}$$

Summing over all grid cells at a time j indeed verifies that we obtain discrete conservation. By the Lax-Wendroff Theorem [62] we can conclude that if this method converges to a solution $\bar{u}$, then $\bar{u}$ is in fact a weak solution to the conservation law.

As mentioned above, satisfying the entropy condition is crucial for a numerical method and it ensures that the method converges to the physically meaningful solution. We introduce another property, called **monotonicity**, to help us develop methods that result in correct weak solutions. A method is called monotone [61] if the following holds for any two numerical solutions u and v:

$$v_i^j \geq u_i^j \ \forall j \Rightarrow v_i^{j+1} \geq u_i^{j+1}. \tag{2.64}$$

For reference, all monotone methods are total variation diminishing (TVD), and an overview of such terminology and application to finite difference methods can be found in [61, 62, 67]. The notion of a monotone method is useful because it satisfies the entropy condition and results in the desired weak solution for nonlinear hyperbolic conservation laws. This is a valuable result, but it does come with a cost. Godunov's Theorem [40] states that explicit schemes that are monotone

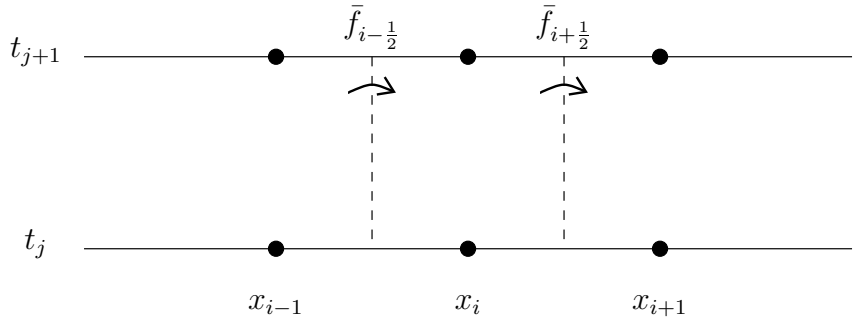


Figure 2.6: An example cell located around the point x_i^j. The approximations to the average fluxes at the boundaries of this cell are given by $\bar{f}^j_{i\pm\frac{1}{2}}$.

are at most first-order accurate. Loss of monotonicity, typically in the form of oscillations, is inherent in resolving discontinuities and steep profiles. This is a result that is widely used to motivate current work in finite difference and finite volume methods. If higher-order monotone methods cannot be devised naturally, then another approach must be taken.

The task of developing high-order methods to give good convergence properties in smooth regions and sharp resolution around discontinuities is commonly achieved by introducing numerical dissipation near the suspected discontinuity and constraining the amount of overshoot. Typically referred to as **limiting** techniques, the idea is to balance the benefits of low-order monotone schemes with the non-diffusive high-order methods. This results in an approximation that is nearly monotone and produces limited oscillations. We consider the Lax-Wendroff method (2.61) for the linear advection equation and rewrite it in a form that corresponds to the conservative form (2.63). Following [61, 62] et al., let $\nu = \frac{ak}{h}$ and rewrite (2.61) as

$$u_i^{j+1} = u_i^j - \nu(u_i^j - u_{i-1}^j) + (\frac{\nu^2 - \nu}{2})(u_{i-1}^j - 2u_i^j + u_{i+1}^j). \tag{2.65}$$

Define

$$\bar{f}_{L,i+\frac{1}{2}} = au_i^j, \tag{2.66}$$

$$\bar{f}_{H,i+\frac{1}{2}} = au_i^j + a(\frac{1-\nu}{2})(u_{i+1}^j - u_i^j). \tag{2.67}$$

We arrive at

$$u_i^{j+1} = u_i^j - \frac{k}{h}(\bar{f}_{i+\frac{1}{2}} - \bar{f}_{i-\frac{1}{2}}), \tag{2.68}$$

where

$$\bar{f}_i = \bar{f}_{L,i} + \phi[\bar{f}_{H,i} - \bar{f}_{L,i}] \tag{2.69}$$

and ϕ is a flux limiting function, which is clearly equal to one for the Lax-Wendroff method. Varying ϕ globally cannot result in improve accuracy according to Godunov's Theorem, so ϕ typically depends on i, the spatial index. The goal is to vary ϕ_i according to the location of the discontinuity. As a result, the method is no longer linear and Godunov's Theorem does

not apply. However, complexity becomes an issue as methods and algorithms become more complicated.

Developing an efficient, high-order, and essentially non-oscillatory method is certainly a goal in current research efforts [5]. As we have seen, higher-order accurate methods and non-oscillatory solutions cannot be achieved by standard techniques. Thus, much of the effort in finite difference and finite volume methods has been focused on limiting the poor solution quality with a cost of higher computation time. A dominating strategy has not yet emerged in this area. We have also only discussed successful techniques for scalar problems of one spatial dimension. The complications are multiplied when 2-D, 3-D, and higher-dimensional problems are considered. Systems of equations also pose significant challenges to these ideas. Computing solutions to many variables results in multiple directions and regions of discontinuities for each of the variables, yet only one computational mesh. This also indicates a challenge for spatial and temporal adaptivity. The methods we develop throughout this dissertation address many of these computational difficulties.

2.4.2 Finite Elements

Finite element discretizations attempt to add generality to the numerical approximation of hyperbolic PDEs. Unlike finite difference methods, which often require special consideration for source terms, systems, etc., many finite element methods can easily be extended to more complicated and general balance equations. A first attempt to discretize conservation laws with finite element methods was based on Galerkin variational principles, which provides an adequate approach for elliptic and parabolic problems and is thus a natural consideration for PDEs of hyperbolic type. To explain this methodology, consider the linear scalar conservation law in differential form:

$$u_t + au_x = f \quad \text{in } \Omega, \tag{2.70}$$

where Ω is a bounded convex space-time domain in $\mathbb{R}^2$ and $f \in L^2(\Omega)$. Write

$$\mathbf{b} = \begin{pmatrix} a \\ 1 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \in \mathbb{R}^2 \quad \text{and} \quad \nabla = \begin{pmatrix} \partial_x \\ \partial_t \end{pmatrix}. \tag{2.71}$$

If we consider a more general balance equation with a reaction term, we can write (2.70), or any steady state hyperbolic PDE of the same general form, as

$$\mathbf{b} \cdot \nabla u + cu = f. \tag{2.72}$$

Without a loss of generality, let $c = 1$ for simplicity. Then the weak problem corresponding to the classical Galerkin formulation is stated as the following.

Problem 2.4 (Galerkin Weak Problem). *Find $u^h \in V^h$ such that*

$$\langle \mathbf{b} \cdot \nabla u^h + u^h, v^h \rangle_{0,\Omega} = \langle f, v^h \rangle_{0,\Omega}, \quad \forall v^h \in V^h, \tag{2.73}$$

where V^h is some finite element space.

Here, $\langle \cdot, \cdot \rangle_{0,\Omega}$ and $\langle \cdot, \cdot \rangle_{\Gamma}$ are defined as

$$\langle u, v \rangle_{0,\Omega} = \iint_{\Omega} uv \, d\mathbf{x}, \tag{2.74}$$

$$\langle u, v \rangle_{\Gamma} = \oint_{\Gamma} |\mathbf{b} \cdot \mathbf{n}| uv \, ds, \tag{2.75}$$

where $\Gamma = \partial\Omega$. From Green's formula, we have

$$\langle \mathbf{b} \cdot \nabla u, v \rangle_{0,\Omega} = \langle u, v \rangle_{\Gamma} - \langle u, \mathbf{b} \cdot \nabla v \rangle_{0,\Omega},$$

which implies that

$$\langle \mathbf{b} \cdot \nabla u, u \rangle_{0,\Omega} = \frac{1}{2} \|u\|^2_{0,\Gamma}. \tag{2.76}$$

Substituting $v^h = u^h$ in (2.73) and using (2.76), we arrive at the stability bound

$$\|u^h\|_{0,\Omega} + \|u^h\|_{0,\Gamma} \leq C \|f\|_{0,\Omega}. \tag{2.77}$$

This suggests that the method may have no control of $\mathbf{b} \cdot \nabla u$, which reveals a significant shortcoming of this formulation. Indeed, this is the case (cf. [11]), as the Galerkin method is shown

to produce oscillations particularly in the case of discontinuous solutions. This has been known for decades and has given rise to the so-called Streamline Upwind Petrov Galerkin methods (SUPG), which attempt to take into account the anisotropic nature of the PDE. Moreover, the SUPG method also improves suboptimal error estimates for the classical Galerkin method. From standard finite element interpolation theory (see [13, 18], e.g.) we have: Given a function $u \in H^{k+1}(\Omega)$, there is an interpolant $\Pi^h u \in V^h$ such that

$$\|u - \Pi^h u\|_{0,\Omega} \leq Ch^{k+1}\|u\|_{k+1,\Omega}, \tag{2.78}$$

whereas the error estimates for the Galerkin formulation are given by

$$\|e^h\|_{0,\Omega} + \|e^h\|_{0,\Gamma} \leq Ch^k\|u\|_{k+1,\Omega}. \tag{2.79}$$

The Galerkin method is then said to be suboptimal with a gap of 1.

SUPG methods have superb stability properties, although smearing remains a concern in the classic implementation of this method. The basic idea behind SUPG is its stream-like property of adding $h\mathbf{b}\cdot\nabla v^h$ to the space of test functions. The SUPG weak formulation is the following.

Problem 2.5 (SUPG Weak Problem). *Find $u^h \in V^h$ such that*

$$\langle \mathbf{b}\cdot\nabla u^h + u^h, v^h + h\mathbf{b}\cdot\nabla v^h\rangle_{0,\Omega} = \langle f, v^h + h\mathbf{b}\cdot\nabla v^h\rangle_{0,\Omega}, \quad \forall v^h \in V^h, \tag{2.80}$$

where V^h is some finite element space.

Using an approach analogous to what we used for Galerkin, we obtain the SUPG stability estimate

$$\|u^h\|_{0,\Omega} + \sqrt{h}\|\mathbf{b}\cdot\nabla u^h\|_{0,\Omega} + \|u^h\|_{0,\Gamma} \leq C\|f\|_{0,\Omega}, \tag{2.81}$$

this shows control over $\mathbf{b}\cdot\nabla u^h$, thus reducing the oscillations. The error estimates are also improved compared with the Galerkin method, now with a gap of $\frac{1}{2}$ (see [55]):

$$\|e^h\|_{0,\Omega} + \|e\|_{0,\Gamma} \leq Ch^{k+\frac{1}{2}}\|u\|_{k+1,\Omega}. \tag{2.82}$$

Although attempts have been made in recent years to further improve SUPG by introducing cross-stream diffusion to reduce oscillations and feedback control parameters to enhance the resolution around discontinuities [37, 34], the generality of the method is limited and computational cost has not yet been fully addressed. Other improvements to the SUPG method, such as shock capturing techniques, result in computationally intensive algorithms because the resulting linear systems are often nonsymmetric and complicated by the presence of grid dependent terms like $h\mathbf{b} \cdot \nabla v^h$.

One of the most studied finite element approach for hyperbolic equations in recent years is the Discontinuous Galerkin method (DG) or stabilized Discontinuous Galerkin method, which is similar to the SUPG method described above. This method, introduced in the 1970s [57], was popularized by the work of Johnson et al. [54] and by Cockburn, Shu et al. [27] with their pioneering work in higher-order methods and extensions to nonlinear equations. There are several advantages to the DG method and its variants, including the broad range of acceptable domains and tessellations. Computational complexity, adaptivity, parallelization, and other high performance computing issues have only recently been addressed for this methodology. Preliminary findings show that efficient adaptive techniques and parallelization of the DG method are attainable for certain problems [8]. Also, the DG method is able to handle non-matching and non-uniform grids, which fit well into an adaptive framework. We address these concerns more when we considers least-squares finite element methods in the following chapters.

Consider (2.72) with boundary conditions:

$$\mathbf{b} \cdot \nabla u + cu = f, \quad \text{in } \Omega, \tag{2.83a}$$

$$u = g, \quad \text{on } \Gamma_I. \tag{2.83b}$$

Here, for any domain K, we define $K_I = \{x \in \partial K : \mathbf{n} \cdot \mathbf{b} < 0\}$ as its inflow boundary, where $\mathbf{n}$ is the outward unit normal on ∂K. Thus, Γ_I is the inflow portion of the boundary and $\partial\tau_I$ is the inflow portion of element τ. Based on the notation introduced by Johnson et al. [54, 55], for a given tessellation $\mathcal{T}^h$ of Ω, let $u^{h,\pm}(\mathbf{x}) = \lim_{\varepsilon \to 0^\pm} u^h(\mathbf{x} + \varepsilon\mathbf{b})$, with $u^{h,-}(\mathbf{x}) = g(\mathbf{x})$ if $\mathbf{x} \in \Gamma_I$.

Also, set

$$\Gamma_h = \left(\bigcup_{\tau \in \mathcal{T}^h} \partial\tau \right) \setminus \partial\Omega, \tag{2.84}$$

$$\langle u^h, v^h \rangle_{\Gamma_*} = \int_{\Gamma_*} |\mathbf{n} \cdot \mathbf{b}| u^h v^h \, ds, \tag{2.85}$$

$$V^h = \text{space of piecewise polynomials}, \tag{2.86}$$

$$= \{ v^h \in L^2(\Omega) : v|_\tau \in \mathcal{P}^k_\tau, \, \tau \in \mathcal{T}^h \}. \tag{2.87}$$

In this case, we denote $\mathcal{P}^k_\tau$ as the set of polynomials of degree k. We now write the (stabilized) DG method as the following.

Problem 2.6 (DG Weak Problem). *Find $u^h \in V^h$ such that*

$$B_{\mathrm{DG}}(u^h, v^h) = \langle f, v^h + \delta \mathbf{b} \cdot \nabla v^h \rangle_{0,\Omega} + \langle g, v^h \rangle_{\Gamma_I} \quad \forall v^h \in V^h, \tag{2.88}$$

where

$$B_{\mathrm{DG}}(u^h, v^h) = \sum_{\tau \in \mathcal{T}^h} \langle \mathbf{b} \cdot \nabla u^h + u^h, v^h + \delta \mathbf{b} \cdot \nabla v^h \rangle_{0,\tau} + \langle u^{h,+} - u^{h,-}, v^{h,+} \rangle_{\Gamma_h} + \langle u^h, v^h \rangle_{\Gamma_I} \tag{2.89}$$

and δ is a stabilization parameter that depends on h.

This method allows freedom in the numerical approximation to exhibit jump discontinuities, which in turn allows for improved stability. Moreover, for $\delta = h$, the theory in [54] shows that

$$\|u^h\|^2_{0,\Omega} + h \sum_{\tau \in \mathcal{T}^h} \|\mathbf{b} \cdot \nabla u^h\|^2_{0,\tau} + \sum_{\tau \in \mathcal{T}^h} \|u^{h,+} - u^{h,-}\|^2_{0,\tau} \leq C \left(\|f\|_{0,\Omega} + \|g\|_{0,\Gamma_I} \right). \tag{2.90}$$

Thus $\mathbf{b} \cdot \nabla u^h$ is controlled as it is with the SUPG method, but so are the inter-element jumps expressed in the $\|u^{h,+} - u^{h,-}\|_{0,\tau}$ term. Johnson et al. also prove error estimates of the form

$$\|e^h\|_{0,\Omega} \leq C h^{k+\frac{1}{2}} \|u\|_{k+1,\Omega}, \tag{2.91}$$

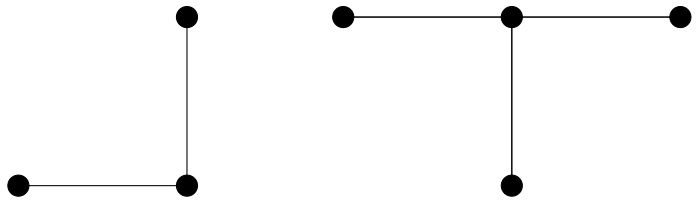
which, compared to (2.79) and (2.82), shows improvement over the Galerkin method and a gap that equals that of the SUPG method.

The approach described here is a **stabilized** DG method depending on the value for δ. Choosing $\delta = 0$, results in a pure (discontinuous) Galerkin variational form, while choosing $\delta = h$ corresponds to SUPG, but DG can be more robust with a better choice for δ. As δ increases, DG becomes more and more stable, although smearing of discontinuous solutions becomes more pronounced. This motivates our introduction of a discontinuous least-squares stabilized finite element method in Chapter 4, which closely follows the stabilized DG method, but in a bona fide least-squares framework.

Stabilized Galerkin methods, or Galerkin Least-Squares methods, are gaining popularity, although practical implementation and other computational issues are not yet fully resolved. In the next chapter, we address pure least-squares formulations for the hyperbolic PDE, which is a highly stable method based on a minimization principle. This methodology has attractive computational benefits and is well developed for other classes of problems.

2.4.3 Time Considerations

Time-stepping techniques are a major focus for all of the methods presented in Sections 2.4.1 and 2.4.2. Leveque [62] states that explicit time-stepping schemes are most widely used in finite difference and finite volume formulations. Others have described many DG finite element implementations based on Implicit Runge-Kutta (IRK) type methods. Questions concerning computational cost, accuracy, implementation, and stability arise in each of these formulations.



(a) First Order Upwind (explicit) (b) Backward Euler (implicit)

Figure 2.7: Finite difference stencils.

Explicit time-stepping schemes, such as First Order Upwind (FOU), are straightforward to implement, since there are no algebraic systems to solve at each time step (the stencil is illustrated in Figure 2.7). However, stability of the method in time is heavily dependent on the size of the time step [49]. The CFL condition dictates how large the time step can be while still ensuring stability, and a large number of small time steps can drastically increase the total computation cost. Moreover, many of the explicit methods are low-order accurate in time, also forcing more time steps in order to achieve adequate resolution. High-order Runge-Kutta methods, however, can overcome this limitation.

Implicit time-stepping schemes, on the other hand, often do not place restrictions on time-step sizes where stability is concerned. Figure 2.7 illustrates the Backward Euler method, a typical (unconditionally stable) finite difference stencil. The ability to take larger time steps and the combination with high-order discretizations make implicit schemes competitive with explicit approaches. However, the total computational cost for implicit schemes is heavily dependent on the solution method used to solve the resulting linear system at each time step.

Both explicit and implicit time-stepping methods require careful attention in a high performance computing environment. Adaptive grid techniques can be difficult to develop for both types of methods. Spatial adaptivity is seemingly achievable, provided a sharp error estimator exists. However, adaptivity in space and time remains a challenge. Due to the nature of the time direction, time-stepping methods are inherently difficult to implement with adaptive meshing. This is also true in parallelizing these methods. Although the solution at each time step is quickly resolved, the number of time steps remains an issue. This motivates the use of space-time discretizations. In this approach, time is often treated simply as another dimension. An obvious disadvantage of this approach is that the linear system to be solved is at least as large as the number of time partitions multiplied by the number of spatial nodes. However, in some algorithms, adaptivity in space-time is easily handled, parallelization is more apparent, and coarsening in both space and time is achievable, resulting in efficient solves of the whole domain.

Chapter 3

Least-Squares Finite Elements

We noted in Sections 2.4.1 and 2.4.2 that finite difference, finite volume, and Galerkin-based finite element methods often lead to excessive oscillations in the numerical approximation, without special treatment. The oscillations are especially prevalent near discontinuities and in several finite element higher-order schemes. The classical Galerkin method is numerically unstable and additional stabilization terms are usually required. Streamline Upwind Petrov Galerkin (SUPG) methods (2.80), which use a least-squares stabilization term, are a common approach. Stabilized Discontinuous Galerkin (DG) methods follow this strategy, but extend in generality with the use of discontinuous finite element spaces. In this chapter, we introduce least-squares methods. The full least-squares method is still a Petrov-Galerkin approach, but differs from other formulations since it is based on a minimization principle that offers many theoretical and computational advantages (see Section 3.3).

Recall the model problem (2.83):

$$\mathbf{b} \cdot \nabla u + cu = f, \quad \text{in } \Omega, \tag{3.1a}$$

$$u = g, \quad \text{on } \Gamma_I, \tag{3.1b}$$

with $\mathbf{b}(\mathbf{x})$ a flow field on $\Omega \subset \mathbb{R}^2$ and

$$\Gamma_I := \{\mathbf{x} \in \partial\Omega \colon \mathbf{n}(\mathbf{x}) \cdot \mathbf{b}(\mathbf{x}) < 0\} \tag{3.2}$$

the inflow part of the boundary of domain Ω, where $\mathbf{n}(\mathbf{x})$ is the outward unit normal on $\partial\Omega$.

We seek to minimize the square of the residual in the L^2-norm: find $u \in V$ such that

$$u = \underset{v \in V}{\operatorname{argmin}} \mathcal{F}(u; f, g), \tag{3.3}$$

where

$$\mathcal{F}(u; f, g) := \|\mathbf{b} \cdot \nabla u + cu - f\|_{0,\Omega}^2 + \|u - g\|_{\Gamma_I}^2. \tag{3.4}$$

Here, $\|\cdot\|_{\Gamma_I}$ is a weighted L^2-norm over the inflow boundary (cf. [50, 11]). We describe more notation in the next section and discuss the least-squares weak problem in further detail in Section 3.2.

3.1 Notation

First, we establish notation that is used throughout the remainder of this dissertation. Although we have briefly introduced the model problem in previous sections, important details have so far been ignored. We limit the majority of our discussions in this dissertation to two dimensions ($d = 2$), referring to possible extensions to higher dimensions where appropriate. We use the standard definition and notation for a Sobolev space $H^s(\Omega)$, for $s \geq 0$, and its associated inner product $\langle \cdot, \cdot \rangle_{s,\Omega}$ and norm $\|\cdot\|_{s,\Omega}$. For the case of $s = 0$, we use $L^2(\Omega)$. Let $H_0^s(\Omega)$ denote the closure in the $\|\cdot\|_s$-norm of the linear space of infinitely differentiable functions with compact support in Ω.

We define the **curl** operator in 2-D by first considering the **curl** in 3-D:

$$\mathbf{curl} = \nabla\times = \begin{bmatrix} 0 & -\partial_z & \partial_y \\ \partial_z & 0 & -\partial_x \\ -\partial_y & \partial_x & 0 \end{bmatrix}. \tag{3.5}$$

From the dotted lines in (3.5), we define for distribution p and vector distribution $\mathbf{w}$ the following 2-D operators:

$$\mathbf{curl}\, p = \nabla^\perp p \quad = \begin{pmatrix} \partial_y p \\ \\ -\partial_x p \end{pmatrix}, \tag{3.6}$$

$$\operatorname{curl} \mathbf{w} = \nabla \times \mathbf{w} = \partial_x w_2 - \partial_y w_1. \tag{3.7}$$

The curl is the formal adjoint of the **curl** operator. We define div $\equiv \nabla\cdot$ in the standard way and let

$$H^1(\Omega) = \{p \in L^2(\Omega) : \nabla p \in L^2(\Omega)^2\}, \tag{3.8}$$

$$H(\mathbf{curl}, \Omega) = \{p \in L^2(\Omega) : \nabla^\perp p \in L^2(\Omega)^2\}, \tag{3.9}$$

$$H(\mathrm{div}, \Omega) = \{\mathbf{w} \in L^2(\Omega)^2 : \nabla \cdot \mathbf{w} \in L^2(\Omega)\}, \tag{3.10}$$

which are Hilbert spaces with the respective norms

$$\|p\|_{H(\Omega,\mathbf{grad})} := \left(\|p\|^2_{0,\Omega} + \|\nabla p\|^2_{0,\Omega}\right)^{\frac{1}{2}}, \tag{3.11}$$

$$\|p\|_{H(\Omega,\mathbf{curl})} := \left(\|p\|^2_{0,\Omega} + \|\nabla^\perp p\|^2_{0,\Omega}\right)^{\frac{1}{2}}, \tag{3.12}$$

$$\|\mathbf{w}\|_{H(\Omega,\mathrm{div})} := \left(\|\mathbf{w}\|^2_{0,\Omega} + \|\nabla \cdot \mathbf{w}\|^2_{0,\Omega}\right)^{\frac{1}{2}}. \tag{3.13}$$

Inner product notation and nonstandard spaces are defined as needed.

3.2 Least-Squares Finite Element Methodology

As summarized in Section 2.4.2, finite element methods for (3.1) have been considered before, in Galerkin, streamline upwind Petrov-Galerkin (SUPG), and residual distribution frameworks [1, 35, 55], for example. Least-squares terms have been added to Galerkin methods for stabilization (see e.g. [4, 48, 47]), and the SUPG method can be written as a linear combination of a Galerkin method and a least-squares term [35]. Full least-squares methods were considered numerically in [24, 52, 50]. A comparison of Galerkin, SUPG, and LSFEM for convection problems can be found in [11] and improved error estimates and convergence analysis for smooth solutions in the least-squares framework were addressed in [9]. Full least-squares formulations for (3.1) with discontinuous solutions are the primary focus of this dissertation. In this section, we motivate the use of the least-squares methodology and highlight some relevant results for the least-squares formulation of hyperbolic PDEs for the case of continuous solutions.

While least-squares finite element methods (LSFEMs) have been investigated extensively for equations of elliptic type [12, 21, 23, 50], their use for hyperbolic PDEs has only been

initiated recently [24, 11, 9]. LSFEMs are inherently attractive variational formulations for which well-posedness of the resulting discrete problems can often be proved rigorously. The main advantage of least-squares methods is that the formulation is based on a minimization principle. A clear optimization strategy is often not apparent in many numerical approaches and this is one of the main advantages that the least-squares functional provides. The minimization principle incorporates the notion of optimality in the method, aiding in the development of appropriate discretizations for problems with a variety of anomalies, including the discontinuities and nonlinearities [66] that arise in hyperbolic equations.

Given a (first-order) system

$$\mathcal{L}u = f, \quad \text{in } \Omega, \tag{3.14a}$$

$$\mathcal{B}u = g, \quad \text{on } \Gamma, \tag{3.14b}$$

we seek to minimize the least-squares functional

$$\mathcal{G}(u; f, g) := \|\mathcal{L}u - f\|_{0,\Omega}^2 + \|\mathcal{B}u - g\|_{\Gamma}^2, \tag{3.15}$$

where $\|\cdot\|_{\Gamma}$ is typically the $H^{\frac{1}{2}}$-norm or a weighted L^2-norm. This yields a weak problem: find $u \in V$ such that

$$\mathcal{F}(u, v) = \ell(v) \quad \forall v \in V, \tag{3.16}$$

where

$$\mathcal{F}(u, v) = \langle \mathcal{L}u, \mathcal{L}v \rangle_{0,\Omega} + \langle \mathcal{B}u, \mathcal{B}v \rangle_{\Gamma}, \tag{3.17}$$

$$\ell(v) = \langle f, \mathcal{L}v \rangle_{0,\Omega} + \langle g, \mathcal{B}v \rangle_{\Gamma}. \tag{3.18}$$

Suppose that the (symmetric) bilinear form is V-elliptic, that is, coercive and continuous with respect to the norm of some Hilbert space, V. Then well-posedness can easily be established and also holds for any finite-dimensional subspace, $V^h \subset V$.

Berndt, Manteuffel, and McCormick [7] suggest additional benefits to this approach. Since the minimum of the functional is known (zero), a sharp **a posteriori** error estimator is

available as a natural measure of the quality of the solution. Let u denote the exact solution to (3.14) and let u^h be a numerical approximation. Then the functional residual equation says

$$\mathcal{G}(u^h; f, g) = \mathcal{G}(u - u^h; 0, 0) = \mathcal{G}(e^h; 0, 0) \tag{3.19}$$

$$= \mathcal{F}(u - u^h, u - u^h). \tag{3.20}$$

Since $\mathcal{F}(\cdot, \cdot)$ is V-elliptic, (3.20) provides an **a posteriori** global error estimator. The functional residual is equivalent to the V-norm of the error, up to a constant:

$$c_0 \|e^h\|_V^2 \leq \mathcal{F}(e^h, e^h) \leq c_1 \|e^h\|_V^2. \tag{3.21}$$

We also have a local error estimator given by the element contributions to $\mathcal{G}(u^h; f, g)$ and sharpness of the least-squares error estimator is established and addressed in detail in [7]. In general, practical **a posteriori** error estimators are not readily available in other variational formulations, although they can often be computed with additional work for many applications, including hyperbolic PDEs [77, 76, 78]. Even with the availability of effective error indicators, however, the implementation and computational difficulties of adaptive mesh-refinement techniques remains a large area of active research [3, 71].

The least-squares approach also also provides a natural setting for the use of nonconforming finite elements. In particular, the least-squares method greatly simplifies Strang's Second Lemma (cf. [13, 6]), which aides in proving well-posedness of the nonconforming method. We develop and analyze a nonconforming least-squares finite element method in the following chapter that utilizes this advantage of the least-squares formulation.

Often, equations are transformed into a first-order system where a least-squares functional is formulated to decouple the system variables. The process is attractive since it avoids the LBB condition in finite elements, as well as the need to use staggered grids and other restrictions [66]. Minimizing the L^2-norm is the most common and generally most effective form of the least-squares method, although successful algorithm have been developed for H^{-1}-norm [14] and for least-squares in an adjoint setting [22]. As mentioned above, least-squares has been particularly successfully for elliptic-type problems [12, 21, 23, 50]. The success is largely due to

H^1 equivalence (H^1 ellipticity) of the first-order systems of equations. With this level of equivalence, both optimal finite element and multigrid convergence is theoretically proved and indeed verified computationally (see [23] and the references therein). The least-squares methodology also provides a natural setting for multigrid iterative methods. With a least-squares discretization, multigrid solution methods are greatly simplified. An appropriate relaxation strategy and adequate interpolation schemes are all that need to be determined, with the coarse-grid and restriction operators provided naturally by the minimization process. We use this motivation throughout the dissertation and comment on various aspects of geometric and algebraic multigrid methods for the linear systems resulting from the least-squares method for hyperbolic PDEs. Computational implementation and complexity have not previously been addressed for numerical methods for hyperbolic conservation laws in this setting, and we intend to integrate these ideas throughout the thesis.

3.3 Least-Squares Finite Elements Methods for Hyperbolic PDEs with Smooth Solutions

Although the solutions to problems we seek to discretize do not possess nearly enough smoothness to make the least-squares bilinear form H^1 equivalent, we nonetheless pose the hyperbolic problem in a least-squares setting to appeal to its other advantages. A least-squares method for smooth solutions of the hyperbolic PDEs introduced by Bochev and Choi [11, 9] shows encouraging results for discontinuous solutions. We consider this approach further below, but begin instead by applying the standard least-squares minimization technique (cf. [11, 50]) to scalar linear partial differential equations of hyperbolic type that are of the form (3.1).

Without loss of generality, we let $c = 1$ for the remainder of this chapter. We restate (3.1) as the minimization of the least-squares functional

$$\mathcal{G}(u; f, g) := \|\mathbf{b} \cdot \nabla u + u - f\|^2_{0,\Omega} + \|u - g\|^2_{\Gamma_I}, \tag{3.22}$$

where

$$\|g\|^2_{\Gamma_I} = \int_{\Gamma_I} |\mathbf{n}\cdot\mathbf{b}| g^2 \, ds. \tag{3.23}$$

Denote by $\langle\cdot,\cdot\rangle_{\Gamma_I}$ the associated inner product. The space of admissible boundary data is formally considered in the next chapter, where theoretical justification of the least-squares formulation is presented. The weak problem associated with this minimization principle is the following.

Problem 3.1. *Find $u \in V$ such that*

$$\mathcal{F}(u,v) = F(v), \quad \forall\, v \in V, \tag{3.24}$$

where

$$\mathcal{F}(u,v) := \langle \mathbf{b}\cdot\nabla u + u, \mathbf{b}\cdot\nabla v + v\rangle_{0,\Omega} + \langle u, v\rangle_{\Gamma_I}, \tag{3.25}$$

$$F(v) := \langle f, \mathbf{b}\cdot\nabla v + v\rangle_{0,\Omega} + \langle g, v\rangle_{\Gamma_I}, \tag{3.26}$$

and V is any subspace of $H^{1+\varepsilon}(\Omega)$.

Compare this problem with Problems 2.4, 2.5, and 2.6, the forms for Galerkin, SUPG, and Discontinuous Galerkin methods, respectively. We consider here conforming finite element subspaces for this problem. In particular, elements of varying polynomial degree are discussed. Let V^h be a space of continuous piecewise polynomials,

$$V^h = \mathcal{M}^h_k \cap \mathcal{C}^0(\Omega), \tag{3.27}$$

$$\mathcal{M}^h_k = \{p\colon p \in \mathcal{P}_k(\tau),\, \forall \tau \in \mathcal{T}^h\}, \tag{3.28}$$

where $\mathcal{P}_k(\tau)$ is the space polynomials of degree k over the domain τ.

3.3.1 Quotation of Results

In this section, we summarize several results that have previously been established. The conclusions derived are for smooth solutions and based on Problem 3.1. Convergence properties

of the least-squares approximation for smooth solutions are important to establish in order to accurately compare with the performance for discontinuous solutions.

Existence and uniqueness of Problem 3.1 can easily be obtained based on the coercivity and continuity of the problem together with the assumption of smoothness of the solution. Define the least-squares energy norm to be

$$\|u\|_V := \|u\|_{0,\Omega}^2 + \|\mathbf{b}\cdot\nabla u\|_{0,\Omega}^2 + \|u\|_{\Gamma_I}^2. \tag{3.29}$$

We state coercivity and continuity with respect to this norm in the following way.

Lemma 3.2 (Continuity and Coercivity). *Let $u, v \in H^1(\Omega)$, $f \in L^2(\Omega)$, and $g \in L^2(\Gamma_I)$. Assume $1 - \frac{1}{2}\nabla\cdot\mathbf{b} \geq \alpha > 0$. Then there exist constants c_0 and c_1 such that*

$$\mathcal{F}(u,v) \leq c_0\|u\|_V\|v\|_V, \tag{3.30}$$

$$\mathcal{F}(v,v) \geq c_1\|v\|_V^2. \tag{3.31}$$

Proof. Using the Cauchy-Schwarz and triangle inequalities, we have

$$\begin{aligned}\mathcal{F}(u,v) &= \langle \mathbf{b}\cdot\nabla u + u, \mathbf{b}\cdot\nabla v + v\rangle_{0,\Omega} + \langle u, v\rangle_{\Gamma_I} \\ &\leq (\|u\|_{0,\Omega}^2 + \|\mathbf{b}\cdot\nabla u\|_{0,\Omega}^2)(\|v\|_{0,\Omega}^2 + \|\mathbf{b}\cdot\nabla v\|_{0,\Omega}^2) + \|u\_\Gamma_I)\|v\|_{\Gamma_I} \\ &\leq c_0\|u\|_V\|v\|_V.\end{aligned} \tag{3.32}$$

By Green's Formula for this problem (Lemma 18.1, in [35]), we have

$$\begin{aligned}\mathcal{F}(v,v) &= \|\mathbf{b}\cdot\nabla v\|_{0,\Omega}^2 + \|v\|_{0,\Omega}^2 + \|v\|_{\Gamma_I}^2 \\ &\quad + \langle \mathbf{b}\cdot\nabla v, v\rangle_{0,\Omega} \\ &= \|\mathbf{b}\cdot\nabla v\|_{0,\Omega}^2 + (1 - \frac{1}{2}\nabla\cdot\mathbf{b})\|v\|_{0,\Omega}^2 \\ &\quad + \|v\|_{\Gamma_I}^2 + \frac{1}{2}\int_\Gamma \mathbf{b}\cdot\mathbf{n}v^2\,ds \\ &= \|\mathbf{b}\cdot\nabla v\|_{0,\Omega}^2 + (1 - \frac{1}{2}\nabla\cdot\mathbf{b})\|v\|_{0,\Omega}^2 \\ &\quad + \frac{1}{2}\|v\|_{\Gamma_I}^2 + \frac{1}{2}\|v^2\|_{\Gamma_O}^2 \\ &\geq c_1\|v\|_V^2.\end{aligned} \tag{3.33}$$

□

Remark 3.3. *The assumption that* $1 - \frac{1}{2}\nabla \cdot \mathbf{b} \geq \alpha > 0$ *is often necessary for Galerkin-based methods and simplifies the proof for the least-squares formulation. We avoid this restriction in Chapter 4 by establishing a Poincaré inequality and a trace Theorem for the general case.*

We can now conclude existence and uniqueness. We denote by V^h a finite dimensional subspace of $V \subset H^1(\Omega)$. One example for V^h is the space of standard continuous bilinear finite elements. In summary, we have

Corollary 3.4 (Existence and Uniqueness). *There exists a unique solution* $u \in H^1(\Omega)$ *to Problem 3.1, and a unique solution* $u^h \in V^h$ *to the corresponding discrete problem.*

Proof. The Lax-Milgram Theorem [13] establishes the result for u. Existence and uniqueness also follow for u^h, since $V^h \subset H^1(\Omega)$. This is also described for this problem in [9, 25]. □

The next result can be found in many standard finite element texts (see [18, 26]). It provides an important benchmark for the quality we obtain in our results.

Lemma 3.5. *For any* $u \in H^{k+1}(\Omega)$*, where* k *is the degree of a piecewise polynomial continuous space,* V^h*, there exists interpolant* $\Pi^h u \in V^h$ *such that*

$$\begin{aligned} \|u - \Pi^h u\|_{s,\Omega} &\leq C h^{k+1-s} \|u\|_{k+1,\Omega}, \\ \|u - \Pi^h u\|_{\Gamma_I} &\leq C h^{k+\frac{1}{2}} \|u\|_{k+1,\Omega}, \end{aligned} \tag{3.34}$$

where C *is a grid independent constant and* $0 \leq s \leq k$*.*

These inequalities do not directly determine the accuracy of the least-squares approximation, although they are used in the following theorem to prove the desired relation.

Theorem 3.6 (Energy Norm a priori Estimate). *Let* $u \in H^{k+1}(\Omega)$ *be the solution to (3.1) and let* $u^h \in V^h$ *be the solution to the least-squares weak problem (3.24). Then there exists a constant* C*, independent of* h*, such that*

$$\|u - u^h\|_V \leq C h^k \|u\|_{k+1,\Omega}. \tag{3.35}$$

Proof. Recall that the least-squares minimization principle implies

$$\begin{aligned} \|u - u^h\|_V &= \inf_{v^h \in V^h} \|u - v^h\|_V \\ &\leq \|u - \Pi^h u\|_V, \end{aligned} \tag{3.36}$$

where $\Pi^h u$ is the interpolant from Lemma 3.5. We then have

$$\begin{aligned} \|u - u^h\|_V &\leq c \left(\|\mathbf{b} \cdot \nabla (u - \Pi^h u)\|_{0,\Omega}^2 + \|u - \Pi^h u\|_{0,\Omega}^2 + \|u - \Pi^h u\|_{\Gamma_I}^2 \right)^{\frac{1}{2}} \\ &\leq c \left(h^{2k} \|u\|_{0,\Omega}^2 + h^{2(k+1)} \|u\|_{0,\Omega}^2 + h^{2(k+\frac{1}{2})} \|u\|_{k+1,\Omega} \right)^{\frac{1}{2}} \\ &\leq c h^k \|u\|_{k+1,\Omega}. \end{aligned} \tag{3.37}$$

□

Convergence in the energy norm is critical in establishing stability of the discretization, while L^2-norm estimates are important because they indicate how well we are resolving certain profiles in the solution, such as steep slopes or discontinuities. Recalling (2.82), we find a gap of $\frac{1}{2}$ for the L^2-norm error estimates. For the LSFEM, a gap of 1 can be found directly from (3.35), which is summarized by the following.

Lemma 3.7 (L^2 a priori estimate). *Under the assumptions of Theorem 3.6, there exists a constant C, independent of h, such that*

$$\|u - u^h\|_0 \leq C h^k \|u\|_{k+1,\Omega}. \tag{3.38}$$

Proof. The result follows immediately from Theorem 3.6. □

For grid aligned flow—e.g., $\mathbf{b} = (1, 0)$—Bochev and Choi [9] improve this bound to find a gap of $\frac{2}{3}$. Moreover, numerically, they find no gap for smooth solutions. We find similar convergence results as reported in [11] and we comment on how these results carry into the study of discontinuous solutions.

We consider two continuous solutions that differ in degree of smoothness, both of which are used in [11] for a comparative study of Galerkin, SUPG, and LS finite element methods. We investigate these results to gain an understanding of the performance we should expect in

the discontinuous case. Consider a constant advection given by $\mathbf{b} = (\cos(\theta), \sin(\theta)) = (b_1, b_2)$, where θ is the angle the flow makes with the x-axis. Let $\theta = \frac{\pi}{6}$ and $\Omega = [0,1]^2$ for this brief study.

Example #1: $u \in H^{k+1}(\Omega)$, where $k \geq 1$ is the polynomial degree of the finite element test space V^h, (3.27):

$$u(x,y) = \frac{1}{(y - (\frac{b_2}{b_1}x + 0.5))^2 + .1}. \tag{3.39}$$

This example allows us to use a zero right side, f, since $\mathbf{b} \cdot \nabla u = 0$. The solution is infinitely smooth and is illustrated in Figure 3.1.

Example #2: $u \in H^{1+\varepsilon}(\Omega)$:

$$u(x,y) = \begin{cases} \frac{1}{(y-(\frac{b_2}{b_1}))^2+.1}, & \text{if } y \geq \frac{b_2}{b_1}x \\ \frac{20}{7}, & \text{if } y < \frac{b_2}{b_1}x. \end{cases} \tag{3.40}$$

This example is based on Example #1, which allows us to use a zero right side. Although the solution is still continuous, we cannot apply our previous theorems because $u \notin H^{k+1}(\Omega)$. The non-smooth edge can be seen emanating from the origin in Figure 3.1.

For completeness, we add the following example, which is similar to the discontinuous solutions considered in the next chapter. The profile is displayed in Figure 3.1.

Example #3: $u \in H^{\frac{1}{2}-\varepsilon}(\Omega)$:

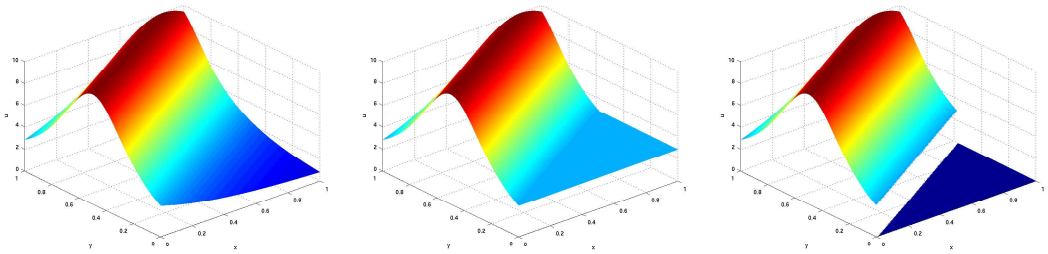
$$u(x,y) = \begin{cases} \frac{1}{(y-(\frac{b_2}{b_1}))^2+.1}, & \text{if } y \geq \frac{b_2}{b_1}x \\ 0, & \text{if } y < \frac{b_2}{b_1}x. \end{cases} \tag{3.41}$$

k	L^2-norm	Energy norm	Expected L^2	Expected Energy
1	1.93	1.00	1.33	1.00
2	3.62	1.99	2.33	2.00
3	4.47	2.85	3.33	3.00

Table 3.1: Finite element convergence for Example #1, (3.39).

Our results complement those obtained by Bochev and Choi [11]. Table 3.1 shows convergence for various polynomials degrees in the L^2 and energy norms for Example #1. We present the convergence rates using a triangular mesh, although using a quadrilateral mesh reveals similar trends. What we obtain matches the expected convergence rates and tends toward or exceeds the optimal rates for several of values of k. An important property is that the magnitude of the error is smaller for higher-order elements per degree of freedom. See Figure 3.2.

As shown in Table 3.3, the convergence for Example #2 is severely degraded. Since u is only in $H^{1+\varepsilon}(\Omega)$, the assumptions for Theorem 3.6 and Lemma 3.7 are not met. However, the performance is still adequate and we see a slight increase in the convergence rates as the



(a) $H^{k+1}(\Omega)$ (b) $H^{1+\varepsilon}(\Omega)$ (c) $H^{\frac{1}{2}-\varepsilon}(\Omega)$

Figure 3.1: Solution examples.

k	L^2-norm	Energy norm
1	.92	.97
2	.98	1.09
3	1.15	1.10

Table 3.2: Finite element convergence for Example #2, (3.40).

k	L^2-norm	Energy norm
1	.24	.26
2	.32	.31
3	.36	.34

Table 3.3: Finite element convergence for Example #3, (3.41).

polynomial degree increases. Moreover, the errors are smaller per degree of freedom for higher-order elements, Figure 3.3.

We briefly present results for Example #3 in Table 3.3. We find similar behavior for the discontinuous profile as we did for Example #3. Convergence is poor, although expected, because we are approximating discontinuous profiles with a continuous finite element space. This type of approximation is the focus of the next chapter.

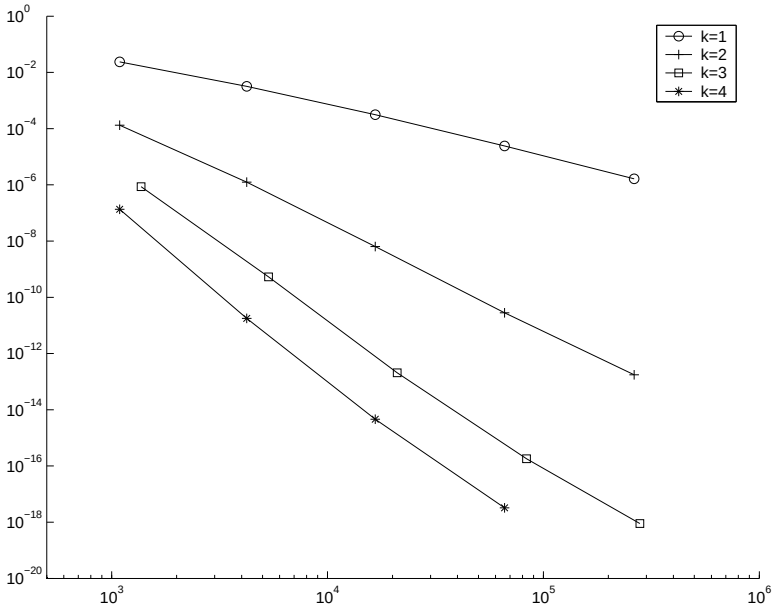
Although we have thus far only presented a theoretical basis for solutions smooth enough to be in $H^{k+1}(\Omega)$, we can draw conclusions about the LSFEM for discontinuous solutions. Recall the stability estimates for the three main finite element approaches. We have

$$\|u_h\|_{0,\Omega} + \|u_h\|_{0,\Gamma} \leq C\|f\|_{0,\Omega}, \tag{3.42a}$$

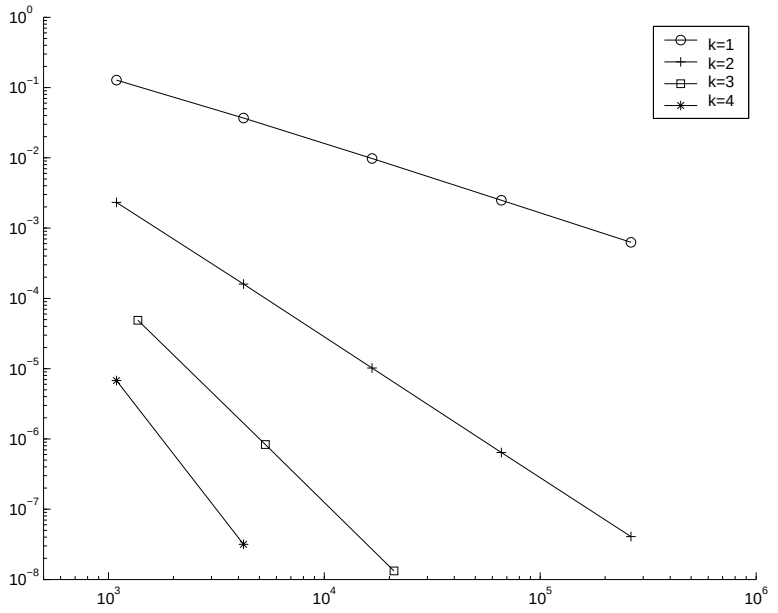
$$\|u_h\|_{0,\Omega} + \sqrt{h}\|\mathbf{b}\cdot\nabla u_h\|_{0,\Omega} + \|u_h\|_{0,\Gamma} \leq C\|f\|_{0,\Omega}, \tag{3.42b}$$

$$\|u_h\|_{0,\Omega} + \|\mathbf{b}\cdot\nabla u_h\|_{0,\Omega} + \|u_h\|_{0,\Gamma} \leq C\|f\|_{0,\Omega}, \tag{3.42c}$$

for Galerkin (2.77), SUPG (2.81), and Least-Squares, respectively. These estimates reveal the superiority of the LSFEM to control streamline derivatives, an advantage of the least-squares approach that motivates studying discontinuous solutions in this framework. Stability is a common roadblock to many methods, particularly for higher-order accurate schemes. A comparative study of the three methods in [11] finds that the least-squares method maintains attractive solution quality over the standard Galerkin and SUPG methods in terms of oscillations, overshoots, and undershoots, yet, smearing of the discontinuity is more pronounced. The results in Table 3.3 are encouraging and lead us to a more comprehensive analysis of the solution quality, convergence per degree of freedom, and theoretical background in the next chapter.

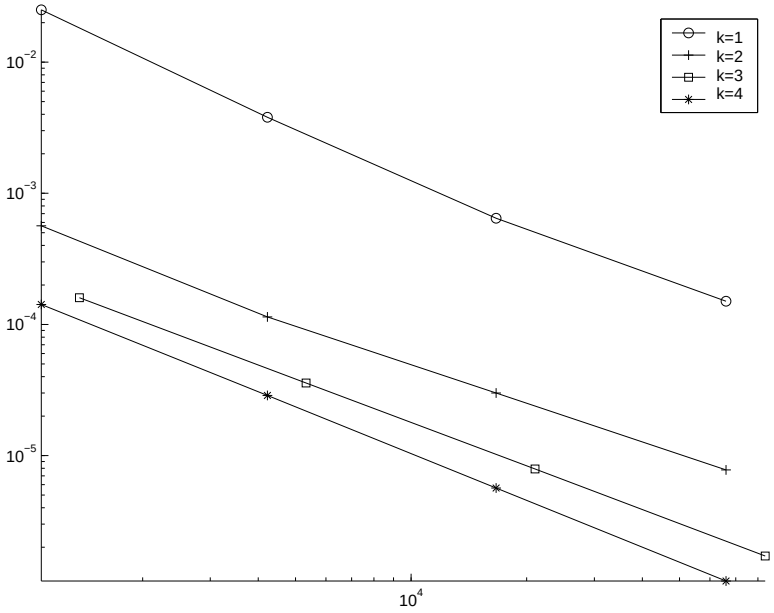


(a) L^2-norm

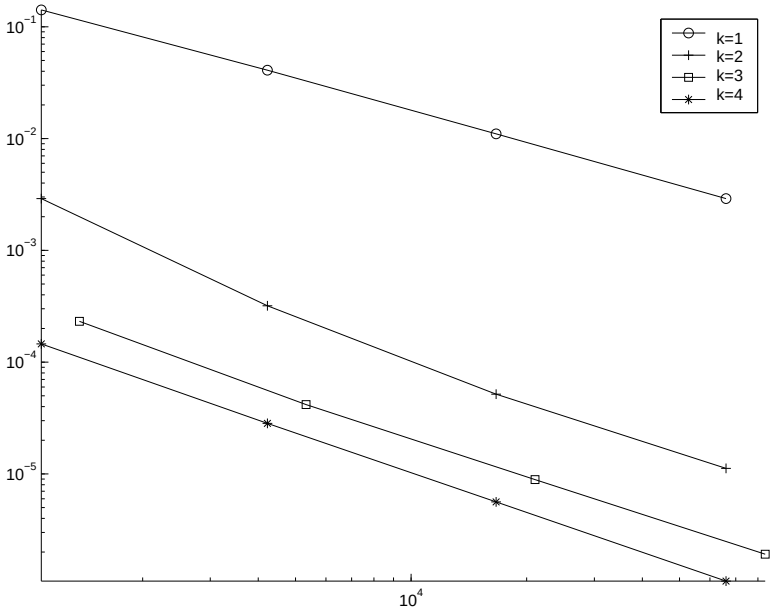


(b) Energy norm

Figure 3.2: Convergence plots for Example #1.



