Institute for Advanced Studies in Basic Sciences

Streamline Diffusion Finite Element Method for Coupling Equations of Nonlinear Hyperbolic Scalar Conservation Laws

M.Sc. Thesis

Author

Mohammad Izadi

Supervisor

Prof. Mohammad Asadzadeh

August 2005

To

My Parents

Acknowledgments

I would like to take this opportunity to extend my deepest gratitude to my supervisor, professor Mohammad Asadzadeh, of the Faculty of Mathematics, Chalmers University of Technology, Sweden. He has devoted me a lot of time to guide me in my research and has continuously stimulated me to investigate further aspects of the problem.

I am very grateful also to Dr. Mohammad Reza Razvan for his assistance and companionship throughout his staying at Institute for Advanced Studies in Basic Sciences (IASBS), as my advisor.

I would like to thank, moreover, Professor Bahman Mehri, who has given me the possibility to learn really a lot of things during my M.Sc. I am particulary grateful to him for the enthusiasm in scientific research he could transmit to me.

My sincere thanks also goes to my former undergraduate advisor, Proffessor Hossein Javadpour, of the faculty of Shahid Bahonar University of Kerman who has introduced me the world of PDE's and has given me the opportunity to work on a very interesting subjects.

Special thanks, finally, to my friends and classmate of the Department of Mathematics at IASBS, in particular, D. M. Moradi, A. Abedinzadeh and M. Soufi with whom I shared many experiences during these three years.

Last but not least, I am greatly indebted to my family for their understanding, patience and support during the entire period of my study.

Abstract

In the present thesis we study the Streamline Diffusion Finite Element method for one-space dimensional time dependent coupling equation of two hyperbolic conservation laws. We derive optimal convergence rates; in particular we prove an a priori error estimate of order $\mathcal{O}(h^{k+1/2})$ in domains where the exact solution is smooth; here h is the mesh width and k is the degree of the piecewise polynomial functions spanning the finite element subspace.

We also consider the a posteriori error estimate for our model problem and perform numerical implementation supporting the theory.

Contents

List of Figures

PREFACE

The main objective in this study is to analyze the *streamline diffusion* finite element approximation for the solution of the interface problem in a system of two, different, partial differential equations describing, e.g., multifluids with different state equations and moving contact discontinuities. More specifically we consider the following coupling of two hyperbolic conservation laws in one dimension:

$$\begin{cases} u_t + f_R(u)_x = 0, & x > 0, \quad t > 0, \\ u_t + f_L(u)_x = 0, & x < 0, \quad t > 0, \\ u(x,0) = u_0, & x \in \mathbb{R}. \end{cases} \tag{1}$$

This type of phenomenon appears for example in an increasing number of problems of fluid mechanics, among others, we emphasize the case of coupled problem involving Euler equation on one side of the interface and Navier-Stokes equation on the other side [17], as well as modelling certain plasma physical problems cf [37]. In this thesis we take a step towards the approximate solution for the problem stated in (1): We construct, analyze and implement *Streamline Diffusion Finite Element* method and prove *a priori* and *a posteriori* error estimates for the system of conservation laws introduced in (1). In our knowledge, the only related previous works for the numerical method for (1) are the studies by Godlewski and Raviart considering the finite difference method for the scalar case in [37] and also for the system in [38].

The Streamline Diffusion method (for short Sd-method below), also referring to as Galerkin/Least Square or SUPG (Streamline Upwind Petrov-Galerkin method) is a general finite element method for hyperbolic type partial differential equations. The Sd-method first was introduced, in the case of stationary problems, by Hughes and Brooks [45, 46], in the beginning of eighties. The mathematical analysis of this method for linear problems, together with extensions to time-dependent problems using space-time elements, was started in Johnson and Nävert [52] and was continued in Johnson et al. [48], and [53] and Nävert [63] in the early eighties and has now successfully been applied to stationary and time-dependent, convection-diffusion problems [47], as well as hyperbolic conservation laws [71, 72, 66, 35], reactive compressible flow [68], second order wave equations [51], incompressible and compressible Euler and Navier-Stokes equations [53, 55, 50, 43]. The convergence analysis of the method is extended to equations of gas dynamics and charged particle transport in [5] for the Vlasov-Poisson equation, for the Fermi and Fokker-Planck pencil beam equations in [6, 7] and for the Vlasov-Poisson-Fokker-Planck system in [8, 9]. Yet a modified version: *the characteristic streamline diffusion method*, invokes the advantages of *the method of characteristics*, for certain problems, in the Sd-method, see, e.g., [41, 42] and [12]. We recall that the conventional finite element methods for hyperbolic problems lack in either stability, like the standard Galerkin method, giving spurious oscillations if the exact solution is nonsmooth, or in accuracy, like the classical artificial diffusion method with considerable smearing of sharp fronts and at most first order accuracy. The outcome of this work is that the Sd-method can be demonstrated to have good stability properties and high accuracy: For the exact solution in the Sobolev space H^{k+1}, the error is of order $\mathcal{O}(h^{k+1/2})$. Below we give a concise description. The Sd-method is a modified Galerkin method based on piecewise polynomial approximation with the following three basic modifications:

- Consistent use of space-time finite element discretization for time dependent problems, where the basis functions can be discontinuous in time but are continuous in space.

- The test functions are modified by adding a small ($\sim$ mesh size) multiple of a linearized form of the hyperbolic operator involved, which results in a particular Petrov-Galerkin method with a weighted least squares control of the residual $\mathcal{R}(u_h)$ of the finite element solution u_h. *The residual is the error (the difference between the left and the right hand side) when in the original continuous equation the exact solution is replaced by the finite element solution.*

- Modification of a given viscosity ε to an artificial viscosity $\hat{\varepsilon}$ of the form

$$\hat{\varepsilon} = \max\left(\varepsilon, Ch|\mathcal{R}(u_h)|/|\nabla u_h|, Ch^{3/2}\right)$$

  or

$$\hat{\varepsilon} = \max\left(\varepsilon, Ch^2|\mathcal{R}(u_h)|/|u_h|, Ch^{3/2}\right)$$

  where h is the mesh size and C denotes a positive constant.