(a) L^2-norm



(b) Energy norm

Figure 3.3: Convergence plots for Example #2.

Chapter 4

Least-Squares Finite Elements for the Advection Equation

We consider scalar linear PDEs of hyperbolic type of the form

$$\mathbf{b} \cdot \nabla u = f, \quad \text{in } \Omega, \tag{4.1a}$$

$$u = g, \quad \text{on } \Gamma_I, \tag{4.1b}$$

with $\mathbf{b}(\mathbf{x})$ a flow field on $\Omega \subset \mathbb{R}^d$,

$$\Gamma_I := \{\mathbf{x} \in \partial\Omega \colon \mathbf{n}(\mathbf{x}) \cdot \mathbf{b}(\mathbf{x}) < 0\} \tag{4.2}$$

the inflow part of the boundary of domain Ω, and, $\mathbf{n}(\mathbf{x})$ is the outward unit normal on $\partial\Omega$. Equations of this type, often called transport equations or linear advection equations, arise in many applications in science and engineering, as noted in Chapter 2. For the last several decades, much effort has been devoted to finding more accurate and efficient numerical solution methods for equations of this form. Not only do these equations have wide applications by themselves, but they also form a prototype for more general hyperbolic problems, including systems of nonlinear conservation laws [61] or transport equations in phase space [63] (see Section 2.1). Successful numerical methods for (4.1a-4.1b) can often be used as building blocks for the numerical solution of more complicated hyperbolic PDEs [61].

As discussed in Chapter 3, least-squares finite element formulations lead to a number of additional computational benefits for hyperbolic PDEs. The method results in stable numerical approximations with attractive solution quality, including limited oscillations and contained

overshoots and undershoots. The least-squares approach also leads to near optimal finite element convergence for smooth solutions. Another advantage of the LSFEM approach is that higher-order accurate methods can easily be constructed to be linear (for linear or linearized PDEs). As is shown in this section, these linear higher-order discretizations do not exhibit spurious oscillations at discontinuities and overshoots do not grow. This is in contrast to most other methods, e.g. DG methods, where nonlinear limiter functions must be employed to maintain monotonicity [27].

Discontinuous finite element methods for hyperbolic PDEs, in particular, DG methods [54], have enjoyed substantial interest in recent years [27, 47]. They have proved to be effective and versatile high-order methods for nonlinear hyperbolic systems with natural conservation properties and good monotonicity properties near discontinuities due to up-winding. They can handle non-matching grids and nonuniform polynomial approximations, with orthogonal bases which lead to diagonal mass matrices, and they are easily parallelized by using block-type preconditioners [27, 47]. We extend the standard LSFEM to a nonconforming discontinuous least-squares finite element method (DLSFEM) in Section 4.4.

The contributions of this chapter are twofold. First, we establish finite element convergence estimates for the continuous and discontinuous LSFEM formulations we propose. In Section 4.1, we start by presenting a trace theorem that precisely identifies the space of admissible boundary data. Our continuous LSFEM is a modification of the LSFEM studied by Bochev and Choi [9] for a problem similar to (4.1a-4.1b). We briefly studied the alternative problem in Chapter 3, equation (3.1), where (4.1a) is replaced by

$$\mathbf{b} \cdot \nabla p + cu = f \quad \text{in } \Omega. \tag{4.3}$$

Unfortunately, the convergence proof in [9] for this modified problem does not carry over to our LSFEM formulation for (4.1a-4.1b). Our discontinuous LSFEM (DLSFEM) is a slight modification of the method proposed by Houston et al. in [47], which does not provide a rigorous finite element convergence proof for this method. So the theory developed here is important in

establishing the efficacy of our approach.

Second, we numerically study the order of convergence of our LSFEM and DLSFEM for discontinuous flow solutions in the general case that the discontinuity is not aligned with the computational mesh. For extensive studies of solution quality and convergence orders for continuous flows, we refer to [11, 9, 47]. Bochev and Choi [11] show in numerical LSFEM experiments that no substantial spurious oscillations arise near discontinuities in the solution. This finding is confirmed in Houston et al. [47] for DLSFEM. Both papers show that, for continuous flows, the accuracy of the (D)LSFEM is comparable to (D)G and (D)SUPG results (especially for higher-order elements and fine grids), while for discontinuous solutions the smearing is substantially larger in the (D)LSFEM results. Unfortunately, [11, 47] do not obtain order of convergence estimates for discontinuous flow solutions. We thus study numerical convergence of discontinuous flow solutions for elements of increasing polynomial degree on triangular and quadrilateral meshes. Our numerical study of discontinuous flow simulation with LSFEM and DLSFEM yields interesting results. The smearing of the discontinuity improves while the overshoots and oscillations remain contained as we increase the order of the polynomial degree of the finite elements. We find an increase in the convergence rate as the polynomial degree increases. We observe similar behavior in the L^2 norm and functional norm for LSFEM and DLSFEM and for different scalar flow fields.

This chapter is organized as follows. In the next section, we examine the space of admissible boundary data (g in (4.1b)) and establish a trace theorem and Poincaré inequality. This leads, in Section 4.2, to the formulation of a least-squares minimization principle from which a weak form is derived. Coercivity and continuity are proved and **a priori** estimates are obtained. Well-posedness is also proved for a slightly modified functional that is suitable for computations. In Section 4.3, we describe conforming finite element methods that are obtained when the least-squares functional is minimized over finite-dimensional subspaces and error bounds for discontinuous solutions are discussed. In Section 4.4, a discontinuous LSFEM is obtained by minimizing a modified functional that incorporates jump terms over a discontinu-

ous finite dimensional space. Section 4.5 presents a numerical study of the convergence behavior of LSFEM and DLSFEM for discontinuous solutions and for elements of increasing polynomial degree on triangular and quadrilateral meshes. Sharpness and monotonicity of the approximate solution in the neighborhood of discontinuities are investigated.

4.1 Admissible Boundary Data

In this section, we examine the space of admissible boundary data for (4.1a-4.1b) and formulate a Poincaré inequality and a trace theorem. Given Ω in (4.1a-4.1b) $\subset \mathbb{R}^d$, let $\mathbf{b}(\mathbf{x}) = (b_1(\mathbf{x}), \ldots, b_d(\mathbf{x})) \in L^2(\Omega)^d$ be a vector field on Ω. We make the following assumptions on $\mathbf{b}$: for any $\hat{\mathbf{x}} \in \Gamma_I$, let $\mathbf{x}(r) = (x_1(r), \ldots, x_d(r))$ be a streamline of $\mathbf{b}$, that is, the solution of

$$\frac{dx_i(r)}{dr} = b_i(\mathbf{x}(r)), \quad i = 1, \ldots, d, \tag{4.4}$$

with initial condition $\mathbf{x}(r_0) = \hat{\mathbf{x}}$. In this thesis, we limit the discussion to the case where $d = 2$, although extensions to higher dimensions can be established [64]. Assume that there exist constants β_0 and β_1 such that $0 < \beta_0 \leq |\mathbf{b}| \leq \beta_1 < \infty$ on Ω. We assume that there exists a transformation to a coordinate system (r, s) such that the streamlines align with the r coordinate direction and that the Jacobian, J, of the transformation is bounded. This implies that no two streamlines intersect and that Ω is the collection of all such streamlines. Further, we assume that every streamline connects Γ_I and Γ_O with a finite length $\ell(\hat{\mathbf{x}})$, where $\hat{\mathbf{x}} \in \Gamma_I$. We require partition $\mathcal{T}^h$ of Ω to be an admissible, quasi-uniform tessellation (see [13, 18]). We assume the same for $\hat{\mathcal{T}}^h$ of $\hat{\Omega}$, the image of $\mathcal{T}^h$ under the transformation. For our numerical tests, we use uniform partitions of triangles and quadrilaterals for $\mathcal{T}^h$. A specific structure for $\hat{\mathcal{T}}^h$ need not be defined, as it is used only theoretically to so simplify certain proofs.

We define the boundary norm

$$\|g\|^2_{B_\ell} := \int_{\Gamma_I} \ell(\mathbf{x}(\sigma)) |\bar{\mathbf{b}} \cdot \mathbf{n}| g^2 d\sigma, \tag{4.5}$$

where $\bar{\mathbf{b}}$ is the unit vector in the direction $\mathbf{b}$ and $\ell(\mathbf{x})$ is the length of the streamline of $\mathbf{b}$ connecting $\mathbf{x}(\sigma) \in \Gamma_I$ to the outflow boundary Γ_O. Define the space B_ℓ to be the closure of

$C_0^\infty(\Gamma_I)$ in the B_ℓ-norm (4.5). Assuming that $f \in L^2(\Omega)$ in (4.1a) and using standard notation for L^2 norms, we define the natural norm (often called the **graph** norm) as

$$\|u\|_{V_\ell}^2 := \|u\|_{0,\Omega}^2 + \|\mathbf{b}\cdot\nabla u\|_{0,\Omega}^2, \tag{4.6}$$

and the solution space as

$$V_\ell := \{u \in L^2(\Omega) \ : \ \|u\|_{V_\ell} < \infty\}. \tag{4.7}$$

Remark 4.1. *Depending on* $\mathbf{b}$ *and* Ω*,* B_ℓ *can be larger than* $L^2(\Gamma_I)$*. As an example, take* $\Omega = [0,1]^2$*. We can construct a vector field* $\mathbf{b}$ *that is small at the origin, while the function is large in comparison. Let*

$$\mathbf{b}(\mathbf{x}) = (y, x) \tag{4.8}$$

and let $g(x,y)$ *be the restriction of*

$$\tilde{g}(x,y) = (x+y)^{-\frac{1}{2}} \tag{4.9}$$

to the in inflow boundary

$$\Gamma_I = [0,1]\times\{0\} \cup \{0\}\times[0,1]. \tag{4.10}$$

It follows that $\|g\|_{0,\Omega} = \infty$*, while* $\|g\|_{B_\ell}$ *is finite. Similar examples hold where* $\ell(x)$ *is small in regions where* $g(x,y)$ *is not square integrable on* Γ_I*.*

Lemma 4.2 (Trace Inequality). *If* $u \in V_\ell$ *and* $u = g$ *on* Γ_I*, then there exists a constant* C*, depending only on* β_0*,* Ω*, and the transformation Jacobian* J*, such that*

$$\|g\|_{B_\ell}^2 \le C\left(\|u\|_{0,\Omega}^2 + \|\mathbf{b}\cdot\nabla u\|_{0,\Omega}^2\right). \tag{4.11}$$

Proof. We first prove (4.11) assuming $\mathbf{b}$ is constant. Let $\bar{\mathbf{b}} = \frac{1}{|\mathbf{b}|}\mathbf{b}$, the unit vector in the direction of $\mathbf{b}$. For every $\hat{\mathbf{x}} \in \Gamma_I$, let

$$\ell(\hat{\mathbf{x}}) = |s_1(\hat{\mathbf{x}})|, \tag{4.12}$$

where $(0, s_1)$ is the largest interval for which $\hat{\mathbf{x}} + s\hat{\mathbf{b}} \in \Omega$ for all $s \in (0, s_1)$. Here, $\hat{\mathbf{b}} = \bar{\mathbf{b}}(\hat{\mathbf{x}})$ generates the unique streamline intersecting the point $\hat{\mathbf{x}}$. Let $d\sigma$ be the differential arc length

along Γ_I and $\mathbf{n}$ the outward unit normal on Γ_I. Then, for any $u \in V_\ell$, we have

$$\iint_\Omega u(\mathbf{x})\, dA = \int_{\Gamma_I} \int_0^{s_1(\hat{\mathbf{x}})} u(\hat{\mathbf{x}} + s\hat{\bar{\mathbf{b}}})\, ds\, |\bar{\mathbf{b}} \cdot \mathbf{n}|\, d\sigma. \tag{4.13}$$

For any $s \in [0, s_1(\hat{\mathbf{x}})]$, we have

$$u^2(\hat{\mathbf{x}} + s\hat{\bar{\mathbf{b}}}) = u^2(\hat{\mathbf{x}}) + \int_0^s \bar{\mathbf{b}} \cdot \nabla u^2(\hat{\mathbf{x}} + t\hat{\bar{\mathbf{b}}})\, dt, \tag{4.14}$$

so

$$u^2(\hat{\mathbf{x}}) \le u^2(\hat{\mathbf{x}} + s\hat{\bar{\mathbf{b}}}) + \int_0^{s_1(\hat{\mathbf{x}})} \left| \bar{\mathbf{b}} \cdot \nabla u^2(\hat{\mathbf{x}} + t\hat{\bar{\mathbf{b}}}) \right| \, dt. \tag{4.15}$$

Integrating over $(0, s_1(\hat{\mathbf{x}}))$ with length element dt and using the relation $\ell(\hat{\mathbf{x}}) = \int_0^{s_1(\hat{\mathbf{x}})} dt$, we thus obtain

$$\ell(\hat{\mathbf{x}}) u^2(\hat{\mathbf{x}}) \le \int_0^{s_1(\hat{\mathbf{x}})} u^2(\hat{\mathbf{x}} + t\hat{\bar{\mathbf{b}}})\, dt + \ell(\hat{\mathbf{x}}) \int_0^{s_1(\hat{\mathbf{x}})} \left| \bar{\mathbf{b}} \cdot \nabla u^2(\hat{\mathbf{x}} + t\hat{\bar{\mathbf{b}}}) \right| \, dt. \tag{4.16}$$

Integrating along Γ_I with length element $|\bar{\mathbf{b}} \cdot \mathbf{n}| d\sigma$ yields

$$\begin{aligned} \int_{\Gamma_I} \ell(\hat{\mathbf{x}}) u^2(\hat{\mathbf{x}}) |\bar{\mathbf{b}} \cdot \mathbf{n}| d\sigma \le & \int_{\Gamma_I} \int_0^{s_1(\hat{\mathbf{x}})} u^2(\hat{\mathbf{x}} + t\hat{\bar{\mathbf{b}}})\, dt\, |\bar{\mathbf{b}} \cdot \mathbf{n}|\, d\sigma \\ & + \int_{\Gamma_I} \ell(\hat{\mathbf{x}}) \int_0^{s_1(\hat{\mathbf{x}})} \left| \bar{\mathbf{b}} \cdot \nabla u^2(\hat{\mathbf{x}} + t\hat{\bar{\mathbf{b}}}) \right| \, dt\, |\bar{\mathbf{b}} \cdot \mathbf{n}|\, d\sigma. \end{aligned} \tag{4.17}$$

Let $D = \operatorname{diam}(\Omega)$. Applying the Cauchy-Schwarz and ε inequalities, we thus have

$$\begin{aligned} \|u\|_{B_\ell}^2 &\le \|u\|_{0,\Omega}^2 + 2D \|(\bar{\mathbf{b}} \cdot \nabla) u\|_{0,\Omega} \|u\|_{0,\Omega} \\ &\le \|u\|_{0,\Omega}^2 + D^2 \|u\|_{0,\Omega}^2 + \|(\bar{\mathbf{b}} \cdot \nabla) u\|_{0,\Omega}^2 \\ &\le C(\|u\|_{0,\Omega}^2 + \|(\mathbf{b} \cdot \nabla) u\|_{0,\Omega}^2). \end{aligned} \tag{4.18}$$

For the general case of variable $\mathbf{b}(\mathbf{x})$, the bound (4.11) follows using the assumed transformation with bounded Jacobian and the fact that $u = g$ on the inflow boundary Γ_I. □

Remark 4.3. *The constants C that appear in Lemma 4.2 and throughout the rest of the dissertation are generic and may change value with each occurrence, but depend only on β_0, Γ_I, and Ω.*

Lemma 4.4 (Poincaré Inequality). *Let $D = diam(\Omega)$. There exists a constant C, depending only on β_0, Ω, and the transformation Jacobian J, such that*

$$\|u\|_{0,\Omega}^2 \leq C(\|u\|_{B_\ell}^2 + D^2\|\mathbf{b}\cdot\nabla u\|_{0,\Omega}^2). \tag{4.19}$$

Proof. As in the preceding proof, we derive this Poincaré inequality for constant $\mathbf{b}$ and rely on the transformation with bounded Jacobian to achieve the general result. Let $\bar{\mathbf{b}} = \frac{1}{|\mathbf{b}|}\mathbf{b}$ and let 0 and $s_1(\mathbf{x})$ be as in the proof of Lemma 4.2. For every $\hat{\mathbf{x}} \in \Gamma_I$, let $\ell(\hat{\mathbf{x}}) = |s_1(\hat{\mathbf{x}})|$. Also, let $\hat{\mathbf{b}} = \bar{\mathbf{b}}(\hat{\mathbf{x}})$ generate the unique streamline intersecting the point $\hat{\mathbf{x}}$. Notice that for $s \in [0, s_1(\mathbf{x})]$, we have

$$u(\hat{\mathbf{x}} + s\hat{\mathbf{b}}) = u(\hat{\mathbf{x}}) + \int_0^s \bar{\mathbf{b}}\cdot\nabla u(\hat{\mathbf{x}} + t\hat{\mathbf{b}})\, dt. \tag{4.20}$$

Squaring both sides and using the ε and Jensen inequalities yields

$$\begin{aligned} |u(\hat{\mathbf{x}} + s\hat{\mathbf{b}})|^2 &\leq 2\left(|u(\hat{\mathbf{x}})|^2 + \left(\int_0^{s_1(\hat{\mathbf{x}})} \left|\bar{\mathbf{b}}\cdot\nabla u(\hat{\mathbf{x}} + t\hat{\mathbf{b}})\right| dt\right)^2\right) \\ &\leq 2\left(|u(\hat{\mathbf{x}})|^2 + \ell(\hat{\mathbf{x}})\int_0^{s_1(\hat{\mathbf{x}})} \left|\bar{\mathbf{b}}\cdot\nabla u(\hat{\mathbf{x}} + t\hat{\mathbf{b}})\right|^2 dt\right). \end{aligned} \tag{4.21}$$

Integrating over $(0, s_1(\hat{\mathbf{x}}))$ with dt and using the relation $\ell(\hat{\mathbf{x}}) = \int_0^{s_1(\hat{\mathbf{x}})} dt$ we thus obtain

$$\int_0^{s_1(\hat{\mathbf{x}})} \left|u(\hat{\mathbf{x}} + t\hat{\mathbf{b}})\right|^2 dt \leq 2\left(\ell(\hat{\mathbf{x}})|u(\hat{\mathbf{x}})|^2 + \ell(\hat{\mathbf{x}})^2\int_0^{s_1(\hat{\mathbf{x}})} \left|\bar{\mathbf{b}}\cdot\nabla u(\hat{\mathbf{x}} + t\hat{\mathbf{b}})\right|^2 dt\right). \tag{4.22}$$

Integrating along Γ_I with $|\bar{\mathbf{b}}\cdot\mathbf{n}|d\sigma$ and using (4.13) then yields

$$\begin{aligned} \|u\|_{0,\Omega}^2 &\leq 2\left(\int_{\Gamma_I} \ell(\hat{\mathbf{x}})u^2(\hat{\mathbf{x}})|\bar{\mathbf{b}}\cdot\mathbf{n}|d\sigma + D^2\|\bar{\mathbf{b}}\cdot\nabla u\|_{0,\Omega}^2\right) \\ &\leq C\left(\|u\|_{B_\ell}^2 + D^2\|\mathbf{b}\cdot\nabla u\|_{0,\Omega}^2\right). \end{aligned} \tag{4.23}$$

□

The following trace theorem establishes B_ℓ as the space of admissible functions for inflow boundary conditions when the right side, f in (4.1a), is in $L^2(\Omega)$.

Theorem 4.5 (Trace Theorem). *For $u \in V_\ell$, let $\gamma(u)$ denote the trace of u on Γ_I. Then the map $\gamma : V_\ell \to B_\ell$ is a bounded surjection.*

Proof. For any $g \in B_\ell$, we can construct a "flat" function u such that $u = g$ on Γ_I and $\mathbf{b} \cdot \nabla u = 0$ in Ω (i.e., u is constant along streamlines). From Poincaré inequality (4.19), it follows that $u \in V_\ell$. Together with the trace inequality, this proves the theorem. □

Remark 4.6. *Our trace theorem is similar to the theorem proved in [64] for the more general case of the neutron transport equation in phase space. A different characterization of the trace space is given in [46] for the general class of Friedrichs systems, of which (4.1a-4.1b) is a special case. The trace operator defined in [46] is not surjective. In this sense, in contrast to Theorem 4.5, the trace space identified in [46] does not provide a sharp trace theorem.*

4.2 Least-Squares Weak Form

In this section, we formulate a least-squares minimization principle, derive the associated weak form, and prove existence of a unique weak solution $u \in V_\ell$. We use the tools developed in the previous section and coercivity and continuity with respect to the natural norm in (4.6) to arrive at these results.

Define the least-squares functional

$$\mathcal{G}_\ell(u;\, f, g) := \|\mathbf{b} \cdot \nabla u - f\|_{0,\Omega}^2 + \|u - g\|_{B_\ell}^2. \tag{4.24}$$

First note that if u satisfies (4.1a-4.1b), then

$$u = \underset{u \in V_\ell}{\operatorname{argmin}}\, \mathcal{G}_\ell(u;\, f, g). \tag{4.25}$$

The bilinear form associated with $\mathcal{G}_\ell$ in (4.24) is

$$\mathcal{F}_\ell(u, v) := \langle \mathbf{b} \cdot \nabla u,\, \mathbf{b} \cdot \nabla v \rangle_{0,\Omega} + \langle u,\, v \rangle_{B_\ell}, \tag{4.26}$$

and the corresponding linear form is

$$F(v) = \langle f,\, \mathbf{b} \cdot \nabla v \rangle_{0,\Omega} + \langle g,\, v \rangle_{B_\ell}, \tag{4.27}$$

with standard notation for scalar products associated with norms. Note: $F(v) \in V_\ell'$, the dual space of V_ℓ.

The weak form of the minimization principle is stated as the following.

Problem 4.7. *Find $u \in V_\ell$ such that*

$$\mathcal{F}_\ell(u, v) = F(v) \qquad \forall\, v \in V_\ell. \tag{4.28}$$

The following theorem establishes coercivity and continuity in the V_ℓ norm of the bilinear form $\mathcal{F}_\ell(\cdot, \cdot)$. With these properties, $\mathcal{F}_\ell(\cdot, \cdot)$ is frequently referred to as being V_ℓ-**elliptic** [13].

Theorem 4.8 (Coercivity and Continuity, Existence and Uniqueness). *There exist constants c_0 and c_1 such that, for every $u, v \in V_\ell$,*

$$c_0 \|u\|_{V_\ell}^2 \leq \mathcal{F}_\ell(u, u) \tag{4.29}$$

and

$$\mathcal{F}_\ell(u, v) \leq c_1 \|u\|_{V_\ell} \|v\|_{V_\ell}. \tag{4.30}$$

Furthermore, for every $f \in L^2(\Omega)$ and $g \in B_\ell$, there exists a unique $u \in V_\ell$ solving weak problem (4.28).

Proof. Using the definition of $\|u\|_{V_\ell}$ from (4.6) and Poincaré Inequality (4.19) yields

$$\begin{aligned} \|u\|_{V_\ell}^2 &\leq C(\|u\|_{B_\ell}^2 + D^2 \|\mathbf{b} \cdot \nabla u\|_{0,\Omega}^2) + \|\mathbf{b} \cdot \nabla u\|_{0,\Omega}^2 \\ &\leq C(\|u\|_{B_\ell}^2 + \|\mathbf{b} \cdot \nabla u\|_{0,\Omega}^2) \\ &= C\mathcal{F}_\ell(u, u), \end{aligned} \tag{4.31}$$

which yields (4.29). Similarly, applying the Cauchy-Schwarz inequality followed by trace inequality (4.11) and Cauchy-Schwarz again, we have

$$\begin{aligned} \mathcal{F}_\ell(u, v) &\leq \|\mathbf{b} \cdot \nabla u\|_{0,\Omega} \|\mathbf{b} \cdot \nabla v\|_{0,\Omega} + \|u\|_{B_\ell} \|v\|_{B_\ell} \\ &\leq \|\mathbf{b} \cdot \nabla u\|_{0,\Omega} \|\mathbf{b} \cdot \nabla v\|_{0,\Omega} + C \left(\|u\|_{0,\Omega}^2 + \|\mathbf{b} \cdot \nabla u\|_{0,\Omega}^2\right)^{\frac{1}{2}} C \left(\|v\|_{0,\Omega}^2 + \|\mathbf{b} \cdot \nabla v\|_{0,\Omega}^2\right)^{\frac{1}{2}} \\ &\leq C \|u\|_{V_\ell} \|v\|_{V_\ell}, \end{aligned} \tag{4.32}$$

which confirms (4.30).

The trace theorem and Cauchy-Schwarz inequality imply that, for every $f \in L^2(\Omega)$ and $g \in B_\ell$,

$$F(v) := \langle f,\, \mathbf{b}\cdot\nabla v\rangle_{0,\Omega} + \langle g,\, v\rangle_{B_\ell} \tag{4.33}$$

is a bounded linear functional on V_ℓ. Thus, we can embed the pair $(f,g) \in L^2(\Omega)\times B_\ell$ into V_ℓ', the dual space of V_ℓ. We now show that the embedding is injective.

To do this, pick $(f,g) \in L^2(\Omega)\times B_\ell$ and suppose

$$F(v) = \langle f,\, \mathbf{b}\cdot\nabla v\rangle_{0,\Omega} + \langle g,\, v\rangle_{B_\ell} = 0, \tag{4.34}$$

for every $v \in V_\ell$. Thus, if $f = 0$ and $g = 0$, the embedding is injective.

We first show that for $(f,g) \in L^2(\Omega)\times B_\ell$, there exists $u \in V_\ell$ such that

$$\mathcal{L}u = (f,g), \tag{4.35}$$

where $\mathcal{L} : V_\ell \to L^2(\Omega)\times B_\ell$ is defined by

$$\mathbf{b}\cdot\nabla u = f \quad \text{in } \Omega, \tag{4.36a}$$

$$u = g \quad \text{on } \Gamma_I. \tag{4.36b}$$

That is, we must show that $\mathcal{L}$ is surjective. Construct u as follows. Let u_ℓ be a flat function such that

$$\mathbf{b}\cdot\nabla u_1 = 0 \quad \text{in } \Omega, \tag{4.37a}$$

$$u_1 = g \quad \text{on } \Gamma_I. \tag{4.37b}$$

For $\mathbf{x} \in \Omega$, let $\hat{\mathbf{x}}$ be the point on Γ_I with the same streamline as $\mathbf{x}$. Let $\beta(\mathbf{x}) = |\mathbf{b}(\hat{\mathbf{x}})|$ and $\bar{\mathbf{b}} = \frac{\mathbf{b}(\mathbf{x})}{\beta(\mathbf{x})}$. Let u_2 be given by

$$u_2(\mathbf{x}) = \int_{\hat{\mathbf{x}}}^{\mathbf{x}} \frac{f(\hat{\mathbf{x}} + s\bar{\mathbf{b}}(\hat{\mathbf{x}}))}{\beta(\hat{\mathbf{x}} + s\bar{\mathbf{b}}(\hat{\mathbf{x}}))}\, ds. \tag{4.38}$$

Then, u_2 satisfies

$$\mathbf{b}\cdot\nabla u_2 = f \quad \text{in } \Omega, \tag{4.39a}$$

$$u_2 = 0 \quad \text{on } \Gamma_I. \tag{4.39b}$$

Writing $u = u_1 + u_2$, we see that u satisfies (4.36a-4.36b). Thus, $\mathcal{L}$ is a surjection.

Now since $F(v) = 0$ and $u \in V_\ell$ satisfies (4.36a-4.36b), we have

$$\langle f, f\rangle_{0,\Omega} + \langle g,\, g\rangle_{B_\ell} = 0. \tag{4.40}$$

Thus, $f = 0$ and $g = 0$. It follows that the embedding is injective. Together, this implies that for every $(f, g) \in L^2(\Omega) \times B_\ell$, there is a unique $u \in V_\ell$ that satisfies (4.28). □

The following **a priori** estimate is a direct consequence of Theorem 4.8. These bounds are often referred to as stability estimates.

Corollary 4.9 (A Priori Estimate). *There exist constants c_3 and c_4 such that, if u satisfies (4.28), then*

$$c_3\|u\|_{V_\ell} \leq (\|f\|_{0,\Omega} + \|g\|_{B_\ell}) \leq c_4\|u\|_{V_\ell}. \tag{4.41}$$

Proof. The proof follows directly from Theorem 4.8. □

For certain problems, $\ell(\mathbf{x})$ in (4.5) may not be easily computed, making the least-squares formulation intractable. To avoid this difficulty we simply eliminate $\ell(\mathbf{x})$ in the definition of the boundary norm that is used in the functional:

$$\|g\|_B^2 := \int_{\Gamma_I} |\bar{\mathbf{b}} \cdot \mathbf{n}| g^2 ds, \tag{4.42}$$

where $\bar{\mathbf{b}}$ is the unit normal in the direction of $\mathbf{b}$. Let $B = \{g : \|g\|_B < \infty\}$ and notice that $B_\ell \cap L^\infty(\Gamma_I) \subseteq B$. If Ω is such that Γ_I and Γ_O remain a bounded distance apart, then this norm is equivalent to the original norm, $\|\cdot\|_{B_\ell}$. If Γ_I and Γ_O touch, then there are functions in B_ℓ that are not in B. However, for bounded functions, the B_ℓ norm and B norm are equivalent. That is, if we restrict our attention to bounded boundary data, then nothing is lost in modifying the functional in this way.

In general, the trace inequality (4.11) does not hold with B_ℓ replaced by B, but the Poincaré inequality (4.19) does. To retain the inequalities and ellipticity results obtained above, we must include the boundary term in the definition of the norm. Define the norm

$$\|u\|_V^2 := \|u\|_{0,\Omega}^2 + \|\mathbf{b} \cdot \nabla u\|_{0,\Omega}^2 + \|u\|_B^2 \tag{4.43}$$

and the space

$$V := \{u \in L^2(\Omega) \ : \ \|u\|_V < \infty\}. \tag{4.44}$$

The modified functional is then defined as follows: let $(f, g) \in L^2(\Omega) \times B$ and define

$$\mathcal{G}(u;\, f, g) := \|\mathbf{b} \cdot \nabla u - f\|_{0,\Omega}^2 + \|u - g\|_B^2. \tag{4.45}$$

If u satisfies (4.1a-4.1b), then

$$u = \operatorname*{argmin}_{u \in V} \mathcal{G}(u;\, f, g). \tag{4.46}$$

The associated bilinear form is

$$\mathcal{F}(u, v) := \langle \mathbf{b} \cdot \nabla u,\, \mathbf{b} \cdot \nabla v \rangle_{0,\Omega} + \langle u,\, v \rangle_B \tag{4.47}$$

and the linear form is

$$F(v) = \langle f,\, \mathbf{b} \cdot \nabla v \rangle_{0,\Omega} + \langle g,\, v \rangle_B. \tag{4.48}$$

The weak form of the minimization principle is stated as the following.

Problem 4.10. *Find $u \in V$ such that*

$$\mathcal{F}(u, v) = F(v) \qquad \forall\, v \in V. \tag{4.49}$$

With this change, we obtain existence, uniqueness, and an **a priori** estimate as before.

Theorem 4.11 (Coercivity and Continuity, Existence and Uniqueness). *There exist constants c_0 and c_1, such that for every $u, v \in V$,*

$$c_0 \|u\|_V^2 \leq \mathcal{F}(u, u), \tag{4.50}$$

$$\mathcal{F}(u, v) \leq c_1 \|u\|_V \|v\|_V. \tag{4.51}$$

Furthermore, for every $f \in L^2$ and $g \in B$, there exists a unique $u \in V$ solving Problem 4.10.

Corollary 4.12 (A Priori Estimate). *There exist constants c_3 and c_4 such that, if u satisfies (4.49), then*

$$c_3 \|u\|_V \leq (\|f\|_{0,\Omega} + \|g\|_B) \leq c_4 \|u\|_V. \tag{4.52}$$

Remark 4.13. *The $\mathcal{G}$ norm defined by*

$$\|u^h\|_{\mathcal{G}}^2 := \mathcal{G}(u^h, 0, 0) \tag{4.53}$$

is a natural and computable ***a posteriori*** *error estimator. To see this, let $e = u^h - u$, where u solves (4.1a-4.1b). Then*

$$\begin{aligned} \|e\|_{\mathcal{G}} &= \mathcal{G}(u^h - u; 0, 0) \\ &= \mathcal{G}(u^h; f, g). \end{aligned} \tag{4.54}$$

In this way, least-squares methods offer the advantage of a convenient ***a posteriori*** *error indicator. Sharpness of this measure is addressed in [7, 12].*

4.3 Conforming Finite Elements

In this section, we discuss the discrete form of the minimization process. We consider an admissible, quasi-uniform tessellation $\mathcal{T}^h$ of Ω (cf. [13]). For a conforming method, we choose the discrete space $V^h \subset V$. For example, in our numerical tests we use uniform partitions of triangles and quadrilaterals and implement piecewise polynomials with continuity imposed across element edges. Let

$$V^h := \mathcal{M}_k^h \cap \mathcal{C}^0(\Omega), \tag{4.55}$$

where

$$\mathcal{M}_k^h := \{u \colon u \in \mathcal{P}_k(\tau), \forall\, \tau \in \mathcal{T}^h\}. \tag{4.56}$$

Here, $\mathcal{P}_k(\tau)$ is the space of polynomials of total degree $\leq k$ when τ is a triangle and tensor product polynomials of degree $\leq k$ in each coordinate direction when τ is a quadrilateral. We now pose the conforming discrete weak form of the minimization.

Problem 4.14. *Find $u^h \in V^h$ such that*

$$\mathcal{F}(u^h, v^h) = F(v^h) \qquad \forall v^h \in V^h, \tag{4.57}$$

where $\mathcal{F}$ is defined by (4.47) and F is defined by (4.48).

By Ceá's Lemma, we have

$$\|u - u^h\|_V \leq \frac{c_0}{c_1} \inf_{\hat{u}^h \in V^h} \|u - \hat{u}^h\|_V, \tag{4.58}$$

where c_1 and c_0 are the constants from the continuity and coercivity bounds.

In this dissertation, we are primarily interested in discontinuous solutions, u. To this end, suppose g is discontinuous but piecewise smooth, that is, $g \in H^{\frac{1}{2}-\varepsilon}(\Gamma_I)$. Then, for smooth f, u has similar behavior so that $u \in H^{\frac{1}{2}-\varepsilon}(\Omega)$. In this case, it can be shown that, for grid-aligned flow, the discrete solution, u^h, satisfies

$$\|u - u^h\|_V \leq Ch^{\frac{1}{2}-\varepsilon} \|u\|_{\frac{1}{2}-\varepsilon}, \tag{4.59}$$

where C is some grid-independent constant. The exact bound for the non-grid-aligned case remains an open question. Still, the theoretical limit for the grid-aligned case and other results offer some insight. Scott and Zhang describe in [73] an interpolation operator $\tilde{\Pi}^h$ for which $\|u - \tilde{\Pi}^h u\|_{0,\Omega} \leq Ch^{\frac{1}{2}-\varepsilon} \|u\|_{\frac{1}{2}-\varepsilon}$. If we assume that $\frac{1}{2}$ is the optimal L^2-rate of convergence for interpolation, then we can expect the L^2-rate of convergence for the finite element method to be no better than $\frac{1}{2}$. Note that Poincaré inequality (4.19) yields $\|u - u^h\|_{0,\Omega} \leq C\|u - u^h\|_V$. Thus, the V-norm rate of convergence cannot be faster than the L^2-norm rate. In Section 4.5, we discuss our numerical findings regarding error estimates and present results consistent with the error bounds proposed. We find that, as we increase the order of the elements, the convergence rate increases and is bounded by $\frac{1}{2}$ in both the L^2 norm and $\mathcal{G}$ norm. For an extensive analysis of error bounds and convergence rates for smooth solutions, see [11, 9].

4.4 Nonconforming Finite Elements

In this section, we describe the use of discontinuous elements motivated by the case of grid-aligned flow. Consider an example when the characteristics follow the grid and the boundary data is prescribed such that the discontinuity in the solution follows the element edges aligned with the characteristics. In (4.1a-4.1b), let $f = 0$ and g be piecewise constant with discontinuities only at the nodes. If we use the discontinuous space $\mathcal{M}_k^h$ defined by (4.56), then the solution

to (4.1a-4.1b) is in this space. However, the grids we consider are generally not aligned with the flow field $\mathbf{b}(\mathbf{x})$ and boundary data is often more general than this special case. If attention is given to the behavior of the jumps with respect to the grid, a well posed formulation of the problem in a discontinuous least-squares setting is attainable. To this end, let $\mathcal{T}^h = \bigcup_j \tau_j$ be a tessellation of Ω and let $S^h := \mathcal{M}_k^h$ be defined as in (4.56). Let $\Gamma_{i,j} := \tau_i \cap \tau_j$ denote the edge common to elements τ_i and τ_j. Since $S^h \not\subset V$, we call S^h a nonconforming space [13].

For $u^h \in S^h + V$, define the element edge functional as

$$\|u^h\|_{E^h}^2 := \sum_{i,j} \omega_{i,j} \int_{\Gamma_{i,j}} |\mathbf{b} \cdot \mathbf{n}_\tau| [\![u^h]\!]^2 ds, \tag{4.60}$$

where $\mathbf{n}_\tau$ is the outward unit normal to edge $\Gamma_{i,j}$, $\omega_{i,j}$ is a weight to be determined, and $[\![u^h]\!]$ is the jump in u^h across $\Gamma_{i,j}$. We use (4.60) in the least-squares functional to make a distinction between element edges that are closely aligned with the flow and edges that are not by tying together neighboring elements. This behavior is consistent with the regularity of the problem. A solution u of (4.1a-4.1b) would be smooth in the direction of the flow while, perpendicular to the flow, u is only L^2. For further motivation, consider a non-grid-aligned flow with a typical discontinuity (see Figure 4.1). When element edges are nearly aligned with the discontinuity (location A), the term $|\mathbf{b} \cdot \mathbf{n}|$ is small in (4.60), allowing larger jumps between the neighboring elements. However, when an element edge is nearly perpendicular to the flow (location B), $|\mathbf{b} \cdot \mathbf{n}|$ is large, which enforces a stronger connection between the elements and results in a smaller jump.

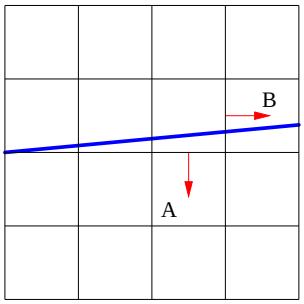


Figure 4.1: Sample of non-grid-aligned flow and outward normals A and B

We can now define a nonconforming least-squares functional similar to (4.45) but now based on so-called "split" norms [22] and edge functional (4.60). With $f \in L^2(\Omega)$, $g \in B$, and $u \in S^h + V$, define the functional

$$\mathcal{G}^h(u; f, g) := \sum_j \|\mathbf{b} \cdot \nabla u - f\|^2_{0,\tau_j} + \|u\|^2_{E^h} + \|u - g\|^2_B, \tag{4.61}$$

and the associated $\mathcal{G}^h$ norm

$$\|u^h\|^2_{\mathcal{G}^h} := \mathcal{G}^h(u^h, 0, 0). \tag{4.62}$$

If $u \in V$, then $\mathcal{G}^h(u; f, g) = \mathcal{G}(u; f, g)$.

Let $e = u^h - u$, where $u^h \in S^h$ and u satisfies (4.1a-4.1b). Notice that

$$\begin{aligned} \|e\|_{\mathcal{G}^h} &= \mathcal{G}^h(u^h - u; 0, 0) \\ &= \mathcal{G}^h(u^h; f, g). \end{aligned} \tag{4.63}$$

Thus, $\mathcal{G}^h$ is a natural and computable **a posteriori** error estimator. The sharpness of LS error estimators is addressed in [7, 12].

We can now describe the discrete variational problem for our discontinuous elements.

Problem 4.15. *Find $u^h \in S^h$ such that*

$$\mathcal{F}(u^h, v^h) = F(v^h) \qquad \forall\, v^h \in S^h, \tag{4.64}$$

where

$$\mathcal{F}(u^h, v^h) := \sum_{\tau_i} \langle \mathbf{b} \cdot \nabla u^h, \mathbf{b} \cdot \nabla v^h \rangle_{0,\tau_i} + \langle u^h,\, v^h \rangle_{E^h} + \langle u^h,\, v^h \rangle_B \tag{4.65}$$

$$F(v^h) := \sum_{\tau_i} \langle f, \mathbf{b} \cdot \nabla v^h \rangle_{0,\tau_i} + \langle g,\, v^h \rangle_B. \tag{4.66}$$

In the following lemma, we find that a uniform Poincaré inequality is satisfied for weights stronger than $\omega = c\frac{1}{h}$, where c is a grid-independent constant. We also show, by example, that weights weaker than $\omega = c\frac{1}{h}$—e.g., $\omega = 1$ or h—result in a violation of the uniform Poincaré inequality. Thus, enforcing the connection between neighboring elements too weakly not only decreases the stability of the solution, but also results in losing a uniform bound on the error in the L^2 norm.

Lemma 4.16 (Uniform Poincaré Inequality). *There exists a constant C, independent of h, such that, for $u^h \in S^h + V$ and $\omega \geq c\frac{1}{h}$, where c is a grid-independent constant, we have*

$$\|u^h\|_{0,\Omega} \leq C\|u^h\|_{\mathcal{G}^h}. \tag{4.67}$$

Furthermore, the above does not hold for $\omega < c\frac{1}{h}$.

Proof. Similar to the proof of Lemma 4.4, we derive the uniform Poincaré inequality for constant $\mathbf{b}$ and rely on the transformation with bounded Jacobian to achieve the general result. As before, let $\bar{\mathbf{b}} = \frac{1}{|\mathbf{b}|}\mathbf{b}$. Let $\hat{\mathbf{x}} \in \Gamma_I$ and let s_k be parameters in $(0, s_m(\hat{\mathbf{x}}))$ such that $\hat{\mathbf{x}}_k = \hat{\mathbf{x}} + s_k\bar{\mathbf{b}}(\hat{\mathbf{x}})$ lies on an element edge, where (s_0, s_m) now plays the role of $(0, s_1)$ in our previous proofs. Since the flow field $\mathbf{b}$ is constant, we have $m(\hat{\mathbf{x}}) = \mathcal{O}(\sqrt{N})$, where N is the number of elements in $\mathcal{T}^h$, the tessellation of Ω, and $m(\hat{\mathbf{x}})$ is the number of element edges encountered by the characteristics generated by $\hat{\mathbf{b}} = \bar{\mathbf{b}}(\hat{\mathbf{x}})$ emanating from $\hat{\mathbf{x}} \in \Gamma_I$. For $0 \leq k < m$, we assume

$$|s_{k+1}(\hat{\mathbf{x}}) - s_k(\hat{\mathbf{x}})| < \tilde{h} \tag{4.68}$$

for all $\hat{\mathbf{x}} \in \Gamma_I$, where

$$\tilde{h} = \max_j\{\text{diam } \tau_j : \tau_j \in \mathcal{T}^h\}. \tag{4.69}$$

Furthermore, assume $\tilde{h} = \mathcal{O}(\frac{1}{\sqrt{N}})$ and let

$$\ell(\hat{\mathbf{x}}) = \sum_{k=1}^{m} |s_k(\hat{\mathbf{x}}) - s_{k-1}(\hat{\mathbf{x}})|. \tag{4.70}$$

Let $[\![u(x)]\!]$ denote the jump in u at x. Using

$$u(\hat{\mathbf{x}} + s\hat{\mathbf{b}}) = u(\hat{\mathbf{x}}) + \sum_{j=1}^{k}\int_{s_{j-1}}^{s_j} \bar{\mathbf{b}} \cdot \nabla u(\hat{\mathbf{x}} + t\hat{\mathbf{b}})\, dt + [\![u(\hat{\mathbf{x}}_j)]\!] + \int_{s_k}^{s} \bar{\mathbf{b}} \cdot \nabla u(\hat{\mathbf{x}} + t\hat{\mathbf{b}})\, dt, \tag{4.71}$$

taking absolute values, extending the range of integration, and then squaring both sides, we arrive at

$$\left|u(\hat{\mathbf{x}} + s\hat{\mathbf{b}})\right|^2 \leq \left(|u(\hat{\mathbf{x}})| + \sum_{j=1}^{m}\int_{s_{j-1}}^{s_j} \left|\bar{\mathbf{b}} \cdot \nabla u(\hat{\mathbf{x}} + t\hat{\mathbf{b}})\right| dt + \sum_{j=1}^{m-1} [\![u(\hat{\mathbf{x}}_j)]\!]\right)^2. \tag{4.72}$$

Using the inequality

$$\left(\sum_{j=1}^{M} a_j\right)^2 \leq M\sum_{j=1}^{M} a_j^2, \tag{4.73}$$

equation (4.68), and Jensen's inequality, we obtain

$$\begin{aligned}\left|u(\hat{\mathbf{x}}+s\hat{\mathbf{b}})\right|^2 &\leq 3\left\{|u(\hat{\mathbf{x}})|^2+\left(\sum_{j=1}^{m}\int_{s_{j-1}}^{s_j}\left|\bar{\mathbf{b}}\cdot\nabla u(\hat{\mathbf{x}}+t\hat{\mathbf{b}})\right|\,dt\right)^2+\left(\sum_{j=1}^{m-1}[\![u(\hat{\mathbf{x}}_j)]\!]\right)^2\right\}\\ &\leq 3\left\{|u(\hat{\mathbf{x}})|^2+m\sum_{j=1}^{m}\tilde{h}\int_{s_{j-1}}^{s_j}\left|\bar{\mathbf{b}}\cdot\nabla u(\hat{\mathbf{x}}+t\hat{\mathbf{b}})\right|^2\,dt+m\sum_{j=1}^{m-1}[\![u(\hat{\mathbf{x}}_j)]\!]^2\right\}.\end{aligned} \tag{4.74}$$

Using the fact that $m\tilde{h}\leq CD$, where $D=\operatorname{diam}(\Omega)$, and integrating over $\int_0^{s_m}dt$ we have

$$\begin{aligned}\sum_{j=1}^{m}\int_{s_{j-1}}^{s_j}\left|u(\hat{\mathbf{x}}+t\hat{\mathbf{b}})\right|^2\,dt \leq 3\Big\{&\ell(\hat{\mathbf{x}})\,|u(\hat{\mathbf{x}})|^2+CD\ell(\hat{\mathbf{x}})\sum_{j=1}^{m}\int_{s_{j-1}}^{s_j}\left|\bar{\mathbf{b}}\cdot\nabla u(\hat{\mathbf{x}}+t\hat{\mathbf{b}})\right|^2\,dt\\ &+m\ell(\hat{\mathbf{x}})\sum_{j=1}^{m-1}[\![u(\hat{\mathbf{x}}_j)]\!]^2\Big\}.\end{aligned} \tag{4.75}$$

We now integrate according to $\int_{\Gamma_I}\cdot|\bar{\mathbf{b}}\cdot\mathbf{n}|\,d\sigma$ to get

$$\begin{aligned}\int_{\Gamma_I}\sum_{j=1}^{m}\int_{s_{j-1}}^{s_j}\left|u(\hat{\mathbf{x}}+t\hat{\mathbf{b}})\right|^2\,dt\left|\bar{\mathbf{b}}\cdot\mathbf{n}\right|\,d\sigma \leq 3\Big\{&\int_{\Gamma_I}\ell(\hat{\mathbf{x}})(u(\hat{\mathbf{x}}))^2\left|\bar{\mathbf{b}}\cdot\mathbf{n}\right|\,d\sigma\\ &+CD^2\int_{\Gamma_I}\sum_{j=1}^{m}\int_{s_{j-1}}^{s_j}\left|\bar{\mathbf{b}}\cdot\nabla u(\hat{\mathbf{x}}+t\hat{\mathbf{b}})\right|^2\,dt\left|\bar{\mathbf{b}}\cdot\mathbf{n}\right|\,d\sigma\\ &+m\int_{\Gamma_I}\ell(\hat{\mathbf{x}})\sum_{j=1}^{m-1}[\![u(\hat{\mathbf{x}}_j)]\!]^2\left|\bar{\mathbf{b}}\cdot\mathbf{n}\right|\,d\sigma\Big\}\\ \leq 3\Big\{&\|u\|_B^2+\frac{CD^2}{\beta_0}\sum_j\|\mathbf{b}\cdot\nabla u\|_{0,\tau_j}^2\\ &+mD\sum_{i,j}\int_{\Gamma_{i,j}}[\![u]\!]^2\left|\mathbf{b}\cdot\mathbf{n}\right|\,ds\Big\}\end{aligned} \tag{4.76}$$

If $\omega\leq c\frac{1}{h}=\mathcal{O}(m)$, then

$$\begin{aligned}\|u\|_{0,\Omega}^2 &\leq C\left\{\|u\|_B^2+\sum_j\|\mathbf{b}\cdot\nabla u\|_{0,\tau_j}^2+\|u\|_{E^h}^2\right\}\\ &= C\|u\|_{\mathcal{G}^h}.\end{aligned} \tag{4.77}$$

For the general case, bound (4.67) now follows using the assumed transformation with bounded Jacobian.

To show that c is not grid independent for $\omega\leq c\frac{1}{h}$, consider the example of a "stair-step function". Let $\Omega=[0,1]\times[0,1]$ and partition $\mathcal{T}^h$ be a uniform tessellation of squares. Let

$\mathbf{b} = (1,0)^T$ and $h = \frac{1}{N}$, where N is the number of elements in each coordinate direction. Define $u(x,y)$ on Ω as

$$u(x,y) = jh \quad \text{for } x \in [(j-1)h,\, jh)\,,\; j = 1,\ldots,N. \tag{4.78}$$

Then

$$\|u\|^2_{0,\Omega} = \mathcal{O}(1) \tag{4.79}$$

and

$$\|u\|^2_{\mathcal{G}^h} = \mathcal{O}(\omega \cdot h). \tag{4.80}$$

So, unless $\omega \geq c\frac{1}{h}$, inequality (4.67) is violated for grid independent c. □

Remark 4.17. *Once the uniform Poincaré inequality is established, Strang's second lemma [13] can be invoked to prove convergence of DLSFEM. In the absence of the uniform Poincaré inequality, one cannot guarantee that convergence in the grid-dependent norm implies finite element convergence, as illustrated by the "stair-step" example described in the proof above.*

Since $V^h \subset S^h$, we can also conclude, for $\hat{u}^h \in V^h$, that

$$\|u - u^h\|_{\mathcal{G}^h} = \inf_{\hat{u}^h \in V^h} \|u - \hat{u}^h\|_{\mathcal{G}^h}. \tag{4.81}$$

Thus, in the $\mathcal{G}^h$-norm, the nonconforming solution is at least as small as the solution from the conforming space. This might lead one to believe that the discretization error in the L^2 norm for the nonconforming solution would be smaller than the L^2 error in the conforming solution. However, our numerical tests show that this is not the case. In fact, using the weight $\omega = \frac{1}{h}$ for non-grid-aligned flow, we show numerically that the convergence rates, for both conforming and nonconforming approximations, appear to be increasing, but to be bounded by $\frac{1}{2}$, in both the L^2-norm and $\mathcal{G}^h$-norm as k, the order of the polynomial, increases.

4.5 Numerical Results

In this section, we present numerical results in support of our theoretical error estimates and conjectures of Sections 4.3 and 4.4, and to demonstrate properties of the least-squares solu-

tion in terms of oscillations and smearing. Convergence rates of the finite element approximation presented here are obtained on sequences of grids ranging from $h = 2^{-4}$ to 2^{-9} in mesh size, depending on the order of the polynomial, k. Let $\|\cdot\|$ denote either the L^2-norm or the functional norm. We assume that

$$\|u - u^h\| \approx ch^\alpha, \tag{4.82}$$

where c is some constant independent of the mesh size. We compute the convergence rate, α, for consecutive grids as

$$\alpha = \frac{\log\left(\frac{\|u^h - u\|}{\|u^{2h} - u\|}\right)}{\log\left(\frac{1}{2}\right)}. \tag{4.83}$$

The convergence rates presented in the following tables are obtained from a comparison of the two finest grid sizes. In each of our test cases, these values reflect the asymptotic convergence rate since the ratios used in computing α are nearly identical to that of the previous grid levels.

Example 1 (Constant Advection.). *Consider (4.1a-4.1b) and let*

$$\Omega = [0,1] \times [0,1], \tag{4.84}$$

$$u(x,y) = \begin{cases} 1, & \text{if } x = 0, \\ 0, & \text{if } y = 0. \end{cases} \tag{4.85}$$

$$\mathbf{b}(\mathbf{x}) = (\cos(\theta), \sin(\theta)), \tag{4.86}$$

where θ is the angle, the flow makes with the first coordinate axis; see Figure 4.2. The inflow boundary defined by (4.2) is $\Gamma_I = (\{0\} \times [0,1]) \cup ([0,1] \times \{0\})$—i.e., the west and south boundaries of the unit square. With $g(x,y)$ discontinuous on the boundary, the exact solution is discontinuous with $u = 1$ above the characteristic emanating from the origin and $u = 0$ below the characteristic. For the tessellation $\mathcal{T}^h$ of Ω, we choose a uniform partition of quadrilaterals and a uniform partition of triangles.

Tables 4.1 and 4.2 show that we achieve consistent convergence rates in the L^2 and $\mathcal{G}^h$ norms for both the quadrilateral and the triangular elements. Furthermore, as the order of the

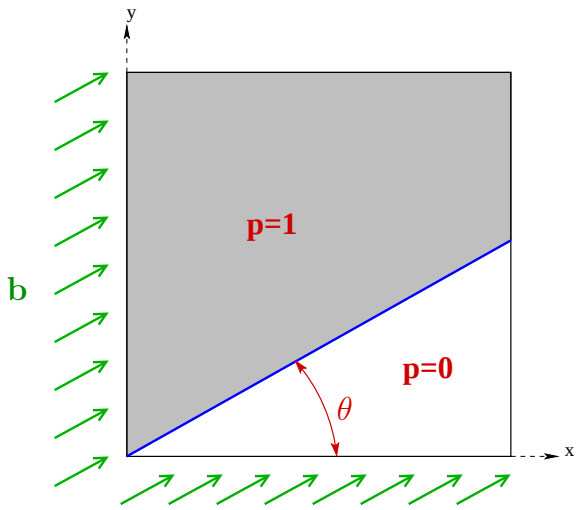


Figure 4.2: Example 1: Constant flow.

polynomials increases, the convergence rates seem to be increasing, but to be bounded by $\frac{1}{2}$. Figure 4.3 shows that, for increasing degree k, the convergence rate (slope) improves slightly, and higher-order methods exhibit smaller error constants per degree of freedom. This suggests that a combination of h and u refinement (where p is the polynomial order) [47] may work well for the kind of discontinuous hyperbolic flows we consider in this dissertation.

The nonconforming space S^h discussed in Section 4.4 offers the ability for the approximation u^h to be discontinuous at the element edges, with the possibility of faster convergence rates for the interior term in the functional. This is indeed the case, as shown in Table 4.3. Since the uniform Poincaré inequality (4.67) does not hold for weaker values of ω, we should expect the $\mathcal{G}^h$-norm to outperform the L^2-norm for weak ω. Moreover, we find that the convergence rates for each term in the functional become less balanced as ω is chosen away from $\frac{1}{h}$.

It is also interesting to study the effect of varying the weight of the boundary functional, e.g., for the continuous LSFEM. Figure 4.4 shows the convergence order for the L^2- and $\mathcal{G}^h$-norms as a function of boundary functional weight. Only for a weight of 1 are the convergence rates in balance, in accordance with our theoretical results in Sections 2 and 3.

The above results were obtained using $\theta = \frac{\pi}{8}$. Table 4.4 reveals that the convergence

rates tended to be relatively independent of the angle θ. Table 4.5 shows that, for **very** small angles—e.g., $\theta \leq .05$—convergence rates are very close to $\frac{1}{2}$ for both the L^2-norm and the $\mathcal{G}^h$-norm using conforming and nonconforming elements. Furthermore, as expected, the convergence rates do not exceed $\frac{1}{2}$ and, for the case of grid-aligned flow, the convergence rates are exactly $\frac{1}{2}$.

	Conforming (4.55)		Nonconforming (4.56) $\omega = \frac{1}{h}$	
k	L^2-norm	$\mathcal{G}$-norm	L^2-norm	$\mathcal{G}^h$-norm
1	.25	.26	.24	.26
2	.34	.33	.32	.33
3	.36	.37	.36	.34
4	.38	.38	.37	.37

Table 4.1: Convergence rates for $\theta = \frac{\pi}{8}$ using quadrilaterals.

	Conforming (4.55)		Nonconforming (4.56) $\omega = \frac{1}{h}$	
k	L^2-norm	$\mathcal{G}$-norm	L^2-norm	$\mathcal{G}^h$-norm
1	.25	.28	.23	.24
2	.33	.32	.33	.33
3	.39	.37	.38	.42

Table 4.2: Convergence rates for $\theta = \frac{\pi}{8}$ using triangles.

	$\omega = \frac{1}{h^2}$		$\omega = \frac{1}{h}$		$\omega = 1$		$\omega = h$	
k	L^2	$\mathcal{G}^h$	L^2	$\mathcal{G}^h$	L^2	$\mathcal{G}^h$	L^2	$\mathcal{G}^h$
1	.25	.28	.24	.26	.25	.47	.24	.57
2	.32	.25	.32	.33	.33	.45	.32	.59
3	.36	.37	.36	.34	.38	.44	.37	.52
4	.38	.39	.37	.37	.40	.46	.40	.52

Table 4.3: Convergence rates for $\theta = \frac{\pi}{8}$ using quadrilaterals and various weights ω.

Smearing of discontinuities is an important consideration for numerical approximation of hyperbolic PDEs. In the exact solution of the model problem, the discontinuity on the inflow boundary Γ_I is advected to the outflow boundary Γ_O without diffusion. However, in a discrete space over a grid that is not flow aligned, we cannot exactly resolve the discontinuity and the finite element solution displays smearing along the characteristic defining the discontinuity.

It is shown in [11, 47] that the LS solution smears the discontinuity substantially more

		Conforming (4.55)		Nonconforming (4.56) $\omega = \frac{1}{h}$	
θ	k	L^2-norm	$\mathcal{G}$-norm	L^2-norm	$\mathcal{G}^h$-norm
$\frac{\pi}{20}$	1	.25	.25	.25	.23
	2	.33	.33	.33	.32
	3	.35	.35	.35	.35
	4	.36	.35	.37	.35
$\frac{\pi}{12}$	1	.25	.26	.25	.25
	2	.32	.33	.32	.32
	3	.36	.36	.35	.36
	4	.39	.37	.39	.37
$\frac{\pi}{8}$	1	.25	.26	.24	.26
	2	.33	.33	.32	.33
	3	.36	.37	.36	.35
	4	.38	.38	.39	.38
$\frac{\pi}{6}$	1	.25	.26	.24	.26
	2	.33	.34	.32	.33
	3	.37	.37	.37	.37
	4	.39	.39	.39	.39
$\frac{\pi}{4}$	1	.24	.26	.23	.26
	2	.32	.34	.32	.32
	3	.36	.38	.34	.36
	4	.38	.40	.36	.40

Table 4.4: Convergence rates for various θ using quadrilaterals.

θ	0	0.01	0.02	0.03	0.04	0.1
L^2-norm	0.500	0.497	0.485	0.466	0.441	0.304
$\mathcal{G}^h$-norm	0.500	0.498	0.492	0.481	0.468	0.389

Table 4.5: Convergence rates (α) for varying θ using nonconforming linear ($k = 1$) elements on triangles.

than the SUPG solution while the Galerkin solution had the least smearing. On the other hand, the Galerkin solution shows the most oscillations, while the SUPG solution shows small oscillations, and the LS solution has almost none. Oscillations are an impediment to accurate local adaptive refinements as they obscure where the adaptivity is most effective. Figure 4.5 confirms these results for our least-squares methods and also indicates that the smearing decreases for higher-order elements. Nearly identical plots were obtained using nonconforming elements, but were omitted for brevity.

Next, we evaluate the oscillations arising in the discrete solution and observe the magnitude of the overshoots. Higher-order elements produce undesirable overshoots and unacceptable

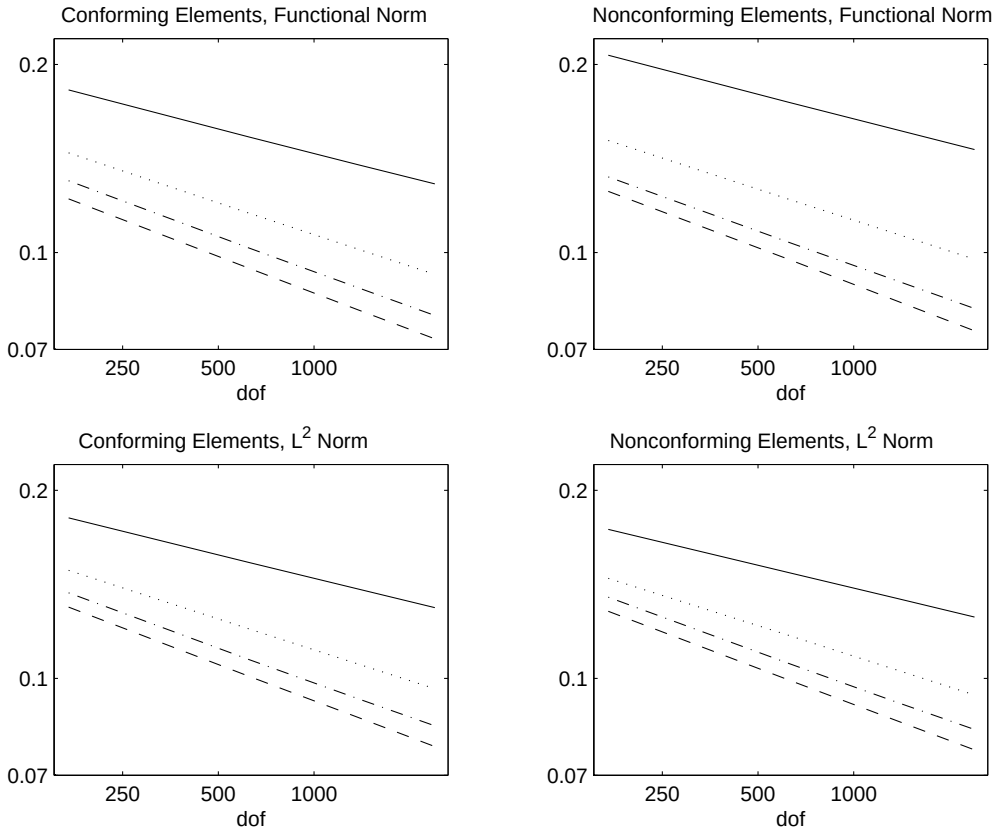


Figure 4.3: Error reductions for (D)LSFEM of various orders, measured per degree of freedom. For degree k increasing from 1 to 4 (solid, dotted, dash-dotted, dashed), the convergence rate (slope) improves slightly, and higher-order methods exhibit smaller error constants per degree of freedom.

oscillations for many finite element methods. However, it was shown in [11, 47] that these negative effects are relatively small in the least-squares formulation. Overshoots for these solutions are displayed in Figure 4.6. Even though the LSFEM solutions are not strictly monotone and overshoots and undershoots do exist, they are contained in a small region near the discontinuity and do not increase in intensity with increasing polynomial order. Nonconforming elements produced nearly identical (less smooth) oscillation and overshoot profiles; see Figure 4.6. In Figures 4.5-4.6, the number of degrees of freedom for the conforming and nonconforming approximations are within 1% of each other.

We also consider an example with a variable flow field.

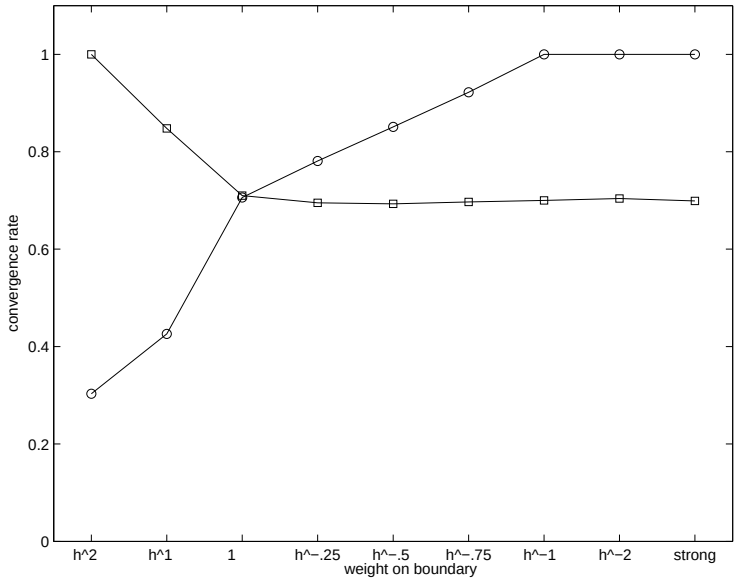


Figure 4.4: Convergence order for the L^2 (squares) and $\mathcal{G}$ (circles) norms as a function of boundary functional weight. For weights stronger than 1, the functional error does not converge well. For weights weaker than 1, the L^2 error does not converge well. Only for a weight equal to 1 are the convergence rates in balance. This agrees with our theoretical results in Sections 2 and 3.

Example 2 (Variable flow field: loop # 1.). *Consider (4.1a-4.1b) and let*

$$\Omega = [0,1] \times [0,1], \tag{4.87}$$

$$u(x,y) = \begin{cases} 0, & \text{if } x = 0, \\ 0, & \text{if } y = 1,\ \frac{1}{2} \le x, \\ 0, & \text{if } y = 0,\ x < \frac{1}{8}, \\ 1, & \text{if } y = 0,\ \frac{1}{8} \le x \ge \frac{3}{8}, \\ 0, & \text{if } y = 0,\ \frac{3}{8} < x. \end{cases} \tag{4.88}$$

$$\mathbf{b}(\mathbf{x}) = (y,\ \frac{1}{2} - x), \tag{4.89}$$

The inflow boundary defined by (4.2) is shown in bold in Figure 4.7. For the tessellation $\mathcal{T}^h$ of Ω, we choose a uniform partition of quadrilaterals.

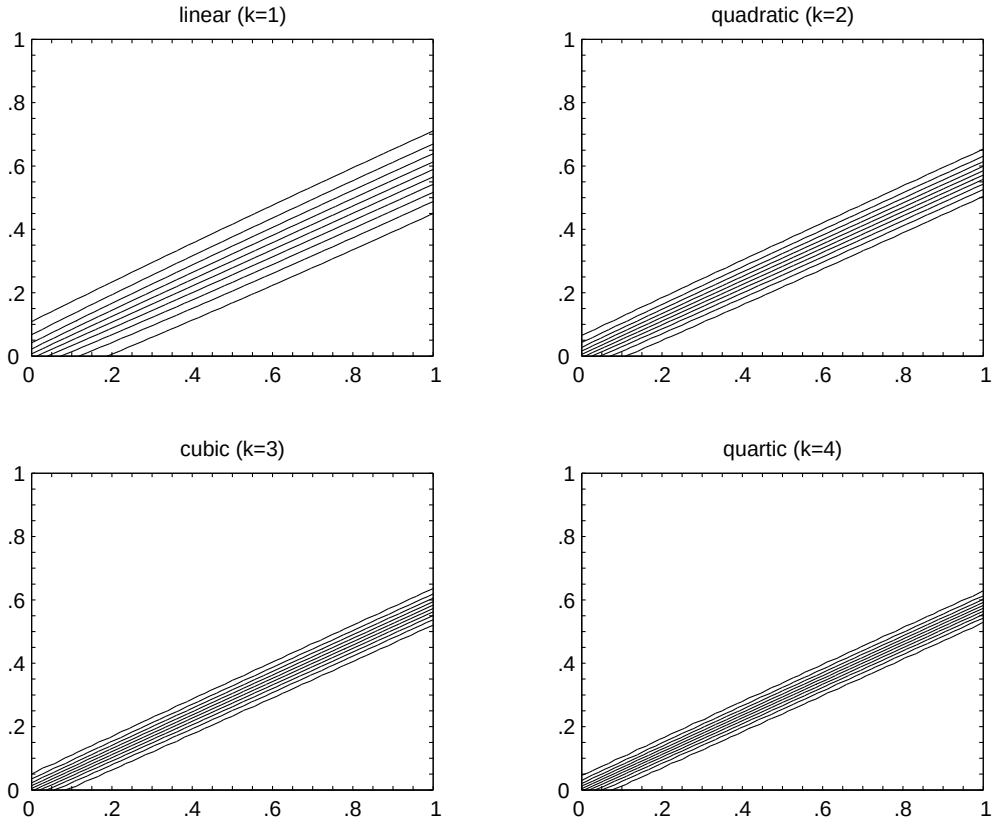


Figure 4.5: Contour plots for various conforming elements. For varying order k, 24, 12, 8, and 6 elements are used in each coordinate direction, respectively.

Example 3 (Variable flow field: loop # 2.). *Consider Example 2 with the following modifications:*

$$u(x,y) = \begin{cases} 0, & \text{if } x = 0, \\ 0, & \text{if } y = 1,\ \frac{1}{2} \le x, \\ 1, & \text{if } y = 0,\ x \le \frac{3}{8}, \\ 0, & \text{if } y = 0,\ \frac{3}{8} < x. \end{cases} \tag{4.90}$$

$$\mathbf{b}(\mathbf{x}) = (\frac{y}{4},\ \frac{1}{2} - x). \tag{4.91}$$

Figure 4.7 illustrates this example.

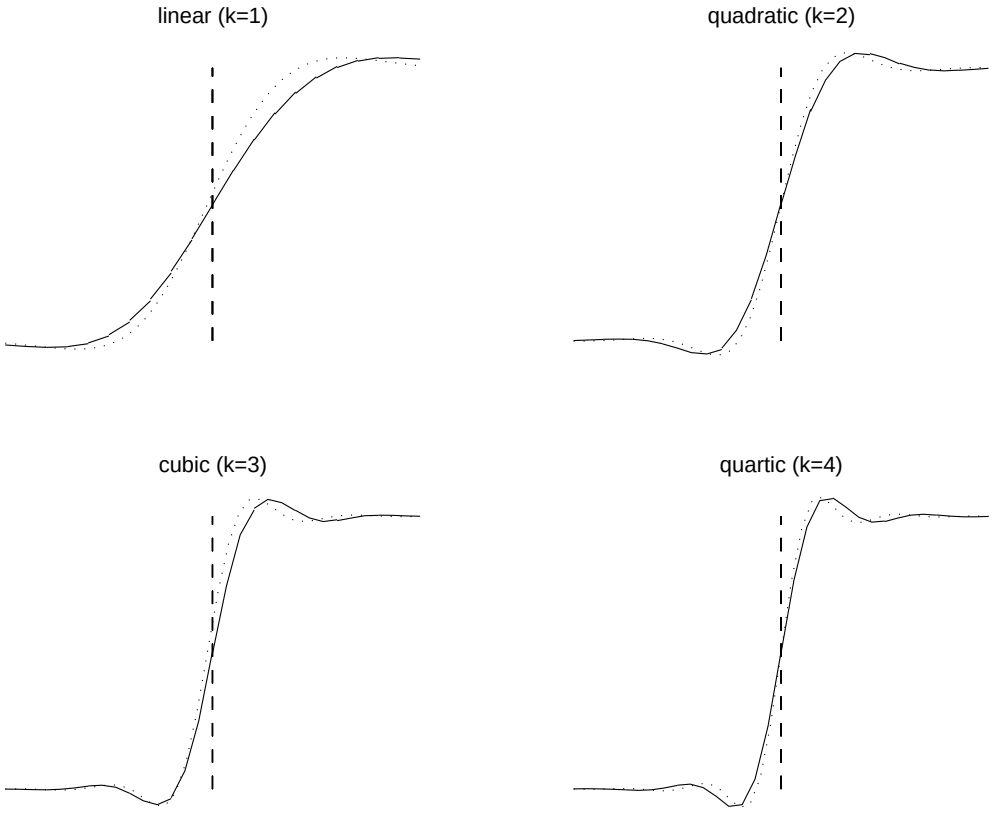


Figure 4.6: Solution profiles for various polynomial degree at slice $x = 0.5$. Dotted line: conforming elements. Solid line: nonconforming elements. Dashed line: location of exact discontinuity.

Example 4 (Variable flow field: loop # 3.). *Consider Example 2 with the following modifications:*

$$u(x,y) = \begin{cases} 0, & \text{if } x = 0, \\ 0, & \text{if } y = 1,\ \frac{1}{2} \le x, \\ 1, & \text{if } y = 0,\ x \le \frac{1}{8}, \\ 0, & \text{if } y = 0,\ \frac{1}{8} < x. \end{cases} \tag{4.92}$$

$$\mathbf{b}(\mathbf{x}) = (\frac{y}{4},\ \frac{1}{2} - x). \tag{4.93}$$

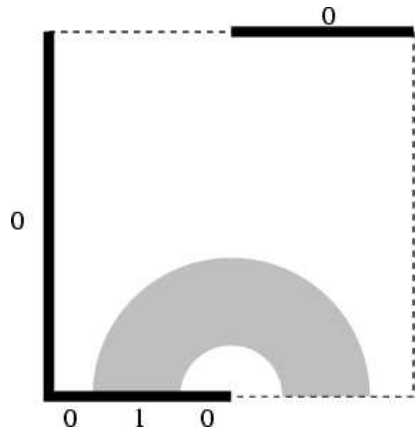
Figure 4.7 illustrates this example.

We limit the convergence analysis for the loop examples (Examples 2, 3, and 4) to bilinear

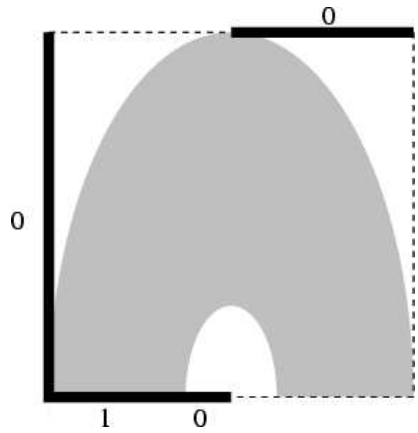
finite elements on a quadrilateral mesh. Figure 4.7 shows the varying loop geometries. Similar convergence results hold for the loop examples. Table 4.6 shows the convergence rates are nearly identical to the rates found for the constant advection test case, Example 1. Moreover, Figures 4.8 and 4.9 verify that all three cases approximate the flow field and that spurious oscillations and overshoots are nearly non-existent. Success for these examples, which include highly varying anisotropies, confirms a level of robustness numerically for the conforming least-squares finite element method developed in this chapter.

	Loop #1	Loop #2	Loop #3
L^2-norm	.23	.23	.23
$\mathcal{G}$-norm	.26	.27	.25

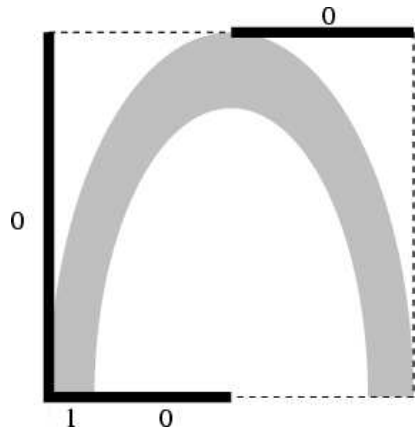
Table 4.6: Convergence rates for Examples 2, 3, and 4



(a) Example 2

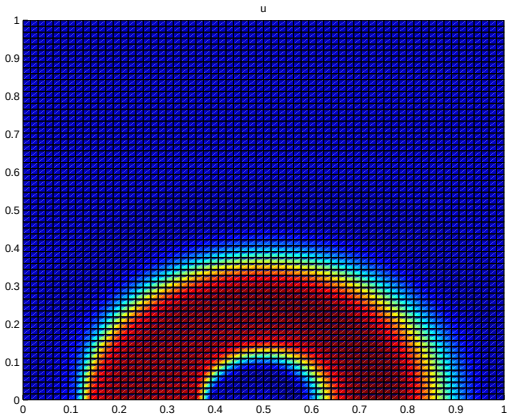


(b) Example 3

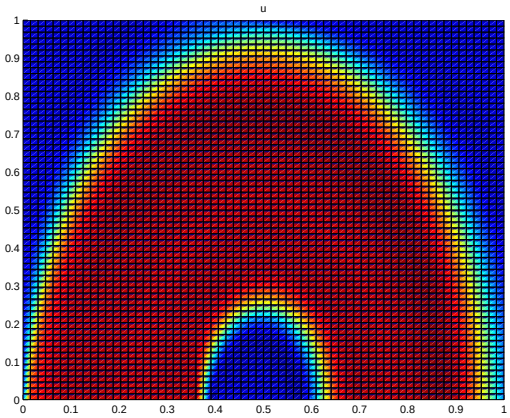


(c) Example 4

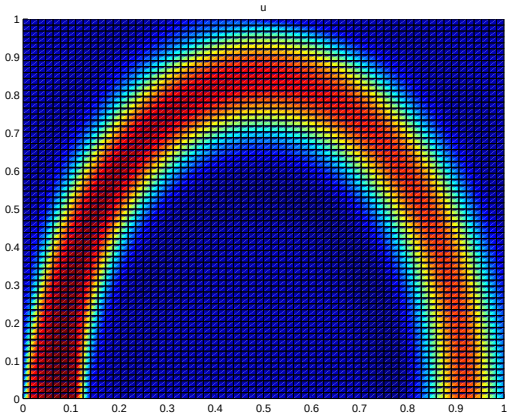
Figure 4.7: Variable flow examples.



(a) Example 2

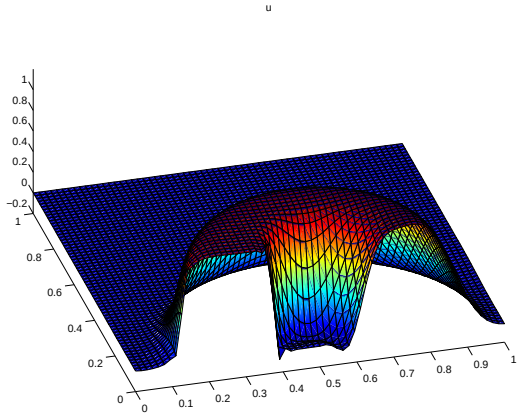


(b) Example 3

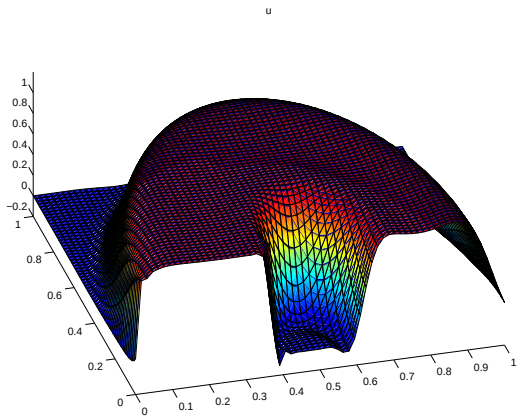


(c) Example 4

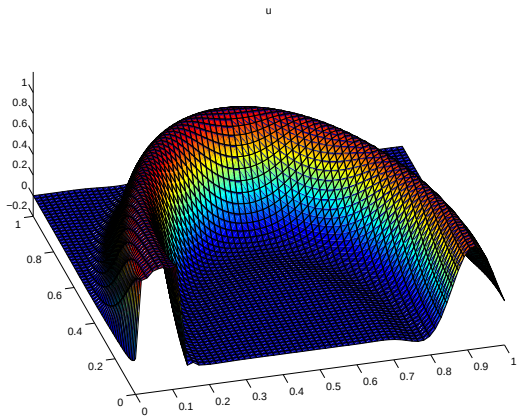
Figure 4.8: 2D view of approximate solution on a 64×64 grid.



(a) Example 2



(b) Example 3



(c) Example 4

Figure 4.9: 3D view of approximate solution on a 64×64 grid.

Chapter 5

Multigrid

This chapter focuses on solution methods for the linear systems arising from the LSFEM developed in Chapter 4. As noted in Chapter 3, the least-squares approach leads to discretizations with symmetric positive definite (SPD) matrices, which provides a natural setting for iterative solution methods. Our focus is on the so-called Ruge-Stüben Algebraic Multigrid (AMG) algorithm [68] because of its ability to handle irregular grids and anisotropic operators. Often, for elliptic-type problems, the least-squares bilinear form can be shown to be H^1 equivalent (see [23] for a complete justification), resulting in not only optimal finite element convergence, but also provably optimal multigrid convergence. It is the goal of this chapter to address the issues of multigrid iterative techniques for problems with suboptimal equivalence.

In Section 5.1, we show some of the difficulties of using standard geometric-based multigrid algorithms for hyperbolic PDEs. The matrix structure is vastly different from a system arising from an isotropic and diffusive PDE. We lose the advantages of an M-matrix and are forced to consider the anisotropic nature of the problem. The result is a matrix with relatively strong connections in one direction. Semicoarsening strategies in a geometric multigrid algorithm are an appropriate technique [79], although they tend to lack robustness when considering variable flow fields, nonlinear problems, or systems of equations, that can result in multiple flow directions on the same computational mesh. The strongest shortcoming of a geometric-based multigrid method, however, is the inability to use standard relaxation methods to adequately smooth oscillatory components for anisotropic problems. We investigate this issue more in

Section 5.1.

In Section 5.2, we consider AMG (cf. [20, 79]). For irregular or unstructured meshes, selection of a coarse grid may be challenging. This is also the case for systems involving highly anisotropic matrix connections. AMG techniques utilize the structure of the discrete operator (matrix), instead of geometric patterns, to select a coarser level. In a sense, the algebraic approach automatically semicoarsens the grid in the appropriate direction. We consider AMG for the linear systems resulting from the least-squares methods we derive in Chapter 4 and relate the formulation to an anisotropic diffusion operator. Algebraic multigrid methods are considered adequate solvers for such an operator [79]. Moreover, AMG is widely used for its robustness on unstructured meshes. Its ability to handle irregular meshes is especially useful for local refinement, which least-squares methods, with their natural error estimator [7], easily support.

Since AMG is intimately coupled with the discrete operator, the choice of discretization naturally plays a large role in the performance of the algorithm. For Stokes's equations, for example, a first-order system can be derived and only with the additional equations can the bilinear form be shown to possess the desired equivalence. Although we do not seek to increase the ellipticity of the hyperbolic problem, we see below that the form of the boundary conditions derived in Chapter 4 do not facilitate a fast multigrid solution. In Section 5.3, we are thus led to amend the equations to improve multigrid convergence, while maintaining desirable solution quality and finite element convergence properties.

In this chapter, we treat simple model problems for which there is a time-like direction that can also be exploited by explicit marching schemes. For optimal global implicit solvers, the number of operations per grid point is bounded, which means that even for time-like problems they may be competitive with explicit marching methods, for which the number of operations per grid point is bounded as well. This is a possibility particularly when local refinement and derefinement are taken into account, both spatially and temporally. Moreover, explicit marching schemes are limited by time step constraints, which can be severe on locally refined grids. Also,

efficient parallel implementations have been developed for AMG solvers [42]. Global implicit solves may be especially competitive for the simulation of problems for which there is no preferred marching direction, e.g., steady flows with rotation or flows of mixed elliptic-hyperbolic type. In this chapter, we make an important first step by investigating whether optimal AMG solvers can be constructed for simple time-like hyperbolic problems.

5.1 Geometric Considerations

In this section, we study geometric multigrid methods for conforming LSFEM discretizations of the model hyperbolic PDE (4.1) presented in Section 4.3. In this setting, coarse grids are chosen based on the geometry of the tessellation. Geometric multigrid is often the starting point for multilevel development because of its low computational setup cost, compared to an algebraic-based multigrid method.

First, we establish notation to be used in this section. A basic understanding of multigrid methods and principles is assumed. Let $Au = f$ be the system of discrete equations we wish to solve. This is assumed to be the matrix equation that arises from the LSFEM for hyperbolic PDEs developed in the previous chapter. Let $u^{(k)}$ be an approximation to u. We now define a multilevel iterative process in the form

$$u^{(k+1)} = u^{(k)} + \text{correction term}, \tag{5.1}$$

where

$$u^{(k)} \to u \text{ as } k \to \infty. \tag{5.2}$$

Let $e^{(k)} = u - u^{(k)}$ denote the error and let $r^{(k)} = f - Au^{(k)}$ denote the residual. We then have

$$u = u^{(k)} + e^{(k)} \tag{5.3}$$

$$= u^{(k)} + A^{-1}r^{(k)}. \tag{5.4}$$

The goal, however, is to efficiently solve the system, so we seek an approximation $\hat{A}$ to A that

is more easily invertible. This results in the iteration

$$u^{(k+1)} = u^{(k)} + \hat{A}^{-1} r^{(k)}. \tag{5.5}$$

Subtracting u from both sides results in

$$e^{(k+1)} = (I - \hat{A}^{-1} A) e^{(k)} \tag{5.6}$$

$$= M e^{(k)}, \tag{5.7}$$

where M is called the error propagation matrix. Convergence is determined by the spectral radius of M, denoted by $\rho = \rho(M)$, since

$$e^{(k+1)} = M^{(k+1)} e^{(0)}. \tag{5.8}$$

Using Local Mode Analysis (LMA), we approximate the convergence factor, ρ, when M is the two-grid multigrid error propagation matrix, which we now describe. First, we define the following multigrid components.

$$A^h = \text{discrete PDE operator on grid } \Omega_h,$$

$$I_h^H = \text{restriction operator from grid } \Omega_h \text{ to grid } \Omega_H,$$

$$I_H^h = \text{interpolation operator from grid } \Omega_H \text{ to grid } \Omega_h.$$

We denote by $S_{\nu_m}^h$ the relaxation error propagation operator applied ν_m times. For initial guess v^h, the two-grid multigrid cycle

$$v^h \leftarrow MG^h(v^h, f^h) \tag{5.9}$$

is given is given as follows.

(1) Presmooth v^h on $A^h u^h = f^h$ ν_1 times. This results in the error propagation

$$e^h \leftarrow S_{\nu_1}^h e^h$$

(2) Restrict the residual and compute the coarse grid solution:

$$r^H \leftarrow I_h^H (f^h - A^h v^h)$$
$$e^H \leftarrow A^{H^{-1}} r^H$$

(3) Compute the corrected approximation with the interpolated coarse grid error:

$$v^h \leftarrow v^h + I_H^h e^H$$

(4) Postsmooth v^h with $A^h u^h = f^h$ ν_2 times. This results in the error propagation

$$e^h \leftarrow S_{\nu_2}^h e^h$$

Steps (2) and (3) of the process result in the two-level coarse-grid error propagation operator

$$K_h^H = I^h - I_H^h A^{H^{-1}} I_h^H A^h, \tag{5.10}$$

where I^h denotes the identity on grid Ω_h. Combined with pre- and post-smoothing steps (1) and (4), the two-level multigrid error propagation operator becomes

$$M_h^H = S_{\nu_1}^h K_h^H S_{\nu_2}^h, \tag{5.11}$$

which is used in the subsequent analysis. The extension to a recursive algorithm follows similarly and we introduce the μ-cycle, for generality. Define

$$v^h \leftarrow \mu MG^h(v^h, f^h) \tag{5.12}$$

to be the following process:

(1) Presmooth v^h with $A^h u^h = f^h$, ν_1 times.

(2) If Ω_h is the coarsest grid, go to step (4). Otherwise, restrict the residual and compute the coarse grid solution with μ recursive calls:

$$r^H \leftarrow I_h^H (f^h - A^h v),$$
$$e^H \leftarrow \mu MG^H(\mathbf{0}^H, r^H), \mu \text{ times}$$

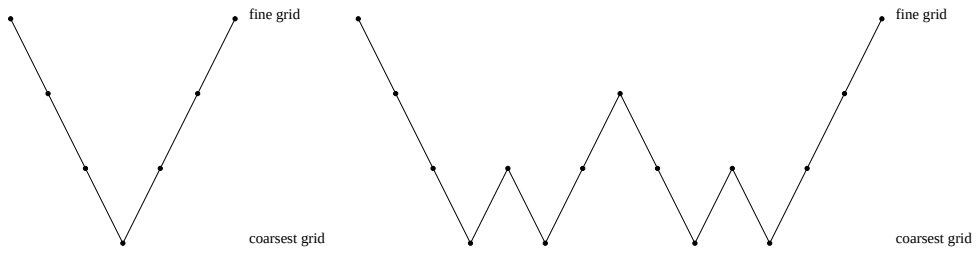
(3) Compute the corrected approximation with the interpolated coarse grid error:

$$v^h \leftarrow v^h + I_H^h e^H$$

(4) Postsmooth v^h with $A^h u^h = f^h$, ν_2 times. This results in the error propagation

$$e^h \leftarrow S_{\nu_2}^h e^h$$

Figure 5.1 illustrates (5.12) for $\mu = 1$ (V-cycle) and $\mu = 2$ (W-cycle) .



(a) V-cycle (b) W-cycle

Figure 5.1: Typical multigrid cycle stencils. Each • represents a level for which work is being done (relaxation sweeps).

5.1.1 Local Mode Analysis

We now use Local Mode Analysis (a.k.a. Local Fourier Analysis, LMA, or LFA) to analyze multigrid processes for solving linear hyperbolic PDEs. The goal of LMA is to predict the smoothing factor for S_h, the smoothing operator, and to predict the two-grid convergence factor for M_h^H, the two-grid error propagation operator. Our model problem is again the linear hyperbolic partial differential equation

$$\mathbf{b} \cdot \nabla u = f \quad \text{in } \Omega, \tag{5.13}$$

$$u = g \quad \text{on } \Gamma_I. \tag{5.14}$$

It is important to note the local aspects of LMA. Although the problem (5.13) is tightly coupled with its boundary data (5.14), mode analysis assumes the problem to be defined on infinite grids. The purpose of this analysis is to obtain a general idea of the performance for a particular smoothing operator, assuming that the boundary is properly treated [79]. Thus, this analysis is a powerful tool, but should be used carefully when assessing a particular multigrid algorithm on a bounded domain.

Local Mode Analysis is based on the grid functions

$$\phi(\boldsymbol{\theta}, \mathbf{x}) = \exp^{i\boldsymbol{\theta} \cdot \frac{\mathbf{x}}{\mathbf{h}}}, \tag{5.15}$$

where

$$\mathbf{G_h} := \{\mathbf{x} = \boldsymbol{\kappa}\mathbf{h} \mid \boldsymbol{\kappa} \in \mathbb{Z}^{\mathbf{2}}\} \tag{5.16}$$

is the (infinite) fine grid. The coarse grid is defined similarly. We use the definition of high and low frequency components (see Figure 5.2) given in [79] that are defined simply as

$$\phi \text{ low frequency component} \Leftrightarrow \boldsymbol{\theta} \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right)^2, \tag{5.17}$$

$$\phi \text{ high frequency component} \Leftrightarrow \boldsymbol{\theta} \in [-\pi, \pi)^2 \setminus \left[-\frac{\pi}{2}, \frac{\pi}{2}\right)^2. \tag{5.18}$$

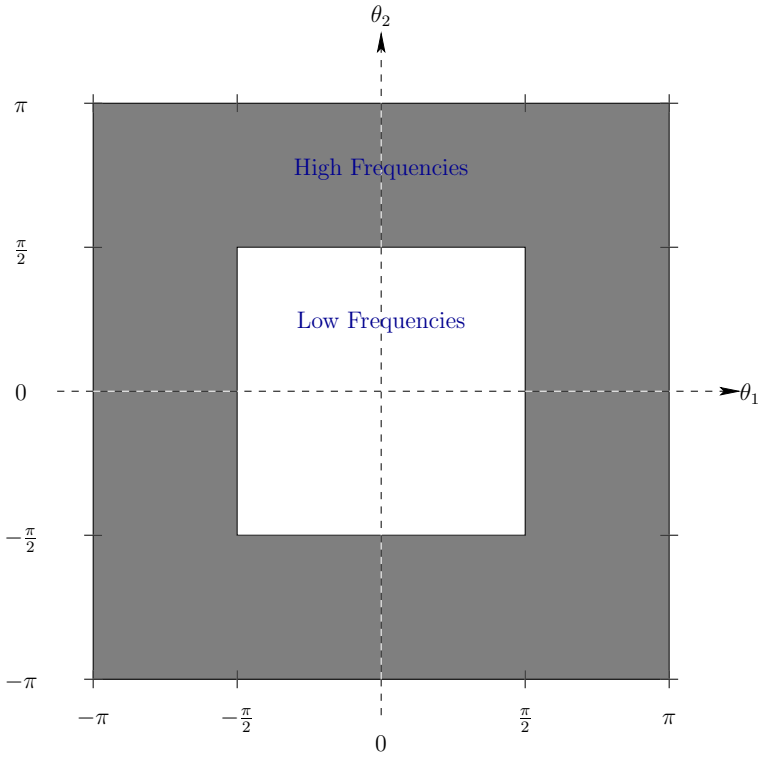


Figure 5.2: Graphical view of low and high frequencies.

Let $A^h u^h = f^h$ be the discretization of (5.13) on grid Ω_h and assume that relaxation can be written in terms of a splitting [79]

$$A^h = A^h_+ + A^h_-. \tag{5.19}$$

That is, the method of smoothing can be written **locally**:

$$A_+^h u^{h,(\text{new})} + A_-^h u^{h,(\text{old})} = f^h. \tag{5.20}$$

The splitting property simplifies the LMA. Standard lexicographic Gauss-Seidel and weighted Jacobi relaxation techniques both adhere to this assumption. For example, the LSFEM discretization of (5.13) with $\mathbf{b} = (\cos(\frac{\pi}{4}), \sin(\frac{\pi}{4}))$ yields the stencil

$$A^h \triangleq \begin{bmatrix} \frac{1}{12} & -\frac{1}{6} & -\frac{5}{12} \\ -\frac{1}{6} & \frac{4}{3} & -\frac{1}{6} \\ -\frac{5}{12} & -\frac{1}{6} & \frac{1}{12} \end{bmatrix}. \tag{5.21}$$

For lexicographic Gauss-Seidel applied to A^h, the splitting is written as

$$A_+^h \triangleq \begin{bmatrix} 0 & 0 & 0 \\ -\frac{1}{6} & \frac{4}{3} & 0 \\ -\frac{5}{12} & -\frac{1}{6} & \frac{1}{12} \end{bmatrix}, \quad A^h \triangleq \begin{bmatrix} \frac{1}{12} & -\frac{1}{6} & -\frac{5}{12} \\ 0 & 0 & -\frac{1}{6} \\ 0 & 0 & 0 \end{bmatrix}. \tag{5.22}$$

To understand the relation between the grid functions $\phi(\boldsymbol{\theta}, \mathbf{x})$ and the operators S^h, A^h, K_h^H, and M_h^H, we introduce formal eigenfunctions and formal eigenvalues. Since the discrete operator A^h can be written as a difference stencil, $\phi(\boldsymbol{\theta}, \mathbf{x})$ is a formal eigenfunction of A^h. That is,

$$A^h \phi(\boldsymbol{\theta}, \mathbf{x}) = \tilde{A}^h(\boldsymbol{\theta}) \phi(\boldsymbol{\theta}, \mathbf{x}), \tag{5.23}$$

where $\tilde{A}^h(\boldsymbol{\theta})$ is the formal eigenvalue of A^h and can be written in terms of the stencil entries and the frequencies, $\boldsymbol{\theta}$. In particular, we have

$$\tilde{A}^h(\boldsymbol{\theta}) = \sum_{i,j} s_{i,j} \phi(\boldsymbol{\theta}, (x_i, x_j)), \tag{5.24}$$

where $s_{i,j}$ are the coefficients in the stencil. We omit further details as they are not important for the comprehension of this analysis. A full explanation can be found in Lemma 4.2.1 of [79]. We define formal eigenvalues for A_+^h and A_-^h in a similar fashion. The importance of the formal eigenvalue can be seen if we write the smoothing operator in terms of the splitting. That is,

define $e^{h,(\text{new})} = u^h - e^{h,(\text{new})}$ and $e^{h,(\text{old})} = u^h - e^{h,(\text{old})}$, and notice that

$$A_+^h e^{h,(\text{new})} + A_-^h e^{h,(\text{old})} = 0 \tag{5.25}$$

and

$$e^{h,(\text{new})} = S^h e^{h,(\text{old})}. \tag{5.26}$$

If we consider e^h to be an eigenfunction ϕ, then (5.25) can be written as

$$e^{h,(\text{new})} = -\frac{\tilde{A}_-^h(\boldsymbol{\theta})}{\tilde{A}_+^h(\boldsymbol{\theta})} e^{h,(\text{old})}, \tag{5.27}$$

where we assume $\tilde{A}_+^h(\boldsymbol{\theta}) \neq 0$. The quantity

$$\tilde{S}^h(\boldsymbol{\theta}) = -\frac{\tilde{A}_-^h(\boldsymbol{\theta})}{\tilde{A}_+^h(\boldsymbol{\theta})} \tag{5.28}$$

is the formal eigenvalue of the smoothing operator and is termed the amplification factor. Note: the terms are labeled as "formal" in the sense that boundary conditions are not taken into account.

The eigenvalue $\tilde{S}^h(\boldsymbol{\theta})$ is thus a measure of the performance of the relaxation sweep on a frequency $\boldsymbol{\theta}$. If the value is small, then the reduction of the error is large. In the multigrid process, we are only concerned that high frequencies are damped by relaxation, and we define the smoothing factor as the following (c.f. [20, 79]):

$$\mu := \sup_{\boldsymbol{\theta} \in \Theta^{\text{high}}} \|\tilde{S}_h(\boldsymbol{\theta})\|. \tag{5.29}$$

We now apply a similar analysis to the two-grid operator M_h^H. Consider the low frequencies $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0,0)} := (\theta_1, \theta_2)$ and their corresponding harmonic frequencies:

$$\boldsymbol{\theta}^{(1,0)} := (\bar{\theta}_1, \theta_2), \tag{5.30a}$$

$$\boldsymbol{\theta}^{(0,1)} := (\theta_1, \bar{\theta}_2), \tag{5.30b}$$

$$\boldsymbol{\theta}^{(1,1)} := (\bar{\theta}_1, \bar{\theta}_2), \tag{5.30c}$$

where

$$\bar{\theta}_i := \begin{cases} \theta_i + \pi, & \text{if } \theta_i < 0, \\ \theta_i - \pi, & \text{if } \theta_i \geq 0. \end{cases} \tag{5.31}$$

Define $\bar{A}^h$ to be the $(4 \times 4)-$matrix

$$\begin{pmatrix} \tilde{A}^h(\boldsymbol{\theta}^{(0,0)}) & & & \\ & \tilde{A}^h(\boldsymbol{\theta}^{(1,0)}) & & \\ & & \tilde{A}^h(\boldsymbol{\theta}^{(0,1)}) & \\ & & & \tilde{A}^h(\boldsymbol{\theta}^{(1,1)}) \end{pmatrix}. \tag{5.32}$$

The $(4\times4)-$matrix representations for K_h^H and S_h are similarly defined. It is shown in Theorem 4.4.1 of [79] that if the space of harmonics defined by

$$E_h^\theta := \text{span}\{\phi(\boldsymbol{\theta}^{(i,j)}, \mathbf{x}) \mid i, j = 0, 1\}, \tag{5.33}$$

is invariant under the smoothing operator S^h, then we can define a $(4 \times 4)-$matrix $\bar{M}_h^H$ with respect to E_h^θ. This matrix representation of M_h^H can be written as

$$\bar{M}_h^H(\boldsymbol{\theta}) = \bar{S}_h(\boldsymbol{\theta})^{\nu_1} \bar{K}_h^H(\boldsymbol{\theta}) \bar{S}_h(\boldsymbol{\theta})^{\nu_2}. \tag{5.34}$$

Since relaxation is responsible for the high-frequency components, we are most interested in the ability of the two-grid operator to handle low-frequency components. Define the asymptotic convergence factor as

$$\rho = \sup_{\substack{\boldsymbol{\theta} \in \boldsymbol{\Theta}^{\text{low}} \\ \tilde{L}_h(\boldsymbol{\theta}) \neq 0, \tilde{L}_H(\boldsymbol{\theta}) \neq 0}} \rho(\bar{M_h^h}(\boldsymbol{\theta})). \tag{5.35}$$

The smoothing and asymptotic convergence factors, μ and ρ, can be written in terms of the stencil coefficients and were computed using a software package written by Jan Metzger [43]. The estimates indicate that we may encounter problems with more than a two-grid multigrid algorithm. The test case is the constant advection problem (5.13-5.14) with $\mathbf{b} = (\cos(\theta_{\mathbf{b}}), \sin(\theta_{\mathbf{b}}))$, where $\theta_{\mathbf{b}}$ is stated with the results. Boundary data g is chosen to be discontinuous, although this algebraic analysis holds for smooth solutions as well. We present LMA results using standard lexicographic Gauss-Seidel relaxation, bilinear interpolation, and full-weighting restriction. Other strategies, such as weighted-Jacobi relaxation and semicoarsening intergrid transfer operators, resulted in similar or degraded estimates for μ and ρ.

Table 5.1 shows results for the case $\nu_1 = \nu_2 = 1$. More relaxation sweeps do not improve these results and the performance degrades considerably as the angle of anisotropy becomes more aligned with the computational mesh. The estimated two-grid convergence factors are even poorer than just smoothing alone, indicating problems on the coarse grid.

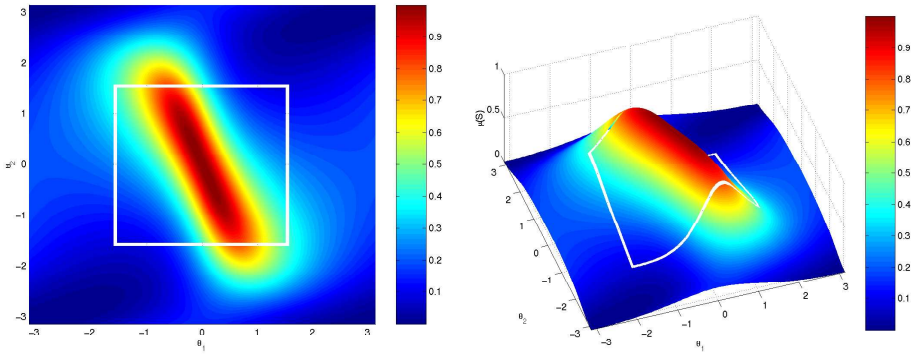
$\theta_{\mathbf{b}}$	μ	ρ
$\frac{\pi}{4}$	.45	.81
$\frac{\pi}{8}$	.80	.86

Table 5.1: Smoothing factors, μ, and two-grid convergence factors, ρ, using 1 pre- and 1 post-smoothing sweep

The effect of smoothing and two-grid correction are perhaps more apparent from a graphical view. Figures 5.3 and 5.4 show the smoothing and two-grid correction performance spectrally for the example with $\theta_{\mathbf{b}} = \frac{\pi}{8}$. The x-y plane is the frequency domain described by Figure 5.2, while the z direction is μ and ρ in the respective figures. These plots reveal difficulties in the cross-stream direction. High and low frequencies are separated by a white box in Figure 5.3. The poor convergence is due to the inadequate smoothing in the high frequency domain. High frequencies that are not effectively damped by relaxation impair convergence because they cannot be approximated well by the coarse grid and thus cannot be eliminated by coarse-grid correction. This is confirmed in Figure 5.4, where the two-grid cycle is shown to fail at precisely the same spacial location as the poorly smoothed high frequency errors.