The last two features enhance the stability of the standard Galerkin method without sacrificing accuracy. Further, the modifications play a crucial role in the *adaptive* Sd-method based on a posteriori error estimates developed in ninties. The improved stability properties of the Sd-method in particular makes it possible to prove sharp a posterior error estimates which may be used to design *reliable* and *efficient* adaptive algorithm of the Sd-method. Here we have not considered adaptive algorithms for our model problem. Our focus has been on *a priori* and *a posteriori* error estimates for the Sd-method. Some studies considering adaptivity can be found in [26]-[31] and [11]. The a posteriori error estimates for the Sd-method for hyperbolic problems

typically have the form:

$$\|u - u_h\|_{L_2} \leq C\|h^2\hat{\varepsilon}^{-1}\mathcal{R}(u_h)\|_{L_2}, \tag{2}$$

where u is the the exact solution, u_h is the computed solution, $\mathcal{R}(u_h)$ is the residual, h is the mesh size and $\hat{\varepsilon}$ is defined as above. This estimate should be compared to the typical corresponding estimate for the standard Galerkin method for hyperbolic problems derived in [30]:

$$\|u - u_h\|_{L_2} \leq C\|\,\mathcal{R}(u_h)\|_{L_2}, \tag{3}$$

and the standard Galerkin method for elliptic problems derived in [27]:

$$\|u - u_h\|_{L_2} \leq C\|h^2\mathcal{R}(u_h)\|_{L_2}. \tag{4}$$

The proof of the a posteriori error estimate (2) has the following structure:

- Representation of the error in the terms of the residual of the finite element solution and the solution of a continuous dual problem.
- Use of the Galerkin orthogonality built in the finite element method.
- Interpolation estimates for the dual solution.
- Strong stability estimates for the continuous dual problem.

A typical a priori error estimate for the Sd-approximation of the hyperbolic type problems with piecewise polynomials of degree k takes the form

$$\|u - u_h\|_{L_2} \leq Ch^{k+1/2}\|u\|_{H^{k+1}}, \tag{5}$$

where u is the exact solution and H^{k+1} is the Sobolev space consisting of all functions having all their partial derivatives of order $\leq k+1$ in L_2, see Folland [33] for the details. The proof of the a priori error estimate typically has the following structure:

- Representation of the error in the terms of the exact solution and a discretized dual problem.
- Use of the Galerkin orthogonality to introduce the truncation error in the error representation.
- Interpolation estimates for the truncation error.
- Strong stability for the discrete dual problem.

We note the similarity in the structure of the proofs of the a priori and a posteriori error estimates, and also the differences: In the a priori case the key roles are played by the truncation error and the strong stability of the discrete problem, and in the a posteriori case these roles are taken by the residual and the stability of the continuous problem. Both the a priori and a posteriori error estimates are fundamental: The a priori error estimate shows that the discretization error (and the residual) will tend to zero with decreasing mesh size, and the a posteriori error estimate is the basis for adaptive quantitative error control. The above approaches for the a priori and a posteriori error estimates are general in nature and was, systematically, used in [26]- [30] for elliptic, parabolic and hyperbolic problems, as well as in [11] for the integral equations. Our approach to a posteriori error analysis is slightly different from (2) and has the following basic form

$$\|u - u_h\|_{L_2} \leq SC_i \| \frac{h^2}{\hat{\varepsilon}} \mathcal{R}(u_h) \|_{L_2}, \tag{6}$$

where h is the mesh size, C_i is an interpolation constant depending only on the angles of the elements of the computational mesh, and, most importantly, S is a stability factor which is related to the regularity properties of an associated linearized dual problem that measures how the error is propagated in time. All quantities, except the stability factor, are given explicitly through the computation. There are two

approaches to estimate the stability factor, analytically or computed a posteriori by solving the dual problem numerically (see e.g., [59, 60]. In the first approach (see [55], [18], [68]), the estimates are usually very pessimistic with exponential growth in time, however, the estimates show that the stability factors are finite, that is the dual problem has the required regularity properties, and also that they do not exhibit blow-up, and in the second approach you can capture the true behavior of the stability factors, at the price of a higher computational cost (see, [18], [68]). In this thesis we extend the work in the above references to the case of coupled problem, focusing on formulating the Sd-method, deriving a priori error estimates of the form (5) and a posteriori error bounds of the form (6) with relevant analytical stability factors. The outline of this thesis is as follows: Chapter 1, contains a brief summary of the finite element method, and set up some notations that will be used in the subsequent chapters. In chapter 2, we introduce the mathematical model, and also summarize without proofs the relevant results on existence and uniqueness of the coupled problem. In chapter 3, we introduce space-time discretization and formulate the streamline diffusion method for the coupled problems. In chapter 4, we derive basic stability estimate for the Sd-method, prove an a priori error estimate and state the a posteriori error estimate for the coupled problems. In chapter 5, we introduce a perturbation (adding a regularizing term of the form: $-\varepsilon u_{xx}$) to the equation which is used to derive a linearized dual problem occurring in the error representation formula, state a posteriori error estimate, and study the strong stability analytically. Finally, Chapter 6 is devoted to numerical experiments testing the reliability of out theoretical results in some concrete examples.

Chapter 1

THE FINITE ELEMENT METHOD

The mathematical models in science and technology mainly take the form of differential or integral equations. With rapid development of high speed computers over the last decades the possibilities of efficiently utilizing these models have dramatically increased. Using computer-implemented mathematical models, one can simulate and analyze complicated systems in science and engineering. This reduces the need for expensive and time-consuming experimental testing and makes it possible to compare many different alternatives for optimization.