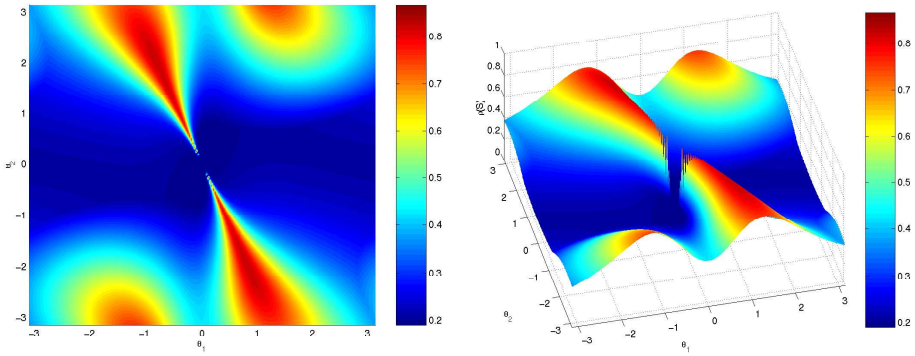
5.1.2 Numerical Study

We first show that using a full set of levels in a standard multigrid scheme does in fact lead to poor convergence factors as our Local Mode Analysis predicts. These results motivate our use of an alternate multigrid approach. The inability of standard relaxation to dampen high frequency error components in the near null-space of the operator $\mathbf{b} \cdot \nabla$ exposes limitations of standard geometric multigrid methods for handling convective flow. The near-null space components and poor coarse-grid approximations clearly play a central role in the inefficiencies



(a) Top view (b) 3-D

Figure 5.3: Smoothing profile for $\mathbf{b} = (\cos(\frac{\pi}{8}), \sin(\frac{\pi}{8}))$ using lexicographic Gauss-Seidel and 2 smoothing passes.



(a) Top view (b) 3-D

Figure 5.4: Two-grid profile for $\mathbf{b} = (\cos(\frac{\pi}{8}), \sin(\frac{\pi}{8}))$ using lexicographic Gauss-Seidel, bilinear interpolation, full-weighting restriction, and 1 pre- and 1 post-smoothing sweep

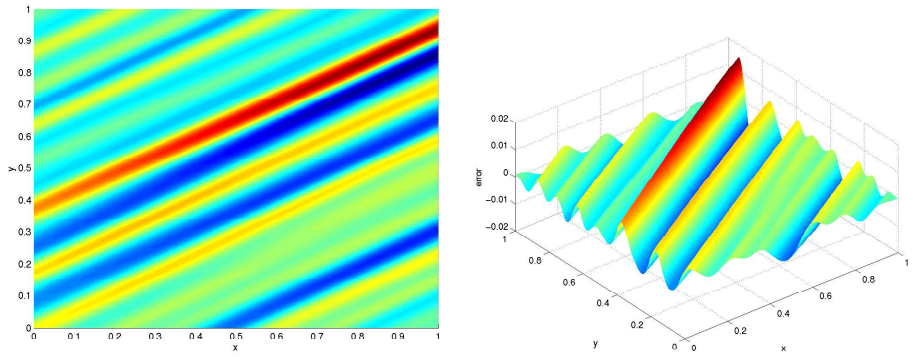
of the solver.

Table 5.2 shows the results of V- and W-cycles (5.12) with lexicographic Gauss-Seidel relaxation, bilinear interpolation, and full-weighting restriction operators, for various pre- and post-smoothing steps. Results for weak and strong implementation of the boundary conditions are presented for completeness and are discussed in the next section. Note that the average

convergence factors approach 1 as h tends to 0, regardless of the number of pre- and post-smoothing steps and the number of visits to the coarse-grids. The average convergence factor ρ is computed as the average of the ratios of residuals between successive cycles.

Increasing the number of smoothing sweeps consistently improves convergence, as does increasing μ, however, performance also degrades in each case as the grid is refined. The results also confirm the LMA predictions with respect to convection angle. Convergence is poorer for the more grid aligned angle, $\theta_{\mathbf{b}} = \frac{\pi}{8}$. Also, since the LMA does not take into account boundary conditions, we also test strong implementation at the boundary, which better reflects the analysis. Although weak treatment of the boundary conditions is required for the discretization (see Section 4.3), strong implementation results in improved performance, as shown in Table 5.2. This difference becomes more pronounced in an algebraic-based multigrid setting and is discussed in more detail in Section 5.2.

Figure 5.5 shows the error in the approximation after 50 V(1,1)-cycles on a 128×128 rectangular mesh using standard bilinear finite elements discussed in Chapter 4. These plots verify that the geometric algorithm is not able to adequately resolve high frequency near-null-space error components.



(a) Top view (b) 3-D

Figure 5.5: Error for $\mathbf{b} = (\cos(\frac{\pi}{6}), \sin(\frac{\pi}{6}))$, zero boundary data, and random initial guess (weak implementation of boundary conditions)

			V(1,1)		V(2,2)		W(1,1)		W(2,2)	
		N	cyc.	$\bar{\rho}$	cyc.	$\bar{\rho}$	cyc.	$\bar{\rho}$	cyc.	$\bar{\rho}$
$\frac{\pi}{4}$	Weak B.C.	16	74	.82	43	.72	51	.73	34	.65
		32	127	.89	78	.84	67	.80	42	.71
		64	232	.94	122	.90	84	.84	54	.77
		128	306	.95	176	.93	109	.87	63	.81
	Strong B.C.	16	46	.71	26	.57	37	.65	26	.57
		32	96	.86	51	.76	54	.75	35	.66
		64	156	.91	91	.86	68	.80	42	.71
		128	269	.95	166	.92	92	.92	59	.79
$\frac{\pi}{8}$	Weak B.C.	16	109	.87	59	.78	65	.78	43	.71
		32	182	.92	105	.87	95	.85	61	.79
		64	307	.95	173	.92	127	.89	71	.82
		128	447	.97	251	.95	147	.90	91	.86
	Strong B.C.	16	68	.79	42	.70	53	.74	35	.64
		32	129	.89	77	.83	77	.81	46	.72
		64	239	.94	125	.89	94	.85	65	.80
		128	319	.95	200	.93	136	.89	80	.84

Table 5.2: Geometric multigrid: The number of cycles needed to reach a relative residual of 1e-8 and the average convergence factor, ρ.

Since the direction of dominance is known from the given flow field $\mathbf{b}(\mathbf{x})$, a natural strategy to improve standard multigrid is simply to reorder the smoothing to adopt a downstream character. Standard lexicographic Gauss-Seidel does not take into account this directionality and it is important to know whether a downstream ordering can improve the smoothing process. Figure 5.6 illustrates a domain that is partitioned into "waves" normal to the flow field. The idea is to relax first on the wave nearest to Γ_I (in any ordering), then to the next wave, and so on, until the last wave at Γ_O has been relaxed.

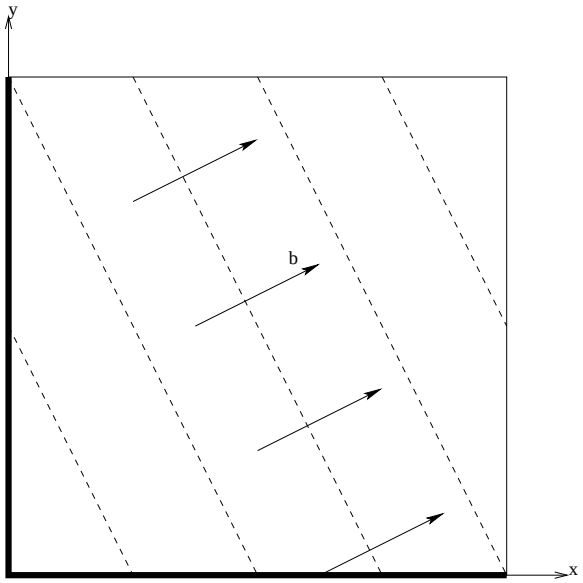


Figure 5.6: Relaxation in waves: typical flow field with partitioned domain.

We implement this algorithm for various numbers of waves and find nearly identical results when compared with standard relaxation strategies. This indicates that a more sophisticated smoothing approach (or coarsening approach) is needed to properly account for the neighbors that influence the downstream nodes. If we consider the $\mathbf{b}\cdot\nabla$ operator in the continuum, we find that solutions are differentially smooth in the direction of flow, while smoothness in the cross-stream direction is dictated globally by what is imposed at the inflow boundary. This is inherent with purely convective problems and we can expect similar characteristics at the discrete level. The principal difficulty with general $\mathbf{b}$ is that neither relaxation nor coarsening can be fully

faithful to the streamline direction, especially for systems of equations and nonlinear problems where flow fields may contain discontinuities. Developing semi-coarsening strategies based on specialized intergrid transfer operators and relaxation techniques that recognize the underlying flow field has been an active area of research [17, 33, 80]. However, robustness and generality of these approaches still hinder the geometric-based multigrid method for convective flow. We therefore turn to an **algebraic** based method, which attempts to identify the anisotropic nature of the discrete problem by considering the "strength" of the matrix entries. One goal of this dissertation is to initiate a study of hyperbolic PDEs in this multilevel context.

5.2 Algebraic Multigrid Considerations

In this section, we focus on iterative solvers in a multilevel framework, employing the techniques of the Ruge-Stüben version of AMG [68]. As we see below, AMG does not yet achieve optimal convergence factors in the sense that they are bounded and independent of h, but the growth in these factors is relatively slow as h tends to zero.

First, consider the problem in the context of the limit case of an anisotropic diffusion operator. More specifically, consider the PDE

$$\mathcal{L}_A p = \tilde{f} \quad \text{in } \Omega, \tag{5.36a}$$

$$p = 0 \quad \text{on } \Gamma_I, \tag{5.36b}$$

$$\mathbf{n} \cdot \nabla p = 0 \quad \text{on } \Gamma \setminus \Gamma_I, \tag{5.36c}$$

where $\tilde{f} \in L^2(\Omega)$ and

$$\mathcal{L}_A p := \nabla \cdot (A \nabla p). \tag{5.37}$$

$A = \mathbf{b}\mathbf{b}^T + \varepsilon \mathbf{d}\mathbf{d}^T$ for some $\varepsilon \in (0, 1)$ with $\mathbf{b} \cdot \mathbf{b} = 1$, $\mathbf{d} \cdot \mathbf{d} = 1$, and $\mathbf{b} \cdot \mathbf{d} = 0$. Note that $A = I$ when $\varepsilon = 1$ and that, for small $\varepsilon > 0$, the operator $\mathcal{L}_A$ is an anisotropic diffusion operator because of the strong connection in the direction $\mathbf{b}$. Promising multigrid algorithms have been developed for this class of PDEs and we would expect to be able to apply similar algorithms to (5.13-5.14).

The Galerkin weak form of (5.36) is: Find $p \in H^1(\Omega)$ with $p = 0$ on Γ_I such that

$$\langle \mathbf{b} \cdot \nabla p,\, \mathbf{b} \cdot \nabla q \rangle_{0,\Omega} + \varepsilon \langle \mathbf{d} \cdot \nabla p,\, \mathbf{d} \cdot \nabla q \rangle_{0,\Omega} = \langle \tilde{f}, q \rangle_{0,\Omega}, \tag{5.38}$$

for every $q \in H^1(\Omega)$ with $q = 0$ on Γ_I. The left side is similar to the left side of the weak form of our LS formulation, which can be written as

$$\langle \mathbf{b} \cdot \nabla p, \mathbf{b} \cdot \nabla q \rangle_{0,\Omega} + \langle p,\, q \rangle_B = \langle f, \mathbf{b} \cdot \nabla q \rangle_{0,\Omega} + \langle g,\, q \rangle_B. \tag{5.39}$$

As mentioned in Section 5.1, a geometric approach is generally not robust enough to handle these rotated anisotropic-type flows.

5.2.1 Non-M-matrix Considerations

As outlined in the previous section, a multigrid method is commonly described using a smoothing operator, intergrid transfer operators (interpolation and restriction), and discrete operators at each grid level. This is also true for algebraic-based multigrid methods. The primary difference between the two methodologies is that geometric-based methods tend to fix a coarsening hierarchy and attempt to design a relaxation strategy **a priori**, while algebraic methods tend to use a given smoother and design the coarse grids and intergrid transfer operators automatically. While the algebraic approach does not rely on the notion of a "physical" grid, the algorithm does define coarse and fine sets of degrees of freedom.

Since the solution to the hyperbolic PDE is only expected to be smooth in one direction that is generally not aligned with the flow, the notion of geometric smoothness is inadequate. AMG ignores this physical sense of smoothness and is based on a so-called algebraic smoothness. This is effective if we properly select coarse grids and interpolation strategies that capitalize on this algebraic property. Our discussion and summary of the basics of AMG closely follows the details presented in [20, 68, 79].

Algebraically, smooth error is defined to be error that is not adequately reduced by relaxation. Such slowly converging error must be handled on a coarser set of points. For

smoothing operator S and algebraically smooth error e, we have (loosely speaking) that

$$Se \approx e. \tag{5.40}$$

Using the discrete energy norm $\|\cdot\|_A = \sqrt{(Ae, e)}$ as a measure, this becomes

$$\|Se\|_A \approx \|e\|_A. \tag{5.41}$$

We now further characterize the idea of algebraically smooth error. Recalling (5.6) and (5.7), where M is now the smoothing operator S, (5.41) becomes

$$\|(I - \hat{A}^{-1}A)e\|_A \approx \|e\|_A. \tag{5.42}$$

For Gauss-Seidel, $\hat{A} = D - L$, where $D = \text{diag}(A)$ and $-L$ is the (strict) lower triangular part of A. Weighted Jacobi relaxation uses $\hat{A} = \omega D$, for some weight $\omega \in \mathbb{R}^+$. Using weighted Jacobi and (5.42) implies

$$\langle D^{-1}r, r\rangle \ll \langle Ae, e\rangle. \tag{5.43}$$

The same relation holds for Gauss-Seidel and many other relaxation techniques, although it is most easily obtained using weighted Jacobi (see [79]). Writing (5.43) componentwise yields

$$\sum_{i=1}^{n} \frac{r_i^2}{a_{ii}} \ll \sum_{i=1}^{n} r_i e_i, \tag{5.44}$$

so that, on average,

$$|r_i| \ll a_{ii}|e_i|, \tag{5.45}$$

for smooth error. Thus, smooth error has small residuals in comparison, which agrees with the behavior of many relaxation strategies. Gauss-Seidel, for example, results in

$$v^{(k+1)} = v^{(k)} + (D - L)^{-1}r^{(k)}, \tag{5.46}$$

while weighted Jacobi has the form

$$v^{(k+1)} = v^{(k)} + \omega D^{-1}r^{(k)}. \tag{5.47}$$

In general, relaxation can be expressed as

$$v^{(k+1)} = v^{(k)} + \hat{A}^{-1}r^{(k)}, \tag{5.48}$$

where $\hat{A}$ is an approximation to A and more easily invertible, as discussed above. When relaxation stalls, we then have $\hat{A}^{-1}r \approx 0$ in comparison to e.

The relaxation method is fixed and the remaining essential components of algebraic-based multigrid that rely on the aforementioned notion of smooth error are

- selection of a coarse set of points on which the smooth error can be accurately represented. For this component, we briefly mention the idea of strong connections particularly for convection dominated flow,
- definition of an interpolation operator I_H^h that accurately transfers smooth error to finer grids. The restriction operator is defined variationally, that is, $I_h^H = {I_H^h}^T$. The coarse-grid operator defined by the Galerkin condition $A^H = I_h^H A^h I_H^h$ then yields an optimal coarse-grid correction in the A^h-norm [20].

Define the two-grid AMG cycle

$$v^h \leftarrow AMG^h(v^h, f^h), \tag{5.49}$$

similar to (5.9), as the following process:

(1) Presmooth v^h on $A^h u^h = f^h$ ν_1 times.

(2) Restrict the residual and compute the coarse grid solution:

$$r^H \leftarrow I_h^H(f^h - A^h v^h)$$
$$e^H \leftarrow {A^H}^{-1} r^H$$

(3) Compute the corrected approximation on the interpolated coarse grid error:

$$v^h \leftarrow v^h + I_H^h e^H$$

(4) Postsmooth v^h with $A^h u^h = f^h$ ν_2 times. This results in the error propagation

$$e^h \leftarrow S_{\nu_2}^h e^h$$

So far we have simply initiated a discussion of the general idea behind the AMG method. Although we primarily use features outlined in the standard Ruge-Stüben algorithm [68], we

need to address issues particular to our problem, including the definition of strong connections and proper interpolation.

The strength of coupling between two degrees of freedom plays a large role in the behavior of this solution method. We say that u_i is **connected** to u_j if $a_{ij} \neq 0$ (recall that A is symmetric as a result of the least-squares formulation), where $A = \{a_{ij}\}_{i,j=1\ldots n}$. The magnitude of a_{ij} dictates how much influence the error at j has on the error at i in relaxation. We select strong connections to the variable u_i by comparing the strength to the maximum in row i of A: u_j is **strongly connected** to u_i if

$$-a_{ij} \geq \varepsilon_{\text{strength}} \max_{k \neq i}(-a_{ik}), \tag{5.50}$$

where $\varepsilon_{\text{strength}} \in \mathbb{R}^+$ is some threshold constant. The following illustrates the structures of the matrix for our discretization for constant flow fields with respect to the angle made with the x-axis, θ.

$$\underset{\theta = \frac{\pi}{4}}{\begin{bmatrix} \frac{1}{12} & -\frac{1}{6} & -\frac{5}{12} \\ -\frac{1}{6} & \frac{4}{3} & -\frac{1}{6} \\ -\frac{5}{12} & -\frac{1}{6} & \frac{1}{12} \end{bmatrix}} \quad \underset{\theta = \frac{\pi}{8}}{\begin{bmatrix} 0.01 & 0.19 & -0.34 \\ -0.52 & 1.33 & -0.52 \\ -0.34 & 0.19 & 0.01 \end{bmatrix}} \quad \underset{\theta = 0}{\begin{bmatrix} -\frac{1}{6} & \frac{1}{3} & -\frac{1}{6} \\ -\frac{2}{3} & \frac{4}{3} & -\frac{2}{3} \\ -\frac{1}{6} & \frac{1}{3} & -\frac{1}{6} \end{bmatrix}}.$$

As we can see, there are positive off-diagonals in locations normal to the direction of flow. To emulate the essential nature of the PDE, we would like there to be minimal cross-stream influence. So, in terms of strong connections, choosing $\varepsilon_{\text{strength}} > .40$ generally adheres to this heuristic, even in the limit cases. The following illustrates the strength of connection for the stencils listed above with $\varepsilon_{\text{strength}} > .40$. Strong connections are given by $\times$ and weak connections by $\cdot$.

$$\underset{\theta = \frac{\pi}{4}}{\begin{bmatrix} \cdot & \cdot & \times \\ \cdot & \times & \cdot \\ \times & \cdot & \cdot \end{bmatrix}} \quad \underset{\theta = \frac{\pi}{8}}{\begin{bmatrix} \cdot & \cdot & \times \\ \times & \times & \times \\ \times & \cdot & \cdot \end{bmatrix}} \quad \underset{\theta = 0}{\begin{bmatrix} \cdot & \cdot & \cdot \\ \times & \times & \times \\ \cdot & \cdot & \cdot \end{bmatrix}}.$$

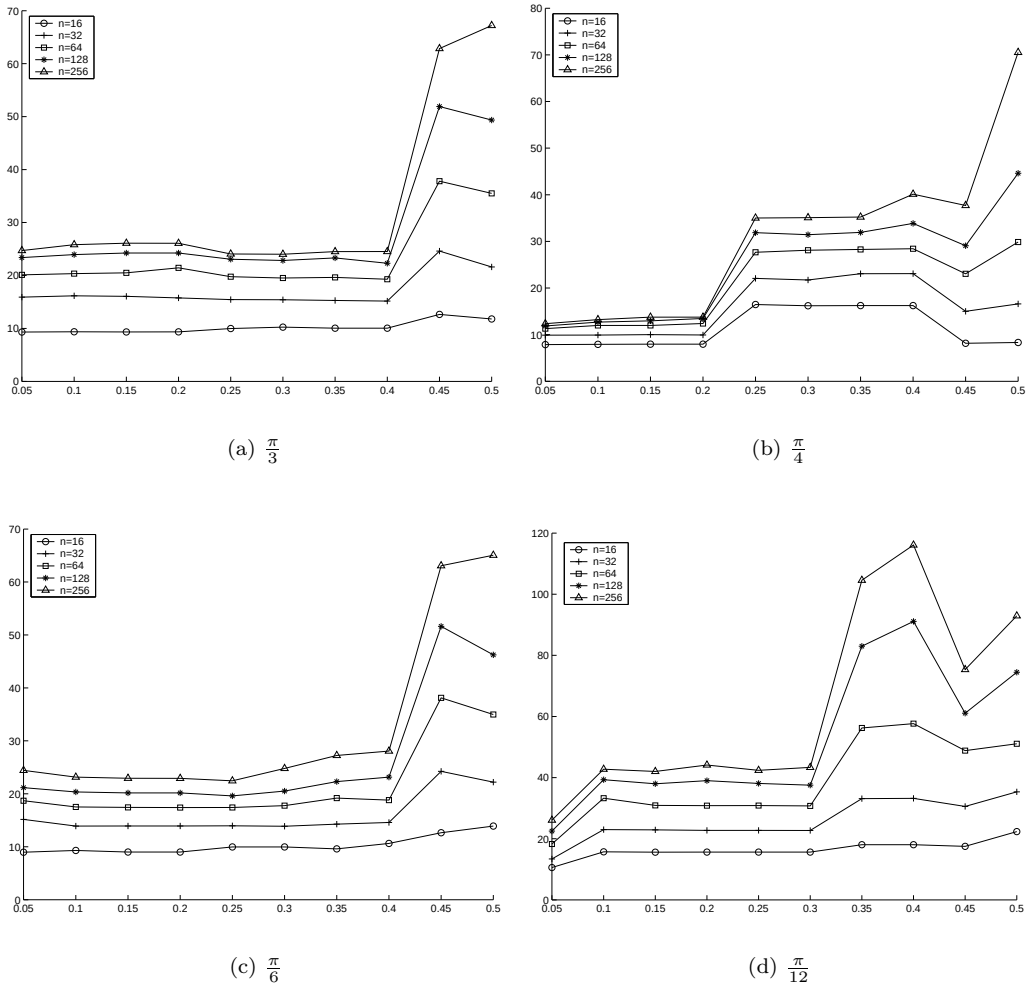
The strength of connection stencils for $\varepsilon_{\text{strength}} > .40$ appear attractive and complement the physics of the PDE, but we are more interested in the nature of the discrete problem. As presented in Chapter 4, the main drawback of the least-squares approach is smearing of the discontinuity. This is evident in the matrix equations where we have off-diagonal terms in the cross-stream direction that are strongly influencing the diagonal. It is then essential that we retain at least some of the off-diagonal terms on the coarse grids in order to ensure accurate interpolation. This is indeed what we find. Figure 5.7 shows total work units per digit accuracy with respect to $\varepsilon_{\text{strength}}$ for grids ranging from 2^4 to 2^8 degrees of freedom in each coordinate direction and using W-cycles. The "work units per digit accuracy" is a measure of the total relative complexity of the algorithm and is a measure work units required to reduce the error by a factor of 10. We define this value in more detail in Section 5.2.2. The plots show the benefits of correctly describing strong influence/dependence. The results presented are for the interior problem, in the sense that strong treatment of the boundary conditions is used. We also discuss treatment of boundary conditions in more detail in the next section. The strong connections for $\varepsilon_{\text{strength}} = 0.1$ are as follows:

$$\begin{bmatrix} \cdot & \times & \times \\ \times & \times & \times \\ \times & \times & \cdot \end{bmatrix} \quad \begin{bmatrix} \cdot & \cdot & \times \\ \times & \times & \times \\ \times & \cdot & \cdot \end{bmatrix} \quad \begin{bmatrix} \times & \cdot & \times \\ \times & \times & \times \\ \times & \cdot & \times \end{bmatrix}.$$

$$\theta = \tfrac{\pi}{4} \qquad \theta = \tfrac{\pi}{8} \qquad \theta = 0$$

Using a small value of $\varepsilon_{\text{strength}}$ defines a large set of influential nodes, but the actual coarse set of points is smaller. The grid complexity, defined as the total number of grid points on all levels divided by the number on the fine level [20], indicates the relative size of the coarse grids. The grid complexities for small values of $\varepsilon_{\text{strength}}$ versus large values are of roughly the same order. More precisely, for $\varepsilon_{\text{strength}}$ ranging from 0.1 to 0.5, the grid complexities vary from 1.7 to 1.96 on a mesh of size 128×128. We can conclude that, for small values of $\varepsilon_{\text{strength}}$, the set of coarse points is slightly smaller, requiring less work. A smaller $\varepsilon_{\text{strength}}$ results in a larger set of points to possibly interpolate from. This enables a more effective definition of interpolation

from a fewer number of points.



(a) $\frac{\pi}{3}$ (b) $\frac{\pi}{4}$

(c) $\frac{\pi}{6}$ (d) $\frac{\pi}{12}$

Figure 5.7: Strength threshold $\varepsilon_{\text{strength}}$ versus work units per digit accuracy.

The efficiencies described above are encouraging, particularly in view of the fact that we do not have an M-matrix. Condition (5.50) ensures that we are not selecting positive off-diagonals as strong connections, but interpolation also needs to account for this.

We noted that smooth error has little variation in the direction of a strong connection. This character of smooth error can be used in defining interpolation. We can accurately find the fine-grid value u_i based on coarse-grid values u_j, assuming i is strongly connected to j and u is smooth. Specifically, following the notation in [20, 79], let i be a fine-grid point to which

we wish to interpolate and let j be a point connected to i. Define the following:

$$C_i^s = \text{coarse-grid strongly connected neighbors of } i,$$

$$C_i^w = \text{coarse-grid weakly connected neighbors of } i,$$

$$C_i^- = \text{coarse-grid connected neighbors of } i \text{ with } sign(a_{ij}) = -sign(a_{ii}),$$

$$F_i^s = \text{fine-grid strongly connected neighbors of } i,$$

$$F_i^w = \text{fine-grid weakly connected neighbors of } i.$$

We wish to describe e_i in terms of e_j, where $j \in C_i^w \cup C_i^s$. The typical approach is to recall that $Ae \approx 0$ for smooth error. This starting point is justified since the fine-grid error is assumed to be algebraically smooth and we can define interpolation based on this intuition. We then have

$$a_{ii}e_i + \sum_{j \in C_i^s} a_{ij}e_j + \sum_{j \in F_i^s} a_{ij}e_j + \sum_{j \in C_i^w \cup F_i^w} a_{ij}e_j \approx 0. \tag{5.51}$$

Since the values of e at F_i^s and F_i^w are undetermined and the points in C_i^w are not points from which we wish to interpolate, we attempt to write such e_j in terms of the e_j at points in C_i^s. First consider strongly connected points $j \in F_i^s$. As a first pass, we could write

$$e_j = \frac{\sum_{k \in C_i^s \cup C_i^w \cup \{i\}} a_{jk}e_k}{\sum_{k \in C_i^s \cup C_i^w \cup \{i\}} a_{jk}}. \tag{5.52}$$

However, with positive and negative connections, the denominator could be near zero. Instead we use

$$e_j = \frac{\sum_{k \in C_i^- \cup \{i\}} a_{jk}e_k}{\sum_{k \in C_i^- \cup \{i\}} a_{jk}}, \tag{5.53}$$

which distributes e_j only along connections with the correct sign. A more common approach is to distribute along only strong connections $C_i^s \subset C_i^-$. We find similar results for most examples using this approach, although for some flows with widely varying anisotropies, the performance becomes erratic and we keep the natural distribution (5.53), which distributes to a potentially larger set of points. Also, distributing to the diagonal results in a very poor approximation, which is consistent with other applications [20].

We now discuss handling of weak connections F_i^w. We may follow a similar approach as above and distribute the values to all of the connections. However, not only does the sign of the

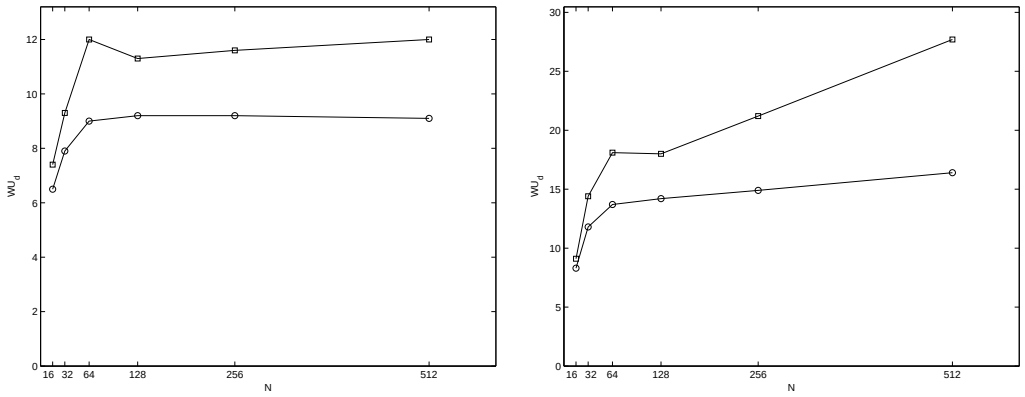
connection continue to be a problem, but complexity becomes an issue. Figure 5.8 confirms this potential trouble. For several flow examples, the total work units to achieve a certain accuracy is larger when distributing to all of the connections as opposed to only connections with the correct sign. We also find similar difficulties in using a distribution to the diagonal, the more common choice in this case. We choose to distribute the weak connections to $C_i^- \cup \{i\}$ since distributing to the smaller set of strong connections C_i^s results in only slight improvement for some cases. The added robustness of considering the sign of the stencil entry is more apparent for flows with widely varying anisotropies. Special consideration for non-M-matrices is often needed, so the added complexity in handling opposite signed connections is to be expected.

The problem of not having properties of an M-matrix has been considered previously. Matrices where the positive off-diagonal is reasonably small is discussed in more detail in [16] with an overview in [79]. Our immediate problem may fit well within this framework, but a detailed analysis is beyond the scope of this dissertation. We seek to use features of the classic Ruge-Stüben algorithm to handle the non-M-matrix problem and we present results for this approach in the following sections.

5.2.2 Numerical Evidence

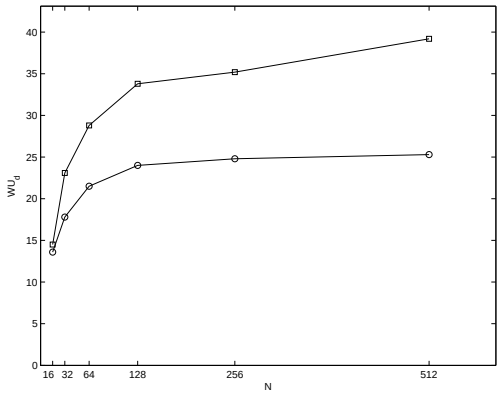
We test the convergence of AMG for the example described in Section 4.5 with several values of θ. Bilinear elements are used for ease of implementation and interpretation of the computational work involved. The AMG convergence results reported in Tables 5.3-5.10 were computed by John Ruge's FOSPACK (First Order Systems Least-Squares Finite Element Software Package) [69] using V and W-cycles based on point Gauss-Seidel relaxation. Each pre-relaxation step performed on the downward leg of the cycle sweeps first over F-points (points not associated with coarse-grid variables), then sweeps over the remaining so-called C-points. Post-relaxation steps reverse the ordering to ensure symmetry of the overall cycle.

The formulation we first consider is the discrete weak form of the LS minimization (4.57). This formulation imposes the boundary conditions weakly, in the sense that the boundary data

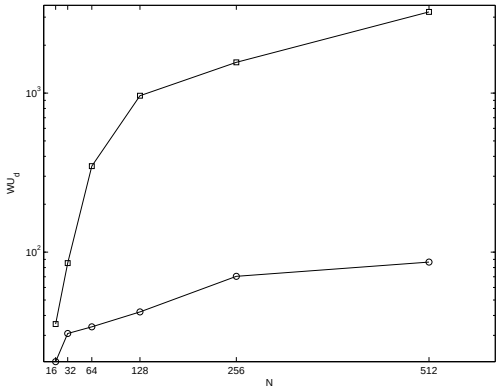


(a) $\frac{\pi}{4}$

(b) $\frac{\pi}{6}$



(c) $\frac{\pi}{8}$



(d) Loop #1

Figure 5.8: Interpolation comparison: Work units using distribution of weak connections along all connections (□) and distribution along connections with the correct sign (○).

is included in the minimization. Alternatively, we consider imposing the boundary conditions strongly, although it is not the proper treatment, as discussed in Section 4.5. Strong treatment of the boundary conditions removes the minimization on the boundary and instead restricts the finite element space of test and trial functions to satisfy the boundary conditions directly. As we discuss in more detail below, this approach isolates the behavior of the algorithm to the interior problem, thus exposing possible difficulties with the boundary-to-interior connections.

Columns (a)-(d) of Table 5.3 show the increase in convergence factors for V(1,1), V(2,2), W(1,1), and W(2,2)-cycles using weak boundary conditions as the mesh is refined. These values

are the factors by which the error is reduced on the finest level by performing one cycle and are the geometric average of convergence factors from one cycle to the next up until the relative residual has reached a prescribed tolerance. For the remainder of the dissertation, a tolerance of 1e-10 is used. We would like these factors to be small—i.e., large reduction in error from one cycle to the next—and bounded as the grid is refined. Increasing the number of smoothing sweeps does not significantly reduce the error. Doubling the relaxation sweeps on every level has little to no affect on the error reduction asymptotically. Comparing columns (a) and (b) in Table 5.3, confirms this for V-cycles, while columns (c) and (d) show similar results for W-cycles.

We now consider strong boundary conditions. For the V-cycle, strongly imposing the boundary conditions has a small affect on the convergence factors compared to weakly imposing boundary conditions. Compare columns (a) and (b) of Table 5.3 with columns (a) and (b) of Table 5.6. An interesting phenomenon is revealed in columns (c) and (d) of Table 5.6, which shows convergence factors for the W-cycle. Columns (c) and (d) reveal that the algorithm is scaling well. Unfortunately, for the strong boundary condition case, $\mathcal{G}(p^h; 0, 0)$ fails to decrease with h (recall $\mathcal{G}$ is a sharp error estimator). The significance of studying this case becomes clear when comparing columns (c) and (d) of Table 5.6 with columns (c) and (d) of Table 5.3. When the boundary term is used in the functional (i.e., weak b.c.), the convergence factors fail to remain constant for the W(1,1)-cycle, although they grow only slowly with respect to the grid size.

The convergence factor alone is often not an appropriate indicator of the overall performance of a multigrid cycle. This is particularly the case for $\mu > 1$, where there are many more visits to coarse grids than in a standard V-cycle. As a measure of the complexity for each cycle, we compute the work per cycle in terms of fine-grid relaxation sweeps (or work units). We compute the complexity by summing the number of nonzero matrix entries on each level multiplied by the number of relaxation sweeps performed on that level, divided by the number of nonzero entries in the fine grid matrix. This complexity is close to a measure of the work units per cycle. Tables 5.4 and 5.7 show that we find little growth in the work per cycle as the

grids become finer (n increases), even in the case of W-cycles.

Combining the work per cycle and the average convergence factor is a better measure of the total work involved to achieve a certain accuracy. In Tables 5.5 and 5.8, we report the number of work units required to reduce the error by a factor of 10. This "work units per digit accuracy" is a measure of the total relative complexity of the algorithm and is computed as

$$W_d = -\frac{W_c}{\log \rho}, \tag{5.54}$$

where W_c is the work units per cycle discussed above and ρ is the convergence factor presented in Tables 5.3 and 5.6.

Columns (a) and (b) in Tables 5.5 and 5.8 show that the number of work units per digit for the V(1,1)-cycle and V(2,2)-cycle appear to be growing slowly with the dimension of the linear system, which increases by a factor of 4 with each row of the table. This is true for both weak and strong treatment of boundary conditions. Likewise, the work units per digit for the W(1,1)-cycle with weak boundary conditions appears to be growing, although more slowly, while the work units per digit for the W(1,1)-cycle with strong boundary conditions appears to be essentially bounded (compare column (c) of Tables 5.5 and 5.8). Strong boundary conditions do not reduce the growth in complexity with grid size for V-cycles, while they do for W(1,1)-cycles. This suggests that W-cycles are necessary and that more work needs to be done on coarse grids as predicted in Section 5.1.

Table 5.8 agrees with Table 5.6. Extra relaxation sweeps have little affect on the error reduction while the work per cycle doubles, as presented in Table 5.7. This has direct impact on the total work per digit accuracy, and although the convergence factors for a W(2,2)-cycle with strong boundary conditions scale with grid size (column (d), Table 5.6), the total work per digit is much more expensive. Comparing columns (c) and (d) of Table 5.8 confirms this observation.

Finally, Tables 5.9 and 5.10 show AMG performance for widely varying anisotropic flow based on Example 2 in Section 4.5. We see similar results as in the constant advection case. Work units are improved for strong boundary conditions and with W-cycles. Again, additional

smoothing has little affect on the error reduction, while work per cycle doubles. With such varying anisotropies, even standard W-cycles fail to scale well for strong boundary conditions. However, column (d) shows that the multigrid W-cycle is still an effective preconditioner for the conjugate gradient method.

5.3 Reformulating the Minimization Principle

The difficulties with the current implementation of weak boundary conditions suggest a reconsideration of the form of the functional. Recall the bilinear form for the least-squares minimization principle (4.47):

$$\mathcal{F}(u,v) := \langle \mathbf{b}\cdot\nabla u,\, \mathbf{b}\cdot\nabla v\rangle_{0,\Omega} + \langle u,\, v\rangle_B. \tag{5.55}$$

Discretization of $\mathcal{F}(\cdot,\cdot)$ yields a discrete matrix $\mathbf{A}$ of the form

$$\mathbf{A} = \overset{\bullet}{\mathbf{A}} + \overset{\circ}{\mathbf{A}}, \tag{5.56}$$

where $\overset{\bullet}{\mathbf{A}}$ comes from the first (interior) term in (5.55) and $\overset{\circ}{\mathbf{A}}$ from the second (boundary) term. Using standard bilinear finite elements and denoting

$$I = \text{set of nodal indices not on the inflow boundary (interior nodes)},$$

$$B = \text{set of nodal indices on the inflow boundary (boundary nodes)},$$

we arrive at the matrix structure

$$\mathbf{A} = \underbrace{\begin{bmatrix} \overset{\bullet}{\mathbf{A}}_{II} & \overset{\bullet}{\mathbf{A}}_{IB} \\ \overset{\bullet}{\mathbf{A}}_{BI} & \overset{\bullet}{\mathbf{A}}_{BB} \end{bmatrix}}_{\mathcal{O}(1)} + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & \overset{\circ}{\mathbf{A}}_{BB} \end{bmatrix}}_{\mathcal{O}(h)}. \tag{5.57}$$

The importance of this form becomes apparent as $h \to 0$. The interior portion, $\overset{\bullet}{\mathbf{A}}$, has entries that are $\mathcal{O}(1)$, while the boundary portion, $\overset{\circ}{\mathbf{A}}$, has $\mathcal{O}(h)$ entries. When the mesh is refined—i.e., h decreases—the matrix $\mathbf{A}$ tends to become, in a sense, dominated by $\overset{\bullet}{\mathbf{A}}$. More importantly, the strength of connection between inflow boundary nodes weakens compared to those of the

		(a)	(b)	(c)	(d)
θ	$N \times N$	V(1,1)	V(2,2)	W(1,1)	W(2,2)
$\frac{\pi}{4}$	16	.34	.25	.27	.20
	32	.48	.38	.30	.21
	64	.57	.49	.29	.23
	128	.69	.63	.33	.25
	256	.74	.70	.42	.36
	512	.79	.75	.40	.34
$\frac{\pi}{6}$	16	.51	.41	.31	.26
	32	.57	.51	.33	.32
	64	.68	.62	.34	.35
	128	.80	.75	.38	.40
	256	.85	.83	.43	.46
	512	.87	.87	.50	.53
$\frac{\pi}{8}$	16	.45	.39	.32	.30
	32	.61	.50	.35	.35
	64	.65	.61	.39	.38
	128	.72	.70	.42	.42
	256	.79	.76	.50	.49
	512	.85	.81	.59	.56

Table 5.3: AMG convergence factors, ρ (weak boundary conditions).

		(a)	(b)	(c)	(d)
θ	$N \times N$	V(1,1)	V(2,2)	W(1,1)	W(2,2)
$\frac{\pi}{4}$	16	3.4	6.7	4.7	9.2
	32	3.6	7.2	5.8	11.5
	64	3.8	7.5	6.5	13.0
	128	3.8	7.6	7.0	14.1
	256	3.9	7.7	7.3	14.5
	512	3.9	7.7	7.4	14.7
$\frac{\pi}{6}$	16	3.8	7.6	5.3	10.4
	32	4.1	8.2	6.9	13.7
	64	4.2	8.5	7.6	15.3
	128	4.3	8.6	8.2	16.4
	256	4.4	8.7	8.7	17.4
	512	4.4	8.8	9.0	18.0
$\frac{\pi}{8}$	16	4.5	9.0	8.0	15.8
	32	5.0	10.0	9.9	19.7
	64	5.3	10.6	11.5	23.0
	128	5.5	10.9	12.6	25.1
	256	5.5	11.1	13.3	26.6
	512	5.6	11.2	13.7	27.4

Table 5.4: Work units per cycle: W_c (weak boundary conditions).

		(a)	(b)	(c)	(d)
θ	$N \times N$	V(1,1)	V(2,2)	W(1,1)	W(2,2)
$\frac{\pi}{4}$	16	7.2	11.1	8.3	13.1
	32	11.4	17.2	11.1	17.0
	64	15.4	24.2	12.1	20.3
	128	23.7	38.1	14.6	23.4
	256	29.5	49.7	19.2	32.6
	512	37.8	62.0	18.5	23.0
$\frac{\pi}{6}$	16	13.0	19.5	10.4	17.8
	32	17.0	28.1	14.3	27.6
	64	25.4	40.9	16.3	33.5
	128	44.6	69.2	19.5	41.2
	256	61.8	107.8	23.8	51.7
	512	72.4	144.9	30.0	65.5
$\frac{\pi}{8}$	16	13.0	22.1	16.1	30.3
	32	23.4	33.3	21.6	43.2
	64	28.4	49.5	28.1	54.6
	128	38.3	70.6	33.4	66.7
	256	54.2	93.1	44.3	86.0
	512	79.1	122.1	59.8	108.9

Table 5.5: Work units per digit of accuracy: W_d (weak boundary conditions).

		(a)	(b)	(c)	(d)
θ	$N \times N$	V(1,1)	V(2,2)	W(1,1)	W(2,2)
$\frac{\pi}{4}$	16	.32	.21	.22	.16
	32	.39	.29	.21	.16
	64	.47	.36	.20	.16
	128	.56	.48	.19	.15
	256	.64	.57	.17	.15
	512	.68	.62	.16	.13
$\frac{\pi}{6}$	16	.31	.27	.22	.18
	32	.46	.39	.26	.25
	64	.56	.49	.27	.29
	128	.63	.59	.26	.30
	256	.69	.67	.26	.31
	512	.74	.73	.28	.33
$\frac{\pi}{8}$	16	.29	.24	.23	.22
	32	.41	.35	.26	.28
	64	.54	.47	.29	.31
	128	.62	.58	.29	.31
	256	.70	.68	.29	.31
	512	.78	.76	.29	.30

Table 5.6: AMG convergence factors, ρ (strong boundary conditions).

		(a)	(b)	(c)	(d)
θ	$N \times N$	V(1,1)	V(2,2)	W(1,1)	W(2,2)
$\frac{\pi}{4}$	16	3.2	6.3	4.3	10.6
	32	3.5	7.0	5.3	13.3
	64	3.7	7.4	6.3	15.7
	128	3.8	7.5	6.6	16.1
	256	3.8	7.6	7.1	17.2
	512	3.9	7.7	7.3	16.4
$\frac{\pi}{6}$	16	3.8	7.5	5.4	10.7
	32	4.1	8.2	6.9	13.7
	64	4.3	8.5	7.8	15.5
	128	4.3	8.6	8.3	16.2
	256	4.4	8.7	8.7	17.4
	512	4.4	8.8	9.1	18.1
$\frac{\pi}{8}$	16	4.6	9.2	8.7	17.3
	32	5.1	10.2	10.4	20.8
	64	5.4	10.2	11.6	23.1
	128	5.5	10.7	13.0	25.8
	256	5.6	11.0	13.3	26.6
	512	5.6	11.2	13.6	27.2

Table 5.7: Work units per cycle: W_c (strong boundary conditions).

		(a)	(b)	(c)	(d)
θ	$N \times N$	V(1,1)	V(2,2)	W(1,1)	W(2,2)
$\frac{\pi}{4}$	16	6.4	9.3	6.5	10.6
	32	8.5	13.0	7.9	13.3
	64	11.2	16.6	8.9	15.7
	128	15.0	23.7	9.2	16.1
	256	19.8	31.4	9.2	17.2
	512	23.0	37.2	9.1	14.9
$\frac{\pi}{6}$	16	7.5	13.3	8.3	14.3
	32	12.2	20.1	11.8	22.8
	64	16.9	27.5	13.7	28.9
	128	21.6	37.7	14.2	31.8
	256	27.1	50.2	14.9	34.3
	512	33.5	64.1	16.4	37.6
$\frac{\pi}{8}$	16	8.6	14.9	13.6	26.3
	32	13.2	22.4	17.8	37.7
	64	20.0	32.6	21.5	45.5
	128	26.5	46.6	24.0	50.8
	256	35.9	66.4	24.8	52.4
	512	51.8	93.8	25.3	52.1

Table 5.8: Work units per digit of accuracy: W_d (strong boundary conditions).

		(a)	(b)	(c)	(d)
	$N \times N$	V(1,1)	V(2,2)	W(1,1)	W(1,1)-PCG
ρ	16	.54	.44	.42	.16
	32	.63	.55	.40	.16
	64	.73	.67	.46	.21
	128	.80	.76	.49	.23
	256	.90	.87	.67	.32
	512	.92	.90	.71	.38
W_c	16	4.8	9.7	9.1	9.2
	32	5.5	11.0	13.0	13.0
	64	5.8	11.6	15.6	15.6
	128	5.8	11.7	17.6	17.6
	256	5.8	11.6	18.0	18.0
	512	5.7	11.4	17.4	17.4
W_d	16	18.2	27.3	24.4	11.5
	32	27.5	42.5	32.7	16.4
	64	42.3	66.4	46.4	23.1
	128	60.4	98.1	56.7	27.5
	256	126.4	191.3	103.6	36.4
	512	156.9	248.3	116.9	41.4

Table 5.9: AMG performance for Example 2 (weak boundary conditions)

		(a)	(b)	(c)	(d)
	$N \times N$	V(1,1)	V(2,2)	W(1,1)	W(1,1)-PCG
ρ	16	.54	.39	.39	.15
	32	.60	.51	.40	.15
	64	.68	.61	.37	.16
	128	.77	.73	.41	.17
	256	.87	.83	.57	.26
	512	.89	.86	.63	.28
W_c	16	4.6	9.2	8.4	8.4
	32	5.3	10.7	12.3	12.3
	64	5.6	11.2	14.7	14.6
	128	5.7	11.4	16.3	16.3
	256	5.7	11.5	17.2	17.2
	512	5.7	11.3	17.4	17.4
W_d	16	17.2	22.5	20.5	10.2
	32	24.1	36.6	30.8	14.9
	64	33.4	52.2	33.9	18.4
	128	50.4	83.7	42.1	21.2
	256	95.0	141.9	70.5	29.4
	512	112.0	173.1	86.6	31.4

Table 5.10: AMG performance for Example 2 (strong boundary conditions)

interior. This directly impacts the performance of the algebraic multigrid method as shown in Tables 5.3-5.10. The interior problem $\overset{\bullet}{\mathbf{A}}_{II}\mathbf{u}_I = \mathbf{f}_I$ (strong treatment of the boundary conditions) is solved much more efficiently than the whole matrix problem (weak treatment of the boundary conditions).

These observations suggest reconsideration of the weighting of the boundary term in the functional. Consider the conforming least-squares functional with a weighted boundary:

$$\mathcal{G}(u;\, f, g) := \|\mathbf{b}\cdot\nabla u - f\|^2_{0,\Omega} + \omega\|u - g\|^2_B, \tag{5.58}$$

where the B-norm is defined by (4.42) and ω is a grid-dependent weight function. In Section 4.1, we prove a trace inequality using $\omega = 1$, noting that a weight of $\omega = 1$ or weaker is sufficient for obtaining this bound. We confirm this numerically in Section 4.1, where we find improved convergence in the functional norm and degraded results in the L^2-norm for weaker weights such as $\omega = h^p$, $p > 0$. Moreover, we find that strengthening the weight, e.g., $\omega = h^{-p}$, results in the opposite effect: functional convergence deteriorates, while convergence in the L^2-norm improves slightly. Only for $\omega = 1$ do we achieve a balance in these two norms (see Figure 4.4).

Thus, it seems likely that we cannot remedy these difficulties by choosing a different wighting of the boundary term. We turn instead to the nature of the boundary conditions in the PDE. Consider a divergence-free flow field $\mathbf{b}(\mathbf{x})$ ($\nabla\cdot\mathbf{b} = 0$) so that (4.1) becomes

$$\nabla\cdot(\mathbf{b}u) = 0, \quad \text{in } \Omega, \tag{5.59a}$$

$$u = g, \quad \text{on } \Gamma_I, \tag{5.59b}$$

where we now assume $\mathbf{b}u \in H(\mathrm{div}, \Omega)$. For a large class of forcing terms, f, we can use a modified lifting argument that results in an auxiliary problem. For the multigrid discussion, however, we focus on conservation laws—i.e., $f = 0$. Then (5.59a) implies that

$$\mathbf{b}u = \nabla^\perp\psi, \tag{5.60}$$

for some $\psi \in H^1(\Omega)$. This results in the boundary conditions

$$\nabla^\perp\psi = \mathbf{b}g, \quad \text{on } \Gamma_I, \tag{5.61}$$

which can be rewritten as

$$\psi = \int_{\Gamma_I} (\mathbf{n}\cdot\mathbf{b})g, \quad \text{on } \Gamma_I, \tag{5.62}$$

$$\mathbf{n}\cdot\nabla\phi = (\tau\cdot\mathbf{b})g, \qquad \text{on } \Gamma_I. \tag{5.63}$$

These relations correspond to Dirichlet and Neumann boundary conditions on the inflow boundary, Γ_I. The modified PDE results in the least-squares functional

$$\begin{aligned}\hat{\mathcal{G}}(u,\psi;\, f,g) := \|\nabla\cdot(\mathbf{b}u)\|^2_{0,\Omega} + \|\nabla^\perp\psi - \mathbf{b}u\|^2_{0,\Omega} \\ + \|u-g\|^2_B + \|\psi - \int_{\Gamma_I}(\mathbf{n}\cdot\mathbf{b})g\|^2_{0,\Gamma_I} + \|\mathbf{n}\cdot\nabla\phi - (\tau\cdot\mathbf{b})g\|^2_{0,\Gamma_I}.\end{aligned} \tag{5.64}$$

The corresponding weak form of the associated minimization problem is stated as the following.

Problem 5.1. *Find* $(u,\psi)\in(V,\Psi)$ *such that*

$$\hat{\mathcal{F}}(u,\psi;v,\phi) = \hat{F}(v,\phi) \qquad \forall\,(v,\phi)\in(V,\Psi), \tag{5.65}$$

where

$$\begin{aligned}\hat{\mathcal{F}}(u,\psi;v,\phi) = \langle\nabla\cdot\mathbf{b}u, \nabla\cdot\mathbf{b}v\rangle_{0,\Omega} + \langle\nabla^\perp\psi - \mathbf{b}u, \nabla^\perp\phi - \mathbf{b}v\rangle_{0,\Omega} \\ + \langle u,v\rangle_B + \langle\nabla^\perp\psi, \nabla^\perp\phi\rangle_{0,\Gamma_I}\end{aligned} \tag{5.66}$$

and

$$\hat{F}(v,\phi) = \langle v,g\rangle_B + \langle\nabla^\perp\phi, \mathbf{b}g\rangle_{0,\Gamma_I}. \tag{5.67}$$

Existence and uniqueness of the least-squares solution to the modified PDE follow similarly to the analysis given in Chapter 4. Define the $\hat{V}$-norm by

$$\begin{aligned}\|(u,\psi)\|^2_{\hat{V}} &:= \|u\|^2_V + \|\nabla^\perp\psi\|^2_{0,\Omega} + \|\nabla^\perp\psi\|^2_{0,\Gamma_I} \\ &= \|u\|^2_{0,\Omega} + \|\nabla\cdot\mathbf{b}u\|^2_{0,\Omega} + \|u\|^2_B \\ &\quad + \|\nabla^\perp\psi\|^2_{0,\Omega} + \|\nabla^\perp\psi\|^2_{0,\Gamma_I}\end{aligned} \tag{5.68}$$

and denote the associated Hilbert space by

$$\hat{V} := \{(u,\psi)\in L^2(\Omega)^2 \;:\; \|(u,\psi)\|_{\hat{V}} < \infty\}. \tag{5.69}$$

The following establishes coercivity and continuity in the $\hat{V}$-norm of the bilinear form, $\hat{\mathcal{F}}$, defined by (5.66).

Theorem 5.2 (Coercivity and Continuity.). *There exist constants c_0 and c_1 such that, for every $(u,\psi) \in (V,\psi)$,*

$$c_0\|(u,\psi)\|_{\hat{V}}^2 \leq \hat{\mathcal{F}}(u,\psi;u,\psi), \tag{5.70}$$

$$\hat{\mathcal{F}}(u,\psi;v,\phi) \leq c_1\|(u,\psi)\|_{\hat{V}}\|(v,\phi)\|_{\hat{V}} \tag{5.71}$$

Proof. From the Poincaré inequality, Lemma 4.4, we have

$$\begin{aligned}\|u\|_{0,\Omega}^2 + \|\nabla\cdot\mathbf{b}u\|_{0,\Omega}^2 + \|u\|_B^2 &\leq c(\|\nabla\cdot\mathbf{b}u\|_{0,\Omega}^2 + \|u\|_B^2)\\ &\leq \hat{\mathcal{F}}(u,\psi;u,\psi).\end{aligned} \tag{5.72}$$

By Cauchy-Schwarz inequality, we also have

$$\begin{aligned}\langle\nabla^\perp\psi,\nabla^\perp\psi\rangle_{0,\Omega} &= \langle\nabla^\perp\psi,\nabla^\perp\psi-\mathbf{b}u\rangle_{0,\Omega} + \langle\nabla^\perp\psi,\mathbf{b}u\rangle_{0,\Omega}\\ &\leq \|\nabla^\perp\psi\|_{0,\Omega}\|\nabla^\perp\psi-\mathbf{b}u\|_{0,\Omega} + \|\nabla^\perp\psi\|_{0,\Omega}\|\mathbf{b}u\|_{0,\Omega},\end{aligned} \tag{5.73}$$

which results in

$$\|\nabla^\perp\psi\|_{0,\Omega} \leq \|\nabla^\perp\psi-\mathbf{b}u\|_{0,\Omega} + \|\mathbf{b}u\|_{0,\Omega} \tag{5.74}$$

Using the ε-inequality and the Poincaré inequality, again yields

$$\begin{aligned}\|\nabla^\perp\psi\|_{0,\Omega}^2 &\leq c(\|\nabla^\perp\psi-\mathbf{b}u\|_{0,\Omega}^2 + \|\mathbf{b}u\|_{0,\Omega}^2)\\ &\leq c(\|\nabla^\perp\psi-\mathbf{b}u\|_{0,\Omega}^2 + \|\nabla\cdot\mathbf{b}u\|_{0,\Omega}^2 + \|u\|_B^2)\\ &\leq c\hat{\mathcal{F}}(u,\psi;u,\psi).\end{aligned} \tag{5.75}$$

Adding $\|\nabla^\perp\psi\|_{0,\Gamma_I}^2$ to both sides and combining with (5.72) yields in the first inequality, (5.70).

Continuity can be found by the triangle inequality:

$$\begin{aligned}\hat{\mathcal{F}}(u,\psi;v,\phi) &\leq \|\nabla\cdot\mathbf{b}u\|_{0,\Omega}^2\|\nabla\cdot\mathbf{b}v\|_{0,\Omega}^2 + \|\nabla^\perp\psi-\mathbf{b}u\|_{0,\Omega}^2\|\nabla^\perp\phi-\mathbf{b}v\|_{0,\Omega}^2\\ &\quad + \|u\|_B^2\|v\|_B^2 + \|\nabla^\perp\psi\|_{0,\Gamma_I}^2\|\nabla^\perp\phi\|_{0,\Gamma_I}^2\\ &\leq c\|(u,\psi)\|_{\hat{V}}\|(v,\phi)\|_{\hat{V}}.\end{aligned} \tag{5.76}$$

This completes the proof. □

Continuity and coercivity of the bilinear form implies existence and uniqueness of the solution by the theorem of Lax-Milgram [13]. Moreover, these results also hold for discrete subspaces $(V^h,\Psi^h) \subset (V,\Psi)$.

We have added differential boundary conditions to the problem, ensuring that the weak connections do not diminish as $h \to 0$ as they did for the original formulation. However, we have not analyzed the entire structure of the problem. For the original bilinear form (5.55), we have the so-called formal normal equations

$$\mathcal{L}^*\mathcal{L} = \left[-\nabla \cdot \mathbf{b}\mathbf{b}^T\nabla \right], \tag{5.77}$$

since

$$\mathcal{L} = \left[\mathbf{b} \cdot \nabla \right] \tag{5.78}$$

and

$$\mathcal{L}^* = \left[-(\nabla\cdot)\mathbf{b} \right]. \tag{5.79}$$

As we described in Section 5.2, this is closely related to the rotated anisotropic diffusion equation. We expect algebraic multigrid solvers to effectively handle this type of equation. This was confirmed for the interior problem in Section 5.2. The formal normal equations for the modified bilinear form in (5.66) are

$$\begin{aligned} \mathcal{L}^*\mathcal{L} &= \begin{bmatrix} -\nabla \times \nabla^\perp & -(\nabla\times)\mathbf{b} \\ -\mathbf{b}^T\nabla^\perp & -|\mathbf{b}|^2 - \mathbf{b}^T(\nabla\nabla\cdot)\mathbf{b} \end{bmatrix} \\ &= \begin{bmatrix} -\Delta & -(\nabla\times)\mathbf{b} \\ -\mathbf{b}^T\nabla^\perp & -|\mathbf{b}|^2 - \nabla \cdot \mathbf{b}\mathbf{b}^T\nabla \end{bmatrix}, \end{aligned} \tag{5.80}$$

since

$$\nabla \cdot \mathbf{b} = 0, \tag{5.81}$$

$$\mathcal{L} = \begin{bmatrix} \nabla^\perp & -\mathbf{b} \\ 0 & (\nabla\cdot)\mathbf{b} \end{bmatrix}, \tag{5.82}$$

$$\mathcal{L}^* = \begin{bmatrix} \nabla\times & 0 \\ -\mathbf{b}^T & -\mathbf{b}^T\nabla \end{bmatrix}. \tag{5.83}$$

Equation (5.80) shows that we retain differential diagonal dominance (i.e., the operator in the off-diagonal term of $\mathcal{L}^*\mathcal{L}$ is first order while the diagonal is second order). Moreover, the structure of the diagonal is perhaps more attractive than simply the anisotropic diffusion operator in the original formulation. An added scaled mass matrix $|\mathbf{b}|^2\mathbf{I}$ and a pure diffusion operator Δ also appear in the diagonal terms. Both of these modifications are attractive qualities for multigrid methods.

Numerically, we first verify that the space of standard bilinear finite elements converge in the L^2-norm and in the modified functional norm $\|\cdot\|_{\hat{\mathcal{G}}}$. The space of bilinear finite elements, V^h, is a natural choice to approximate $\psi \in H^1(\Omega)$, since $V^h \subset H^1(\Omega)$. Table 5.11 confirms that we retain the finite element convergence properties of the original formulation. We again assume

$$\|u - u^h\| \approx ch^\alpha, \tag{5.84}$$

where c is some constant independent of the mesh size. We compute the convergence rate, α, according to (4.83). The convergence rates closely match those presented in Tables 4.3, 4.4, and 4.6, for $k = 1$. Moreover, spurious oscillations and overshoots are nearly non-existent. The solution profiles are visually identical to those of the original formulation (see Figures 4.8 and 4.9).

	$\|e\|_0$	$\|e\|_{\hat{\mathcal{G}}}$
$\theta = \frac{\pi}{4}$	.24	.27
$\theta = \frac{\pi}{6}$	.25	.28
$\theta = \frac{\pi}{8}$	.25	.28
Loop # 1	.27	.22
Loop # 2	.26	.23
Loop # 3	.29	.24

Table 5.11: Convergence rates for standard bilinear finite elements.

The difficulty with the original least-squares functional $\mathcal{G}$ was that we could not simultaneously achieve attractive functional convergence and optimal AMG convergence. In the previous section, we also show the need for W(1,1)-cycles for the interior problem. From (5.80), we see

that the anisotropic interior problem again represents much of the discrete problem, so we study the performance for W(1,1) with the parameters discussed in Section 5.2. Table 5.12 shows that the convergence factor, ρ, remains nearly constant with decreasing h for the constant advection case at various flow angles, θ. The total work units per digit accuracy, W_d, also scale quite well with grid size. The loop example, discussed in Section 4.5, results in slow growth in the total work. In contrast to the results in Table 5.9, the modified approach achieves much better performance, particularly when it is compared without the aid of a preconditioned conjugate gradient (PCG) wrapper.

	ρ	W_c	W_d
$\theta = \frac{\pi}{4}$	.28	4.6	8.4
	.27	5.2	9.1
	.24	5.5	8.9
	.26	5.8	9.9
	.27	6.0	10.6
$\theta = \frac{\pi}{6}$	.28	4.9	8.9
	.28	5.6	10.1
	.26	6.1	10.4
	.29	6.8	12.7
	.33	6.8	14.2
$\theta = \frac{\pi}{8}$	.33	5.8	12.1
	.31	6.9	13.5
	.31	7.5	14.8
	.30	8.1	15.5
	.29	8.6	16.1
Loop # 1	.38	5.9	14.0
	.36	7.8	18.1
	.42	9.0	24.0
	.44	9.9	27.7
	.61	9.8	45.7
Loop # 1 with PCG	.16	5.9	7.4
	.17	7.8	10.2
	.21	9.0	13.3
	.25	9.9	16.4
	.32	9.8	19.8

Table 5.12: AMG data for the modified functional (5.64) with weak treatment of boundary conditions. W(1,1)-cycles are used. ρ: convergence factor, W_c: work units per cycle, W_d: work units per digit accuracy.

Supplying additional terms to improve the overall nature of the problem is a common approach in the least-squares methodology. This is particularly the case for first-order systems

least-squares (FOSLS) [23], where extra variables are added to create a first-order system and redundant equations are supplied to ensure H^1-equivalence of the least-squares bilinear form. Although the goal in this section was not to achieve a certain level of equivalence, the principle is the same. We modified the PDE to maintain convergence properties of the least-squares functional but to obtain a linear system that is more amenable to AMG solves. The improved performance does come with some cost, however. We introduce a new equation (5.60), which increases the number of variables by one. This new variable, ψ, increases the degrees of freedom at each node to two. We thus obtain a system of size $2N \times 2N$, which results in roughly four times the amount of work per cycle compared to the original $N \times N$ system. However, this new approach appears to provide an optimal solution method, which is critical to performance as the problem size increases.

Chapter 6

Least-Squares Finite Element Methods for Nonlinear Hyperbolic PDEs

This chapter introduces a new least-squares approach for nonlinear hyperbolic conservation laws. The focus of Chapters 3, 4, and 5 was on LSFEMs for the linear advection problem given by (4.1), where the solutions contained contact discontinuities. Here, we shift the focus from linear to nonlinear equations. Scalar nonlinear hyperbolic conservation laws of the form

$$\nabla \cdot \mathbf{F}(u) = 0, \quad \text{in } \Omega, \tag{6.1a}$$

$$u = g, \quad \text{on } \Gamma_I, \tag{6.1b}$$

are considered. Here, $\Omega \subset \mathbb{R}^2$ is a convex domain and Γ_I is defined by

$$\Gamma_I := \{x \in \partial\Omega \,:\, d\mathbf{F}(u) \cdot \mathbf{n} < 0\}, \tag{6.2}$$

where $d\mathbf{F}(u)$ is the Fréchet derivative of $\mathbf{F}(u)$. This is consistent with the definition for the the linear case, (4.2), since, for smooth u, $\nabla \cdot \mathbf{F}(u) = d\mathbf{F}(u) \cdot \nabla u$ and vector $d\mathbf{F}(u)$ is tangent to the characteristic curves at the boundary. The outflow boundary is similarly defined by

$$\Gamma_O := \{x \in \partial\Omega \,:\, d\mathbf{F}(u) \cdot \mathbf{n} > 0\}, \tag{6.3}$$

and the boundary that is aligned with flow characteristics is defined by

$$\Gamma_C := \{x \in \partial\Omega \,:\, d\mathbf{F}(u) \cdot \mathbf{n} = 0\}. \tag{6.4}$$

Variable u is the conserved quantity, while $\mathbf{F}(u)$ is the flux vector, as outlined in Chapter 2. The flux vector, $\mathbf{F}(u)$, is a nonlinear function of u, and we require the components, $F_i(u)$, of

$\mathbf{F}(u)$ to be Lipschitz continuous, i.e., there exists a constant $K \in \mathbb{R}^+$ such that

$$|F_i(u_1) - F_i(u_2)| \leq K\,|u_1 - u_2| \tag{6.5}$$

for all $u_1, u_2 \in \mathbb{R}$, and for $i = 1, 2$.

The standard least-squares approach used in Chapters 3, 4, and 5 has a significant shortcoming in its application to nonlinear hyperbolic conservation laws in the form of (6.1): the equations are not Newton-linearizable around a discontinuous solution for nonlinear flux functions, in the sense that the Fréchet derivative of $\nabla \cdot \mathbf{F}$ is unbounded. This is described in more detail in Appendix A.

Following the approach of Section 5.3, the goal is then to properly reformulate (6.1) so that the resulting equations are linearizable around discontinuous solutions and so that the the correct weak solution to the problem is obtained. As discussed in Chapter 2, the nonlinear hyperbolic conservation law allows for weak solutions in which discontinuities arise in the form of shocks. Adequately resolving shocks in the numerical approximation is a focus in our development of the least-squares method throughout this chapter.

We first reformulate the nonlinear conservation law by introducing the flux vector explicitly as an additional dependent variable. This reformulation highlights the smoothness of the flux vector, $\mathbf{F}(u) \in H(div, \Omega)$. The reformulation enables a choice of finite element spaces for the unknown associated with the flux vector that conforms closely to the $H(div)$-regularity of the problem. The standard least-squares finite element method is then applied using $H(div)$-conforming finite elements in a Gauss-Newton linearization setting. We thus label the approach as an $H(div)$ least-squares finite element method (LSFEM). This formulation also yields an additional similar set of equations in terms of a scalar potential, which is the De Rham-dual of the flux vector [70]. The resulting LSFEM is also equivalent to an H^{-1} minimization, and we refer to it as the H^{-1} least-squares finite element method. In the latter approach, the resulting system of equations is perhaps more attractive since it is more naturally discretized using standard bilinear finite elements. Moreover, this approach is closely related to the reformulation of

Section 5.3.

Extensive numerical results are presented for the new least-squares formulations. The numerical results show not only attractive solution quality, but also convergence to the weak solution. Obtaining the weak solution often poses a challenge in the development of numerical methods for nonlinear hyperbolic conservation laws (see Section 2.2.2). We prove weak conservation properties of our methods, namely, that if u^h converges to some $\hat{u}$, then $\hat{u}$ is a weak solution of (6.1). This theoretical result for the LSFEM is equivalent to the Lax-Wendroff conservation result for numerical schemes that satisfy an exact discrete conservation property [58, 59]. An important observation in this respect is that our $H(div)$ and H^{-1} methods do not impose the conservation property. The two approaches naturally converge to the weak solution even though they do not enforce exact discrete conservation.

The LSFEMs we derive in this chapter are different from more standard numerical methods for nonlinear hyperbolic conservation laws. As discussed in Section 2.4, finite volume methods and discontinuous Galerkin methods have been successful in recent years in simulating nonlinear hyperbolic flows, achieving acceptable solution quality in the numerical approximation [62, 27]. The least-squares methods introduced here do not follow these methodologies.

Our approach does have disadvantages, which have already been discussed in the linear setting. The reformulation in Section 5.3 introduces additional variables, increasing the number of degrees of freedom at the discrete level. Also, smearing of shocks is evident in the LSFEM. However, we are again motivated by the computational benefits offered in the LS setting. Minimization of the least-squares functional offers a sense of optimality and provides a natural sharp **a posteriori** error estimator, which can be used in adaptive refinement to resolve the smearing at shocks. We present numerical evidence in Section 6.2 to validate this claim. Also, in Section 5.3, we report on promising numerical results using an AMG method for a similar formulation of the linear problem. Although we do not explore the multigrid solver for the nonlinear case in this dissertation, the numerical results we present in Section 6.2 are computed using multigrid solvers. Preliminary results are encouraging, particularly in a Full Multigrid (FMG) setting [20].

The chapter is structured as follows. In the next section, the $H(div)$-smoothness of the flux vector is discussed in the context of weak solutions, and an $H(div)$ reformulation of nonlinear hyperbolic conservation laws is introduced. A similar H^{-1}-like formulation is also presented. The equations are then posed as a least-squares minimization principle and the corresponding weak problems are derived in a nonlinear context. We follow this with a discussion of the formulations with respect to the H^{-1}-norm. In Section 6.2, we present numerical results that indicate convergence of the LSFEMs to an entropy weak solution of the conservation law. We present finite element convergence rates for shocks and solutions with rarefaction waves. Convergence of the Newton method is also investigated and numerical evidence for convergence of a shock simulation on a locally adapted space-time domain is shown. In Section 6.3, we prove weak conservation theorems for the LSFEMs that are equivalent to the Lax-Wendroff theorem for numerical schemes that satisfy an exact discrete conservation property. In the final section, we explore convergence of the finite element approximations and coercivity of the functional with respect to the graph norm in more detail.

6.1 Reformulation of the Conservation Law

Here, we formalize the definition of a weak solution. We propose new formulations of the conservation law that possess regularity better suited for standard finite element spaces (e.g., Raviart-Thomas and bilinear finite elements). Finally, we draw a connection between the least-squares minimizations of these new formulations and H^{-1} Sobolev norms.

6.1.1 Weak Solutions

We first recall the definition of weak solutions. From Section 2.2.2, note that the integral definition of (6.1) allows shocks and discontinuities as valid solutions. The differential form,

(6.1), also allows such (weak) solutions in a distribution sense [62]. Let

$$V = H^1_{0,\Gamma_O} \quad = \{u \in H^1(\Omega) \,:\, u = 0 \,\text{on}\, \Gamma_O\}, \tag{6.6}$$

$$W = H^1_{0,\Gamma_I \cup \Gamma_C} = \{u \in H^1(\Omega) \,:\, u = 0 \,\text{on}\, \Gamma_I \cup \Gamma_C\}. \tag{6.7}$$

Denote by $\|\cdot\|_V$ the norm on V and W, which is the H^1-seminorm. That is, for $u \in V$(or W)

$$\|u\|_V = \|\nabla u\|_{0,\Omega}. \tag{6.8}$$

We define a weak solution of (6.1) as follows [39, 58, 62]:

Definition 6.1 (Weak Solutions). *Assume $u \in L^\infty(\Omega)$. Then*

$$u \text{ is a weak solution of (6.1)} \tag{6.9}$$

$$\Leftrightarrow$$

$$-\langle \mathbf{F}(u), \nabla\phi\rangle_{0,\Omega} + \langle \mathbf{n}\cdot\mathbf{F}(g), \phi\rangle_{0,\Gamma_I} = 0 \quad \forall\, \phi \in V, \tag{6.10}$$

For this discussion on weak solutions, following [39], we restrict our consideration to the class of piecewise continuously differentiable functions, $\hat{C}^1(\Omega)$. That is, $u \in \hat{C}^1(\Omega)$ if there exists a finite partition, $\{\Omega_i\}_{i=1}^m$, consisting of open sets such that

$$\bar{\Omega} = \bigcup_{i=1}^{m} \Omega_i, \tag{6.11}$$

$$\Omega_i \cap \Omega_j = \emptyset \qquad \text{for}\, i \neq j, \tag{6.12}$$

$$u|_{\Omega_i} \in C^1(\Omega_i) \quad \text{for each } i. \tag{6.13}$$

This accounts for many cases of practical interest and we can formally say that $u \in \hat{C}^1(\Omega) \subset H^{\frac{1}{2}-\varepsilon}(\Omega)$ for every $\varepsilon > 0$ (see [2] for a complete description of Sobolev spaces of fractional order).

We now pose the idea of weak solutions to (6.1) in the context of $H(div, \Omega)$, the space of all square integrable functions with divergence that is square integrable. Let $[\![\mathbf{w}]\!]_\gamma$ be the jump in $\mathbf{w}$ across a smooth curve γ and let $\mathbf{n}$ be the normal to this curve. We then define $H(div, \Omega)$

in the following two equivalent ways [38]:

$$H(div,\Omega) = \{\mathbf{w} \in L^2(\Omega)^2 \,:\, \nabla \cdot \mathbf{w} \in L^2(\Omega)\} \tag{6.14}$$

$$= \{\mathbf{w} \in L^2(\Omega)^2 \,:\, \mathbf{n} \cdot [\![\mathbf{w}]\!]_\gamma = 0 \text{ a.e. for any smooth curve, } \gamma, \text{ in } \Omega\}. \tag{6.15}$$

Condition

$$\mathbf{n} \cdot [\![\mathbf{w}]\!]_\gamma = 0 \tag{6.16}$$

implies that normal vector components of the vector field are continuous across γ. However, this allows the tangential components to be discontinuous. Thus, weak solutions of (6.1) in $\hat{C}^1(\Omega)$ satisfy (6.16) along any curve, γ [67]. This approach also establishes the so-called Rankine-Hugoniot relation [62]. If we consider a straight shock, as illustrated in Figure 6.1 (b), then $\mathbf{n} = (1, -s)$, where s is the shock speed. For flux vector $\mathbf{F}(u) = (F(u), u)$, (6.16) implies

$$s = \frac{[\![F(u)]\!]_\gamma}{[\![u]\!]_\gamma}, \tag{6.17}$$

where γ is now the shock curve. This suggests that weak solutions may be characterized in terms of flux vector $\mathbf{F}(u)$ and Sobolev space $H(div,\Omega)$. We pose the following:

Theorem 6.2 (Weak Solutions and $H(div,\Omega)$). *Assume $u \in L^\infty(\Omega) \cap \hat{C}^1(\Omega)$. Then*

$$u \text{ is a weak solution of (6.1)} \tag{6.18}$$

$$\Leftrightarrow$$

$$\|\nabla \cdot \mathbf{F}(u)\|^2_{0,\Omega} = 0 \text{ and } \|u - g\|^2_{0,\Gamma_I} = 0. \tag{6.19}$$

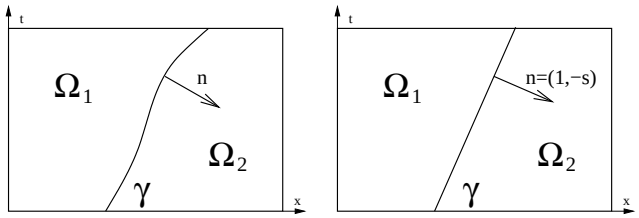
Proof. Begin by assuming that u is a weak solution of (6.1). Godlewski and Raviart [39] (Theorem 2.1) show that this is equivalent to

(I) u satisfies (6.1) pointwise away from shock curves s,

(II) Rankine-Hugoniot condition $\mathbf{n} \cdot [\![\mathbf{F}(u)]\!]_s = 0$ holds along shocks.

From **(I)** and **(II)** and (6.15), we have that $\mathbf{F}(u) \in H(div,\Omega)$. From **(I)**, it then follows that $\|\nabla \cdot \mathbf{F}(u)\|_{0,\Omega} = 0$. Since (6.1b) is satisfied a.e. on Γ_I by **(I)**, we also have $\|u - g\|_{0,\Gamma_I} = 0$.

Conversely, assume that $\|\nabla \cdot \mathbf{F}(u)\|_{0,\Gamma_I} = 0$ and $\|u - g\|_{0,\Omega} = 0$. We then have that $\mathbf{F}(u) \in H(div, \Omega)$ and, by (6.15), $\mathbf{n} \cdot [\![\mathbf{F}(u)]\!]_\gamma = 0$ for any smooth curve, γ, in Ω. Thus, $\mathbf{n} \cdot [\![\mathbf{F}(u)]\!]_s = 0$ for shock curves s and, again by Theorem 2.1 in [39], we have that u is a weak solution of (6.1). □



(a) (b)

Figure 6.1: Illustration of smooth shock curve γ for which $\mathbf{n} \cdot [\![\mathbf{F}(u)]\!]_\gamma = 0$.

6.1.2 Reformulations

With the idea of weak solutions in the context of $H(div, \Omega)$, provided by Theorem 6.2, we are now able to properly reformulate (6.1). We do this problem in terms of the flux vector variable $\mathbf{w} = \mathbf{F}(u)$ to avoid difficulties with the linearization in Newton's method (see Appendix A) and to expose the $H(div)$-smoothness of the generalized flux function. We are thus able to employ standard finite element subspaces that are $H(div)$-conforming and arrive at a more flexible discretization of the problem.

To begin, let $\mathbf{w} = \mathbf{F}(u)$. Problem (6.1) can then be written as

$$\nabla \cdot \mathbf{w} = 0 \qquad \text{in } \Omega, \tag{6.20a}$$

$$\mathbf{w} = \mathbf{F}(u) \qquad \text{in } \Omega, \tag{6.20b}$$

$$\mathbf{n} \cdot \mathbf{w} = \mathbf{n} \cdot \mathbf{F}(g) \quad \text{on } \Gamma_I, \tag{6.20c}$$

$$u = g \qquad \text{on } \Gamma_I. \tag{6.20d}$$

For the new boundary conditions, we find both numerically and theoretically that it is sufficient

to impose only the normal component in (6.20c) versus $\mathbf{w} = \mathbf{F}(g)$. The nonlinear system of equations given by (6.20) is solved using a Newton approach. We address the Newton method in more detail in Section 6.2, although it is important to comment here on the behavior of the Fréchet derivative of nonlinear operator $\mathcal{L}(\mathbf{w}, u)$ defined by (6.20). One of the difficulties with using the original form of conservation law (6.1) is that the Fréchet derivative is unbounded (see Appendix A). With this reformulation, however, the Fréchet derivative operator at $(\mathbf{w}, u)$,

$$d\mathcal{L}|_{(\mathbf{w},u)}[(\hat{\mathbf{w}}, \hat{u})] = \begin{bmatrix} \nabla\cdot & 0 \\ I & -d\mathbf{F}|_u \end{bmatrix} \begin{bmatrix} \hat{\mathbf{w}} \\ \hat{u} \end{bmatrix}, \tag{6.21}$$

is a bounded operator from $H(div, \Omega) \times L^2(\Omega) \rightarrow L^2(\Omega)^2$. This issue is also investigated in more detail in Appendix A.

We briefly study formulation (6.20) numerically in Section 6.2. $H(div)$-conforming finite elements are used, namely, the lowest order Raviart-Thomas finite elements, RT_0. It is for this reason that we label (6.20) the $H(div)$ formulation. Our primary focus throughout this chapter, however, is on a closely related formulation, which we now introduce.

Since $\nabla \cdot \mathbf{F}(u) = 0$, we can conclude that $\mathbf{F}(u) = \nabla^\perp \phi$ for some $\phi \in H^1(\Omega)$ [38]. We then rewrite (6.1) as

$$\nabla^\perp \phi - \mathbf{F}(u) = 0 \qquad \text{in } \Omega, \tag{6.22a}$$

$$\mathbf{n} \cdot \nabla^\perp \phi = \mathbf{n} \cdot \mathbf{F}(g) \quad \text{on } \Gamma_I, \tag{6.22b}$$

$$u = g \qquad \text{on } \Gamma_I. \tag{6.22c}$$

Comparing this formulation with (6.20), we have merely made the substitution $\mathbf{w} = \nabla^\perp \phi$. This is equivalent to choosing $\mathbf{w}$ from the divergence free subspace of RT_0. Each element of this subspace is $\nabla^\perp$ applied to a bilinear finite element basis function. Numerically, however, we find that the results closely agree. For the remainder of the chapter, we focus our theoretical results on the latter formulation, (6.22), only mentioning the corresponding theory for (6.20) when it closely follows. In Section 6.1.4, we prove that the least-squares minimization associated with (6.22) is, in a sense, H^{-1} equivalent. For this reason, we label (6.22) the H^{-1} formulation.

It is important to mention that the Fréchet derivative for this formulation is also a bounded as an operator from $H^1(\Omega) \times L^2(\Omega) \rightarrow L^2(\Omega)^2$. The Fréchet derivative at (ϕ, u) in the direction $(\hat{\phi}, \hat{u})$ is given by

$$d\mathcal{L}|_{(\phi,u)}[(\hat{\phi}, \hat{u})] = \begin{bmatrix} \nabla^\perp & -d\mathbf{F}|_u \end{bmatrix} \begin{bmatrix} \hat{\phi} \\ \hat{u} \end{bmatrix}, \tag{6.23}$$

and is also discussed in more detail in Appendix A.

Remark 6.3. *Recall the modified linear convection problem from Section 5.3:*

$$\nabla \cdot (\mathbf{b}u) = 0 \quad \text{in } \Omega, \tag{6.24a}$$

$$\nabla^\perp \phi - \mathbf{b}u = 0 \quad \text{in } \Omega, \tag{6.24b}$$

plus boundary conditions. There is obviously a strong connection between this formulation and new formulations (6.20) and (6.22). This close relation is interesting since both approaches were motivated from dramatically different directions. The additional equations supplied in Section 5.3 arose from an attempt to address poor multigrid convergence, whereas the formulations presented in this section were derived based on regularity of the problem. Again, we are reminded of the intimate coupling between the discretization and solver for the corresponding discrete linear system. The results of Section 5.3 also indicate that we may find success in solving the new formulations in a multilevel least-squares context. Multigrid solvers for (6.22) are beyond the scope of this dissertation, however.

6.1.3 Least-Squares Finite Element Methods

Formulation (6.22) is now posed as a least-squares minimization principle. We define the nonlinear least-squares functional as

$$\mathcal{G}(\phi, u; g) := \|\nabla^\perp \phi - \mathbf{F}(u)\|^2_{0,\Omega} + \|\mathbf{n} \cdot (\nabla^\perp \phi - \mathbf{F}(g))\|^2_{0,\Gamma_I} + \|u - g\|^2_{0,\Gamma_I}. \tag{6.25}$$

The functional is minimized over subspaces $\Phi^h \subset H^1(\Omega)$ and $\mathcal{U}^h \subset H^1(\Omega)$. That is, we seek (ϕ^h_*, u^h_*) such that

$$(\phi^h_*, u^h_*) = \underset{\phi^h \in \Phi^h, u^h \in \mathcal{U}^h}{\operatorname{argmin}} \mathcal{G}(\phi^h, u^h; g). \tag{6.26}$$

The nonlinear functional in (6.25) is minimized by linearizing equations (6.22) around (ϕ_i^h, u_i^h) and then minimizing the least-squares functional that is based on the linearized equations. Iteratively repeating this procedure until convergence is called the Gauss-Newton method for nonlinear least-squares minimization [32]. We can now formally state the weak problem.

Problem 6.4 (Gauss-Newton H^{-1} LSFEM). *Given (ϕ_i^h, u_i^h), find $\phi_{i+1}^h \in \Phi^h$ and $u_{i+1}^h \in \mathcal{U}^h$ such that*

$$\langle \nabla^\perp \phi_{i+1}^h - \mathbf{F}(u_i^h) - d\mathbf{F}|_{u_i^h}[u_{i+1}^h - u_i^h], \nabla^\perp \psi^h \rangle_{0,\Omega} + \langle \mathbf{n} \cdot (\nabla^\perp \phi_{i+1}^h - \mathbf{F}(g)), \mathbf{n} \cdot \nabla^\perp \psi^h \rangle_{0,\Gamma_I} = 0 \tag{6.27a}$$

$$\langle \nabla^\perp \phi_{i+1}^h - \mathbf{F}(u_i^h) - d\mathbf{F}|_{u_i^h}[u_{i+1}^h - u_i^h], -d\mathbf{F}|_{u_i^h}[v^h] \rangle_{0,\Omega} + \langle u_{i+1}^h - g, v^h \rangle_{0,\Gamma_I} = 0, \tag{6.27b}$$

for all $\psi^h \in \Phi^h$ and $v^h \in \mathcal{U}^h$.

In Section 6.2, we choose standard continuous bilinear finite elements on quadrilaterals for both Φ^h and $\mathcal{U}^h$.

The Gauss-Newton LSFEM for (6.20) follows similarly. Define the associated nonlinear least-squares functional as

$$\mathcal{F}(\mathbf{w}, u; g) := \|\nabla \cdot \mathbf{w}\|_{0,\Omega}^2 + \|\mathbf{w} - \mathbf{F}(u)\|_{0,\Omega}^2 + \|\mathbf{n} \cdot (\mathbf{w} - \mathbf{F}(g))\|_{0,\Gamma_I}^2 + \|u - g\|_{0,\Gamma_I}^2. \tag{6.28}$$

We seek $(\mathbf{w}_*^h, u_*^h)$ such that

$$(\mathbf{w}_*^h, u_*^h) = \underset{\mathbf{w}^h \in \mathcal{W}^h, u^h \in \mathcal{U}^h}{\operatorname{argmin}} \mathcal{F}(\mathbf{w}^h, u^h; g). \tag{6.29}$$

The linearized discrete weak problem is as follows:

Problem 6.5 (Gauss-Newton $H(div)$ LSFEM). *Given $(\mathbf{w}_i^h, u_i^h)$, find $\mathbf{w}_{i+1}^h \in \mathcal{W}^h$ and $u_{i+1}^h \in \mathcal{U}^h$ such that*

$$\langle \nabla \cdot \mathbf{w}_{i+1}^h, \nabla \cdot \boldsymbol{\chi}^h \rangle_{0,\Omega} + \langle \mathbf{w}_{i+1}^h - \mathbf{F}(u_i^h) - d\mathbf{F}|_{u_i^h}[u_{i+1}^h - u_i^h], \boldsymbol{\chi}^h \rangle_{0,\Omega} + \langle \mathbf{n} \cdot (\mathbf{w}_{i+1}^h - \mathbf{F}(g)), \mathbf{n} \cdot \boldsymbol{\chi}^h \rangle_{0,\Gamma_I} = 0 \tag{6.30a}$$

$$\langle \mathbf{w}_{i+1}^h - \mathbf{F}(u_i^h) - d\mathbf{F}|_{u_i^h}[u_{i+1}^h - u_i^h], -d\mathbf{F}|_{u_i^h}[v^h]\rangle_{0,\Omega} + \langle u_{i+1}^h - g, v^h\rangle_{0,\Gamma_I} = 0, \tag{6.30b}$$

for all $\boldsymbol{\chi}^h \in \mathcal{W}^h$ *and* $v^h \in \mathcal{U}^h$.

In our numerical investigations, for $\mathcal{W}^h$, we use RT_0, which are $H(div)$-conforming, and, for $\mathcal{U}^h$, standard continuous bilinear finite elements.

Remark 6.6. *For* $\phi \in H^1(\Omega)$*, we know that* $\mathbf{n}\cdot\nabla\phi \in H^{-\frac{1}{2}}(\Omega)$*, which is generally not in* $L^2(\Omega)$*. Thus*

$$\|\mathbf{n}\cdot(\nabla^\perp\phi - \mathbf{F}(g))\|_{0,\Gamma_I} \tag{6.31}$$

may not be bounded in general. However, this term is in fact bounded for the finite element subspaces that we consider and is simpler than other forms we could use for (6.22b).

6.1.4 H^{-1} Theory

We now consider (6.1) in the context of weak Sobolev-type norms, namely, $H^{-1}(\Omega)$ [81]. We first define notation. Recall that V in (6.6) is the closure of

$$C^\infty_{\Gamma_O,\Omega}(\bar{\Omega}) = \{u \in C^\infty(\bar{\Omega}) \,:\, u = 0 \,\text{on}\, \Gamma_O\}, \tag{6.32}$$

under the H^1-norm, $\|\cdot\|_{1,\Omega}$. The associated **dual** or negative norm of $u \in L^2(\Omega)$ is defined to be

$$\|u\|_{-1,\Gamma_O,\Omega} = \sup_{p\in V} \frac{|\langle u, p\rangle_{0,\Omega}|}{\|\nabla p\|_{0,\Omega}}. \tag{6.33}$$

Define the dual space, V', to be the closure of $L^2(\Omega)$ under dual norm (6.33).

For simplicity, consider the following linear conservation law:

$$\nabla\cdot\mathbf{b}u = f \quad \text{in } \Omega, \tag{6.34a}$$

$$u = g \quad \text{on } \Gamma_I, \tag{6.34b}$$

which, as discussed in Chapter 2, is more precisely termed a balance law. Further, assume that we have applied a lifting argument so that u is homogeneous on the boundary and, thus, $g = 0$.

We are interested in characterizing the weak solution of (6.1), so we consider u in the following normed linear space:

$$U = \{u \in V' \,:\, \nabla \cdot \mathbf{b}u \in V', u = 0 \,\text{on}\, \Gamma_I\}, \tag{6.35}$$

with norm

$$\|u\|_U = \|\nabla \cdot \mathbf{b}u\|_{-1,\Gamma_O,\Omega} = \sup_{p \in V} \frac{|\langle \mathbf{b}u, \nabla p\rangle_{0,\Omega}|}{\|\nabla p\|_{0,\Omega}}. \tag{6.36}$$

That is, $u \in U$ if $\langle \mathbf{b}u, \nabla \cdot \rangle_{0,\Omega} \in V'$. The boundary terms, which are typically present in dual norms, are zero, since $u = 0$ on Γ_I and Γ_O. Solving (6.34) in a weak sense becomes: find $u \in U$ such that

$$\|\nabla \cdot \mathbf{b}u - f\|_{-1,\Gamma_O,\Omega} = 0. \tag{6.37}$$

Equivalently, find $u \in U$ such that

$$a(u, v) = \ell(v) \quad \forall\, v \in V, \tag{6.38}$$

where

$$a(u, v) = \langle \mathbf{b}u, \nabla v\rangle_{0,\Omega} \tag{6.39}$$

$$\ell(v) = \langle f, v\rangle_{0,\Omega}. \tag{6.40}$$

Remark 6.7. *We define the weak solution in terms of (6.37) formally in Section 6.3 (Theorem 6.17).*

The goal is to show that for every $f \in V'$, there exists a unique $u \in U$ that satisfies (6.34)(with $g = 0$). That is, we need to prove that the linear mapping $\mathcal{L} : U \to V'$ defined by (6.34) is an isomorphism. Clearly, V is a Hilbert space, so we must in effect show that U is also a Hilbert space. We do this explicitly, then proceed to establish the assumptions of a Generalized Lax-Milgram Theorem (Theorem III.3.6, [13]). We first prove the following lemma.

Lemma 6.8. *Let* $\mathbf{b} \in H(div, \Omega)$ *and* $u \in U \cap H^1(\Omega)$. *Assume that* $\mathbf{b}$ *is such that* $d(\Gamma_I, \Gamma_O) > 0$. *Then there exists a constant,* $c > 0$, *depending only on the diameter of the domain, such that*

$$\|u\|_{-1,\Gamma_O,\Omega} \le c\|\nabla \cdot \mathbf{b}u\|_{-1,\Gamma_O,\Omega}. \tag{6.41}$$

Proof. Since Γ_I and Γ_O are separated by a nontrivial distance, we can map (with bounded Jacobian) the flow field, $\mathbf{b}$, to one which is aligned with the x-axis, where Γ_I and Γ_O become the left and right boundaries. With this transformation, we state the following proof. Mapping back to the original domain generates a constant that depends only on the diameter of Ω and the Jacobian map. Without loss of generality, we consider $\mathbf{b} = (1, 0)$ for the reminder of the proof. Denote by y_b and y_t the respective bottom and top limits of the transformed domain.

Let $p \in V$ and write u as

$$u(x, y) = \int_{x_I(y)}^{x} \frac{d}{ds} u(s, y)\, ds, \tag{6.42}$$

where $x_I(y) \in \Gamma_I$. Similarly, let $x_O(y) \in \Gamma_O$. Integrating from $x_I(y)$ to $x_O(y)$, we have

$$\int_{x_I(y)}^{x_O(y)} u(x, y)p(x, y)\, dx = \int_{x_I(y)}^{x_O(y)} \left(\int_{x_I(y)}^{x} \frac{d}{ds} u(s, y)\, ds \right) p(x, y)\, dx. \tag{6.43}$$

Changing the order of integration yields

$$\int_{x_I(y)}^{x_O(y)} u(x, y)p(x, y)\, dx = \int_{x_I(y)}^{x_O(y)} \left(\int_{s}^{x_O(y)} p(x, y)\, dx \right) \frac{d}{ds} u(s, y)\, ds. \tag{6.44}$$

Define

$$q(x, y) = \int_{x}^{x_O(y)} p(\eta, y)\, d\eta, \tag{6.45}$$

which is in V since $q(x, y) = 0$ on Γ_O. Substituting into (6.44) results in

$$\int_{x_I(y)}^{x_O(y)} u(x, y)p(x, y)\, dx = \int_{x_I(y)}^{x_O(y)} q(x, y) \frac{d}{dx} u(x, y)\, dx. \tag{6.46}$$

Integrating in y from y_b to y_t, we arrive at

$$\int_{y_b}^{y_t} \int_{x_I(y)}^{x_O} u(x, y)p(x, y)\, dx\, dy = \int_{y_b}^{y_t} \int_{x_I(y)}^{x_O} q(x, y) \frac{\partial}{\partial x} u(x, y)\, dx\, dy, \tag{6.47}$$

which becomes

$$\begin{aligned} \langle u, p \rangle_{0,\Omega} &= \langle \frac{\partial u}{\partial x}, q \rangle_{0,\Omega} \\ &= \langle \nabla \cdot \mathbf{b}u, q \rangle_{0,\Omega}. \end{aligned} \tag{6.48}$$

We now show that there exists a constant c such that

$$c = \sup_{p \in V} \frac{\|\nabla q\|_{0,\Omega}}{\|\nabla p\|_{0,\Omega}}. \tag{6.49}$$

Notice that

$$-\partial_x q = p \tag{6.50}$$

$$\begin{aligned}\partial_y q &= \frac{\partial}{\partial y}\int_x^{x_O(y)} p(\eta, y)\, d\eta \\ &= \int_x^{x_O(y)} \frac{\partial p(\eta, y)}{\partial y}\, d\eta + \underbrace{p(x_O(y), y)}_{0} \frac{dx_O(y)}{dy}.\end{aligned} \tag{6.51}$$

The Poincaré inequality applies to give

$$\|\partial_x q\|_{0,\Omega} = \|p\|_{0,\Omega} \leq D(\Omega)\|\nabla p\|_{0,\Omega}, \tag{6.52}$$

where $D(\Omega)$ is the diameter of the domain, and Jensen's inequality yields

$$\|\partial_y q\|_{0,\Omega}^2 \leq (D(\Omega))^2 \|\frac{\partial p}{\partial y}\|_{0,\Omega}^2. \tag{6.53}$$

Thus, $c = \sqrt{2}D(\Omega)$.

Together with (6.48), we have

$$\begin{aligned}\|u\|_{-1,\Gamma_O,\Omega} &= \sup_{p\in V} \frac{\langle u, p\rangle_{0,\Omega}}{\|\nabla p\|_{0,\Omega}} \\ &= \sup_{p\in V} \left(\frac{\langle \nabla\cdot \mathbf{b}u, q\rangle_{0,\Omega}}{\|\nabla q\|_{0,\Omega}} \frac{\|\nabla q\|_{0,\Omega}|}{\|\nabla p\|_{0,\Omega}} \right) \\ &\leq \sup_{q\in V} \frac{|\langle \nabla\cdot \mathbf{b}u, q\rangle_{0,\Omega}|}{\|\nabla q\|_{0,\Omega}} \sup_{p\in V} \frac{\|\nabla q\|_{0,\Omega}|}{\|\nabla p\|_{0,\Omega}} \\ &\leq c\|\nabla\cdot \mathbf{b}u\|_{-1,\Gamma_O,\Omega}.\end{aligned} \tag{6.54}$$

This completes the proof. □

Corollary 6.9. *Let* $\mathbf{b}$ *satisfy the hypotheses of Lemma 6.8. Then there exists a constant* $c > 0$*, depending only on the diameter of the domain, such that*

$$\|u\|_{-1,\Gamma_O,\Omega} \leq c\|\nabla\cdot \mathbf{b}u\|_{-1,\Gamma_O,\Omega} \tag{6.55}$$

for $u \in U$*.*

Proof. The result follows immediately from Lemma 6.8 using a standard closure argument, since $U \cap H^1(\Omega)$ is dense in U. □

We now imbue U with an inner product. Define inverse Laplace operator $\Delta^{-1} : V' \to H^1(\Omega)$ by the following: $-\Delta^{-1}u = p \in V$ such that

$$\langle \nabla p, \nabla q \rangle_{0,\Omega} = \langle u, q \rangle_{0,\Omega} \quad \forall\, q \in V. \tag{6.56}$$

That is,

$$-\Delta p = u \quad \text{in } \Omega, \tag{6.57a}$$

$$\mathbf{n} \cdot \nabla p = 0 \quad \text{on } \Gamma_I \cup \Gamma_C, \tag{6.57b}$$

in an H^{-1} sense. Space U is characterized by the following lemma.

Lemma 6.10. *U in (6.35) is a Hilbert space with inner product*

$$\langle u, v \rangle_U = \langle \nabla \cdot \mathbf{b}u, -\Delta^{-1}\nabla \cdot \mathbf{b}v \rangle_{0,\Omega}. \tag{6.58}$$

Proof. We show that (6.58), which is symmetric, is an inner product on U that induces norm $\|u\|_U$. Since V is a Hilbert space, we use the Riesz representation theorem [13]: for $\langle u, \cdot \rangle_{0,\Omega} \in V'$, there exits a unique $p \in V$ such that

$$\langle \nabla p, \nabla q \rangle_{0,\Omega} = \langle u, q \rangle_{0,\Omega} \quad \forall\ q \in V. \tag{6.59}$$

From definition (6.36) of the U-norm and relation (6.59), we have

$$\begin{aligned} \|u\|_{-1,\Gamma_O,\Omega} &= \sup_{q \in V} \frac{\langle u, q \rangle_{0,\Omega}}{\|\nabla q\|_{0,\Omega}} \\ &= \sup_{q \in V} \frac{\langle \nabla p, \nabla q \rangle_{0,\Omega}}{\|\nabla q\|_{0,\Omega}} \\ &= \|\nabla p\|_{0,\Omega}. \end{aligned} \tag{6.60}$$

Using definition (6.56) of the inverse Laplace operator, we then have

$$\|u\|_{-1,\Gamma_O,\Omega} = \langle u, -\Delta^{-1}u \rangle_{0,\Omega}. \tag{6.61}$$

Inner product (6.58) and norm (6.36) on U then satisfy

$$\langle u, u \rangle_U = \|u\|_U^2. \tag{6.62}$$

With Corollary 6.9, we then conclude that U is a Hilbert space. ∎

We now show that the hypotheses of the Generalized Lax-Milgram Theorem (Theorem III.3.6, [13]) are satisfied.

Lemma 6.11. *Bilinear form $a(u,v) : U \times V \to \mathbb{R}$ in (6.39) satisfies the following:*

(Continuity). There exists $c_0 > 0$ such that

$$|a(u,v)| \leq c_0 \|u\|_U \|v\|_V \quad \forall\ u \in U,\ v \in V. \tag{A}$$

(Inf-sup condition). There exists $c_1 > 0$ such that

$$\sup_{v \in V} \frac{a(u,v)}{\|v\|_V} \geq c_1 \|u\|_U \quad \forall\ u \in U. \tag{B}$$

(Surjectivity). For every nontrivial $v \in V$, there exists $u \in U$ with

$$a(u,v) \neq 0. \tag{C}$$

Proof. Continuity follows easily. We have

$$\begin{aligned} |a(u,v)| &= |\langle \mathbf{b}u, \nabla v \rangle_{0,\Omega}| \cdot \frac{\|\nabla v\|_{0,\Omega}}{\|\nabla v\|_{0,\Omega}} \\ &\leq \sup_{v \in V} \frac{|\langle \mathbf{b}u, \nabla v \rangle_{0,\Omega}|}{\|\nabla v\|_{0,\Omega}} \cdot \|\nabla v\|_{0,\Omega} \\ &\leq \|\nabla \cdot \mathbf{b}u\|_{-1,\Gamma_O,\Omega} \|v\|_{1,\Omega} \\ &= \|u\|_U \|v\|_V, \end{aligned} \tag{6.63}$$

Assumption (A) is then satisfied with $c_0 = 1$.

The inf-sup condition follows similarly. By definition, we have

$$\begin{aligned} \sup_{v \in V} \frac{a(u,v)}{\|v\|_V} &= \sup_{v \in V} \frac{\langle \mathbf{b}u, \nabla v \rangle_{0,\Omega}}{\|\nabla v\|_{0,\Omega}} \\ &= \|u\|_U, \end{aligned} \tag{6.64}$$

which proves assumption (B) with $c_1 = 1$.

To prove (C), choose nontrivial $v \in V$. That is, $0 < \|v\|_{1,\Omega} < \infty$. Since $\mathbf{b} \in H(div, \Omega)$, we have $\mathbf{b}v \in H(div, \Omega)$. Furthermore, $\mathbf{b} \cdot \nabla v \in L^2(\Omega)$. Letting $u = \mathbf{b} \cdot \nabla v$, we have $u \in L^2(\Omega)$. Suppose

$$\|u\|_{0,\Omega}^2 = \|\mathbf{b} \cdot \nabla v\|_{0,\Omega}^2 = 0. \tag{6.65}$$

Then, by the Poincaré inequality, Lemma 4.4, we have that

$$\|v\|_{0,\Omega}^2 = 0, \tag{6.66}$$

which is a contradiction. Thus, $\|u\|_{0,\Omega}^2 > 0$ and

$$\begin{aligned} a(u,v) &= \langle \mathbf{b}u, \nabla v \rangle_{0,\Omega} \\ &= \|\mathbf{b}\cdot\nabla v\|_{0,\Omega}^2 = \|u\|_{0,\Omega}^2 > 0, \end{aligned} \tag{6.67}$$

which proves (C). □

We now have an isomorphism by the following theorem.

Theorem 6.12. *The linear mapping, $\mathcal{L} : U \to V'$, defined by (6.34) is an isomorphism. That is, given $f \in V'$, there exists a unique $u \in U$ such that $\mathcal{L}u = f$. Also,*

$$a(u,v) = \ell(v) \tag{6.68}$$

for all $v \in V$.

Proof. The result immediately follows from the Generalized Lax-Milgram Theorem (Theorem III.3.6, [13]) since its assumptions are satisfied by Lemmas 6.10 and 6.11 above. □

To investigate the importance of the above discussion in the context of our H^{-1} least-squares functional (6.25), consider nonlinear conservation law (6.1). Norm (6.36) can be written as

$$\begin{aligned} \|\nabla\cdot\mathbf{F}(u)\|_{-1,\Gamma_O,\Omega} &= \sup_{v\in V} \frac{\langle \mathbf{F}(u), \nabla v\rangle_{0,\Omega}}{\|\nabla v\|_{0,\Omega}} \\ &= \|\Pi_G \mathbf{F}(u)\|_{0,\Omega}, \end{aligned} \tag{6.69}$$

where Π_G is the L^2 orthogonal projection onto the space of gradients of V given by

$$G = \{\mathbf{w} \in L^2(\Omega)^2 \,:\, \mathbf{w} = \nabla p,\, p \in V\}. \tag{6.70}$$

We would like to minimize the norm of this projection, which is precisely the goal of solving least-squares minimization problem (6.26) presented in the previous section. The following lemma justifies this viewpoint.