To use numerical models on a computer one needs numerical methods. Only in the very simplest cases is it possible to find exact analytical solutions of the equations in the model, and in general one has to rely on numerical techniques for finding approximate solutions. The Finite Element Method (FEM) is a general technique for numerical solution of ordinary and partial differential equations as well as integral equations arising in science and engineering, both in deterministic and stochastic versions. The equation system under consideration (1) is deterministic and the Streamline Diffusion method used in this thesis is a modified form of the

finite element method with certain improving features for the hyperbolic PDE's. Below we give an introduction to the finite element method as a general technique for the numerical solution of partial differential and integral equations.

1.1 Introduction to the Finite Element Method

The basic idea in any numerical method for a differential equation is to discretize the given continuous problem with infinitely many degrees of freedom to obtain a discrete problem or system of equations with only finitely many unknowns that may be solved by using a computer. A classical numerical method for partial differential equations is the finite difference method where the discrete problem is obtained by replacing derivatives with difference quotients involving the values of the unknown at finitely many points. The discretization process using a finite element method is different. The Finite Element Method (FEM) is a very powerful and flexible numerical approach for solving partial differential equations. Its flexibility means that it can be used to solve complicated equations in domains whose geometries range from a simple polygon or polyhedron such as a square or a cube to more complex shapes with curved boundaries. It is also easy to construct higher-order approximations. However, the programming of finite element methods is more complicated than that of finite differences, and hence in general requires standard software packages. The topics is introduced, in different settings, by several authors; see, e.g., [20], [16], [40], [24], [47, 48], [58], [61] and [76, 77, 78].

1.2 A Short History

The finite element method was first proposed in 1943 by the German Mathematician Richard Courant [23] who solved the Poisson equation based on minimization of a

functional over piecewise linear approximations on subregions. However, it was not until the 1950's that work really began on finite element methods when it was rediscovered by engineers working in the aircraft industry; some of the early papers were by Argris and Kelsey [3, 4] and Turner, Clough, Martin and Topp [73] although the name finite element method was not introduced until 1960 when it was proposed by Clough [21, 22]. A short while later the first book on the subject was published by Zienkiewicz [75]. The mathematical analysis of these methods began in the 1960's. In 1962 Friedrichs [34] used piecewise linear function on triangles to derive a system of equations for solving problems on a general domain. He was also able to prove convergence in H^1 and L_2 (see Section 1.3 for definitions) although not the rate of convergence. In 1963 Oganesjan [64] proved the first *a priori* estimate for the error in H^1 norm for Laplace's and more second order elliptic equations, e.g., for plates. The a priori error estimate for quadratic elements on triangles was produced by Zlámal in 1968 [74]. There was then much more research into the mathematical theory of the finite element method and many of the *a priori* error estimation technique which are widely used today were derived. By the 1970's this theory was well developed, at least for linear elliptic problems. In parallel to this engineers had produced sophisticated codes and the method was becoming widely used. In the late 1970's, work on *a posteriori* error analysis began. Previously all error analysis had been of the a priori type; these are bounds on the error in the finite element solution dependent upon the data of the problem, the mesh size, the degree of the approximation polynomial and the (unknown) exact solution, thus they are not computable, and are useful only as an indication of the rate of convergence of the finite element approximation to the exact solution. In contrast an *a posteriori* error bound provides a computable upper bound on the error in some norm using the computed finite element solution. Such error bounds were first introduced by Babuška and his co-workers in 1978 [13, 14]; they were used in adaptive finite element

computations, the error bound being used to guide mesh refinement.

Although the finite element method was first proposed in 1943, more recent advances in the theory, i.e., *a posteriori* have made the method increasingly powerful; such that there are various approaches to *a posteriori* error estimation and it has now successfully applied to many problems by several authors, see e.g., [2], [36], [49, 31].

1.3 Preliminary Notation and Function Space

One of the basic tools in the study of finite element methods is the theory of functional analysis. Below we recall and introduce some definitions and notations which will be used frequently in the sequel. For a complete presentation and proofs we refer the reader to, e.g, Adams [1].

1.3.1 Lebesgue spaces

Let Q be an open set contained in $\mathbb{R}^n, n \geq 1$; we denote boundary of Q by ∂Q. For $1 \leq p < \infty$, let $L_p(Q)$ denote the set of real-valued Lebesgue measurable functions u defined on Q such that $|u|^p$ is integrable on Q with respect to the Lebesgue measure $dx = dx_1 \dots dx_n$. We define the L_p norm, $\|.\|_{L_p(Q)}$, by

$$\|u\|_{L_p(Q)} = \Big(\int_Q |u(x)|^p dx \Big)^{1/p}.$$

For $p = 2$, the space $L_2(Q)$, the space of square integrable functions will be seen to be of particular importance. For real-valued functions $u, v \in L_2(Q)$, we define the L_2-inner product by

$$(u, v) = \int_Q u(x)v(x)dx.$$

Then for each $u, v \in L_2(Q)$ the Cauchy-Schwartz inequality states that $uv \in L_1(Q)$ and

$$|(u, v)| \leq \|u\|_{L_2(Q)} \|v\|_{L_2(Q)}.$$

For $p = \infty$, $L_\infty(Q)$ denotes the set of all real-valued Lebesgue measurable functions u defined on Q such that $|u|$ has finite essential supremum; the essential supremum of $|u|$ is defined as

$$ess. \sup_{x \in Q} |u(x)| = \inf\{M \geq 0 : |u(x)| \leq \text{M almost everywhere in Q}\} < \infty,$$

and

$$\|u\|_{L_\infty(Q)} = ess. \sup_{x \in Q} |u(x)|,$$

where "almost everywhere in Q" means "except on a subset of Q having zero measure".