Lemma 6.13. *Assume that $\Gamma_I \cup \Gamma_C$ is connected. Given $\mathbf{F}(u) \in L^2(\Omega)^2$, we have*

$$\|\nabla \cdot \mathbf{F}(u)\|_{-1,\Gamma_O,\Omega} = \inf_{\phi \in W} \|\nabla^\perp \phi - \mathbf{F}(u)\|_{0,\Omega}. \tag{6.71}$$

Proof. We begin with a Helmholtz decomposition [38]. For $\mathbf{F}(u) \in L^2(\Omega)^2$, there exist unique $p \in V$ and $q \in W$ such that

$$\mathbf{F}(u) = \nabla p + \nabla^\perp q. \tag{6.72}$$

We then have $\langle \nabla p, \nabla^\perp q \rangle_{0,\Omega} = 0$ and thus

$$\begin{aligned} \|\nabla \cdot \mathbf{F}(u)\|_{-1,\Gamma_O,\Omega} &= \sup_{v \in V} \frac{\langle \mathbf{F}(u), \nabla v \rangle_{0,\Omega}}{\|\nabla v\|_{0,\Omega}} \\ &= \sup_{v \in V} \frac{\langle \nabla p + \nabla^\perp q, \nabla v \rangle_{0,\Omega}}{\|\nabla v\|_{0,\Omega}} \\ &= \sup_{v \in V} \frac{\langle \nabla p, \nabla v \rangle_{0,\Omega}}{\|\nabla v\|_{0,\Omega}} \\ &= \|\nabla p\|_{0,\Omega}. \end{aligned} \tag{6.73}$$

But the right side of (6.71) also equals $\|\nabla p\|_{0,\Omega}$:

$$\begin{aligned} \inf_{\phi \in W} \|\nabla^\perp \phi - \mathbf{F}(u)\|_{0,\Omega}^2 &= \inf_{\phi \in W} \|\nabla^\perp \phi - \nabla p - \nabla^\perp q\|_{0,\Omega}^2 \\ &= \inf_{\phi \in W} \left(\|\nabla^\perp \phi - \nabla^\perp q\|_{0,\Omega}^2 + \|\nabla p\|_{0,\Omega}^2 \right) \\ &= \|\nabla p\|_{0,\Omega}^2, \end{aligned} \tag{6.74}$$

which completes the proof. □

We minimize (6.25) over discrete subspaces $\Phi^h \subset H^1(\Omega)$ and $\mathcal{U}^h \subset H^1(\Omega)$. Assume that we have suitable FE spaces $\mathcal{U}^h$ and $\mathit{\Phi}^h$ with the following approximation properties [18, 73]: there exist interpolants $\Pi^h u \in \mathcal{U}^h, \Pi^h \phi \in \mathit{\Phi}^h$, and constant c, independent of the mesh size h, such that

$$\|\nabla^\perp(\phi - \Pi^h \phi)\|_{0,\Omega} \leq c h^\beta \|\nabla^\perp \phi\|_{\beta,\Omega}, \tag{6.75a}$$

$$\|u - \Pi^h u\|_{0,\Omega} \leq c h^\beta \|u\|_{\beta,\Omega}, \tag{6.75b}$$

$$\|\nabla^\perp(\phi - \Pi^h \phi)\|_{0,\Gamma_I} \leq c h^\beta \|\nabla^\perp \phi\|_{\beta,\Gamma_I}, \tag{6.75c}$$

$$\|u - \Pi^h u\|_{0,\Gamma_I} \leq c h^\beta \|u\|_{\beta,\Gamma_I}. \tag{6.75d}$$

For example, standard bilinear finite elements on quadrilaterals fulfill these assumptions [73].

Assume that Φ^h is the finite element space of continuous piecewise polynomials of degree k as in (4.55). Let $\mathcal{U}^h \subset L^2(\Omega)$ be a finite element space that is piecewise continuous, such as piecewise constants or bilinears. Note that approximation property (6.75) holds. Further assume that boundary conditions (6.22b-6.22c) are directly imposed on the spaces. With this, we state and prove the following weak coercivity and continuity results with respect to the dual norm, (6.33), of $\nabla \cdot \mathbf{F}(u)$.

Theorem 6.14. *Let $u^h \in \mathcal{U}^h \subset U$. Then there exist constants c and ε such that $0 < \varepsilon \leq \beta < k$ and*

$$\|\nabla \cdot \mathbf{F}(u^h)\|_{-1,\Gamma_O,\Omega} \leq \inf_{\phi^h \in \Phi^h} \|\nabla^\perp \phi^h - \mathbf{F}(u^h)\|_{0,\Omega} \tag{6.76}$$

$$\inf_{\phi^h \in \Phi^h} \|\nabla^\perp \phi^h - \mathbf{F}(u^h)\|_{0,\Omega} \leq \|\nabla \cdot \mathbf{F}(u^h)\|_{-1,\Gamma_O,\Omega} + ch^\beta \|\phi_*(u^h)\|_{1+\beta}, \tag{6.77}$$

where

$$\phi_*(u^h) = \operatorname*{argmin}_{\phi \in W} \|\nabla^\perp \phi - \mathbf{F}(u^h)\|_{0,\Omega}. \tag{6.78}$$

Proof. Given $u^h \in \mathcal{U}^h$, the first inequality follows immediately from Lemma 6.13. For $\phi_*(u^h)$, we have

$$\begin{aligned} \inf_{\phi^h \in \Phi^h} \|\nabla^\perp \phi^h - \mathbf{F}(u^h)\|_{0,\Omega} &\leq \inf_{\phi^h \in \Phi^h} \left(\|\nabla^\perp \phi_* - \mathbf{F}(u^h)\|_{0,\Omega}^2 + \|\nabla^\perp \left(\phi_*(u^h) - \nabla^\perp \phi^h \right) \|_{0,\Omega}^2 \right) \\ &= \|\nabla \cdot \mathbf{F}(u^h)\|_{-1,\Gamma_O,\Omega}^2 + \inf_{\phi^h \in \Phi^h} \|\nabla^\perp \left(\phi_*(u^h) - \nabla^\perp \phi^h \right) \|_{0,\Omega}^2. \end{aligned} \tag{6.79}$$

By (6.75) we have

$$\inf_{\phi^h \in \Phi^h} \|\nabla^\perp \phi^h - \mathbf{F}(u^h)\|_{0,\Omega} \leq \|\nabla \cdot \mathbf{F}(u^h)\|_{-1,\Gamma_O,\Omega}^2 + ch^\beta \|\phi_*(u^h)\|_{1+\beta}, \tag{6.80}$$

which completes the proof. □

Remark 6.15. *Thus far we have assumed that $u = 0$ on Γ_I. In practice, however, we enforce this boundary condition weakly. The extension of this theory to weak boundary conditions is left to future work.*

Remark 6.16. *Although convergence in spaces such as U is useful in an abstract setting to formalize the convergence to a weak solution, it is not entirely convincing from a visual standpoint. Thus, in Section 6.4, we address convergence of u^h in $L^2(\Omega)$ for certain finite element spaces. We also discuss solution quality and L^2 convergence numerically in the next section.*

6.2 Numerical Results

Here, we present numerical results for the Gauss-Newton LSFEMs presented above on test problems for the Burgers equation that involve shocks and rarefaction waves. We investigate the solution quality in terms of smearing, oscillations, and overshoots and undershoots at discontinuities. We numerically study convergence in the L^2 sense of u^h to a function $\hat{u}$ and convergence of the nonlinear functionals, $\mathcal{F}(\mathbf{w}, u; g)$ defined in (6.28) and $\mathcal{G}(\phi, u; g)$ defined in (6.25). That is, we confirm that $\|u^h - u\|_{0,\Omega} \to 0$, $\mathcal{F} \to 0$, and $\mathcal{G} \to 0$ as $h \to 0$, and we determine the rates of convergence of each. Convergence properties of the Gauss-Newton procedure are also discussed.

For solutions with discontinuities, it is important to establish that u^h converges to a weak solution of (6.1), which implies that it has the correct shock speed. For problems with non-unique weak solutions—e.g., rarefaction wave problems—we study whether u^h converges to the so-called weak entropy solution, which is the unique weak solution that is stable against arbitrarily small perturbations. This is also the weak solution that satisfies an entropy inequality and can be obtained as the vanishing viscosity limit of a parabolic regularization of (6.1) with a viscosity term, as discussed in Chapter 2. Finally, we show that adaptive refinement, based on the LS error estimator, is an effective mechanism to counter smearing at shocks and we discuss adaptivity on space-time domains. We combine adaptivity with grid continuation for the Gauss-Newton procedure and show numerically that one Newton iteration is sufficient on each grid level to obtain convergence up to discretization error.

Denote by α the rate of convergence of the squared L^2 error. That is,

$$\|u^h - u\|_{0,\Omega}^2 \approx \mathcal{O}(h^\alpha). \tag{6.81}$$

Convergence rate α can be approximated between successive levels of refinement by

$$\frac{\|u^h - u\|^2}{\|u^{2h} - u\|^2} \approx \left(\frac{1}{2}\right)^\alpha. \tag{6.82}$$

Convergence of nonlinear functionals $\mathcal{F}(\mathbf{w}^h, u^h; g)$ and $\mathcal{G}(\phi^h, u^h; g)$ is quantified in a similar way. The convergence rates between grid levels are reported throughout this section and we discuss apparent trends toward an asymptotic rate. In Chapter 4.5, the rate of convergence was reported for the norm, while the rates in this Chapter are for the norm squared. Recall that the theoretical optimal convergence rate for a numerical approximation by continuous finite element spaces to a discontinuous solution in $H^{\frac{1}{2}-\varepsilon}(\Omega)$ is $\frac{1}{2}$ in the L^2-norm. This corresponds to a rate of 1.0 in the L^2-norm squared, which is what we obtain in our numerical tests. Throughout this section, we denote by N the number of finite elements in each coordinate direction and, thus, $h = \frac{1}{N}$.

The nonlinear strategy we use is a Gauss-Newton-type method [32] in a grid continuation setting. A Gauss-Newton method first linearizes the system of equations, (6.20) and (6.22), then forms the linearized least-squares functional. The nonlinear problem is solved in this manner on the coarsest level and the solution is then interpolated to the next finer grid level (which may be locally refined) as an initial guess to solve the nonlinear problem there. This scheme continues until the finest level has been processed. The least-squares problem on each level is solved using AMG for each Newton iteration. This "nested iteration" strategy is closely related to the full multigrid method [20]. Figure 6.2 illustrates the Gauss-Newton approach in this multilevel context. The thick lines indicate interpolation of the solution to be used as an initial guess in the Newton iteration at the finer level. The Newton method is marked by •, which may include many iterations, while the thin lines represent intergrid transfers in the AMG algorithm. Note that the thin lines simply illustrate visits to coarser grids and do not reflect the actual AMG cycling strategy. For example, we employ W-cycles (see Figure 5.1) for the results in this

section. These different phases of the nonlinear process are described in more detail in [28].

For our problem, we let $\mathcal{L}(\phi, u) = 0$ denote the system of equations, (6.22). We recall the Newton procedure

$$\mathcal{L}(\phi_i, u_i) + d\mathcal{L}(\phi_i, u_i)[(\phi_{i+1} - \phi_i, u_{i+1} - u_i)] = 0, \tag{6.83}$$

where $d\mathcal{L}$, given by (6.23), is the Fréchet derivative of $\mathcal{L}$.

6.2.1 $H(div)$ formulation

We begin by presenting numerical results for the $H(div)$ LSFEM, described in Problem 6.5, applied to the inviscid Burgers equation, for which $\mathbf{F}(u) = (\frac{u^2}{2}, u)$ in (6.1). We consider the following model flow:

Example 5 (Single Shock # 1). *The space-time flow domain is given by* $\Omega = [0, 1] \times [0, 1]$, *with initial and boundary conditions*

$$u(x,t) = \begin{cases} 0.5 & \text{if } t = 0, \\ \\ 1.0 & \text{if } x = 0. \end{cases} \tag{6.84}$$

The unique weak solution of this problem consists of a shock propagating with shock speed $s = \frac{3}{4}$ *from the origin,* $(x, t) = (0, 0)$. *The conserved quantity is* $u(x, t) = 1.0$ *to the left of the shock and* $u(x, t) = 0.5$ *to the right.*

Figure 6.3 shows contours of the numerical solution, u^h, to Example 5 on grids with decreasing h. The correct shock speed is obtained and the solution does not show excessive spurious oscillations. Small overshoots and undershoots appear to be generated where the shock interacts with the outflow boundary. In our globally coupled space-time solution, these slight oscillations seem to propagate in the direction of the characteristic curves, in accordance with the signal propagation properties of hyperbolic PDEs. These effects are reduced as the grid is refined. Figure 6.4 shows the solution profile of u^h, which illustrates the well-known fact that LS methods introduce substantial smearing at shocks [11, 9, 50, 31]. However, Table 6.1 shows

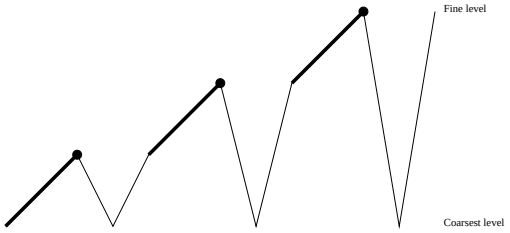


Figure 6.2: The Gauss-Newton, grid continuation, and AMG nonlinear process.

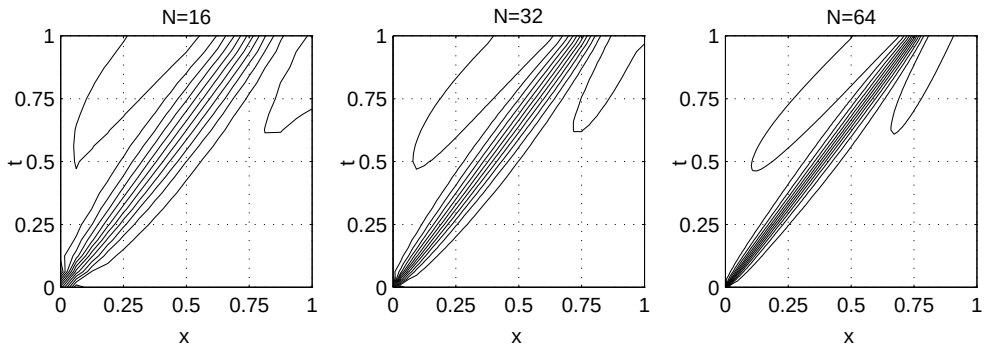


Figure 6.3: $H(div)$ formulation, Example 5: u^h contours on grids with 16^2, 32^2, and 64^2 quadrilateral elements.

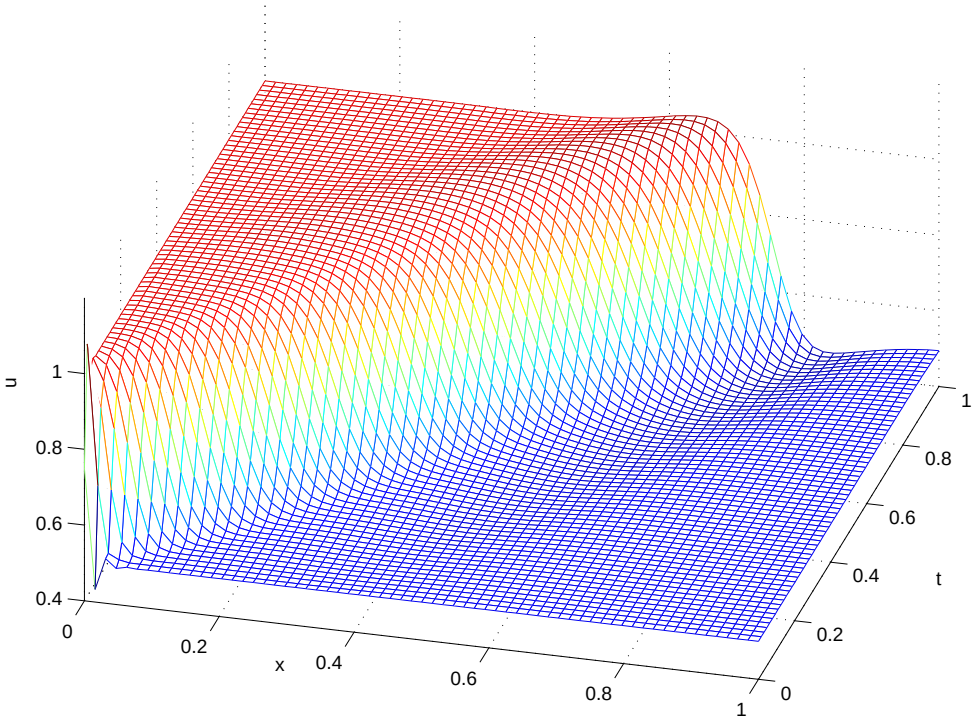
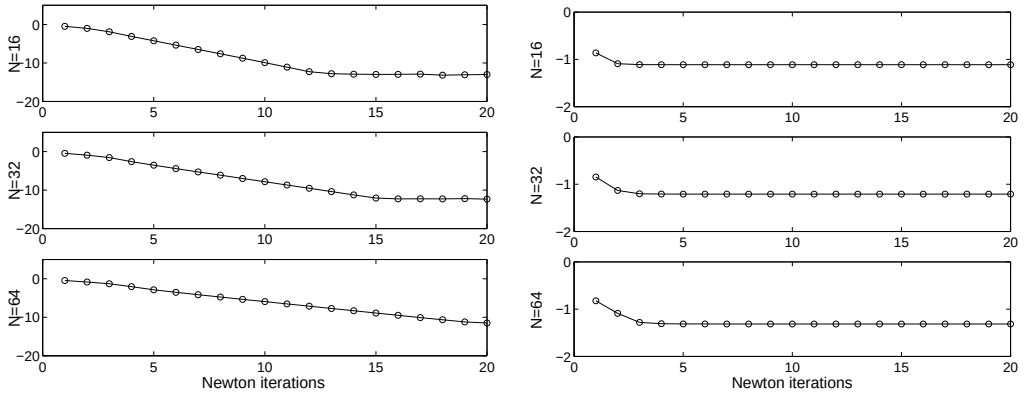


Figure 6.4: $H(div)$ formulation, Example 5: u^h profile on a grid with 32^2 quadrilateral elements.



(a) update convergence

(b) error convergence

Figure 6.5: $H(div)$ formulation, Example 5: Log error versus Newton iterations. Left: $\|u^h_{i+1} - u^h_i\|^2_{0,\Omega}$ Newton update convergence. Linear convergence can be observed. Right: $\|u^h - u\|^2_{0,\Omega}$ error convergence. Discretization error is reached after few Newton iterations.

that both the overshoots and undershoots, and the smearing, disappear in the L^2 sense as the grid is refined. The numerical approximation converges to exact solution u as $h \to 0$, and the rate of convergence that appears to approach $\alpha = 1.0$.

Table 6.1 also shows that the nonlinear functional, $\mathcal{F}(\mathbf{w}^h, u^h; g)$ in (6.28), converges as $h \to 0$, with α approaching 1.0. The left panel of Figure 6.5 shows that the Gauss-Newton method converges linearly on each grid, in accordance with theory [32]. Quadratic Newton convergence can be obtained if a full Newton procedure is employed instead of the Gauss-Newton approach [32]. The right panel of Figure 6.5 shows that the discretization error on each grid level is reached after only a few Newton iterations, which indicates that grid continuation strategies require only a few Newton iterations per grid level, as illustrated below.

Figure 6.6 shows that $\nabla \cdot \mathbf{w}^h$ does not vanish exactly in a discrete sense for our $H(div)$ LSFEM, which illustrates that our method does not impose discrete exact conservation in the sense of Lax-Wendroff [58]. Note, however, that $\nabla \cdot \mathbf{w}^h$ is very small; the scale on Figure 6.6 is $\mathcal{O}(10^{-3})$ and convergence of nonlinear functional $\mathcal{F}(\mathbf{w}^h, u^h; g)$ implies that $\|\nabla \cdot \mathbf{w}^h\|_{0,\Omega} \to 0$ as $h \to 0$ at the optimal rate. Note also that $\nabla \cdot \mathbf{w}^h$ is constant in each cell since the RT_0 vector finite elements are used for flux variable $\mathbf{w}$.

6.2.2 H^{-1} formulation

We now focus on the H^{-1} formulation, given by (6.22), for the remainder of our numerical experiments. The use of standard bilinear finite element spaces make this formulation particularly attractive. Nearly identical numerical results are obtained as for $H(div)$ formulation (6.20), as expected, since the two approaches are closely related: formulation (6.22) is the restriction of $\mathbf{w}$ in (6.20) to a divergence-free subspace. We first confirm this for Example 5, which is also considered for this formulation in the context of local adaptive refinement later in this section.

Figure 6.7 shows contours of the numerical approximation on various grid levels. As in the first formulation considered, spurious oscillations, overshoots, and undershoots diminish in

N	$\|\cdot\|^2_{0,\Omega}$	α	$\mathcal{F}$	α
16	5.96e-3		1.89e-2	
		0.58		1.03
32	3.81e-3		9.25e-3	
		0.69		1.02
64	2.36e-3		4.56e-3	
		0.77		1.01
128	1.38e-3		2.26e-3	
		0.85		1.01
256	7.66e-4		1.12e-3	

Table 6.1: $H(div)$ formulation, Example 5: convergence rates.

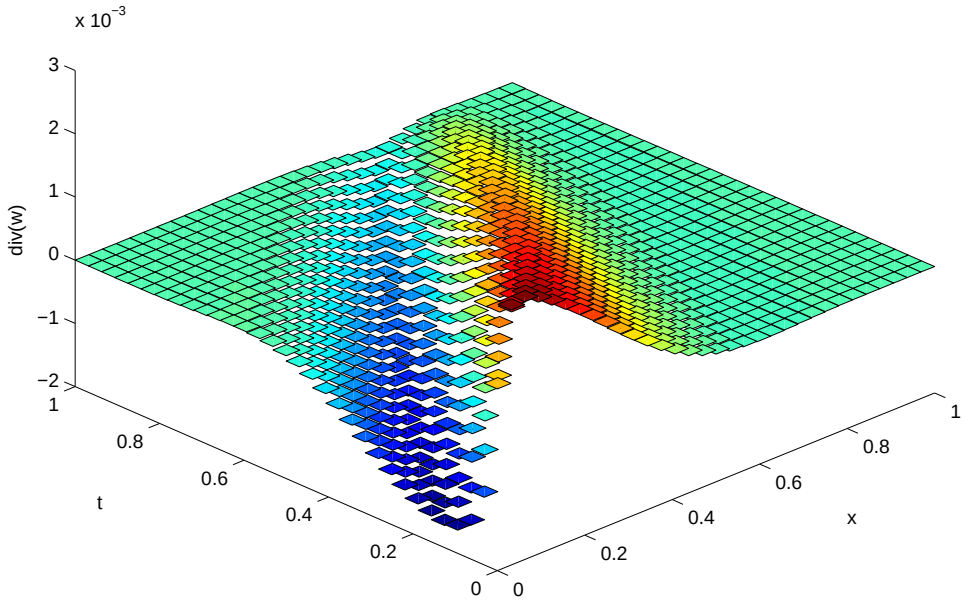


Figure 6.6: $H(div)$ formulation, Example 5: $\nabla \cdot \mathbf{w}^h$ on a grid with 32^2 quadrilateral elements.

the L^2-norm as the grid is refined. The solution profile is nearly identical to the profile obtained with the $\nabla \cdot \mathbf{w}$ formulation, as Figure 6.8 illustrates. Table 6.2 shows that the asymptotic convergence rate is approximately 1.0 for both the L^2-norm squared and the functional. It is interesting to note that the convergence in L^2 improves as the grid is refined, while functional convergence starts off greater than the asymptotic limit, especially in the interior. We denote by $\mathcal{G}_{\text{int}}$ and $\mathcal{G}_{\text{bdy}}$ the interior and boundary terms of the functional, (6.25). That is,

$$\mathcal{G}_{\text{int}}(\phi, u; g) = \|\nabla^{\perp}\phi - \mathbf{F}(u)\|^2_{0,\Omega}, \tag{6.85}$$

$$\mathcal{G}_{\text{bdy}}(\phi, u; g) = \|u - g\|^2_{0,\Gamma_I} + \|\mathbf{n} \cdot (\nabla^{\perp}\phi - \mathbf{F}(g))\|^2_{0,\Gamma_I}. \tag{6.86}$$

Considering the functional in this manner is important since it verifies that we are properly weighting each term. Also, the interior term is particularly important since shocks can emerge while the boundary data is smooth. We thus study the convergence of each term separately.

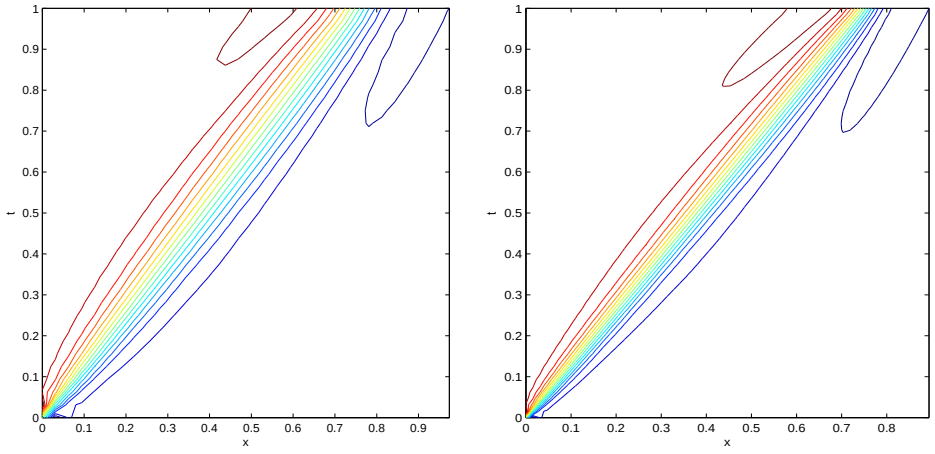
Figures 6.9 and 6.10 show that the nonlinear grid continuation strategy is very efficient: nonlinear convergence can be reached with only one or two Newton iterations per grid level. We compare the error in the L^2-norm squared as well as the functional and find only a small advantage in using more than one Newton iteration. The use of one Newton step per grid level is obviously the most efficient approach.

We now consider Example 5, slightly modified, so that the characteristics are aligned with the grid in a portion of the domain.

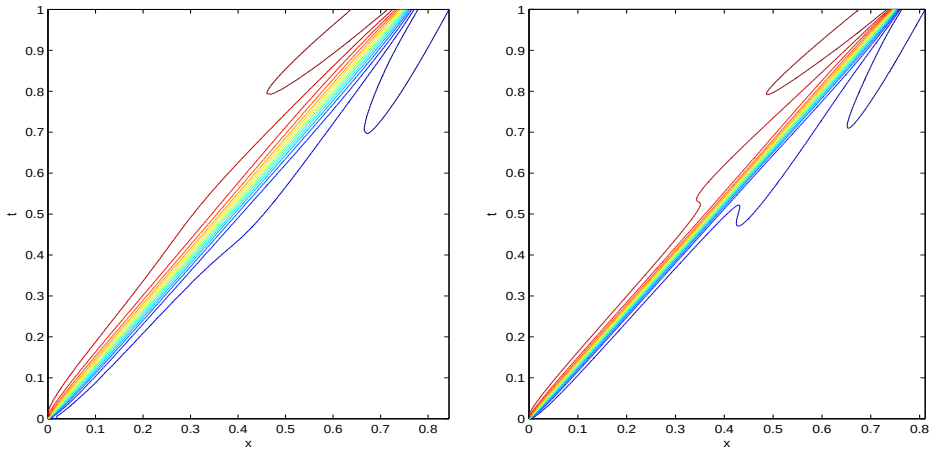
Example 6 (Single Shock # 2). *The space-time flow domain is given by* $\Omega = [0,1] \times [0,1]$, *with initial and boundary conditions*

$$u(x,t) = \begin{cases} 0.0 & \text{if } t = 0, \\ 1.0 & \text{if } x = 0. \end{cases} \tag{6.87}$$

The unique weak solution of this problem contains a shock propagating with shock speed $s = \frac{1}{2}$ *from the origin,* $(x,t) = (0,0)$. *The conserved quantity is* $u(x,t) = 1.0$ *to the left of the shock and* $u(x,t) = 0.0$ *to the right. Characteristics emanating from the* x*-axis are aligned with the grid in the* y*-direction.*



(a) N=32 (b) N=64



(c) N=128 (d) N=256

Figure 6.7: H^{-1} formulation, Example 5: u^h contours on grids with 32^2, 64^2, 128^2, and 256^2 quadrilateral elements.

N	$\|\cdot\|^2_{0,\Omega}$	α	$\|\cdot\|^2_{\mathcal{G}}$	α	$\|\cdot\|^2_{\mathcal{G}_{\text{int}}}$	α	$\|\cdot\|^2_{\mathcal{G}_{\text{bdy}}}$	α
16	5.46e-3		2.82e-3		5.61e-4		2.26e-3	
		0.61		1.06		1.30		1.00
32	3.57e-3		1.36e-3		2.28e-4		1.13e-3	
		0.71		1.04		1.21		1.00
64	2.19e-3		6.63e-4		9.84e-5		5.64e-4	
		0.80		1.02		1.13		1.00
128	1.25e-3		3.27e-4		4.50e-5		2.82e-4	
		0.90		1.01		1.06		1.00
256	6.72e-4		1.63e-4		2.16e-5		1.41e-4	

Table 6.2: H^{-1} formulation, Example 5: convergence rates.

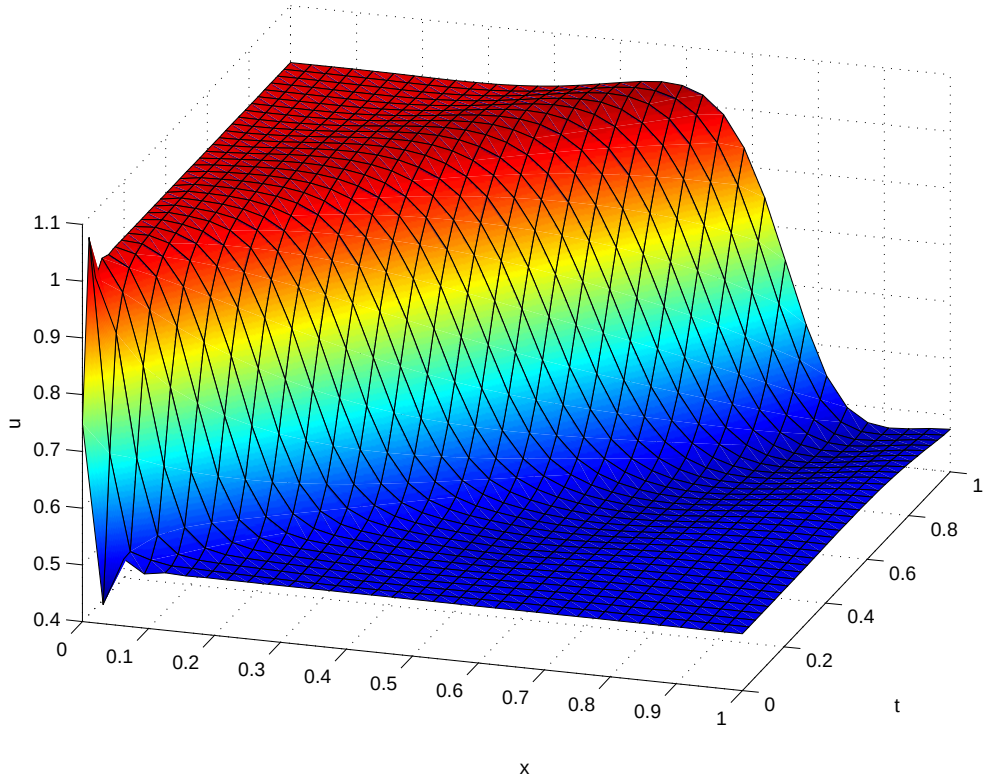


Figure 6.8: H^{-1} formulation, Example 5: u^h profile on a grid with 32^2 quadrilateral elements.

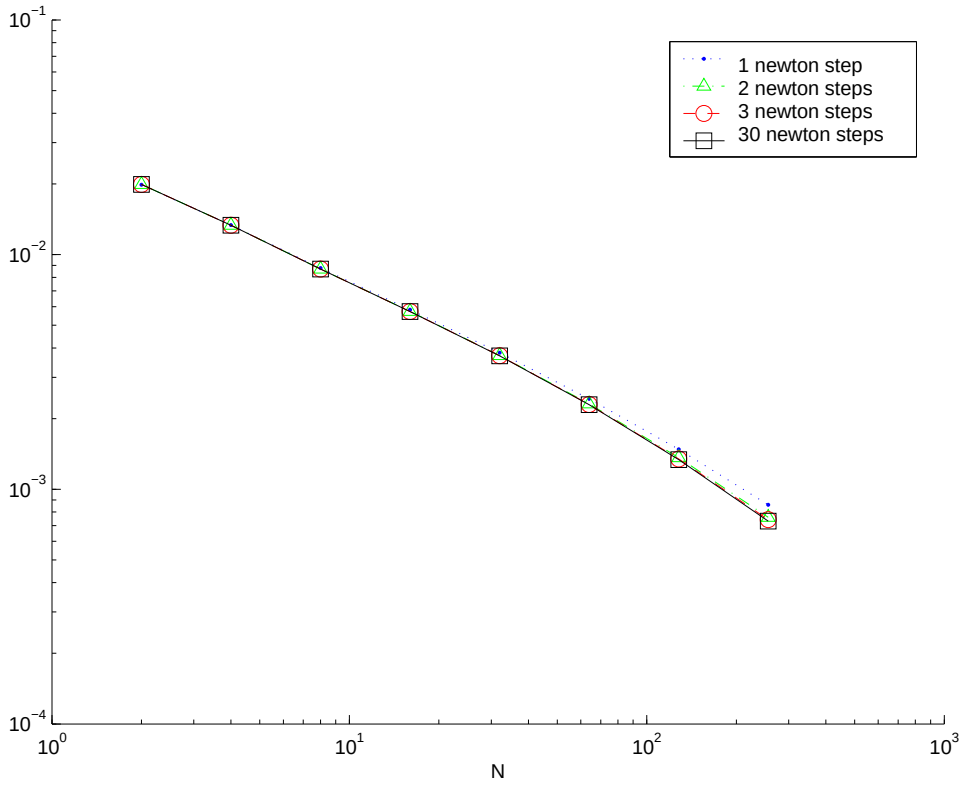


Figure 6.9: H^{-1} formulation, Example 5: $\|u^h - u\|^2_{0,\Omega}$ error convergence on $N \times N$ grids of size $N = 2^k$, where $k = 2, \ldots, 8$, with grid continuation. Results are presented for 1, 2, 3 and 30 Newton steps per grid level.

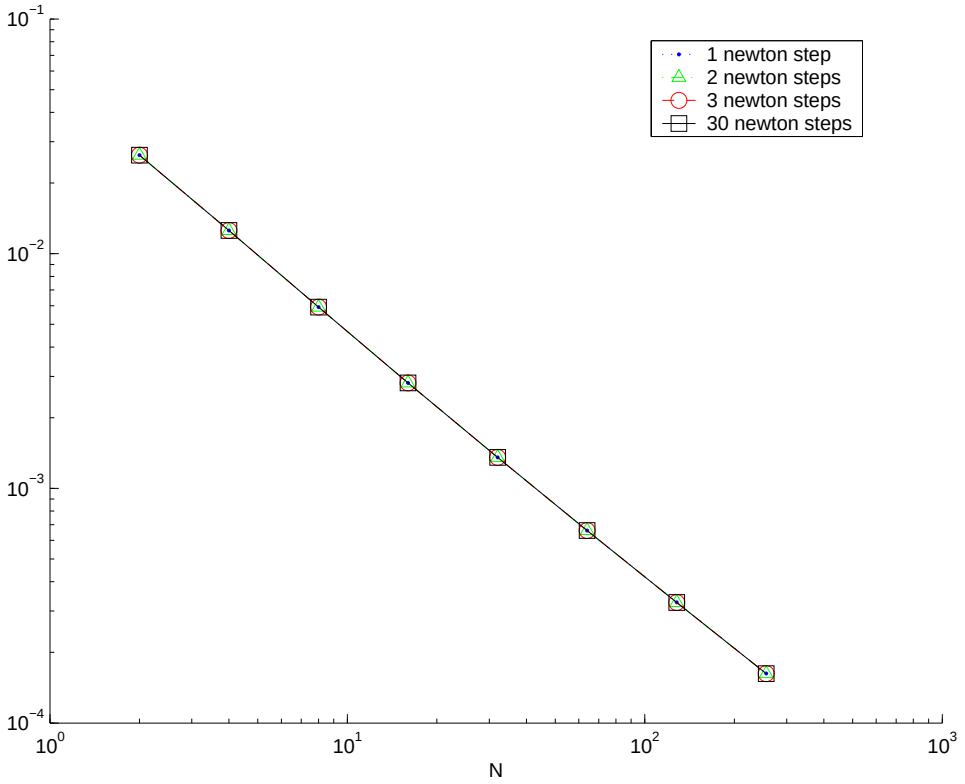


Figure 6.10: H^{-1} formulation, Example 5: functional convergence on $N \times N$ grids of size $N = 2^k$, where $k = 2, \ldots, 8$.

Table 6.3 shows results that are similar to those of Example 5 while using bilinears for both ϕ^h and u^h. Convergence rates approach 1.0 in the L^2-norm squared for the functional. Moreover, the rates are evenly balanced between the L^2-norm squared and the interior functional. Figure 6.11 confirms that the solution profile has similar qualities as the numerical approximation for Example 5 (see Figure 6.7) and also attains the correct shocks speed, $s = \frac{1}{2}$.

Next, consider a problem with two shocks merging into one. This numerical example helps demonstrate the robustness of our LSFEM and again validates our claims of convergence to a weak solution. It also confirms the rates of convergence that we determined numerically for Examples 5 and 6.

Example 7 (Double Shock). *The space-time flow domain is given by* $\Omega = [0, 2] \times [0, 1]$ *with the following initial and boundary conditions*

$$u(x,t) = \begin{cases} 1.5 & \text{if } t = 0,\ x < 0.5, \\ 0.5 & \text{if } t = 0,\ x > 0.5, \\ 2.5 & \text{if } x = 0. \end{cases} \tag{6.88}$$

The boundary data result in two shocks that merge into one shock. The first two shocks emanate from $(x,t) = (0,0)$ *and* $(x,t) = (0,0.5)$ *and travel at speeds* $s = 2$ *and* $s = 1$*, respectively. The shocks merge at* $(x,t) = (1,0.5)$*, and the resulting shock exits the domain at* $(x,t) = (1.75,1)$ *with shock speed* $s = \frac{3}{2}$*.*

Table 6.4 shows convergence of the error in the L^2-norm squared and convergence of the functional, i.e., $\|u^h - u\|^2_{0,\Omega} \to 0$ and $\mathcal{G}(\phi^h, u^h; 0) \to 0$ as $h \to 0$. Again, we find that the convergence rates, α, approach 1.0. This agrees with our findings for the single shock case. Figure 6.12 confirms convergence to the correct weak solution. The shocks merge at the correct location (indicated by the dotted lines) and the resulting single shock exits at the correct location. The contour lines reveal solution quality similar to the single shock case. Spurious oscillations, overshoots, and undershoots are minimal. However, overshoots and undershoots are slightly more pronounced at the shock location on the outflow boundary. Again, these effects

N	$\|\cdot\|^2_{0,\Omega}$	α	$\|\cdot\|^2_{\mathcal{G}}$	α	$\|\cdot\|^2_{\mathcal{G}_{\text{int}}}$	α	$\|\cdot\|^2_{\mathcal{G}_{\text{bdy}}}$	α
16	1.65e-2		1.21e-2		3.04e-3		9.08e-3	
		0.82		1.03		1.10		1.00
32	9.32e-3		5.94e-3		1.42e-3		4.53e-3	
		0.89		1.01		1.04		1.00
64	5.02e-3		2.95e-3		6.90e-4		2.26e-4	
		0.93		1.00		1.01		1.00
128	2.62e-3		1.47e-3		3.43e-4		1.13e-4	
		0.94		1.00		1.00		1.00
256	1.37e-3		7.36e-4		1.72e-4		5.64e-5	

Table 6.3: H^{-1} formulation, Example 6: convergence rates.

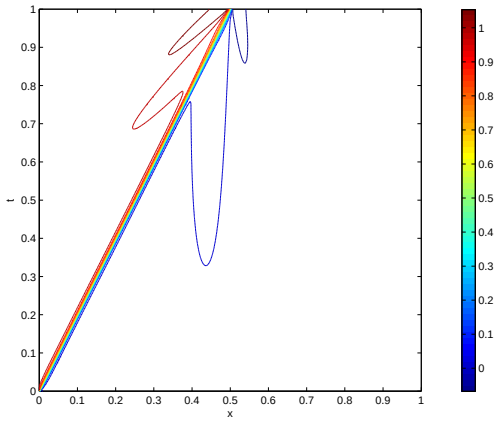


Figure 6.11: H^{-1} formulation, Example 6: u^h contours on a grid of 256^2 quadrilateral elements.

are reduced in the L^2-norm as the grid is refined.

We now consider a rarefaction problem.

Example 8 (Transonic Rarefaction). *Consider the space-time flow domain given by* $\Omega = [-1.0, 1.5] \times [0.0, 1.0]$ *with the following initial and boundary conditions*

$$u(x,t) = \begin{cases} -0.5 & \text{if } t = 0, x < 0; \\ 1.0 & \text{if } t = 0, x \geq 0; \\ -0.5 & \text{if } x = -1. \end{cases} \tag{6.89}$$

In this example, the initial discontinuity at the origin rarefies in time. There are infinitely many weak solutions to Example 8, but the unique entropy solution is a rarefaction wave. For given time t_c, $u(x, t_c)$ *increases linearly from* $u = -0.5$ *to* $u = 1.0$, *between the straight characteristic lines* $t = -\frac{x}{2}$ *and* $t = x$.

This rarefaction is called transonic [62] because the characteristic velocity, $f'(u) = u$, transits through zero within the rarefaction. Many higher-order numerical schemes that are based on approximate Riemann solvers or, equivalently, upwind or numerical dissipation ideas, fail to obtain the entropy weak solution for transonic rarefactions. Higher-order approximate Riemann solvers often do not capture the transonic rarefaction wave at cell interfaces appropriately and so-called entropy fixes are necessary, for example, the Roe scheme [62]. Therefore, it is important to demonstrate that our LSFEM schemes obtain the entropy solution in the case of transonic rarefactions.

Figure 6.13 shows the u^h solution profile for the transonic rarefaction case. The entropy solution is obtained. Table 6.5 shows that the squared L^2 error, $\|u^h - u\|^2_{0,\Omega}$, and nonlinear functional $\mathcal{G}(\phi, u; g)$ in (6.25) converge to zero as $h \to 0$. The convergence rates for the squared L^2 error and for the interior functional appear to be approaching $\alpha = 1.5$ and $\alpha = 2.0$, respectively. Higher convergence rates are obtained because the solution is no longer discontinuous in the interior, although it is in $H^{\frac{1}{2}-\varepsilon}$ on the boundary.

N	$\|\cdot\|^2_{0,\Omega}$	α	$\mathcal{G}$	α	$\mathcal{G}_{\text{int}}$	α	$\mathcal{G}_{\text{bdy}}$	α
16	2.50e-1		6.26e-2		3.06e-2		3.09e-2	
		0.82		1.10		1.19		1.02
32	1.42e-1		2.87e-2		1.34e-2		1.52e-2	
		0.86		1.05		1.10		1.01
64	7.82e-2		1.38e-2		6.27e-3		7.57e-3	
		0.91		1.02		1.03		1.00
128	4.17e-2		6.85e-3		3.07e-3		3.77e-3	
		0.93		1.00		1.00		1.00
256	2.19e-2		3.42e-3		1.54e-4		1.88e-3	

Table 6.4: H^{-1} formulation, Example 7: convergence rates.

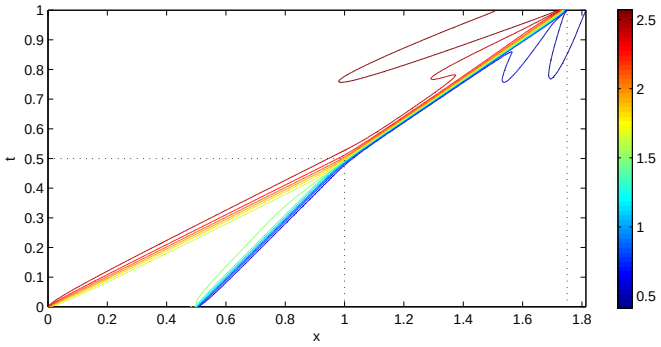
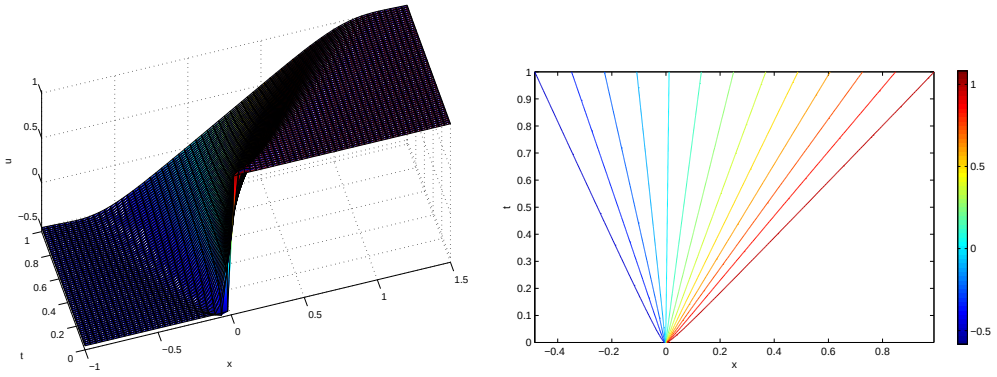


Figure 6.12: H^{-1} formulation, Example 7: u^h contours on a grid of 256^2 quadrilateral elements.



(a) Solution profile.

(b) Solution contours.

Figure 6.13: H^{-1} formulation, Example 8: u^h solution on a grid with 64^2 quadrilateral elements.

Finally, we return to Example 5 to investigate the effectiveness of local adaptive refinement. Adaptive refinement is applied in an FMG framework along with the grid continuation strategy outlined at the beginning of the section. We use the nonlinear least-squares functional as the **a posteriori** error estimator (see (3.19-3.20)). Although the least-squares error estimator is developed in a linear setting, we use nonlinear functional (6.25), since Newton convergence is achieved on each grid level. Convergence of Newton's method implies that the value of the functional of the linearized equations closely approximates the full nonlinear functional, and we are thus able to apply the general strategies developed in [7]. In all of our test cases, we found the functional based on the linearized equations and the nonlinear functional to yield identical values at convergence.

An element is marked for refinement if the functional density over an element is greater than a fixed threshold of the functional density over the full domain. The functional density is the functional value over an element divided by the area. For this test we use a threshold of 0.3. That is, element τ is refined if

$$\text{functional density on } \tau \geq 0.3 * \text{total functional density.} \tag{6.90}$$

Adaptive refinement based on a density argument is inexpensive, straightforward to implement, and amenable to parallelism. Although we achieve good performance using this approach, a detailed study of the adaptive algorithm is beyond the scope of this dissertation. A perhaps more optimal strategy is outlined in [7].

Figure 6.14 shows that the adaptive solution strategy effectively identifies the shock location, thus resolving the least-squares smearing. Moreover, this is done with limited overshoots and undershoots. Table 6.6 shows that squared L^2 error $\|u^h - u\|^2_{0,\Omega}$ and nonlinear functional $\mathcal{G}(\psi^h, u^h; g)$ in (6.25) converge to zero as $h \to 0$. The convergence rates, α, are asymptotically 1.0 in both cases, with the L^2-norm approaching this value from below and the functional norm from above. Table 6.6 also shows that the convergence rates and error magnitudes for the locally refined grids are comparable to those of the uniformly refined grids. Comparing column 2 in

the top and bottom sections of the table reveals that the locally refined grids are using many fewer nodes than the uniformly refined grids, while achieving similar accuracy. Surprisingly, the L^2 error is slightly less for the locally adapted grids. The functional does not show this since we are obtaining the best approximation in the functional norm on a given grid. Fewer basis functions (local refinement) would not yield a better approximation, although it is nearly the same in this case. Figure 6.15 shows that the numerical approximation is improved using fewer nodes as the grid is refined. Figure 6.16 shows a more detailed view of the number of nodes used in the adaptive algorithm versus the uniform refinement algorithm. In this figure, we first present a direct comparison at each grid level in the left figure. A comparison is also presented in the right figure, where the ratios of the number of nodes used in locally refined grids to the number of nodes used in uniformly refined grids at each level are given by $\Diamond$. Notice that the ratios decrease as the grid is refined. This is another indicator of the success of the adaptive approach. Moreover, if we compare the total number of nodes used up to some grid level k, we find a similar decrease in ratios (see line $\blacktriangledown$). This is also true for the total number of nodes used in the adaptive algorithm up to grid level k compared with the number of nodes used in the uniform algorithm **at** grid level k (see line $\triangledown$). This shows that we are indeed attaining the same level of accuracy with a decreasing number of points as the grids are refined.

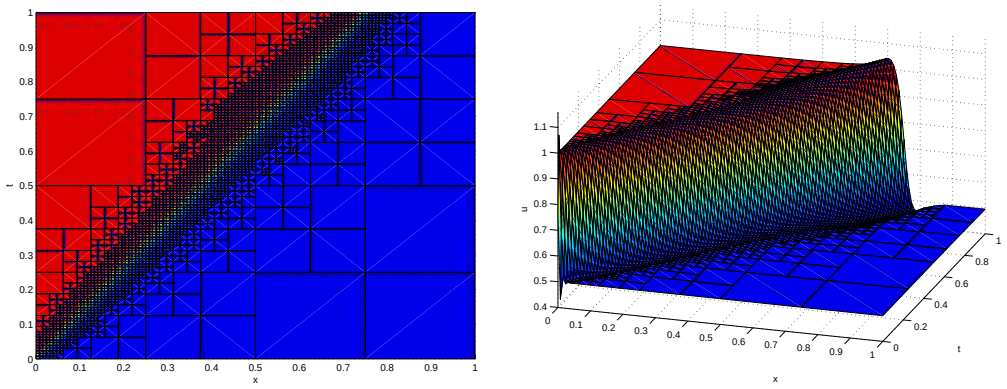
6.2.3 Discussion

In conclusion, we have shown numerically that the new LSFEMs proposed in the previous section do not exhibit large overshoots and undershoots, and that excessive oscillations are not introduced.

Because of the undershoots and overshoots, our methods are not fully monotone. The monotonicity property is often taken as a guideline for the design of numerical schemes for hyperbolic conservation laws because solutions of scalar hyperbolic equations with a convex flux function have a monotonicity property, as is discussed in Chapter 2. However, monotonicity is not a strict criterion for numerical approximation schemes, as long as the overshoots and

N	$\|\cdot\|^2_{0,\Omega}$	α	$\mathcal{G}$	α	$\mathcal{G}_{\text{int}}$	α	$\mathcal{G}_{\text{bdy}}$	α
16	1.27e-2		4.56e-2		4.99e-3		4.07e-2	
		1.41		1.06		1.69		1.00
32	4.79e-3		2.19e-2		1.55e-3		2.03e-2	
		1.46		1.04		1.73		1.00
64	1.74e-3		1.06e-2		4.66e-4		1.02e-2	
		1.49		1.03		1.77		1.00
128	6.21e-4		5.20e-3		1.37e-4		5.07e-3	
		1.50		1.02		1.80		1.00
256	2.20e-4		2.58e-3		3.94e-5		2.44e-3	

Table 6.5: H^{-1} formulation, Example 8: convergence rates.