1.3.2 Weak (Generalized) L_2-derivatives

We want to generalize the concept of classical derivative to define what we refer to as *weak* or *generalized derivative* and do it in such a way that if everything is "smooth enough" the classical and weak derivatives coincide. The concept of a weak derivative is an extension of the classical concept in which we want to maintain the validity of integration by parts formulas. Our generalization will allow functions such as $u(x) = |x|$ on $[-1, 1]$ to have a derivative in the weak sense (for more details we refer to [16]). As usual, we let Q be an open subset of $\mathbb{R}^n$ and let $x = (x_1, \ldots, x_n) \in Q$ denote an arbitrary point in Q. The set of all real-valued functions $u(x) = u(x_1, \ldots, x_n)$ which are defined and continuous in Q is denoted by $C(Q)$ and the set having all classical derivatives of order $\leq k$ continuous in Q is denoted by $C^k(Q)$. To simplify the notation for differentiation we introduce the *multi-index* α which is defined by $\alpha = (\alpha_1, \ldots, \alpha_n)$, where the α_i, $i = 1, \ldots, n$, are

nonnegative integers. The length of the multi-index α is defined to be

$$|\alpha| = \sum_{i=1}^{n} \alpha_i.$$

In this way we can rewrite the $|\alpha|-$th order differential operator D^α by

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x^{|\alpha|}} = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}}.$$

Using this notation we can define $C^k(Q)$ as

$$C^k(Q) = \{u : D^\alpha u \in C(Q), |\alpha| \leq k\}.$$

Now we define the set of locally integrable functions L_1^{loc} by

$$L_1^{loc}(\Omega) = \{u : u \in L_1(\Lambda) \quad \text{for all measurable compact subset} \quad \Lambda \subset\subset \Omega\}.$$

Now we are in a position to define the concept of the weak (or generalized or distributional) L_2-derivatives of a function. Let $u \in Ł_1^{loc}(Q)$; we say that u has a *weak derivative* of order α in the weak L_2-sense if there exist a function $v \in Ł_1^{loc}(Q)$ such that

$$\int_Q v(x)\phi(x)dx = (-1)^{|\alpha|} \int_Q u(x) D^\alpha \phi(x) dx,$$

holds for all $\phi \in C_0^\infty(Q)$ which is the space of functions having continuous derivatives of all order and compact support in Q. If ϕ has compact support in Q then $\phi = 0$ outside a closed and bounded subset of Q; generally the support of a function $\phi(x)$ (Suppϕ) is the closure of the set $\{x \in Q : \phi(x) \neq 0\}$. We call $v = D^\alpha u$ the weak L_2-derivative of u of order α.

1.3.3 Sobolev spaces

We finally introduce another class of Hilbert spaces that will be used to formulate our weak problem. A comprehensive presentation of these spaces can be found in

Adams [1]. Equipped with the definition of weak derivative we may define, for a positive integer s, the *Sobolev space* $H^s(Q)$ as the set of functions $u \in L_2(Q)$ which possess generalized L_2-derivatives $D^\alpha u \in L_2(Q)$ for $0 \leq |\alpha| \leq s$; i.e.,

$$H^s(Q) = \{u \in L_2(Q) : D^\alpha u \in L_2(Q) \quad \text{for} \quad 0 \leq |\alpha| \leq s\}.$$

Clearly, $H^s(Q)$ is a subspace of $L_2(Q)$ and $H^0(Q) = L_2(Q)$. On $H^s(Q)$ we define the inner product

$$\begin{aligned}(u,v)_s &= \sum_{|\alpha|\leq s} \int_Q D^\alpha u D^\alpha v dQ \\ &= \sum_{|\alpha|\leq s} (D^\alpha u, D^\alpha v), \qquad \forall u, v \in H^s(Q),\end{aligned}$$

where $(.,.)$ denotes the standard inner product in $L_2(Q)$. Using the definition of inner product, we define the norm on $H^s(Q)$ as

$$\|u\|_{s,Q} = (u,u)_s^{1/2} = \Big(\sum_{|\alpha|\leq s} \|D^\alpha u\|_Q^2 \Big)^{1/2}, \qquad \forall u \in H^s(Q),$$

where $\|.\|_Q$ denotes the standard norm on $L_2(Q)$. Clearly, $\|.\|_{0,Q} = \|.\|_Q$ so in the sequel we will denote the L_2-norm by $\|.\|_Q$. We will make extensive use of the space $H^1(Q)$; if $Q \subset \mathbb{R}$ then the norm on $H^1(Q)$ is explicitly given by

$$\|u\|_{1,Q} = \Big(\|u\|_Q^2 + \|u'\|_Q^2 \Big)^{1/2}.$$

We will also make use of the constrained space $H_0^1(Q)$ which will be introduced by two following Theorems:

Theorem 1.1 (Trace Theorem). *Assume that Q is bounded and ∂Q is Lipschitz continuous* [1]. *Then there exist a bounded operator*

$$\gamma : H^1(Q) \longrightarrow L_2(\partial Q),$$

[1] A function $f : \mathbb{R}^n \supset D \longrightarrow \mathbb{R}^m$ is called Lipschitz continuous provided that for some number c, $\|f(x) - f(y)\| \leq c\|x - y\|$ for all $x, y \in D$. A hypersurface in $\mathbb{R}^n$ is a graph whenever it can

such that

$$\gamma u = u\big|_{\partial Q} \quad \textit{if} \quad u \in H^1(Q) \cap C(\bar{Q}),$$

and

$$\|\gamma u\|_{L_2(\partial Q)} \leq C\|u\|_{1,Q},$$

for each $u \in H^1(Q)$, with the constant C depending only on Q.

Proof. See [32]. □

We call γu the trace of u on the boundary, i.e., "loosely speaking" the restriction of u to the boundary. The above Theorem asserts that the restriction of $u \in H^1$ to the boundary is at least an L_2 function. Next we examine more closely what it means for a function to have a zero trace.

Theorem 1.2 (Trace-zero functions in H^1). *Assume Q is bounded and ∂Q is Lipschitz continuous. Suppose furthermore that $u \in H^1(Q)$. Then*

$$u \in H_0^1(Q) \quad \textit{if and only if} \quad \gamma u = 0 \quad \textit{on} \quad \partial Q.$$

Proof. See [32]. □

Thus we can rewrite

$$H_0^1(Q) = \{u \in H^1(Q) : \gamma u = 0 \quad \textit{on} \quad \partial Q\}.$$

1.4 Weak Formulation

Before applying the FEM to solve given equation, it is necessary to transform the equation into a more suitable form. Below we present two common approaches:

represented in the form $x_k = f(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)$, with $1 \leq k \leq n$ and some suitable domain in $\mathbb{R}^{n-1}$. A domain $\Omega \subset \mathbb{R}^n$ is called a Lipschitz domain provided that for every $x \in \partial\Omega$, there exist a neighborhood $\mathcal{N}_x \subset \partial\Omega$ of x which can be represented as the graph of a Lipschitz continuous function.

- One can derive an equivalent minimization problem, which has exactly the same solution as the differential equation.
- One can derive a so called weak formulation.

Both methods leads finally to exactly the same results, however, we shall restrict ourselves to second method and to give an idea we describe this approach for a typical linear elliptic equation with Dirichlet boundary condition:

$$\begin{cases} \mathcal{L}u &= f, & \text{in} \quad \Omega, \\ u &= u_\Gamma, & \text{on} \quad \partial\Omega = \Gamma. \end{cases} \tag{1.1}$$

1.4.1 Trial solution and weighting functions

To define the weak, or variational, form of the boundary value problems, we need to define two classes, or collections, of functions: the *test* or weighting functions and the *trial* or admissible solutions. Here these spaces are defined in the context of the standard Galerkin formulation. For the Dirichlet problem (1.1), the first collection of functions, denoted by $\mathcal{V}$, is composed of test functions and consist of all functions which are square integrable, have square integrable first derivatives over the computational domain Ω, and vanish on the boundary Γ. It is defined as follows:

$$\mathcal{V} = \{w \in H^1(\Omega) : w = 0 \quad \text{on} \quad \Gamma\} \equiv H^1_\Gamma(\Omega).$$

This is as previously noted a Sobolev space and its inner product and norm coincide with those of $H^1(\Omega)$.

The second collection of functions is called the *trial solutions*. This collection is similar to the test function except that these admissible functions are required to satisfy the Dirichlet conditions on Γ. This second collection is denoted by $\mathcal{S}$ and is defined by

$$\mathcal{S} = \{u \in H^1(\Omega) : u = u_\Gamma \quad \text{on} \quad \Gamma\} \equiv \mathcal{V} \oplus \{\bar{u}_\Gamma\},$$

where $\bar{u}_\Gamma$ is any function in $H^1(\Omega)$ such that $\bar{u}_\Gamma = u_\Gamma$ on Γ. Thus, $\mathcal{S}$ can be viewed as a translation of $\mathcal{V}$ and, consequently, it is an affine space. Note, for instance, that, for $u_\Gamma \neq 0$, the sum of two element of $\mathcal{S}$ is not an element of $\mathcal{S}$. However, for homogeneous boundary conditions, $u_\Gamma = 0$, trial and test spaces coincide, $\mathcal{S} = \mathcal{V} = H_0^1(\Omega)$.

The sets $\mathcal{S}$ and $\mathcal{V}$ clearly contain infinitely many functions. In the finite element method, $\mathcal{S}$ and $\mathcal{V}$ are approximated by convenient finite dimensional subsets of these collections which will be denoted by $\mathcal{S}_h$ and $\mathcal{V}_h$, respectively. These finite element spaces are characterized, among other things, by a partition of domain.

1.4.2 One-dimensional boundary value problem

The first step in the weighted residual formulation (Galerkin method) leading to the finite element discretization of given boundary value problem consists of formulating a weak (or variational) form of the boundary value problem. This is achieved by multiplying the governing equation by a weighting function and integrating over the computational domain Ω. To be more precise we consider a typical weak form, e.g., for Poisson's equation in one dimension with zero Dirichlet boundary conditions for simplicity, by taking $\mathcal{L} = -\frac{d}{dx}\Big(a(x).\frac{d}{dx}\Big)$ in (1.1):

$$\begin{aligned} -\frac{d}{dx}\Big(a(x).\frac{d}{dx}u(x)\Big) = f(x) \quad x \in \Omega, \\ u = 0 \quad x \in \Gamma, \end{aligned} \tag{1.2}$$

where $\Omega = (0,1)$, $f \in L_2(\Omega)$ and $a(x)$ is piecewise continuous function on Ω. By letting $a(x) = 1$ and integrating the equation $-u'' = f$ twice, it is easy to see that this problem has a unique classical solution u. A classical solution of this problem would satisfy the equation and boundary condition and would lie in the space $C^2(\Omega)$. Such a smooth solution may not exist if the data for the problem are not sufficiently

smooth. Hence, we relax the requirement $u \in C^2(\Omega)$ using weak derivatives, define

$$\mathcal{S} = \mathcal{V} = H_0^1(\Omega) = \{u \in H^1(\Omega) : u(0) = u(1) = 0\},$$

then multiplying any function $w \in \mathcal{V}$ on both sides of (1.2) we have

$$-\int_\Omega u'' w dx = \int_\Omega f w dx,$$

using integrating by parts and boundary conditions we get

$$\int_0^1 u' w' dx = \int_0^1 f w dx.$$

Now a sufficient requirement for both sides to be meaningful is that $u \in H^1(\Omega)$, and thus the weak (or variational) formulation of the problem is to find $u \in H^1(\Omega)$ such that

$$\int_0^1 u' w' dx = \int_0^1 f w dx, \qquad \forall w \in \mathcal{V} \equiv \mathcal{S}. \tag{1.3}$$