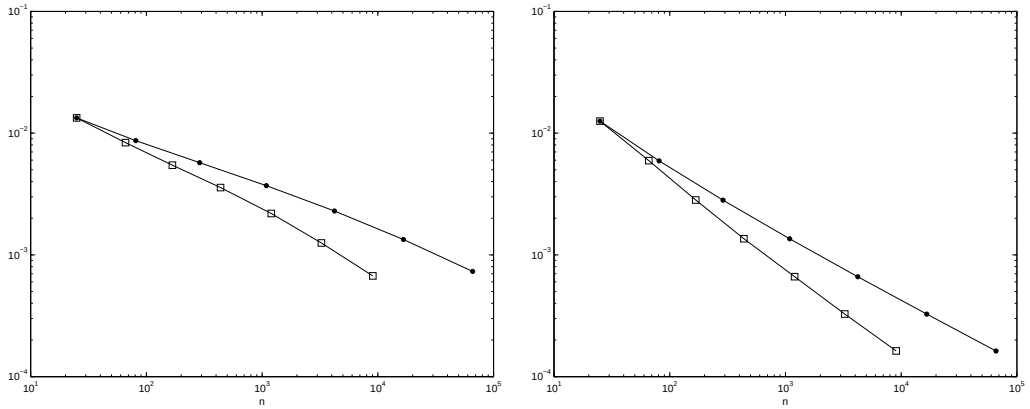
(a) 2D

(b) 3D

Figure 6.14: H^{-1} formulation, Example 5: solution u^h on an locally refined grid with a finest resolution of $h = \frac{1}{128}$.

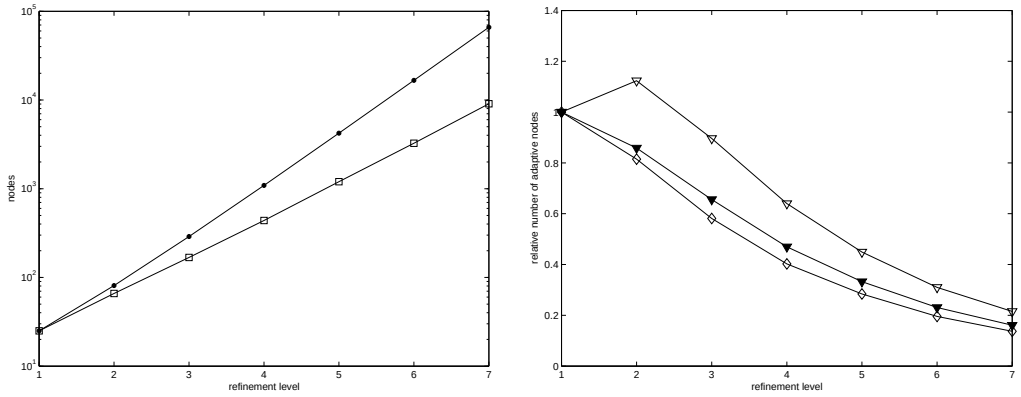
N	Nodes	$\|\cdot\|^2_{0,\Omega}$	α	$\mathcal{G}$	α	$\mathcal{G}_{\text{int}}$	α	$\mathcal{G}_{\text{bdy}}$	α
4	25	1.33e-2		1.26e-2		3.41e-3		9.15e-3	
			0.62		1.08		1.30		1.01
8	81	8.68e-3		5.92e-3		1.38e-3		4.54e-3	
			0.60		1.07		1.33		1.01
16	289	5.72e-3		2.82e-3		5.50e-4		2.26e-3	
			0.63		1.05		1.29		1.00
32	1089	3.70e-3		1.35e-3		2.25e-4		1.13e-3	
			0.69		1.03		1.22		1.00
64	4225	2.30e-3		6.61e-4		9.68e-5		5.64e-4	
			0.78		1.02		1.13		1.00
128	16641	1.34e-3		3.26e-4		4.43e-5		2.82e-4	
			0.87		1.01		1.06		1.00
256	66049	7.32e-4		1.62e-4		2.13e-5		1.41e-4	
4	25	1.33e-2		1.26e-2		3.41e-3		9.15e-3	
			0.67		1.08		1.27		1.01
8	66	8.36e-3		5.95e-3		1.41e-3		4.54e-3	
			0.62		1.08		1.33		1.01
16	168	5.46e-3		2.82e-3		5.61e-4		2.26e-3	
			0.61		1.06		1.30		1.00
32	438	3.57e-3		1.36e-3		2.28e-4		1.13e-3	
			0.71		1.04		1.21		1.00
64	1200	2.19e-3		6.63e-4		9.84e-5		5.64e-4	
			0.80		1.02		1.13		1.00
128	3258	1.25e-3		3.27e-4		4.50e-5		2.82e-4	
			0.90		1.01		1.06		1.00
256	9058	6.72e-4		1.63e-4		2.16e-5		1.41e-4	

Table 6.6: H^{-1} formulation, Example 5: convergence rates. Top section: uniform grid refinement. Bottom section: local grid refinement. Here, N is the size of the finest element with $h = \frac{1}{N}$.



(a) $\|\cdot\|^2_{0,\Omega}$ (b) Functional

Figure 6.15: H^{-1} formulation, Example 5: error versus nodes for locally refined and uniformly refined grids. • corresponds to uniform refinement and □ to local refinement.



(a) per level (b) composite

Figure 6.16: H^{-1} formulation, Example 5: Node usage on the local grids compared with the uniform grid. Left: Direct comparison of the nodes used at each level. • corresponds to uniform refinement and □ to local refinement. Right: ◊, direct ratio of nodes used on level k; ▼, ratio of the total number of locally refined nodes to the total number of uniform nodes up to level k; ▽, ratio of the total number of locally refined nodes up to level k to the number of uniform nodes on level k.

undershoots are small and disappear in an appropriate norm (the L^2 norm in our case) as the grid is refined. The mechanism that controls the overshoots and undershoots in our schemes is the LS minimization. Since the LS minimization approach extends naturally to systems of equations and multiple dimensions, it can be expected that overshoots and undershoots remain small.

The LSFEMs have many attractive properties despite lack of strict monotonicity, including the natural error estimator and resulting SPD linear systems. The smearing at shocks is one of the main shortcomings of LSFEMs. We illustrated one way to address this, namely, local adaptive refinement based on the least-squares error estimator. We efficiently capture the shock as shown in Figure 6.14. Convergence is maintained using a significantly smaller number of nodes than for a uniform refinement of the grid.

The optimal convergence rate in the L^2-norm squared is $\alpha = 1.0$, i.e., $\|u-u^h\|^2_{0,\Omega} = \mathcal{O}(h)$. The convergence rate obtained with our schemes approaches optimality as the grid is refined (see Tables 6.1, 6.2, 6.3, and 6.4). In Chapter 4, we observe that convergence rates improve with the use of high-order elements, while overshoots and undershoots do not increase. This suggests that the convergence order for the methods proposed in this section can be improved by employing higher-order elements. The resulting higher-order LSFEM schemes are fully linear, that is, nonlinear limiter functions do not need to be used, which make the higher-order LSFEM schemes attractive for iterative solution methods. Thus, the use of higher-order elements seems to be another viable approach to better resolving the discontinuity in a least-squares setting.

Our numerical results indicate that our LSFEMs converge to the entropy weak solution of (6.1). This is an interesting result, since these schemes do not satisfy an exact discrete conservation property in the sense of Lax-Wendroff [58], and because the entropy solution is obtained without additional constraints. Our numerical experiments show that the correct shock speed is obtained. We prove convergence for our finite element method in several steps. In the next section, we prove that convergence of u^h in the L^2 sense implies convergence to a weak solution. This result is related to the Lax-Wendroff theorem for conservative finite

difference methods [58]. We also prove convergence of the LS functional, assuming that the Newton procedure succeeds in minimizing the functional on each grid level. In Section 6.4, we then argue that, with an appropriate choice of finite element spaces, we have an adequate amount of coercivity of the functional, resulting in convergence in L^2.

6.3 Weak Convergence Theory

In this section, we prove that if u^h converges to a function, $\hat{u}$, in the L_2 sense as $h \to 0$ for our H^{-1} LSFEM, then $\hat{u}$ is a weak solution of the conservation law. The assumption that u^h converges to $\hat{u}$ is made plausible in the numerical results of Section 6.2 and in the heuristics developed in Section 6.4.

We first relate the definition of weak solution, Definition 6.1, to an H^{-1}-like norm. Define the $H^{-1}_{0,\Gamma_O,\Omega}$-norm of $\nabla \cdot \mathbf{F}(u)$ by

$$\|\nabla \cdot \mathbf{F}(u)\|_{-1,\Gamma_O,\Omega} = \sup_{\phi \in V} \left| \frac{-\langle \mathbf{F}(u), \nabla\phi \rangle_{0,\Omega} + \langle \mathbf{n} \cdot \mathbf{F}(u), \phi \rangle_{0,\Gamma_I}}{\|\nabla\phi\|_{0,\Omega}}, \right| \tag{6.91}$$

where V is defined in (6.6). This norm is more general than the $\|\cdot\|_U$-norm in that weak boundary conditions are allowed. The following theorem relates the H^{-1}-norm to weak solutions:

Theorem 6.17. *If*

$$\|\nabla \cdot \mathbf{F}(u)\|_{-1,\Gamma_O,\Omega} = 0 \tag{6.92}$$

and

$$\|u - g\|_{0,\Gamma_I} = 0. \tag{6.93}$$

Then u is a weak solution of (6.1).

Proof. If $\|\nabla \cdot \mathbf{F}(u)\|_{-1,\Gamma_O,\Omega} = 0$ and $\|u - g\|_{0,\Gamma_I} = 0$, then, by (6.5), for all $\psi \in V$, we have

$$\begin{aligned}
|-\langle \mathbf{F}(u), \nabla\psi\rangle_{0,\Omega} + \langle \mathbf{n}\cdot\mathbf{F}(g), \psi\rangle_{0,\Gamma_I}| &= |-\langle \mathbf{F}(u), \nabla\psi\rangle_{0,\Omega} + \langle \mathbf{n}\cdot(\mathbf{F}(g) - \mathbf{F}(u) + \mathbf{F}(u)), \psi\rangle_{0,\Gamma_I}| \\
&\leq |-\langle \mathbf{F}(u), \nabla\psi\rangle_{0,\Omega} + \langle \mathbf{n}\cdot\mathbf{F}(u), \psi\rangle_{0,\Gamma_I}| \\
&\quad + \|\mathbf{n}\cdot(\mathbf{F}(g) - \mathbf{F}(u))\|_{0,\Omega}\|\psi\|_{0,\Omega} \\
&\leq \|\nabla\cdot\mathbf{F}(u)\|_{-1,\Gamma_O,\Omega} + K\|u-g\|_{0,\Omega}\|\psi\|_{0,\Omega} \\
&= 0.
\end{aligned} \tag{6.94}$$

Thus, according to Definition 6.1, u is a weak solution of (6.1). □

Using approximation properties for appropriate finite element spaces, we can prove that for solutions of least-squares minimization problem (6.26) on successively refined grids, the functional goes to zero as the grid is refined. That is, if Φ^h and $\mathcal{U}^h$ satisfy approximation property (6.75), we can prove that $\mathcal{G}(\phi^h, u^h; g) \to 0$ as $h \to 0$, where (ϕ^h, u^h) solves (6.26).

Lemma 6.18. *Assume that there exists unique solution $(\phi, u) \in H^1(\Omega) \times H^{\frac{1}{2}-\varepsilon}$ to (6.25) for $\varepsilon > 0$. Let $(\phi^h, u^h) \in \Phi^h \times \mathcal{U}^H$ be the solution of least-squares minimization problem (6.26) and let Φ^h and $\mathcal{U}^h$ satisfy approximation property (6.75). Then nonlinear LS functional $\mathcal{G}(\phi^h, u^h; g) \to 0$ as $h \to 0$.*

Proof. We have

$$\begin{aligned}
\mathcal{G}(\phi^h, u^h; g) &\leq \mathcal{G}(\Pi^h\phi, \Pi^h u; g) \\
&= \|\nabla^\perp \Pi^h\phi - \mathbf{F}(\Pi^h u)\|^2_{0,\Omega} \\
&\quad + \|\mathbf{n}\cdot(\nabla^\perp\Pi^h\phi - \mathbf{F}(g))\|^2_{0,\Gamma_I} + \|\Pi^h u - g\|^2_{0,\Gamma_I} \\
&= \|\nabla^\perp(\Pi^h\phi - \phi) - (\mathbf{F}(\Pi^h u) - \mathbf{F}(u))\|^2_{0,\Omega} \\
&\quad + \|\mathbf{n}\cdot(\nabla^\perp(\Pi^h\phi - \phi))\|^2_{0,\Gamma_I} + \|\Pi^h u - u\|^2_{0,\Gamma_I} \\
&\leq \|\nabla^\perp(\Pi^h\phi - \phi)\|^2_{0,\Omega} + K\|\Pi^h u - u\|^2_{0,\Omega} \\
&\quad + \|\nabla^\perp(\Pi^h\phi - \phi)\|^2_{0,\Gamma_I} + \|\Pi^h u - u\|^2_{0,\Gamma_I} \\
&\leq ch^\beta,
\end{aligned} \tag{6.95}$$

where $\beta > 0$ comes from approximation property (6.75). This shows that $\mathcal{G}(\phi^h, u^h; g) \to 0$ as $h \to 0$, which completes the proof. □

We can now state and prove the following weak convergence theorem.

Theorem 6.19 (H^{-1} LSFEM Weak Conservation). *Assume that there exists unique solution $(\phi, u) \in H^1(\Omega) \times H^{\frac{1}{2}-\varepsilon}$ to (6.25) for $\varepsilon > 0$. Let $(\phi^h, u^h) \in \Phi^h \times \mathcal{U}^H$ be the solution of LS minimization problem (6.26). If finite element approximation u^h converges to $\hat{u}$ in the L^2 sense as $h \to 0$, then $\hat{u}$ is a weak solution of (6.1). That is, if*

$$\|u^h - \hat{u}\|_{0,\Omega} \to 0, \tag{6.96}$$

$$\|u^h - \hat{u}\|_{0,\Gamma_I} \to 0, \tag{6.97}$$

for some $\hat{u} \in L^2(\Omega)$, then $\hat{u}$ is a weak solution.

Proof. According to Theorem 6.17, it suffices to prove that

$$\|\nabla \cdot \mathbf{F}(\hat{u})\|_{-1,\Gamma_O,\Omega} = 0, \tag{6.98}$$

$$\|\hat{u} - g\|_{0,\Gamma_I} = 0. \tag{6.99}$$

This can be obtained as follows. From (6.91), we have

$$\|\nabla \cdot \mathbf{F}(\hat{u})\|_{-1,\Gamma_O,\Omega} = \sup_{\phi \in V} \left| \frac{-\langle \mathbf{F}(\hat{u}), \nabla\phi \rangle_{0,\Omega} + \langle \mathbf{n} \cdot \mathbf{F}(\hat{u}), \phi \rangle_{0,\Gamma_I}}{\|\nabla\phi\|_{0,\Omega}} \right|. \tag{6.100}$$

Adding and subtracting $\nabla\phi^h$ and $\mathbf{F}(u^h)$ in the interior term results in

$$\begin{aligned} \|\nabla \cdot \mathbf{F}(\hat{u})\|_{-1,\Gamma_O,\Omega} = \sup_{\phi \in V} \Bigg| & \frac{-\langle \nabla^\perp\phi^h + \mathbf{F}(u^h) - \nabla^\perp\phi^h + \mathbf{F}(\hat{u}) - \mathbf{F}(u^h), \nabla\phi \rangle_{0,\Omega}}{\|\nabla\phi\|_{0,\Omega}}, \\ & + \frac{\langle \mathbf{n} \cdot \mathbf{F}(\hat{u}), \phi \rangle_{0,\Gamma_I}}{\|\nabla\phi\|_{0,\Omega}} \Bigg| \end{aligned} \tag{6.101}$$

and, by Green's Formula,

$$\begin{aligned} \|\nabla \cdot \mathbf{F}(\hat{u})\|_{-1,\Gamma_O,\Omega} = \sup_{\phi \in V} \Bigg| & \frac{-\langle \nabla \cdot \nabla^\perp\phi^h, \phi \rangle_{0,\Omega}}{\|\nabla\phi\|_{0,\Omega}} \\ & - \frac{\langle \mathbf{F}(u^h) - \nabla^\perp\phi^h + \mathbf{F}(\hat{u}) - \mathbf{F}(u^h), \nabla\phi \rangle_{0,\Omega}}{\|\nabla\phi\|_{0,\Omega}} \\ & + \frac{\langle \mathbf{n} \cdot \nabla^\perp\phi^h, \phi \rangle_{0,\Gamma} + \langle \mathbf{n} \cdot \mathbf{F}(\hat{u}), \phi \rangle_{0,\Gamma_I}}{\|\nabla\phi\|_{0,\Omega}} \Bigg|. \end{aligned} \tag{6.102}$$

Since $\phi = 0$ on Γ_O and $\nabla \cdot \nabla^\perp \phi = 0$, we have

$$\begin{aligned}\|\nabla \cdot \mathbf{F}(\hat{u})\|_{-1,\Gamma_O,\Omega} = \sup_{\phi \in V} \Bigg| & \frac{-\langle \mathbf{F}(u^h) - \nabla^\perp \phi^h, \nabla\phi\rangle_{0,\Omega} + \langle \mathbf{F}(\hat{u}) - \mathbf{F}(u^h), \nabla\phi\rangle_{0,\Omega}}{\|\nabla\phi\|_{0,\Omega}} \\ & + \frac{\langle \mathbf{n} \cdot (\mathbf{F}(\hat{u}) - \nabla^\perp \phi^h), \phi\rangle_{0,\Gamma_I}}{\|\nabla\phi\|_{0,\Omega}} \Bigg|.\end{aligned} \tag{6.103}$$

By adding and subtracting $\mathbf{F}(u^h)$ and $\mathbf{F}(g)$ in the boundary terms, this results in

$$\begin{aligned}\|\nabla \cdot \mathbf{F}(\hat{u})\|_{-1,\Gamma_O,\Omega} = \sup_{\phi \in V} \Bigg| & \frac{-\langle \mathbf{F}(u^h) - \nabla^\perp \phi^h, \nabla\phi\rangle_{0,\Omega} + \langle \mathbf{F}(\hat{u}) - \mathbf{F}(u^h), \nabla\phi\rangle_{0,\Omega}}{\|\nabla\phi\|_{0,\Omega}} \\ & + \frac{\langle \mathbf{n} \cdot (\mathbf{F}(\hat{u}) - \mathbf{F}(u^h) + \mathbf{F}(u^h) - \mathbf{F}(g) - \nabla^\perp \phi^h + \mathbf{F}(g)), \phi\rangle_{0,\Gamma_I}}{\|\nabla\phi\|_{0,\Omega}} \Bigg|.\end{aligned} \tag{6.104}$$

We now apply the generalized Cauchy-Schwarz inequality [2],

$$\langle \phi, \phi \rangle_{0,\Gamma} \leq \|\phi\|_{-\frac{1}{2}} \|\phi\|_{\frac{1}{2}}, \tag{6.105}$$

to arrive at

$$\begin{aligned}\|\nabla \cdot \mathbf{F}(\hat{u})\|_{-1,\Gamma_O,\Omega} \leq \sup_{\phi \in V} & \frac{\|\mathbf{F}(u^h) - \nabla^\perp \phi^h\| \|\nabla\phi\|_{0,\Omega} + \|\mathbf{F}(\hat{u}) - \mathbf{F}(u^h)\| \|\nabla\phi\|_{0,\Omega}}{\|\nabla\phi\|_{0,\Omega}} \\ & + \frac{\|\mathbf{n} \cdot (\mathbf{F}(\hat{u}) - \mathbf{F}(u^h))\|_{-\frac{1}{2},\Gamma_I} \|\phi\|_{\frac{1}{2},\Gamma_I}}{\|\nabla\phi\|_{0,\Omega}} \\ & + \frac{\|\mathbf{n} \cdot (\mathbf{F}(u^h) - \mathbf{F}(g))\|_{-\frac{1}{2},\Gamma_I} \|\phi\|_{\frac{1}{2},\Gamma_I}}{\|\nabla\phi\|_{0,\Omega}} \\ & + \frac{\|\mathbf{n} \cdot (\mathbf{F}(g) - \nabla^\perp \phi^h)\|_{-\frac{1}{2},\Gamma_I} \|\phi\|_{\frac{1}{2},\Gamma_I}}{\|\nabla\phi\|_{0,\Omega}} \\ \leq & \|\mathbf{F}(u^h) - \nabla^\perp \phi^h\| + \|\mathbf{F}(\hat{u}) - \mathbf{F}(u^h)\| \\ & + \sup_{\phi \in V} \frac{\|\mathbf{n} \cdot (\mathbf{F}(\hat{u}) - \mathbf{F}(u^h))\|_{-\frac{1}{2},\Gamma_I} \|\phi\|_{\frac{1}{2},\Gamma_I}}{\|\nabla\phi\|_{0,\Omega}} \\ & + \frac{\|\mathbf{n} \cdot (\mathbf{F}(u^h) - \mathbf{F}(g))\|_{-\frac{1}{2},\Gamma_I} \|\phi\|_{\frac{1}{2},\Gamma_I}}{\|\nabla\phi\|_{0,\Omega}} \\ & + \frac{\|\mathbf{n} \cdot (\mathbf{F}(g) - \nabla^\perp \phi^h)\|_{-\frac{1}{2},\Gamma_I} \|\phi\|_{\frac{1}{2},\Gamma_I}}{\|\nabla\phi\|_{0,\Omega}}.\end{aligned} \tag{6.106}$$

Using the trace theorem on $H^1_{0,\Gamma_O,\Omega}$, we arrive at

$$\begin{aligned}\|\nabla \cdot \mathbf{F}(\hat{u})\|_{-1,\Gamma_O,\Omega} \leq c \, (& \|\mathbf{F}(u^h) - \nabla^\perp \phi^h\| + \|\mathbf{F}(\hat{u}) - \mathbf{F}(u^h)\| \\ & + \|\mathbf{n} \cdot (\mathbf{F}(\hat{u}) - \mathbf{F}(u^h))\|_{-\frac{1}{2},\Gamma_I} + \|\mathbf{n} \cdot (\mathbf{F}(u^h) - \mathbf{F}(g))\|_{-\frac{1}{2},\Gamma_I} \\ & + \|\mathbf{n} \cdot (\mathbf{F}(g) - \nabla^\perp \phi^h)\|_{-\frac{1}{2},\Gamma_I}).\end{aligned} \tag{6.107}$$

Since $\mathbf{F}$ is Lipschitz continuous in the interior and on the boundary, we find that

$$\begin{aligned}\|\nabla\cdot\mathbf{F}(\hat{u})\|_{-1,\Gamma_O,\Omega} \leq c\,(&\|\mathbf{F}(u^h)-\nabla^\perp\phi^h\| + \|\hat{u}-u^h\| \\ &+ \|\hat{u}-u^h\|_{-\frac{1}{2},\Gamma_I} + \|u^h-g\|_{-\frac{1}{2},\Gamma_I} \\ &+ \|\mathbf{n}\cdot\mathbf{F}(g)-\mathbf{n}\cdot\nabla^\perp\phi^h)\|_{-\frac{1}{2},\Gamma_I})\end{aligned} \tag{6.108}$$

for every $\phi^h \in \Phi^h$, $u^h \in \mathcal{U}^h$. Recalling that $\mathcal{G}(\phi^h, u^h; g) \to 0$, $\|\hat{u}-u^h\|_{0,\Omega} \to 0$, and $\|\hat{u}-u^h\|_{0,\Gamma_I} \to 0$, we arrive at

$$\|\nabla\cdot\mathbf{F}(\hat{u})\|_{-1,\Gamma_O,\Omega} = 0. \tag{6.109}$$

Similarly, we also have

$$\begin{aligned}\|\hat{u}-g\|_{0,\Gamma_I} &= \|\hat{u}-u^h+u^h-g\|_{0,\Gamma_I} \\ &\leq \|\hat{u}-u^h\|_{0,\Gamma_I} + \|u^h-g\|_{0,\Gamma_I}\end{aligned} \tag{6.110}$$

for every $\phi^h \in \Phi^h$, $u^h \in \mathcal{U}^h$, which yields $\|\hat{u}-g\|_{0,\Gamma_I} = 0$. This completes the proof. □

A weak conservation theorem for the $H(div)$ formulation, given by (6.20), is obtained in a similar way. We omit the details because the proofs are nearly identical.

Lemma 6.20. *Let $(\mathbf{w}^h, u^h)$ be the solution of least-squares minimization problem (6.29). Then nonlinear LS functional $\mathcal{F}(\mathbf{w}^h, u^h; g) \to 0$ as $h \to 0$.*

Proof. The proof is analogous to the proof of Lemma 6.18. □

Similar to Theorem 6.19, we have

Theorem 6.21 ($H(div)$ LSFEM Weak Conservation). *Let $(\mathbf{w}^h, u^h)$ be the solution of LS minimization (6.29). If finite element approximation u^h converges to $\hat{u}$ in the L^2 sense as $h \to 0$, then $\hat{u} \in L^2(\Omega)$ is a weak solution of (6.1). That is, if*

$$\|u^h-\hat{u}\|_{0,\Omega} \to 0, \tag{6.111}$$

$$\|u^h-\hat{u}\|_{0,\Gamma_I} \to 0, \tag{6.112}$$

for some $\hat{u}$, then $\hat{u}$ is a weak solution.

Proof. The proof is analogous to the proof of Lemma 6.19. □

As discussed in Chapter 2, convergence of a numerical approximation does not necessarily imply that the approximation converges to a weak solution of the conservation law, (6.1). This was studied in depth for finite difference methods [62]. For non-conservative finite difference schemes, Hou and LeFloch [45] explained, in part, the deviation from the correct weak solution. The incorrect solution is found to be a solution to an inhomogeneous conservation law that contains a Borel source term. If the finite difference scheme has exact discrete conservation and if the scheme converges, then the theory of Lax and Wendroff [58] implies that the approximation converges to a weak solution. Since this result was established, the exact discrete conservation property has **de facto** been considered a strong requirement for numerical schemes for hyperbolic conservation laws [62]. Numerical schemes that satisfy such an exact discrete conservation property have since been called **conservative schemes**. However, exact discrete conservation is only a **sufficient** condition to ensure convergence to a weak solution, not a **necessary** condition. Moreover, the application to finite element methods has not been fully detailed. In [1], it is shown, in the context of residual distribution finite element schemes, that exact discrete conservation need not be imposed strictly over the whole domain, but can be limited to regions of discontinuity using an adaptive quadrature procedure.

The $H(div)$ finite element method and the H^{-1} method that we develop do not satisfy an exact discrete conservation property. This is verified numerically in Figure 6.6, which shows $\nabla\cdot\mathbf{w}$ on a 32×32 grid. Even so, our numerical and theoretical results show that this method converges to a weak solution, in accordance with Theorems 6.19 and 6.21. This illustrates that the discrete conservation property of Lax-Wendroff is not a necessary condition for convergence to a weak solution. The exact discrete conservation requirement can be replaced by a minimization principle in a suitable continuous norm.

Remark 6.22. *In a certain sense, the LSFEM we develop based on H^{-1} formulation (6.22) does impose exact discrete conservation. Indeed, $\nabla\cdot\nabla^{\perp}\phi^h\equiv 0$ in the whole domain Ω. Thus, in*

a sense, the approximate flux vector, $\nabla^{\perp}\phi^h$, is necessarily conserved discretely, in a pointwise sense, while the approximate flux vector, $\mathbf{F}(u^h)$, is not.

6.4 Finite Element Convergence

In this section, we investigate convergence of the numerical approximation, u^h, in the L^2-norm. Section 6.3 focused on convergence to a weak solution, assuming that the method converges in L^2. We now show that it is plausible, with the appropriate selection of finite element spaces, for the least-squares numerical approximation to converge in $L^2(\Omega)$. First, an example with instabilities is presented. Oscillatory error is not reduced by the functional in this case. We then argue that, with compatible spaces, we achieve sufficient coercivity of the functional. We define space compatibility in Section 6.4.2. Finally, we present numerical evidence for the linear case that confirms our theoretical observations.

6.4.1 Oscillatory Example

For simplicity, consider the linear conservation law, where $\mathbf{F}(u) = \mathbf{b}u$, and assume $|\mathbf{b}| = 1$. $H(div)$ formulation (6.20) becomes

$$\nabla \cdot w = 0, \qquad \text{in } \Omega, \tag{6.113a}$$

$$\mathbf{w} = \mathbf{b}u, \qquad \text{in } \Omega, \tag{6.113b}$$

$$\mathbf{n} \cdot \mathbf{w} = \mathbf{n} \cdot \mathbf{b}g, \quad \text{on } \Gamma_I, \tag{6.113c}$$

$$u = g, \qquad \text{on } \Gamma_I. \tag{6.113d}$$

The associated least-squares functional is

$$\mathcal{F}(\mathbf{w}, u; g) := \|\nabla \cdot \mathbf{w}\|^2_{0,\Omega} + \|\mathbf{w} - \mathbf{b}u\|^2_{0,\Omega} + \|\mathbf{n} \cdot (\mathbf{w} - \mathbf{F}(g))\|^2_{0,\Gamma_I} + \|u - g\|^2_{0,\Gamma_I}. \tag{6.114}$$

This leads to the following minimization problem: find $(\mathbf{w}^*, u^*)$ such that

$$(\mathbf{w}^*, u^*) = \underset{u \in L^2(\Omega),\, \mathbf{w} \in H(div,\Omega)}{\operatorname{argmin}} \mathcal{F}(\mathbf{w}, u; g). \tag{6.115}$$

If uniform coercivity of the associated bilinear form could be shown with respect to the graph norm, then we would automatically achieve convergence in L^2 and for any conforming finite element subspace. The functional is coercive if there exists constant C, independent of the grid size, such that

$$\begin{aligned}\|u\|_{0,\Omega}^2 + \|\mathbf{w}\|_{0,\Omega}^2 + \|\nabla\cdot\mathbf{w}\|_{0,\Omega}^2 \\ \leq C\left(\|\nabla\cdot\mathbf{w}\|_{0,\Omega}^2 + \|\mathbf{w}-\mathbf{b}u\|_{0,\Omega}^2 + \|\mathbf{n}\cdot(\mathbf{w}-\mathbf{F}(g))\|_{0,\Gamma_I}^2 + \|u-g\|_{0,\Gamma_I}^2\right)\end{aligned} \tag{6.116}$$

for all $u \in H^{\frac{1}{2}-\varepsilon}(\Omega)$ and $\mathbf{w} \in H(div,\Omega)$. We would then have that $\|u-u^h\|_{0,\Omega}$ converges when $\mathcal{F}(\mathbf{w}^h, u^h; 0)$ converges.

However, functional (6.114) is not uniformly coercive. Consider the following example on a grid with mesh size h. Let $\Omega = [0,1]^2$ and $\mathbf{b} = (1,0)$. Define

$$\phi^h = \frac{h}{2\pi}x^2(1-x)^2\sin^2(\frac{\pi y}{h}), \tag{6.117a}$$

$$\mathbf{w}^h = \nabla^\perp\phi^h = \begin{pmatrix} x^2(1-x)^2\cos(\frac{\pi y}{h})\sin(\frac{\pi y}{h}) \\ \frac{h}{\pi}x(2x^2-1)\sin^2(\frac{\pi y}{h}) \end{pmatrix}, \tag{6.117b}$$

$$u^h = \partial_y\phi^h = x^2(1-x)^2\cos(\frac{\pi y}{h})\sin(\frac{\pi y}{h}). \tag{6.117c}$$

It is easily verified that

$$\|\mathbf{w}^h\|_{0,\Omega}^2 \geq \|u^h\|_{0,\Omega}^2 \geq c_0 > 0 \tag{6.118}$$

for some constant $c_0 \in \mathbb{R}^d$ independent of h, while

$$\mathcal{F}(\mathbf{w}^h, u^h; 0) \leq c_1 h^2. \tag{6.119}$$

Remark 6.23. *So far we have merely shown that $H(div,\Omega)$ is not compact in L^2. For this reason, we should then not expect uniform coercivity of our functional with respect to the graph norm.*

Error components such as (6.117) can be found in a finite element space that is not chosen properly. That is, there are components that are small in the functional but are large in an L^2 sense. This is, indeed, the case if we choose standard Raviart-Thomas (RT_0) finite elements for

w and piecewise constants for u. Figure 6.17 verifies that this kind of oscillatory error dominates the solution, u^h, to Burger's equation with $u = 1.0$ on the left and $u = 0.0$ on the bottom inflow boundaries (Example 5). Newton's method fails to converge in this case and the oscillations continue to grow as the grid is refined. However, numerical results and further analysis show that, with a different set of finite element spaces, the highly oscillatory error is suppressed.

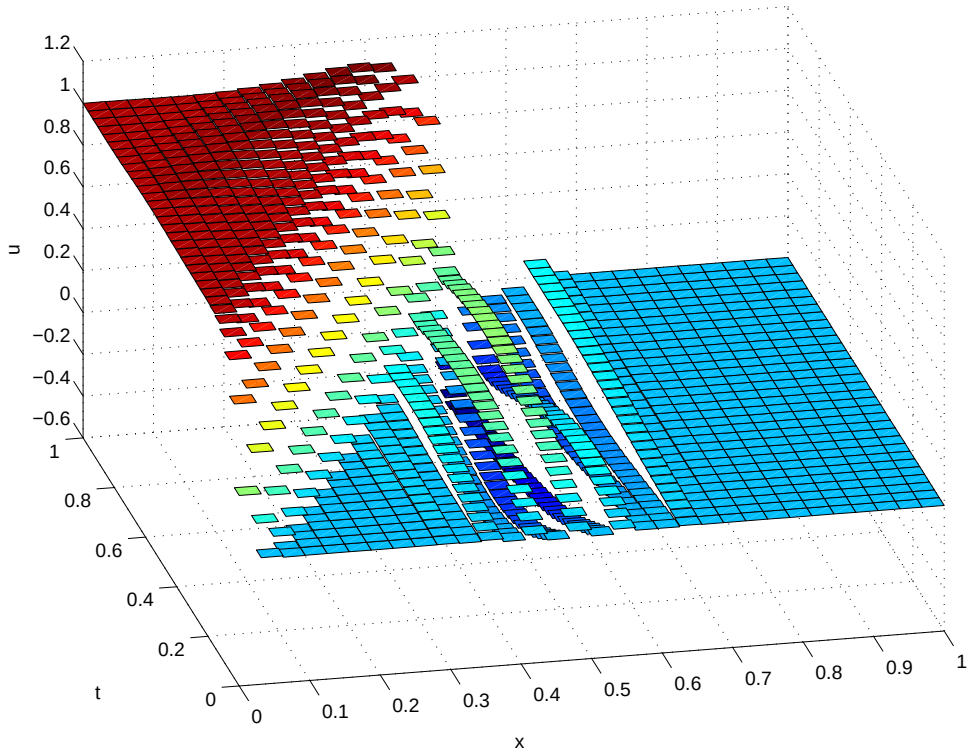


Figure 6.17: Oscillatory error components dominate the solution to Burger's equation for piecewise constant u^h. Solution u^h is shown on a 32×32 grid.

6.4.2 Compatible Finite Element Spaces

We begin by defining what is meant by compatible finite element spaces.

Definition 6.24. *Let $(\phi^h, u^h) \in \Phi^h \times \mathcal{U}^h$ be the finite element approximation to (6.26). Finite element spaces Φ^h and $\mathcal{U}^h$ are defined to be* ***compatible*** *if there exists constants $c > 0$ and $\mu > 0$*

such that

$$ch^{\mu}\|u-u^h\|_{0,\Omega}^2 \leq \mathcal{G}(\phi^h, u^h; 0), \tag{6.120}$$

and if

$$\|u-u^h\|_{0,\Omega} \to 0, \tag{6.121}$$

$$\mathcal{G}(\phi^h, u^h; g) \to 0, \tag{6.122}$$

as $h \to 0$.

Remark 6.25. *Compatibility is similarly defined for the* $H(div)$ *problem, (6.29).*

If finite element spaces are not chosen carefully (for example, bilinears for ϕ^h and piecewise constants for u^h) then the functional may converge to zero while the L^2-norm of the error does not. With a compatible set of finite elements spaces, which is explained in more detail below, sufficient coercivity can be obtained. We have confirmed this numerically in Section 6.2 for the $H(div)$ formulation, using RT_0 for $\mathbf{w}$ and continuous bilinear finite elements for u. This was also true for the H^{-1} LSFEM using continuous bilinear finite elements for both ϕ^h and u^h. We now further motivate that convergence in L^2 occurs for these spaces by considering a heuristic analysis for grid-aligned flow.

Let $\mathcal{T}^h$ be a quadrilateral tessellation of Ω and let

$$\mathcal{U}^h = \text{the space of continuous piecewise bilinear finite elements with } u = 0 \text{ on } \Gamma_I \tag{6.123}$$

$$= \{u \in C^0_{\Gamma_I}(\Omega) \,:\, u|_\tau \in \mathcal{P}_1(\tau),\, \forall\, \tau \in \mathcal{T}^h\},$$

$$\mathcal{W}^h = RT_0 \text{ with } \mathbf{n}\cdot\mathbf{w}^h = 0 \text{ on } \Gamma_I \tag{6.124}$$

$$= \left\{ \mathbf{w}^h \in \begin{pmatrix} \partial_y \phi^h \\ 0 \end{pmatrix} \oplus \begin{pmatrix} 0 \\ \partial_x \phi^h \end{pmatrix} \,:\, \phi^h \in \mathcal{U}^h,\, \mathbf{n}\cdot\mathbf{w}^h = 0,\, \text{on}\, \Gamma_I \right\},$$

$$\mathcal{C}^h = \text{null-space of } \mathcal{W}^h \text{ under the divergence operator} \tag{6.125}$$

$$= \{\mathbf{w}^h \in \mathcal{W}^h \,:\, \nabla\cdot\mathbf{w}^h = 0\},$$

$$= \{\mathbf{w}^h \in \mathcal{W}^h \,:\, \mathbf{w}^h = \nabla^\perp \phi^h \text{ for some } \phi^h \in \mathcal{U}^h\}$$

$$\mathcal{S}^h = \text{piecewise constant finite elements,} \tag{6.126}$$

$$= \{s^h \in L^2(\Omega) \,:\, s^h|_\tau \in \mathcal{P}_0(\tau),\, \forall\, \tau \in \mathcal{T}^h\}$$

$$\mathcal{C}^{h\perp} = G^h = \{\mathbf{w}^h \in \mathcal{W}^h \,:\, \langle \mathbf{w}^h, \mathbf{c}^h\rangle_{0,\Omega} = 0,\, \forall\, \mathbf{c}^h \in \mathcal{C}^h\}. \tag{6.127}$$

The first step toward obtaining discrete coercivity is to show that, for $(\mathbf{w}^h, u^h) \in \mathcal{W}^h \times \mathcal{U}^h$, we have

$$K^h \|\mathbf{w}^h\|^2_{0,\Omega} \le \mathcal{F}(\mathbf{w}^h, u^h; 0), \tag{6.128}$$

where $K^h \in \mathbb{R}^d$ may depend on mesh size h. To follow Definition 6.24, we write $K^h = \mathcal{O}(h^\delta)$, for some $\delta > 0$. If $ch^\delta \le K^h$, $\mathcal{F}(\mathbf{w}^h, u^h, 0) \le ch^\beta$ and (6.128) holds, then $\|u - u^h\|^2_{0,\Omega} \le ch^{\beta-\delta}$. This is summarized in more detail in Theorem 6.26 at the end of the section.

Define weak gradient $\nabla^h : \mathcal{W}^h \to \mathcal{S}^h$, which is injective up to a constant, by the following: given $s^h \in \mathcal{S}^h$, let $\nabla^h s^h \in \mathcal{C}^{h\perp}$ such that

$$\langle \nabla^h s^h, \mathbf{w}^h\rangle_{0,\Omega} = -\langle s^h, \nabla\cdot\mathbf{w}^h\rangle_{0,\Omega} \quad \forall\, \mathbf{w}^h \in \mathcal{W}^h. \tag{6.129}$$

Since $\mathbf{w}^h \in \mathcal{W}^h$, we use a discrete Helmholtz decomposition [13, 38]

$$\mathbf{w}^h = \nabla^\perp \phi^h + \nabla^h s^h, \tag{6.130}$$

where $\phi^h \in \mathcal{U}^h$ and $s^h \in \mathcal{S}^h/\mathbb{R}$.

From a Poincaré inequality [13] on weak gradient ∇^h, let C be a constant such that $\|q^h\|_{0,\Omega} \leq C\|\nabla^h q^h\|_{0,\Omega}$, $\forall\, q^h \in \mathcal{S}^h/\mathbb{R}$. Using (6.130), we then have that

$$\begin{aligned}
\|\nabla \cdot \mathbf{w}^h\|_{0,\Omega} &= \|\nabla \cdot \nabla^h s^h\|_{0,\Omega} \\
&= \sup_{q^h \in \mathcal{S}^h} \frac{|\langle \nabla \cdot \nabla^h s^h, q^h \rangle_{0,\Omega}|}{\|q^h\|_{0,\Omega}} \\
&= \sup_{q^h \in \mathcal{S}^h} \frac{|\langle \nabla^h s^h, \nabla^h q^h \rangle_{0,\Omega}|}{\|q^h\|_{0,\Omega}} \\
&\geq \frac{|\langle \nabla^h s^h, \nabla^h q^h \rangle_{0,\Omega}|}{\|q^h\|_{0,\Omega}} \\
&\geq \frac{1}{C}\|\nabla^h s^h\|_{0,\Omega}.
\end{aligned} \tag{6.131}$$

Using the triangle inequality results in

$$\|\mathbf{b}u - (\nabla^\perp \phi^h + \nabla^h s^h)\|_{0,\Omega} \geq \|\mathbf{b}u - \nabla^\perp \phi^h\|_{0,\Omega} - \|\nabla^h s^h\|_{0,\Omega}. \tag{6.132}$$

Squaring both sides, applying the triangle and epsilon inequalities, and adding $\|\nabla \cdot \mathbf{w}^h\|^2_{0,\Omega}$ results in

$$\mathcal{F}(\mathbf{w}^h, u^h, 0) \geq \frac{1}{2}\|\mathbf{b}u - \nabla^\perp \phi^h\|^2_{0,\Omega} + C\|\nabla^h s^h\|^2_{0,\Omega}. \tag{6.133}$$

Since

$$\|\mathbf{w}^h\|^2 = \|\nabla^\perp \phi^h + \nabla^h s^h\|^2_{0,\Omega} = \|\nabla^\perp \phi^h\|^2_{0,\Omega} + \|\nabla^h s^h\|^2_{0,\Omega}, \tag{6.134}$$

we now only need to show that, for a given $\phi^h \in \mathcal{U}^h$, there exists a K^h_ϕ such that

$$\inf_{u^h \in \mathcal{U}^h} \|\mathbf{b}u^h - \nabla^\perp \phi^h\|_{0,\Omega} \geq K^h_\phi \|\nabla^\perp \phi^h\|_{0,\Omega} \quad \forall\, \phi^h \in \Phi^h. \tag{6.135}$$

This also shows convergence for H^{-1} least-squares method (6.26) since the functional in this case is

$$\|\mathbf{b}u^h - \nabla^\perp \phi^h\|_{0,\Omega}, \tag{6.136}$$

for finite element spaces with boundary conditions (6.22b-6.22c) enforced strongly.

As indicated above, uniform coercivity of the functional does not hold. That is, K^h_ϕ cannot be a constant independent of the grid size h in

$$\|\mathbf{b}u^h - \nabla^\perp \phi^h\|_{0,\Omega} \geq K^h_\phi \|\nabla^\perp \phi^h\|_{0,\Omega} \quad \forall\, u^h \in \mathcal{U}^h. \tag{6.137}$$

We thus look for a K_ϕ^h that varies as a function of h. With such a bound, convergence can be proved if the functional converges sufficiently faster than the L^2-norm required in (6.116).

We can write (6.135) as

$$\inf_{u^h \in \mathcal{U}^h} \|u^h - \mathbf{b}\cdot\nabla^\perp\phi^h\|_{0,\Omega}^2 + \|\mathbf{b}^\perp\cdot\nabla^\perp\phi^h\|_{0,\Omega}^2 \geq K_\phi^h \left(\|\mathbf{b}\cdot\nabla^\perp\phi^h\|_{0,\Omega}^2 + \|\mathbf{b}^\perp\cdot\nabla^\perp\phi^h\|_{0,\Omega}^2\right), \quad (6.138)$$

since $|\mathbf{b}| = 1$.

We now show, for a specific flow, that $K^h = \mathcal{O}(h)$, and that this is sufficient coercivity in the sense discussed above. For grid-aligned flow with $\mathbf{b} = (1,0)$, we have $\mathbf{b}^\perp = (0,-1)^T$, and (6.138) becomes

$$\inf_{u^h \in \mathcal{U}^h} \|u^h - \partial_y\phi^h\|_{0,\Omega}^2 + \|\partial_x\phi^h\|_{0,\Omega}^2 \geq K_\phi^h \left(\|\partial_y\phi^h\|_{0,\Omega}^2 + \|\partial_x\phi^h\|_{0,\Omega}^2\right). \quad (6.139)$$

We thus focus on finding K_ϕ^h such that

$$K_\phi^h = \inf_{u^h \in \mathcal{U}^h} \frac{\langle u^h, u^h\rangle_{0,\Omega} - 2\langle u^h, \partial_y\phi^h\rangle_{0,\Omega} + \langle \partial_y\phi^h, \partial_y\phi^h\rangle_{0,\Omega} + \langle \partial_x\phi^h, \partial_x\phi^h\rangle_{0,\Omega}}{\langle \partial_y\phi^h, \partial_y\phi^h\rangle_{0,\Omega} + \langle \partial_x\phi^h, \partial_x\phi^h\rangle_{0,\Omega}}. \quad (6.140)$$

Let $\{\psi_j\}_{j=1}^N$ be the standard 1-D linear basis in each coordinate direction, where N is the number of nodes in one direction, that are not located on the inflow boundary. Writing u^h in terms of this basis we have

$$\phi^h = \sum_{i,j=1}^N \gamma_{i,j}\psi_i(x)\psi_j(y), \quad (6.141)$$

$$u^h = \sum_{i,j=1}^N \alpha_{i,j}\psi_i(x)\psi_j(y), \quad (6.142)$$

where $\gamma_{i,j}$ and $\alpha_{i,j}$ are unknown coefficients. Let $\stackrel{\wedge}{=}$ be the equality that sets the (i,j,k,l) component of the corresponding tensor. That is, $A \otimes B \stackrel{\wedge}{=} c_{i,j,k,l}$ means $A_{i,j}B_{k,l} = c_{i,j,k,l}$.

Define for $i, j, k, l = 1, \ldots, N$

$$M_x \otimes M_y \triangleq \langle \psi_i \psi_j, \psi_k \psi_l \rangle_{0,\Omega}, \tag{6.143}$$

$$M_x \otimes D_y \triangleq \langle \psi_i \psi_j, \partial_y(\psi_k \psi_l) \rangle_{0,\Omega}, \tag{6.144}$$

$$M_x \otimes D_{yy} \triangleq \langle \partial_y(\psi_i \psi_j), \partial_y(\psi_k \psi_l) \rangle_{0,\Omega}, \tag{6.145}$$

$$M_y \otimes D_{xx} \triangleq \langle \partial_x(\psi_i \psi_j), \partial_x(\psi_k \psi_l) \rangle_{0,\Omega}, \tag{6.146}$$

$$\boldsymbol{\gamma}_x \otimes \boldsymbol{\gamma}_y \triangleq \gamma_{i,k}, \tag{6.147}$$

$$\boldsymbol{\alpha}_x \otimes \boldsymbol{\alpha}_y \triangleq \alpha_{i,k}, \tag{6.148}$$

where $\psi_i = \psi_i(x)$, $\psi_j = \psi_j(y)$, $\psi_k = \psi_k(x)$, and $\psi_l = \psi_l(y)$. Writing $\boldsymbol{\gamma} = \boldsymbol{\gamma}_x \otimes \boldsymbol{\gamma}_y$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}_x \otimes \boldsymbol{\alpha}_y$, (6.140) becomes

$$K_\phi^h = \min_{\boldsymbol{\alpha}} \frac{\langle M_x \otimes M_y \boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle - 2\langle \boldsymbol{\alpha}, M_x \otimes D_y \boldsymbol{\gamma} \rangle + \langle M_x \otimes D_{yy} \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle + \langle D_{xx} \otimes M_y \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle}{\langle M_x \otimes D_{yy} \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle + \langle D_{xx} \otimes M_y \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle}, \tag{6.149}$$

where $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{\ell^2}$, the discrete L^2 inner product. From minimization problem (6.149), for a given ϕ^h, we have

$$M_x \otimes M_y \boldsymbol{\alpha}_x \otimes \boldsymbol{\alpha}_y = M_x \otimes D_y \boldsymbol{\gamma}_x \otimes \boldsymbol{\gamma}_y. \tag{6.150}$$

This is essentially the weak form of the least-squares minimization problem. Substituting into (6.149), rearranging terms, and minimizing over all possible $\phi^h \in \mathcal{U}^h$, we arrive at

$$K^h = \min_{\boldsymbol{\gamma}} \frac{\langle D_{xx} \otimes M_y \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle + \langle M_x \otimes D_{yy} \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle - \langle M_x \otimes D_y \boldsymbol{\gamma}, (M_x \otimes M_y)^{-1} M_x \otimes D_y \boldsymbol{\gamma} \rangle}{\langle M_x \otimes D_{yy} \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle + \langle D_{xx} \otimes M_y \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle}, \tag{6.151}$$

which becomes

$$\begin{aligned} K^h &= \min_{\boldsymbol{\gamma}} \frac{\langle M_x \otimes \left(D_{yy} - (M_x \otimes D_y)^T (M_x \otimes M_y)^{-1} (M_x \otimes D_y) \right) \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle + \langle D_{xx} \otimes M_y \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle}{\langle M_x \otimes D_{yy} \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle + \langle D_{xx} \otimes M_y \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle} \\ &= \min_{\boldsymbol{\gamma}} \frac{\langle M_x \otimes \left(D_{yy} - D_y^T M_y^{-1} D_y \right) \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle + \langle D_{xx} \otimes M_y \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle}{\langle M_x \otimes D_{yy} \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle + \langle D_{xx} \otimes M_y \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle}. \end{aligned} \tag{6.152}$$

Viewing (6.152) as a Rayleigh quotient, we find that K^h is the minimum eigenvalue of the associated generalized eigenvalue problem. Since the eigenvalues and eigenvectors of D_{xx} and D_{yy} are known in closed form, we continue with a local mode analysis. The primary

difficulty in directly considering generalized eigenvalue problem (6.152) is that the eigenvalues of the composite tensors are not easily found in closed form. Using the technique of local mode analysis, however, we can investigate the effect of complicated terms such as

$$M_x \otimes \left(D_{yy} - D_y^T M_y^{-1} D_y\right). \tag{6.153}$$

By assumption, we impose Dirichlet boundary conditions $u^h = 0$ and $\phi^h = 0$ directly on the finite element space on the inflow boundary. For ease of computation, we also impose these constraints on boundary $x = 0$ of domain $\Omega = [0,1]^2$, although no boundary conditions are required since $\mathbf{b} = (1,0)$. We then write $M_y = M_x$, $D_y = D_x$, and $D_{yy} = D_{xx}$ explicitly as

$$M_y = h \begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & 0 & \dots & & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & & & \vdots \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & & & \\ \vdots & & & & \ddots & & \\ 0 & & & & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & & \dots & & 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix}, \tag{6.154}$$

$$D_y = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & \dots & & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 & & & \vdots \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & & & \\ \vdots & & & & \ddots & & \\ 0 & & & & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & & \dots & & 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}, \tag{6.155}$$

$$D_{yy} = \frac{1}{h} \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & & 0 \\ -1 & 2 & -1 & 0 & & & \vdots \\ 0 & -1 & 2 & -1 & & & \\ \vdots & & & & \ddots & & \\ 0 & & & & -1 & 2 & -1 \\ 0 & & \dots & & 0 & -1 & 1 \end{bmatrix}. \tag{6.156}$$

The $(1,1)$ coefficients of M_y, D_y, and D_{yy} are the same as the diagonal entries since homogeneous Dirichlet boundary conditions are enforced on the left and bottom boundaries. Likewise, the last entry (N,N) is adjusted in each of these matrices since no boundary conditions are imposed. The eigenvectors $\mathbf{v}_{yy}$ of hD_{yy} are easily verified to be

$$\mathbf{v}_{yy}^{(j)} = \begin{bmatrix} \vdots \\ \sin\left(\frac{i(2j-1)\pi}{2N}\right) \\ \vdots \end{bmatrix}, \tag{6.157}$$

with corresponding eigenvectors

$$\lambda^{(j)} = 2 - 2\cos\left(\frac{(2j-1)\pi}{2N}\right), \tag{6.158}$$

where $i, j = 1, \ldots, N$. Eigenvectors $\mathbf{v}_{xx}$ of hD_{xx} are similarly defined since $D_{xx} = D_{yy}$.