The solution u is then known as a weak solution to the problem (1.2). Let

$$\begin{aligned} a(u,v) &= \int_0^1 u' v' dx, \\ L(v) &= \int_0^1 f v dx, \end{aligned} \tag{1.4}$$

we then reformulate the variational formulation (1.3) in an abstract form as follows: Find $u \in H^1(\Omega)$ such that

$$a(u,v) = L(v), \quad \forall v \in \mathcal{S}. \tag{1.5}$$

Clearly if u is a classical solution to (1.2), then it is also a weak solution of (1.2). However, the converse may not hold in general (for example if f has discontinuity) although, if the solution is sufficiently smooth in a subdomain of Ω, reversing the above argument shows that the solution satisfies the differential equation in that subdomain in a classical sense. Thus the advantage of seeking a weak solution

may be summerized as follows: through relaxing the regularity requirement (for the solution) we may hope to find weak solutions for problems where classical solutions may not exist and in the case where classical solutions do exist the two should coincide. In addition it is helpful to address the issue of well-posedness of the problem through a weak formulation.

1.5 Abstract Variational Formulation

We formulate the general theorems for existence and uniqueness in Hilbert space framework and state the conditions that spaces and bilinear form should satisfy. These results are applied to investigate solvability of particular partial differential equations.

1.5.1 Existence and uniqueness theorem

The existence and uniqueness of a solution to the weak formulation of the problem can be proved using the Lax-Milgram Theorem. This states that the weak formulation admits a unique solution.

Theorem 1.3. *(**Lax-Milgram lemma**). Let $a(.,.)$ be a bilinear form on a Hilbert space $\mathcal{H}$ equipped with norm $\|.\|_{\mathcal{H}}$ and the following properties:*

I. $a(.,.)$ is continuous, that is

$$\exists \gamma_1 > 0 \quad \textit{such that} \quad |a(w,v)| \leq \gamma_1 \|w\|_{\mathcal{H}} \|v\|_{\mathcal{H}} \quad \forall w, v \in \mathcal{H}, \tag{1.6}$$

II. $a(.,.)$ coercive (or $\mathcal{H}$-elliptic), that is

$$\exists \alpha > 0 \quad \textit{such that} \quad a(v,v) \geq \alpha \|v\|^2_{\mathcal{H}}, \quad \forall v \in \mathcal{H}. \tag{1.7}$$

Further

III. L is a linear mapping on $\mathcal{H}$ (thus L is continuous), that is

$$\exists \gamma_2 > 0 \quad \textit{such that} \quad |L(w)| \leq \gamma_2 \|w\|_{\mathcal{H}}, \quad \forall w \in \mathcal{H}.$$

Then there exists a unique $u \in \mathcal{H}$ such that

$$a(w,u) = L(w), \quad \forall w \in \mathcal{H}.$$

A proof can be found for instance in Ciarlet [20] or Brenner and Scott [16].

Definition 1.4. The *energy* norm on $\mathcal{H}$ is defined by $\|v\|_a = \sqrt{a(v,v)}, v \in \mathcal{H}$.

Recalling the relations (1.6) and (1.7) above, the energy norm satisfies

$$\alpha \|v\|_{\mathcal{H}}^2 \leq a(v,v) = \|v\|_a^2 \leq \gamma_1 \|v\|_{\mathcal{H}}^2.$$

Therefore the energy norm $\|v\|_a$ is equivalent to the abstract norm $\|v\|_{\mathcal{H}}$.

1.5.2 Applications to elliptic boundary value problems

The purpose now is to verify that properties of the Lax-Milgram lemma are fulfilled in an example of a boundary value problem. To this end we need two inequalities, the Cauchy-Schwarz inequality,

$$\begin{aligned} |(v,w)_{L_2(\Omega)}| &\leq \|v\|_{L_2(\Omega)} \|v\|_{L_2(\Omega)}, \qquad v, w \in L_2(\Omega) \\ |(v,w)_{H^1(\Omega)}| &\leq \|v\|_{1,\Omega} \|w\|_{1,\Omega}, \qquad v, w \in H^1(\Omega), \end{aligned} \tag{1.8}$$

and the Poincaré's inequality,

$$\|v\|_{L_2(\Omega)}^2 \leq C_\Omega \|v'\|_{L_2(\Omega)}^2, \quad 0 < C_\Omega < \infty \quad \text{and} \quad v \in H_0^1(\Omega). \tag{1.9}$$

Example 1.5. *We consider again $-u''(x) = f(x)$ in $\Omega = (0,1)$, with $u(0) = u(1) = 0$ and the relevant Hilbert space is $\mathcal{H} = H_0^1(\Omega)$. The weak (varaitional) form is now*

$$\int_0^1 u'(x)v'(x)dx = a(u,v) = L(v) = \int_0^1 f(x)v(x)dx, \qquad \forall v \in H_0^1(\Omega).$$

It is clear that, due to the linearity of the integral operator, $a(.,.)$ and $L(.)$ are bilinear and linear forms, respectively. It is sufficient to show that $a(.,.)$ is continuous, coercive and $L(.)$ is bounded.

- *To show that $L(v)$ is bounded, we use the Cauchy-Schwarz inequality and the fact that $\|v\|_{L_2(\Omega)} \leq \|v\|_{1,\Omega}$:*

$$|L(v)| = \Big| \int_0^1 fv dx \Big| = |(f,v)_{L_2(\Omega)}| \leq \|f\|_{L_2(\Omega)} \|v\|_{L_2(\Omega)} \leq \|f\|_{L_2(\Omega)} \|v\|_{1,\Omega}.$$

  Hence, we can choose $\gamma_2 = \|f\|_{L_2(\Omega)}$. We now see why it is natural to demand $f \in L_2(\Omega)$.