Let $\boldsymbol{\gamma}_x \otimes \boldsymbol{\gamma}_y = \mathbf{w}_{xx}^{(j)} \otimes \mathbf{w}_{yy}^{(k)}$, where

$$\mathbf{w}_{yy} = \sqrt{\frac{2}{N}}\mathbf{v}_{yy}. \tag{6.159}$$

The eigenvectors are rescaled to properly account for the mesh dependent weights on M_y and D_{yy}. These scales cancel in the following computations, except when (6.153) is transformed into the eigenspace of D_{yy}. We use j and k to index the spectrum in the x and y-directions, respectively. To simplify notation, we define

$$m_x^{(j)} = \langle M_y \mathbf{v}_{xx}^{(j)}, \mathbf{v}_{xx}^{(j)} \rangle, \tag{6.160}$$

$$d_x^{(j)} = \langle D_{xx} \mathbf{v}_{xx}^{(j)}, \mathbf{v}_{xx}^{(j)} \rangle, \tag{6.161}$$

$$m_y^{(k)} = \langle M_y \mathbf{v}_{yy}^{(k)}, \mathbf{v}_{yy}^{(k)} \rangle, \tag{6.162}$$

$$d_y^{(k)} = \langle D_{yy} \mathbf{v}_{yy}^{(k)}, \mathbf{v}_{yy}^{(k)} \rangle. \tag{6.163}$$

Substituting into (6.152) results in

$$K^h = \min_{j,k} \frac{m_x^{(j)} \cdot \nu_y^{(k)} + d_x^{(j)} \cdot m_y^{(k)}}{m_x^{(j)} \cdot d_y^{(k)} + d_x^{(j)} \cdot m_y^{(k)}}, \tag{6.164}$$

where

$$\nu_y^{(k)} = \operatorname{diag}\left(\mathbf{v}_{yy}{}^T (D_{yy} - D_y^T M_y^{-1} D_y) \mathbf{v}_{yy}\right) \tag{6.165}$$

and the columns of $\mathbf{v}_{yy}$ are $\mathbf{v}_{yy}^{(k)}$. Numerically, from mode analysis, we find

$$\min_k \nu_y^{(k)} = \mathcal{O}(h^2). \tag{6.166}$$

The eigenvalues of matrices M_x, D_{xx}, M_y, and D_{yy} are known. Specifically, we can easily obtain the following bounds:

$$\frac{1}{h} m_x^{(j)} \in [\frac{1}{4}, 1], \tag{6.167}$$

$$h d_x^{(j)} \in [\mathcal{O}(h^2), 4], \tag{6.168}$$

$$\frac{1}{h} m_y^{(k)} \in [\frac{1}{4}, 1], \tag{6.169}$$

$$h d_y^{(k)} \in [\mathcal{O}(h^2), 4]. \tag{6.170}$$

Together with (6.166), a proof by numerical computation yields

$$K^h = \mathcal{O}(h). \tag{6.171}$$

This allows us to interpret the value in (6.164) for different frequencies as is done below. For the matrices and eigenvectors defined above, Table 6.7 reports the rate at which K^h in (6.164) decreases between each grid level. We report grid sizes $N = 2^4, \ldots, 2^9$, where $h = \frac{1}{N}$. The

N	σ
16	
	0.998
32	
	0.995
64	
	1.000
128	
	0.996
256	
	1.000
512	

Table 6.7: Rates σ, where $K^h = \mathcal{O}(h^\sigma)$.

decrease in K is much slower than $\mathcal{O}(h^2)$, which means that the high-frequency error modes are filtered out by a proper choice of finite element spaces. Figure 6.18 illustrates the quantity in (6.164) over the spectrum for a grid with $N = 32$. Nearly identical plots can be obtained for larger grid sizes. It is clear that the minimum occurs for low frequency modes in the cross-stream

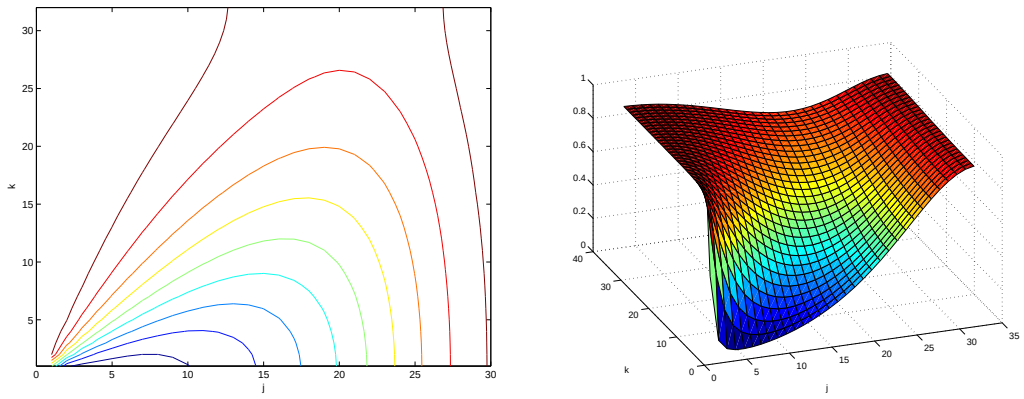
(k) and streamline (j) directions. Moreover, in the high frequency regions, particularly in the streamline direction, the values are nearly constant, indicating that our functional maintains coercivity for highly oscillatory modes. In summary, we have obtained in this analysis that there exists constant $c > 0$, independent of h, such that

$$ch\|\nabla^{\perp}\phi^h\|^2 \leq \|\mathbf{b}u^h - \nabla^{\perp}\phi^h\|^2_{0,\Omega}. \tag{6.172}$$

This implies that

$$ch\|\mathbf{w}^h\|^2 \leq \mathcal{F}(\mathbf{w}^h, u^h, 0) \quad \text{(cf. (6.128))}. \tag{6.173}$$

With this result, we are able to relate the error, $u - u^h$, in the L^2-norm to the functional and prove convergence with some additional assumptions. This is summarized in the following theorem.



(a) Contours. (b) Profile.

Figure 6.18: Spectral view of the quantity (6.164) with $N = 32$. The minimum value is K.

Theorem 6.26 (L^2 Convergence of the H^{-1} formulation). *Let $\mathcal{U}^h$ be a finite element space such that boundary conditions on u and ϕ are directly imposed. Let $\mathcal{U}^{h_j}$ be a nested sequence of*

grids with $h_{j+1} = \frac{1}{2}h_j$. *Let* $(\phi_*^{h_j}, u_*^{h_j})$ *be the solution to the minimization problem*

$$\begin{aligned}(\phi_*^{h_j}, u_*^{h_j}) &= \underset{\substack{\hat{u}^{h_j}\in\mathcal{U}^{h_j}\\ \hat{\phi}^{h_j}\in\mathcal{U}^{h_j}}}{\operatorname{argmin}} \mathcal{G}(\hat{\phi}^{h_j}, \hat{u}^{h_j}; g) \\ &= \underset{\substack{\hat{u}^{h_j}\in\mathcal{U}^{h_j}\\ \hat{\phi}^{h}\in\mathcal{U}^{h}}}{\operatorname{argmin}} \|\mathbf{b}u^{h_j} - \nabla^\perp\phi^{h_j}\|_{0,\Omega}^2\end{aligned} \tag{6.174}$$

Assume that there exists constant c_0 *such that*

$$\mathcal{G}(\phi_*^{h_j}, u_*^{h_j}; g) \le c_0 h_j^\beta. \tag{6.175}$$

Let (ϕ^{h_j}, u^{h_j}) *satisfy (6.175) up to a constant. That is,*

$$\mathcal{G}(\phi^{h_j}, u^{h_j}; g) \le c_0\hat{c}_0 h_j^\beta \tag{6.176}$$

for some $\hat{c}_1$ *independent of* h_j. *Furthermore, assume there exists constant* c_1 *such that*

$$c_1 h^\delta \|\nabla^\perp\left(\phi^{h_j} - \phi\right)\|_{0,\Omega}^2 \le \min_{\hat{u}^{h_j}\in\mathcal{U}^{h_j}} \|\mathbf{b}\hat{u}^{h_j} - \nabla^\perp\phi^{h_j}\|_{0,\Omega}^2, \qquad \forall\, \phi^{h_j}\in\mathcal{U}^{h_j} \tag{6.177}$$

and that $\beta > \delta \ge 0$. *Then*

$$\|u^{h_j} - u\| \le c h_j^\mu, \tag{6.178}$$

where $\mu \ge \beta - \delta > 0$ *and* (ϕ, u) *is the solution to linear advection problem (6.113) with* $\mathbf{w} = \nabla^\perp\phi$.

Proof. We start by establishing a bound on $\|u^{h_j}\|_{0,\Omega}$ given (6.177). By the triangle inequality, we have

$$\begin{aligned}\|u^{h_j} - u\|_{0,\Omega}^2 &\le 2\left(\|u^{h_j} - \mathbf{b}\cdot\nabla^\perp\phi^{h_j}\|_{0,\Omega}^2 + \|\mathbf{b}\cdot\nabla^\perp\phi^{h_j}\|_{0,\Omega}^2\right) \\ &\le 2\left(\|u^{h_j} - \mathbf{b}\cdot\nabla^\perp\phi^{h_j}\|_{0,\Omega}^2 + \|\mathbf{b}\cdot\nabla^\perp\phi^{h_j}\|_{0,\Omega}^2 + \|\mathbf{b}^\perp\cdot\nabla^\perp\phi^{h_j}\|_{0,\Omega}^2\right). \\ &= 2\left(\|u^{h_j} - \mathbf{b}\cdot\nabla^\perp\phi^{h_j}\|_{0,\Omega}^2 + \|\nabla^\perp\phi^{h_j}\|_{0,\Omega}^2\right).\end{aligned} \tag{6.179}$$

Together with (6.177), we arrive at

$$\begin{aligned}\|u^{h_j} - u\|_{0,\Omega}^2 &\le 2\left(\|u^{h_j} - \mathbf{b}\cdot\nabla^\perp\phi^{h_j}\|_{0,\Omega}^2 + \frac{1}{c_1 h_j^\delta}\left(\|u - \mathbf{b}\cdot\nabla^\perp\phi^{h_j}\|_{0,\Omega}^2 + \|\mathbf{b}^\perp\cdot\nabla^\perp\phi^{h_j}\|_{0,\Omega}^2\right)\right) \\ &\le 2\left(1 + \frac{1}{c_1 h_j^\delta}\right)\|\mathbf{b}u^{h_j} - \nabla^\perp\phi^{h_j}\|_{0,\Omega}^2.\end{aligned} \tag{6.180}$$

From (6.176), we have

$$\begin{aligned} \|u^{h_j} - u\|_{0,\Omega}^2 &\leq 2c_0\hat{c}_0 h^\beta \left(1 + \frac{1}{c_1 h_j^\delta}\right) \\ &= ch^{\beta-\delta} = ch^\mu, \end{aligned} \tag{6.181}$$

which completes the proof. □

6.4.3 Numerical Results

We now support the theoretical observations presented above with numerical evidence for the linear advection equation. We first illustrate the essence of Theorem 6.26 by considering an interpolant. Again, let $\mathbf{F}(u) = \mathbf{b}u$, with $\mathbf{b} = (1, 0)$. This simplifies the following heuristic reasoning, because a discontinuity on the left boundary is propagated only in the x-direction. We are then left to study

$$\|\partial_y \phi - u\|_0^2 \tag{6.182}$$

around the discontinuity. Let u be a discontinuous solution in 1-D, which is illustrated by the solid red line in Figure 6.19. Denote by $\Pi^h u$ the linear interpolant of u with width $\chi = ch^\sigma$, where $\sigma \leq 1$ and $c \in \mathbb{R}^+$ (dashed blue line in Figure 6.19). We choose a $\Pi^h \phi$ piecewise linear such that $\partial_y(\Pi^h \phi)$ is piecewise constant, which is shown as a dotted green line in Figure 6.19.

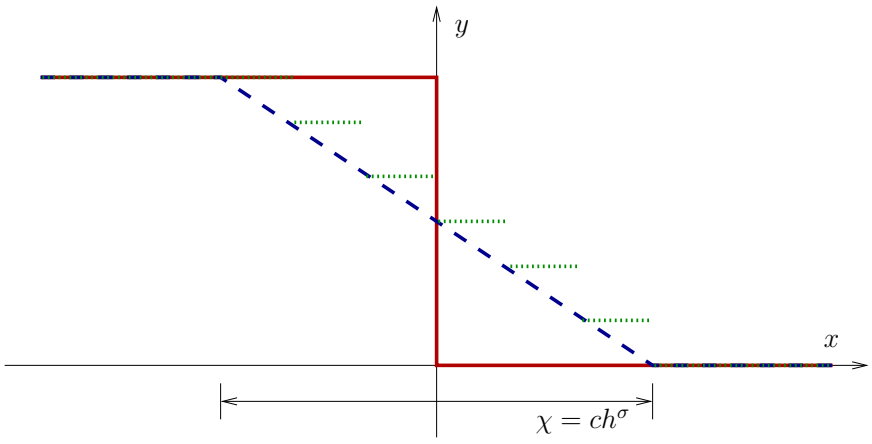


Figure 6.19: Interpolants $\Pi^h u$ and $\partial_y(\Pi^h \phi)$ (dashed blue and dotted green, respectively) with width $\chi = ch^\sigma$.

We now show that the functional error dominates the L^2-norm squared of the error. Let

$$e_u = u - \Pi^h u, \tag{6.183}$$

$$e_\phi = \phi - \Pi^h \phi, \tag{6.184}$$

be the error for the interpolants of the respective variables. It is then easily verified that

$$\|e_u\|_{0,\Omega}^2 = \frac{ch^\sigma}{12}. \tag{6.185}$$

With a width of $\chi = ch^\sigma$, the number of elements within the interpolated shock region is approximately

$$n_\chi = \lceil ch^{\sigma-1} \rceil, \tag{6.186}$$

where $\lceil \cdot \rceil$ is the ceiling function. With this, we then know that there are n_χ elements for which $\Pi^h \phi$ is piecewise constant. We arrive at

$$\begin{aligned} \mathcal{G}(e_u, e_\phi; g) &\leq \|\nabla^\perp(\phi - \Pi^h\phi) - \mathbf{b}(u - \Pi^h u)\|_{0,\Omega}^2, \\ &= \|\partial_y \Pi^h \phi - \Pi^h u\|_{0,\Omega}^2 + \underbrace{\|\partial_x(\phi - \Pi^h\phi)\|_{0,\Omega}^2}_{0} \\ &= n_\chi \|\partial_y \Pi^h \phi - \Pi^h u\|_{0,[0,h]}^2 \\ &= \frac{h^{2-\sigma}}{3c}. \end{aligned} \tag{6.187}$$

For example, if we choose Φ^h to have width $\chi = ch$, then the functional error is $\mathcal{O}(h)$ and the L^2-norm squared of the error is also $\mathcal{O}(h)$. These are the rates we observe for the nonlinear problem, as shown in Tables 6.2, 6.3, 6.4, and 6.6. This indicates that, in the nonlinear problem, smearing of the shock in the numerical approximation is of width $\mathcal{O}(h)$. If, on the other hand, we choose Φ^h to have width $\chi = ch^{\frac{1}{2}}$, then the heuristic closely matches the following constant advection example.

Example 9 (Constant Advection). *We consider (4.1), where* $\mathbf{b} = (\cos(\frac{\pi}{6}), \sin(\frac{\pi}{6}))$*, which is divergence-free. The space-time flow domain is given by* $\Omega = [0,1] \times [0,1]$*, with initial and*

boundary conditions

$$u(x,t) = \begin{cases} 0.0 & \text{if } t = 0, \\ 1.0 & \text{if } x = 0. \end{cases} \tag{6.188}$$

The solution of this problem contains a discontinuity propagating with speed $s = \tan(\frac{\pi}{6})$ *from the origin,* $(x,t) = (0,0)$*. The conserved quantity is* $u(x,t) = 1.0$ *to the left of the discontinuity and* $u(x,t) = 0.0$ *to the right (cf. Example 1).*

Figure 6.20 shows an adaptively refined solution, similar to Figure 6.14. Solution quality closely resembles the profiles obtained for the least-squares formulations developed in Chapter 4. The convergence rates in Table 6.8 approach 0.5 for the L^2-norm squared of the error, while the functional error converges at a rate of 1.5. Similar to the nonlinear example, the error in the locally refined solution is almost equal to the error magnitude of the uniformly refined approach.

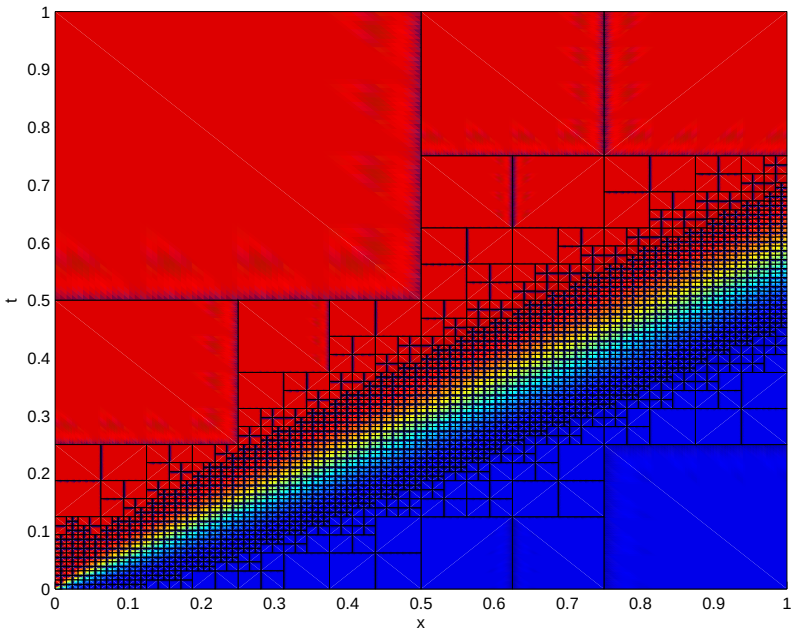


Figure 6.20: H^{-1} formulation, Example 9: solution u^h on an adaptively refined grid with a finest resolution of $h = \frac{1}{128}$.

N	Nodes	$\|\cdot\|^2_{0,\Omega}$	α	$\|\cdot\|^2_{\mathcal{G}}$	α	$\|\cdot\|^2_{\mathcal{G}_{\text{int}}}$	α	$\|\cdot\|^2_{\mathcal{G}_{\text{bdy}}}$	α
16	289	2.61e-02		1.02e-02		1.19e-03		9.04e-03	
			0.47		1.05		1.51		1.00
32	1089	1.89e-02		4.93e-03		4.20e-04		4.52e-03	
			0.47		1.04		1.52		1.00
64	4225	1.36e-02		2.40e-03		1.47e-04		2.26e-03	
			0.46		1.03		1.52		1.00
128	16641	9.72e-03		1.18e-03		5.12e-05		1.13e-03	
			0.48		1.02		1.52		1.00
256	66049	6.93e-03		5.82e-04		1.79e-05		5.64e-04	
16	148	2.40e-02		1.03e-02		1.24e-03		9.04e-03	
			0.42		1.06		1.54		1.00
32	428	1.80e-02		4.94e-03		4.27e-04		4.51e-03	
			0.47		1.04		1.52		1.00
64	1303	1.31e-02		2.41e-03		1.49e-04		2.26e-03	
			0.48		1.03		1.52		1.00
128	3937	9.33e-03		1.18e-03		5.21e-05		1.13e-03	
			0.48		1.02		1.52		1.00
256	12052	6.70e-03		5.82e-04		1.82e-05		5.64e-04	

Table 6.8: H^{-1} formulation, Example 9: convergence rates. Top section, uniform grid refinement. Bottom section: adaptive grid refinement.

Letting smearing width $\chi = ch^{\frac{1}{2}}$ in the interpolation heuristic described above yields $\mathcal{O}(h^{\frac{3}{2}})$ for the functional interpolation error and $\mathcal{O}(h^{\frac{1}{2}})$ for the L^2-norm squared of the interpolation error. This agrees closely with the values in Table 6.8, indicating that the smearing width is $\mathcal{O}(h^{\frac{1}{2}})$ for the linear advection problem. From the results in Table 6.8, we can also confirm Theorem 6.26. The gap of $\mathcal{O}(h)$ found between the convergence rate of the functional error and the L^2-norm of the error in (6.173) precisely matches the numerical results in Table 6.8 and the gap following from the interpolant with shock width $\chi = ch^{\frac{1}{2}}$.

Remark 6.27. *The values presented in Table 6.8 also agree with the convergence rates obtained in Section 4.5 for standard bilinear finite elements. It is conceivable that increasing the polynomial order for the H^{-1} formulation improves the convergence rate in a similar fashion as in Section 4.5. With this, the functional error and L^2-norm squared of the error would both approach convergence rates of 1.0, resulting in less smearing with a width of $\mathcal{O}(h)$, thus agreeing with the nonlinear results presented in this section.*

Chapter 7

Concluding Remarks

The focus of this thesis has been on least-squares finite elements methods (LSFEMs) for hyperbolic partial differential equations (PDEs) in a multilevel context. Solutions containing contact discontinuities and shocks are difficult to resolve numerically because many methods yield approximations with spurious oscillations and excessive overshoots. Moreover, methods for approximating such solutions are often computationally expensive and fail to offer a satisfactory approach when considering the total cost and accuracy of the algorithm. We find the least-squares methodology to be a promising strategy that effectively addresses both of these concerns. The least-squares finite element method results in some smearing of the discontinuity that can be remedied by adaptive refinement without introducing further overshoots in the approximation. Adaptively refining the numerical solution is naturally facilitated by the least-square minimization principle. The LSFEM yields algebraic systems that are symmetric positive definite and that have been found to be efficiently handled by multigrid solution methods. We provide a theoretical footing and numerical evidence to validate the performance of the least-squares methods proposed for linear and nonlinear hyperbolic equations. Also, a detailed study of multigrid iterative methods is presented for the linear case.

7.1 Thesis Contributions

Chapter 4 considered linear hyperbolic PDEs and, in particular, a comprehensive study of the numerical approximation of discontinuous solutions. Existence and uniqueness of the

least-squares minimization problem was proved. A new discontinuous LSFEM was proposed that can be considered as a least-squares equivalent to popular discontinuous Galerkin (DG) methods for hyperbolic problems. We used nonconforming finite elements that allow jumps in the numerical approximation and thus more closely match the smoothness of the solution to the PDE. A thorough numerical study for constant and variable flow fields confirmed convergence of both LSFEMs. Numerical results showed improved convergence rates for the LSFEM and DLSFEM as the polynomial order of the finite elements increased. Per degree of freedom, the smearing in the numerical solution was reduced as the polynomial order increased, without introducing excessive oscillations. Our results indicated no apparent advantage in using the nonconforming LS method, which is interesting since DG has proved superior to the Streamline Upwind Petrov Galerkin (SUPG) method.

Chapter 5 initiated a study of multigrid for the algebraic systems resulting from the LSFEM studied in Chapter 4. Standard geometric multigrid methods were found to yield poor convergence due to the strong anisotropies in the problem. Thus, algebraic based methods were considered, namely the Ruge-Stüben Algebraic Multigrid (AMG) method. With the least-squares method proposed in Chapter 4, AMG was found to yield adequate performance. With modification of the PDE, the new LS formulation included boundary terms that allowed more efficient AMG solves. For constant advection, we obtained optimal AMG convergence. For problems with widely varying anisotropies, we found significant improvement with the new formulation and near-optimal performance of AMG, particularly when coupled with a preconditioned conjugate gradient (PCG) wrapper.

Finally, Chapter 6 investigated LSFEMs for nonlinear conservation laws. New formulations were introduced that enable numerical approximation in a Gauss-Newton setting. The solution to the associated LS minimization problems were shown to converge to the weak solution of the nonlinear conservation law. We argued that the methods are footed in an H^{-1} Sobolev setting and, with appropriate choices of finite element spaces, convergence of the numerical approximation to the weak solution in an L^2 sense was shown. Numerically, we confirmed

these methods for Burger's equation and optimal convergence was observed resolving a shock simulation with continuous finite elements. Correct weak solutions were found for the several examples containing shocks that we considered and the entropy solution was found for a rarefaction flow. Finally, we examined an adaptive approach in a grid continuation setting. The error in the adaptively refined approximation equaled the error in the uniform case, while using far fewer nodes. Moreover, the ratio of the number of nodes in the adaptive case to the uniform case was found to decrease with grid size.

7.2 Ongoing Work

Least-squares methods have only recently been gaining popularity for elliptic problems. Such methods for hyperbolic conservation laws are also slowly gaining acceptance as viable approaches. We exposed many possibilities in this dissertation for future development.

One direction of importance is to develop a firmer understanding of shock and discontinuity smearing. We have addressed this briefly in Chapters 4 and 6 and found that the smearing is reduced per degree of freedom with the use of higher order elements. Moreover, we observed less smearing in the case of nonlinear equations: $\mathcal{O}(h^{\frac{1}{2}})$ versus $\mathcal{O}(h)$ for linear equations (Section 6.4). Convergence in regions both near to and far away from the discontinuity remains largely unstudied and a more thorough investigation is needed.

More robust AMG solvers need to be developed that more effectively handle widely varying anisotropies. We considered AMG for the linear problem and found satisfactory results for mildly varying flows. Preliminary findings also show that optimal solvers may be attainable for the nonlinear problem.

Finally, we are continuing to extend the theoretical assertions and generality of the methods proposed in Chapter 6. The convergence heuristics appear to hold for weak treatment of boundary conditions, which is important for the theoretical justification of the method. As we extend the methods to higher spatial dimensions and systems of equations, we also need to address the idea of compatible finite element spaces with a more general theory.

Bibliography

[1] R. Abgrall. Toward the ultimate conservative scheme: following the quest. J. Comput. Phys., 167(2):277–315, 2001.

[2] Robert A. Adams. Sobolev spaces. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65.

[3] Mark Ainsworth and J. Tinsley Oden. A posteriori error estimation in finite element analysis. Pure and Applied Mathematics. Wiley-Interscience [John Wiley & Sons], New York, 2000.

[4] Timothy J. Barth. Numerical methods for gasdynamic systems on unstructured meshes. In An introduction to recent developments in theory and numerics for conservation laws (Freiburg/Littenweiler, 1997), volume 5 of Lect. Notes Comput. Sci. Eng., pages 195–285. Springer, Berlin, 1999.

[5] Timothy J. Barth and Herman Deconinck, editors. High-order methods for computational physics, volume 9 of Lecture Notes in Computational Science and Engineering. Springer-Verlag, Berlin, 1999.

[6] Markus Berndt. Adaptive Refinement and the Treatment of Discontinuous Coefficients for Multilevel First-Order System Least Squares (FOSLS). Ph.d. dissertation, University of Colorado at Boulder, Boulder, CO, 1999.

[7] Markus Berndt, Thomas A. Manteuffel, and Stephen F. McCormick. Local error estimates and adaptive refinement for first-order system least squares (FOSLS). Electron. Trans. Numer. Anal., 6(Dec.):35–43 (electronic), 1997. Special issue on multilevel methods (Copper Mountain, CO, 1997).

[8] Kim S. Bey, J. Tinsley Oden, and Abani Patra. A parallel *hp*-adaptive discontinuous Galerkin method for hyperbolic conservation laws. Appl. Numer. Math., 20(4):321–336, 1996. Adaptive mesh refinement methods for CFD applications (Atlanta, GA, 1994).

[9] P. B. Bochev and J. Choi. Improved least-squares error estimates for scalar hyperbolic problems. Comput. Methods Appl. Math., 1(2):115–124, 2001.

[10] Pavel Bochev, Jonathan Hu, Allen C. Robinson, and Ray Tuminaro. Towards robust 3d z-pinch simulations: Discretization and fast solvers for magnetic diffusion in heterogeneous conductors. Technical report, Sandia National Laboratories, 2001. SAND Report 2001 8363J.

[11] Pavel B. Bochev and Jungmin Choi. A comparative study of least-squares, SUPG and Galerkin methods for convection problems. Int. J. Comput. Fluid Dyn., 15(2):127–146, 2001.

[12] Pavel B. Bochev and Max D. Gunzburger. Finite element methods of least-squares type. SIAM Rev., 40(4):789–837 (electronic), 1998.

[13] Dietrich Braess. Finite elements. Cambridge University Press, Cambridge, second edition, 2001. Theory, fast solvers, and applications in solid mechanics, Translated from the 1992 German edition by Larry L. Schumaker.

[14] James H. Bramble, Raytcho D. Lazarov, and Joseph E. Pasciak. A least-squares approach based on a discrete minus one inner product for first order systems. Math. Comp., 66(219):935–955, 1997.

[15] A. Brandt, S. McCormick, and J. Ruge. Algebraic multigrid (AMG) for sparse matrix equations. In Sparsity and its applications (Loughborough, 1983), pages 257–284. Cambridge Univ. Press, Cambridge, 1985.

[16] Achi Brandt. Algebraic multigrid theory: the symmetric case. Appl. Math. Comput., 19(1-4):23–56, 1986. Second Copper Mountain conference on multigrid methods (Copper Mountain, Colo., 1985).

[17] Achi Brandt and Boris Diskin. Multigrid solvers for nonaligned sonic flows. SIAM J. Sci. Comput., 21(2):473–501 (electronic), 1999.

[18] Susanne C. Brenner and L. Ridgway Scott. The mathematical theory of finite element methods, volume 15 of Texts in Applied Mathematics. Springer-Verlag, New York, 1994.

[19] Alberto Bressan. Hyperbolic systems of conservation laws, volume 20 of Oxford Lecture Series in Mathematics and its Applications. Oxford University Press, Oxford, 2000. The one-dimensional Cauchy problem.

[20] William L. Briggs, Van Emden Henson, and Steve F. McCormick. A multigrid tutorial. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2000.

[21] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick. First-order system least squares for second-order partial differential equations. I. SIAM J. Numer. Anal., 31(6):1785–1799, 1994.

[22] Z. Cai, T. A. Manteuffel, S. F. McCormick, and J. Ruge. First-order system $\mathcal{LL}^*$ (FOSLL*): scalar elliptic partial differential equations. SIAM J. Numer. Anal., 39(4):1418–1445 (electronic), 2001.

[23] Zhiqiang Cai, Thomas A. Manteuffel, and Stephen F. McCormick. First-order system least squares for second-order partial differential equations. II. SIAM J. Numer. Anal., 34(2):425–454, 1997.

[24] Graham F. Carey and Bo Nan Jiang. Least-squares finite elements for first-order hyperbolic systems. Internat. J. Numer. Methods Engrg., 26(1):81–93, 1988.

[25] Jungmin Choi. The Least-Squares Method for Hyperbolic Problems. Ph.d. dissertation, University of Texas at Arlington, Arlington, TX, 2000.

[26] Philippe G. Ciarlet. The finite element method for elliptic problems, volume 40 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original [North-Holland, Amsterdam; MR **58** #25001].

[27] Bernardo Cockburn. Discontinuous Galerkin methods for convection-dominated problems. In High-order methods for computational physics, volume 9 of Lect. Notes Comput. Sci. Eng., pages 69–224. Springer, Berlin, 1999.

[28] Andrea L. Codd. Elasticity-Fluid Coupled Systems and Elliptic Grid Generation (EGG) based on First-Order System Least Squares (FOSLS). Ph.d. dissertation, University of Colorado at Boulder, Boulder, CO, 2001.

[29] Constantine M. Dafermos. Hyperbolic conservation laws in continuum physics, volume 325 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 2000.

[30] Hans De Sterck. Numerical simulation and analysis of magnetically dominated MHD bow shock flows with applications in space physics. PhD thesis, Katholieke Universiteit Leuven (Belgium), and NationalCenter for Atmospheric Research, Boulder, Colorado (USA), 1999.

[31] Hans De Sterck, Thomas A. Manteuffel, Stephen F. McCormick, and Luke Olson. Least-squares finite element methods for linear hyperbolic pdes. SIAM J. Sci. Comput., 2002. submitted.

[32] J. E. Dennis, Jr. and Robert B. Schnabel. Numerical methods for unconstrained optimization and nonlinear equations, volume 16 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. Corrected reprint of the 1983 original.

[33] Boris Diskin. Efficient multigrid methods for solving upwind-biased discretizations of the convection equation. Appl. Math. Comput., 123(3):343–379, 2001.

[34] Eduardo Gomes Dutra do Carmo and Augusto Cesar Galeão. Feedback Petrov-Galerkin methods for convection-dominated problems. Comput. Methods Appl. Mech. Engrg., 88(1):1–16, 1991.

[35] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. Computational differential equations. Cambridge University Press, Cambridge, 1996.

[36] Lawrence C. Evans. Partial differential equations, volume 19 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 1998.

[37] Augusto Cesar Galeão and Eduardo Gomes Dutra do Carmo. A consistent approximate upwind Petrov-Galerkin method for convection-dominated problems. Comput. Methods Appl. Mech. Engrg., 68(1):83–95, 1988.

[38] Vivette Girault and Pierre-Arnaud Raviart. Finite element methods for Navier-Stokes equations, volume 5 of Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 1986. Theory and algorithms.

[39] Edwige Godlewski and Pierre-Arnaud Raviart. Numerical approximation of hyperbolic systems of conservation laws, volume 118 of Applied Mathematical Sciences. Springer-Verlag, New York, 1996.

[40] S. K. Godunov. A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. Mat. Sb. (N.S.), 47 (89):271–306, 1959.

[41] Ronald B. Guenther and John W. Lee. Partial differential equations of mathematical physics and integral equations. Dover Publications Inc., Mineola, NY, 1996. Corrected reprint of the 1988 original.

[42] V. Henson and U. Meier Yang. BoomerAMG: a parallel algebraic multigrid solver and preconditioner. Technical Report UCRL-JC-141495, Lawrence Livermore National Laboratory, 2001.

[43] Ralf Hiptmair and Jan Metzger. Automated local mode analysis. Technical Report 174, Universität Tübingen, February 2002. SFB 382.

[44] G. Horton and S. Vandewalle. A space-time multigrid method for parabolic partial differential equations. SIAM J. Sci. Comput., 16(4):848–864, 1995.

[45] Thomas Y. Hou and Philippe G. LeFloch. Why nonconservative schemes converge to wrong solutions: error analysis. Math. Comp., 62(206):497–530, 1994.

[46] P. Houston, J. A. Mackenzie, E. Süli, and G. Warnecke. A posteriori error analysis for numerical approximations of Friedrichs systems. Numer. Math., 82(3):433–470, 1999.

[47] Paul Houston, Max Jensen, and Endre Süli. *hp*-discontinuous Galerkin finite element methods with least-squares stabilization. In Proceedings of the Fifth International Conference on Spectral and High Order Methods (ICOSAHOM-01) (Uppsala), volume 17, pages 3–25. 2002.

[48] Paul Houston, Christoph Schwab, and Endre Süli. Stabilized *hp*-finite element methods for first-order hyperbolic problems. SIAM J. Numer. Anal., 37(5):1618–1643 (electronic), 2000.

[49] Arieh Iserles. A first course in the numerical analysis of differential equations. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 1996.

[50] Bo Nan Jiang. The least-squares finite element method. Scientific Computation. Springer-Verlag, Berlin, 1998. Theory and applications in computational fluid dynamics and electromagnetics.

[51] Bo Nan Jiang and Graham F. Carey. A stable least-squares finite element method for nonlinear hyperbolic problems. Internat. J. Numer. Methods Fluids, 8(8):933–942, 1988.

[52] Bo Nan Jiang and Graham F. Carey. Least-squares finite element methods for compressible Euler equations. Internat. J. Numer. Methods Fluids, 10(5):557–568, 1990.

[53] Fritz John. Partial differential equations, volume 1 of Applied Mathematical Sciences. Springer-Verlag, New York, fourth edition, 1991.

[54] C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. Math. Comp., 46(173):1–26, 1986.

[55] Claes Johnson, Uno Nävert, and Juhani Pitkäranta. Finite element methods for linear hyperbolic problems. Comput. Methods Appl. Mech. Engrg., 45(1-3):285–312, 1984.

[56] Dietmar Kröner. Numerical schemes for conservation laws. Wiley-Teubner Series Advances in Numerical Mathematics. John Wiley & Sons Ltd., Chichester, 1997.

[57] P. Lasaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In Mathematical aspects of finite elements in partial differential equations (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1974), pages 89–123. Publication No. 33. Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974.

[58] Peter Lax and Burton Wendroff. Systems of conservation laws. Comm. Pure Appl. Math., 13:217–237, 1960.

[59] Peter D. Lax. Hyperbolic systems of conservation laws and the mathematical theory of shock waves. Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1973. Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 11.

[60] Philippe G. LeFloch. Hyperbolic systems of conservation laws. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2002. The theory of classical and nonclassical shock waves.

[61] Randall J. LeVeque. Numerical methods for conservation laws. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 1990.

[62] Randall J. LeVeque. Finite volume methods for hyperbolic problems. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2002.

[63] E. E. Lewis and J. W. F. Miller. Computational methods of neutron transport. American Nuclear Society, La Grange Park, IL, 1993.

[64] Thomas A. Manteuffel, Klaus J. Ressel, and Gerhard Starke. A boundary functional for the least-squares finite-element solution of neutron transport problems. SIAM J. Numer. Anal., 37(2):556–586 (electronic), 2000.

[65] Stephen F. McCormick. Multilevel adaptive methods for partial differential equations, volume 6 of Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1989.

[66] Stephen F. McCormick. First-order system least squares philosophy. unpublished notes, 1995.

[67] Andreas Meister and Jens Struckmeier, editors. Hyperbolic partial differential equations. Friedr. Vieweg & Sohn, Braunschweig, 2002. Theory, numerics and applications, Papers from the Summer School held at the Technical University of Hamburg-Harburg, Hamburg, March 2001.

[68] J. Ruge and K. Stüben. Efficient solution of finite difference and finite element equations. In Multigrid methods for integral and differential equations (Bristol, 1983), pages 169–212. Oxford Univ. Press, New York, 1985.

[69] John Ruge. FOSPACK Users Manual, 2001. Version 1.0.

[70] Lewis H. Ryder. Quantum field theory. Cambridge University Press, Cambridge, second edition, 1996.

[71] Rüdiger Schüller. A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques. John Wiley & Sons Inc., Chichester, 1996.

[72] Christoph Schwab. *hp*-FEM for fluid flow simulation. In High-order methods for computational physics, volume 9 of Lect. Notes Comput. Sci. Eng., pages 325–438. Springer, Berlin, 1999.

[73] L. Ridgway Scott and Shangyou Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. Math. Comp., 54(190):483–493, 1990.

[74] Denis Serre. Systems of conservation laws. 1. Cambridge University Press, Cambridge, 1999. Hyperbolicity, entropies, shock waves, Translated from the 1996 French original by I. N. Sneddon.

[75] Walter A. Strauss. Partial differential equations. John Wiley & Sons Inc., New York, 1992. An introduction.

[76] E. Süli. A posteriori error analysis and global error control for adaptive finite volume approximations of hyperbolic problems. In Numerical analysis 1995 (Dundee, 1995), volume 344 of Pitman Res. Notes Math. Ser., pages 169–190. Longman, Harlow, 1996.

[77] Endre Süli and Paul Houston. Adaptive finite element approximation of hyperbolic problems. In Error estimation and adaptive discretization methods in computational fluid dynamics, volume 25 of Lect. Notes Comput. Sci. Eng., pages 269–344. Springer, Berlin, 2003.

[78] Endre Süli, Paul Houston, and Christoph Schwab. *hp*-finite element methods for hyperbolic problems. In The mathematics of finite elements and applications, X, MAFELAP 1999 (Uxbridge), pages 143–162. Elsevier, Oxford, 2000.

[79] U. Trottenberg, C. W. Oosterlee, and A. Schüller. Multigrid. Academic Press Inc., San Diego, CA, 2001. With contributions by A. Brandt, P. Oswald and K. Stüben.

[80] Irad Yavneh, Cornelis H. Venner, and Achi Brandt. Fast multigrid solution of the advection problem with closed characteristics. SIAM J. Sci. Comput., 19(1):111–125 (electronic), 1998. Special issue on iterative methods (Copper Mountain, CO, 1996).

[81] Kōsaku Yosida. Functional analysis. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the sixth (1980) edition.

Appendix A

Nonlinear Conservation Laws and Linearization

In this appendix, we investigate the linearizability of the nonlinear conservation law

$$\nabla \cdot \mathbf{F}(u) = 0. \tag{A.1}$$

The formulation presented in Chapter 4, in the context of nonlinear conservation laws, is considered as well as the reformulations of Chapter 6. We discuss implications for the convergence of the Newton method.

A.1 1-D Scalar Function Example

Consider the 1-D scalar function

$$f(x) = \begin{cases} x^\alpha, & \text{if } x > 0, \\ |x|^\alpha, & \text{if } x < 0, \end{cases} \tag{A.2}$$

where $\alpha \in (0, \frac{1}{2})$. The solution to $f(x) = 0$ is clearly $x^* = 0$. Recall that the Newton method for solving the equation $f(x) = 0$ is given by

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}. \tag{A.3}$$

In this case, we have $f'(x) = \alpha|x|^{\alpha-1}$, which is unbounded at $x^* = 0$. The Newton iteration becomes

$$\begin{aligned} x_{i+1} &= x_i - \frac{x_i^\alpha}{\alpha|x_i|^{\alpha-1}} \\ &= (1 - \frac{1}{\alpha})x_i. \end{aligned} \tag{A.4}$$

The iteration diverges since $0 < \alpha < \frac{1}{2}$ and the quantity $|(1-\frac{1}{\alpha})| > 1$. Thus, if $f'(x^*) \to \infty$, then the basin of attraction can be empty. This cannot happen when $f'(x)$ is Lipschitz continuous in a neighborhood of x^*, which is a typical assumption in convergence proofs for Newton's method [32].

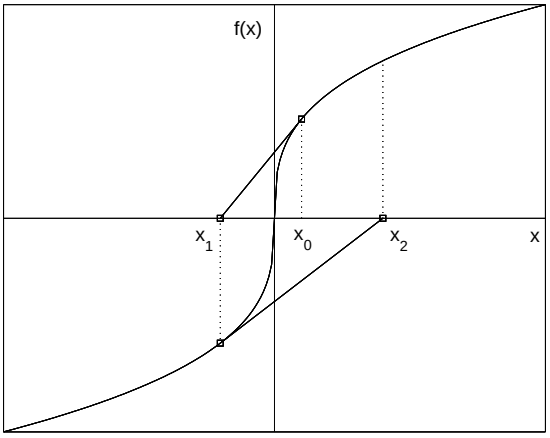


Figure A.1: Newton method applied to the problem $f(x) = 0$.

A.2 Standard Scalar Conservation Law Operator

To solve

$$\mathcal{L}(u) = \nabla \cdot \mathbf{F}(u) = 0, \tag{A.5}$$

we use Newton's method given by

$$\mathcal{L}(u_j) + d\mathcal{L}|_{u_j}[u_{j+1} - u_j] = 0, \tag{A.6}$$

where $d\mathcal{L}$ is the Fréchet derivative of $\mathcal{L}$. The Fréchet derivative at u in the direction $\hat{u}$ is

$$\begin{aligned} d\mathcal{L}|_u[\hat{u}] &= \lim_{\varepsilon \to 0} \frac{\mathcal{L}(u + \varepsilon\hat{u}) - \mathcal{L}(u)}{\varepsilon} \\ &= \nabla \cdot (d\mathbf{F}|_u[\hat{u}]), \end{aligned} \tag{A.7}$$

Proposition A.1. *The Fréchet derivative operator $d\mathcal{L}|_u : H^{\frac{1}{2}-\varepsilon} \to L^2(\Omega)$ for (A.5), given by (A.7), is unbounded:*

$$\|d\mathcal{L}|_u\| = \infty, \tag{A.8}$$

where $\|\cdot\|$ *denotes the associated operator norm.*

Proof. The operator norm squared is

$$\begin{aligned}\|d\mathcal{L}|_u[\hat{u}]\|^2 &= \sup_{\substack{\hat{u}\in H^{\frac{1}{2}-\varepsilon},\\ \|\hat{u}\|_{\frac{1}{2}-\varepsilon}=1}} \|d\mathcal{L}|_u[\hat{u}]\|^2_{0,\Omega} \\ &= \sup_{\substack{\hat{u}\in H^{\frac{1}{2}-\varepsilon},\\ \|\hat{u}\|_{\frac{1}{2}-\varepsilon}=1}} \|\nabla\cdot(d\mathbf{F}|_u[\hat{u}])\|^2_{0,\Omega} \\ &= \infty,\end{aligned} \tag{A.9}$$

since, in general, $d\mathbf{F}|_u[v] \notin H(div,\Omega)$ for $u, \hat{u} \in H^{\frac{1}{2}-\varepsilon}(\Omega)$. □

This observation implies that we may encounter problematic Newton convergence when applying the Newton procedure directly to the standard conservation law formulation. For example, define the least-squares functional

$$\mathcal{H}(u;g) := \|\nabla\cdot\mathbf{F}(u)\|^2_{0,\Omega} + \|u-g\|^2_{0,\Gamma_I}, \tag{A.10}$$

and formulate the minimization over a finite-dimensional subspace

$$u^h_* = \underset{u^h\in\mathcal{U}^h}{\operatorname{argmin}}\,\mathcal{H}(u;g), \tag{A.11}$$

with bilinear elements on quadrilaterals. The least-squares approximation fails to converge for a discontinuous solution when $\mathbf{F}(u)$ is nonlinear. We illustrate this difficulty with the Burger's equation, with boundary data $u = 1$ on the left boundary and $u = 0$ on the right. Figure A.2 shows u^h contours for the shock. The approximation is clearly not a weak solution and fails to resolve the shock, particularly at the inflow boundary. On each level, the Newton procedure converges, but to an incorrect solution. The L_2 error and the nonlinear functional fail to converge as the grid is refined. It is plausible that this behavior of the non-$H(div)$-conforming Gauss-Newton LSFEM is caused by the unboundedness of the Fréchet derivative. The method seems to converge to a spurious solution, which may be a spurious stationary point of the functional (A.10). It is interesting to note that the numerical approximation resembles a rarefied solution emanating from the outflow boundary.

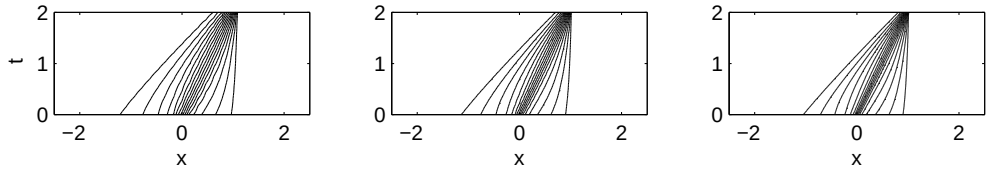


Figure A.2: Non-$H(div)$-conforming LSFEM: shock problem on grids with 32^2, 64^2, and 128^2 elements, which illustrates that u^h does not converge to the weak solution.

It may be possible to obtain a convergent LSFEM for the functional $\mathcal{H}(u;g)$ (A.10) by employing other nonlinear solution procedures than the Gauss-Newton method that do not rely on linearization and the Fréchet derivative. This remains the subject of further study.

Remark A.2. *It is interesting to note that by adding numerical dissipation to the standard conservation law formulation, the unboundedness of the Fréchet derivative in the Newton procedure can be remedied. This may be the reason why convergence problems do not immediately show up when Newton methods are used for implicit schemes that rely on discretizations of (A.5) in which numerical dissipation is added. In the LSFEM presented here, numerical dissipation is not added, and the Newton convergence problem appears.*

A.3 H^{-1} Reformulation of the Conservation Law

We now consider operator $\mathcal{L}(\phi, u)$ of the form

$$\nabla^{\perp}\phi - \mathbf{F}(u) = 0. \tag{A.12}$$

The Fréchet derivative at (ϕ, u) in the direction $(\hat{\phi}, \hat{u})$ is then

$$d\mathcal{L}|_{(\phi,u)}[(\hat{\phi}, \hat{u})] = \nabla^{\perp}\hat{\phi} - d\mathbf{F}|_u[\hat{u}], \tag{A.13}$$

Lemma A.3. *The Fréchet derivative operator* $d\mathcal{L}|_{(\phi,u)} : H^1(\Omega) \times L^2(\Omega) \to L^2(\Omega)$ *for (A.12), given by (A.13), is bounded:*

$$\|\mathcal{L}|_{(\phi,u)}\| = \sqrt{1 + K^2}, \tag{A.14}$$

where $\|\cdot\|$ *denotes the associated operator norm and* K *is the Lipschitz constant. That is,* K *satisfies*

$$|F_i(u_1) - F_i(u_2)| \leq K|u_1 - u_2|, \tag{A.15}$$

for all u_1, u_2, *and* $i = 1, 2$, *where* F_i *are the components of* $\mathbf{F}$.

Proof. The operator norm squared becomes

$$\begin{aligned}
\|d\mathcal{L}_{(\phi,u)}\|^2 &= \sup_{\substack{\hat{\phi}\in H^1(\Omega),\, \hat{u}\in L^2(\Omega),\\ \|\hat{\phi}\|^2_{H^1(\Omega)}+\|\hat{u}\|^2_{0,\Omega}=1}} \|d\mathcal{L}|_{(\phi,u)}[(\hat{\phi},\hat{u})]\|^2_{0,\Omega} \\
&= \sup_{\substack{\hat{\phi}\in H^1(\Omega),\, \hat{u}\in L^2(\Omega),\\ \|\hat{\phi}\|^2_{H^1(\Omega)}+\|\hat{u}\|^2_{0,\Omega}=1}} \left(\|\nabla^\perp\hat{\phi}\|^2_{0,\Omega} + \|d\mathbf{F}|_u[\hat{u}]\|^2_{0,\Omega}\right) \\
&\leq \sup_{\substack{\hat{\phi}\in H^1(\Omega),\, \hat{u}\in L^2(\Omega),\\ \|\hat{\phi}\|^2_{H^1(\Omega)}+\|\hat{u}\|^2_{0,\Omega}=1}} \left(\|\hat{\phi}\|^2_{H^1(\Omega)} + K^2\|\hat{u}\|^2_{0,\Omega}\right) \\
&= 1 + K^2,
\end{aligned} \tag{A.16}$$

where K is the Lipschitz constant. Here we have used

$$|d\mathbf{F}|_u[\hat{u}]| \leq K|\hat{u}|, \quad \text{a.e..} \tag{A.17}$$

□

This result indicates that, for this formulation, Newton's method will not encounter the same difficulties as for the formulation presented above. This was indeed confirmed numerically in Section 6.2.

A.4 $H(div)$-Conforming Reformulation of the Conservation Law

Finally, for completeness, we consider operator $\mathcal{L}(\mathbf{w}, u)$ of the form

$$\nabla \cdot \mathbf{w} = 0 \tag{A.18a}$$

$$\mathbf{w} - \mathbf{F}(u) = 0. \tag{A.18b}$$

The Fréchet derivative at $(\mathbf{w}, u)$ in the direction $(\hat{\mathbf{w}}, \hat{u}$ is then

$$d\mathcal{L}|_{(\mathbf{w},u)}[(\hat{\mathbf{w}}, \hat{u})] = \begin{bmatrix} \nabla \cdot (\hat{\mathbf{w}}) \\ \\ \hat{\mathbf{w}} - d\mathbf{F}|_u[\hat{u}] \end{bmatrix}, \tag{A.19}$$

Lemma A.4. *The Fréchet derivative operator* $d\mathcal{L}|_{(\mathbf{w},u)} : H(div, \Omega) \times L^2(\Omega) \to L^2(\Omega)$ *for (A.18), given by (A.19), is bounded:*

$$\|\mathcal{L}|_u\| \leq \sqrt{1 + K^2}. \tag{A.20}$$

Proof. The proof is analogous to the proof of Lemma A.3. □