- *Continuity of $a(.,.)$ follows from the Cauchy-Schwarz inequality (this time in $L_2(\Omega)$) and the fact that $\|v'\|_{L_2(\Omega)} \leq \|v\|_{1,\Omega}$,*

$$|a(v,w)| = |(v',w')_{L_2(\Omega)}| \leq \|v'\|_{L_2(\Omega)} \|w'\|_{L_2(\Omega)} \leq \|v\|_{1,\Omega} \|w\|_{1,\Omega}.$$

  Hence, $\gamma_1 = 1$.

- *The $\mathcal{H}-$ellipticity follows from Poincaré's inequality,*

$$\|v\|_{1,\Omega}^2 = \int_0^1 (v^2 + (v')^2) dx \leq (C_\Omega + 1) \int_0^1 (v')^2 dx = (c_\Omega + 1) a(v,v).$$

  This means that $\alpha = \frac{1}{C_\Omega + 1}$.

Therefore from the Lax-Milgram lemma it follows that the above variational problem has an unique solution $u \in H_0^1(\Omega)$ satisfying the bound $\|u\|_{1,\Omega} \leq (1+c_\Omega)\|f\|_{L_2(\Omega)}$.

1.6 Finite Element Spaces

In practice, the spaces over which we solve variational problems associated with boundary value problems are called *finite element spaces*. We partition the given

domain Ω into finitely many sub-domains, and consider functions which reduce to a polynomial on each sub-domain. The sub-domains are called *elements*.

There is no change conceptually in going from one dimension to two or three dimensions. The main practical difference is that instead of subintervals in one dimension, the element become triangles or quadrilaterals in two dimensions, and tetrahedra, cubes, rectangular parallelepiped, etc. in three dimensions. For simplicity, we restrict our discussion primarily to one and two dimensional cases.

1.6.1 Finite element triangulations

As we mentioned, the idea of finite element method is to first split the computational domain Ω into a set of elements K. The set of elements is known as the *mesh* and the vertices of the elements are known as the *nodes*. For an admissible triangulation (or subdivision) $\mathcal{T}_h = \{K\}$ of Ω we require that:

- $\bar{\Omega} = \bigcup_{K \in T_h} K$,
- The intersection of any two distinct elements in $\mathcal{T}_h$ either consists of a common face, common side or common vertex, or is empty.

For any $K \in \mathcal{T}_h$, let

$$h_K = diam(K), \quad h = \max_{K \in \mathcal{T}_h} h_K,$$

and

$$\rho_K = \sup\{diam(B) : B \text{ is a ball contained in } K\}.$$

We assume that $\mathcal{T}_h$ is quasi-uniform in sense of Ciarlet [20], i.e., the following two conditions are satisfied:

(i) There exist a number $\varrho > 0$, independent of h and K such that

$$\frac{\rho_K}{h_K} \geq \varrho, \qquad \forall K \in \mathcal{T}_h,$$

(ii) The parameter h approaches zero

This condition means that the triangles $K \in \mathcal{T}_h$ are not allowed to be arbitrary thin, or equivalently, the angels of the triangles K are not allowed to be arbitrary small. The constant ϱ is a measure of the smallest angle in any $K \in \mathcal{T}_h$.

We introduce the finite element spaces $\mathcal{S}_h$, $\mathcal{V}_h$; these defined to be finite dimensional subspaces of the trial functions $\mathcal{S}$, and test functions $\mathcal{V}$, (see Section 1.2.1) respectively, which consist of functions that are polynomials of degree $\leq k$ in each element K, i.e.,

$$\mathcal{V}_h := \{w \in H^1(\Omega) : w\big|_K \in \mathcal{P}_k(K), \quad \forall K \in \mathcal{T}_h \quad \text{and} \quad w = 0 \quad \text{on} \quad \Gamma\},$$

$$\mathcal{S}_h := \{u \in H^1(\Omega) : u\big|_K \in \mathcal{P}_k(K), \quad \forall K \in \mathcal{T}_h \quad \text{and} \quad u = u_\Gamma \quad \text{on} \quad \Gamma\}.$$

With these definitions, the finite element method for (1.5) can now be formulated as follows:

$$\begin{cases} \text{Find} \quad u_h \in \mathcal{S}_h (= \mathcal{V}_h), \quad \text{such that} \\ a(u_h, v) = L(v), \qquad\qquad \forall v \in \mathcal{V}_h, \end{cases}$$

1.6.2 Piecewise linear finite elements in one space dimension

In the piecewise linear version of finite element method we take $k = 1$. We will now construct a finite element space $\mathcal{V}_h$ consisting of piecewise linear functions. Let $0 = x_0 < x_1 ... < x_M < x_{M+1} = 1$, be a partition of the interval $(0, 1)$ into subintervals $\Omega_j = (x_{j-1}, x_j)$ of length $h_j = x_j - x_{j-1}$, where the set of points x_j is called nodes, $j = 1, ..., M+1$, and set $h = \max h_j$. The quantity h is then a measure of how fine the partition is. One possible choice of the nodes is to place them at the boundary of each element. One of the strengths of the finite element method is the flexibility in the choice of elements. In regions where the solution is rapidly

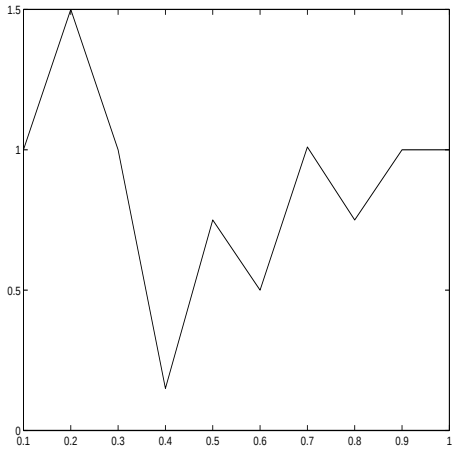

Figure 1.1: Example of a function $\varphi \in \mathcal{V}_h$.

varying, one can have small elements. The smoother parts of the solution can have an associated grid with larger elements and perhaps higher polynomial degrees for φ_j's. This flexibility is particularly important in 2 and 3 dimensions. We now let $\mathcal{V}_h$ be the set of all functions φ, such that φ is linear on each subinterval Ω_j and φ is continuous on $[0, 1]$. An example of a function in $\mathcal{V}_h$ is shown in the Figure 1.1.

Apparently $\mathcal{V}_h$ is a linear space of dimension M and each internal nodal point corresponds to a degree of freedom. $\mathcal{V}_h$ has a natural basis, known as nodal basis (shape) functions, which is given by, for $j = 1, ..., M$

$$\psi_j(x) = \begin{cases} \frac{x - x_{j-1}}{h_j}, & x \in [x_{j-1}, x_j], \\ \frac{x_{j+1} - x}{h_{j+1}}, & x \in [x_j, x_{j+1}], \\ 0 & \text{elsewhere.} \end{cases}$$

Note that

$$\psi_j(x_i) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \quad i, j = 1, ..., M, \end{cases}$$

i.e., ψ_j is the continuous piecewise linear function that takes the value 1 at node point x_j and the value 0 at other node points. See Figure 1.2. One often refers to piecewise linear basis functions as *linear elements*.

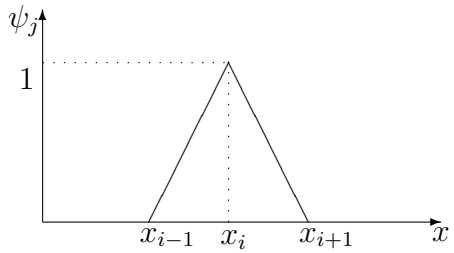

Figure 1.2: The basis function ψ_j in $1D$ for one element.

A function $\varphi \in \mathcal{V}_h$ then has the representation

$$\varphi(x) = \sum_{i=1}^{M} \varphi_i \psi_i(x), \quad \varphi_i = \varphi(x_i), \quad x \in [0,1],$$

i.e., each $\varphi \in \mathcal{V}_h$ can be written in a unique way as a linear combination of the basis functions ψ_i. In particular it follows that $\mathcal{V}_h$ is a linear space of dimension M with basis $\{\psi_i\}_{i=1}^{M}$.

The finite element method can now be formulated as follows: Find $u_h \in \mathcal{V}_h = \mathcal{S}_h$ such that

$$a(u_h, \varphi) = (f, \varphi), \qquad \forall \varphi \in \mathcal{V}_h, \tag{1.10}$$

where $(v, w) = \int_\Omega vw dx$. This problem is usually referred to as Galerkin method. We observe that if $u_h \in \mathcal{V}_h$ satisfies (1.10), then in particular

$$a(u_h, \psi_j) = (f, \psi_j), \qquad j = 1, ..., M, \tag{1.11}$$

and if these equations hold, then by taking linear combinations, we see that u_h satisfies (1.10). Since

$$u_h(x) = \sum_{i=1}^{M} \mu_i \psi_i(x), \quad \mu_i = u_h(x_i),$$

by linearity we can write (1.11) as

$$\sum_{i=1}^{M} \mu_i a(\psi_i, \psi_j) = (f, \psi_j), \qquad j = 1, \ldots, M, \tag{1.12}$$

which is a linear system of equations with M equations and M unknowns: $\mu_i := u_h(x_i)$, $i = 1, \ldots, M$. In matrix form the linear system (1.12) can be written as

$$\mathbf{A}\mu = \mathbf{f}, \tag{1.13}$$

where $\mathbf{A} = (a_{ij})$ is the $M \times M$ matrix with elements $a_{ij} = a(\psi_i, \psi_j) := \int_\Omega \psi_i' \psi_j' \, dx$, and where $\mu = (\mu_1, \ldots, \mu_M)$ and $\mathbf{f} = (f_1, \ldots, f_M)$ with $f_i = (f, \psi_i)$ are M-vectors:

$$\mathbf{A} = \begin{pmatrix} a_{11} & \ldots & a_{1M} \\ . & & . \\ . & & . \\ . & & . \\ a_{M1} & \ldots & a_{MM} \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_1 \\ . \\ . \\ . \\ \mu_M \end{pmatrix}, \qquad \mathbf{f} = \begin{pmatrix} f_1 \\ . \\ . \\ . \\ f_M \end{pmatrix}.$$

The matrix $\mathbf{A}$ is known as *stiffness* matrix with respect to the nodal basis $\{\psi_j\}_{j=1}^M$ and $\mathbf{f}$ the *load* vector.

The elements $a_{ij} = (\psi_i, \psi_j)$ in the matrix $\mathbf{A}$ can easily be computed: We first observe that $(\psi_i, \psi_j) = 0$ if $|i - j| > 1$ since in this case for all $x \in [0, 1]$ either $\psi_i(x)$ or $\psi_j(x)$ is equal to zero (for more details see [10]). Therefor we have the following Lemma:

Lemma 1.6. *The Stiffness matrix $\boldsymbol{A}$ has the following properties*

1. *$\boldsymbol{A}$ is symmetric,*

2. *$\boldsymbol{A}$ is positive definite.*

Proof. Symmetry is obvious, because $a(u, v) = a(v, u)$. To show that $\mathbf{A}$ is positive definite, we first recall that an $n \times n$ matrix $\mathbf{A}$ is positive definite if $v^T \mathbf{A} v > 0$ for all nonzero $v \in \mathbb{R}^n$. Write $v = (v_1, \ldots, v_n)$ and $v = \sum_{j=1}^n v_j \psi_j$. Then by bilinearity

of $a(\cdot,\cdot)$ it follows that

$$v^T\mathbf{A}v = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i a_{ij} v_j = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i a(\psi_i,\psi_j) v_j$$
$$= a\big(\sum_{i=1}^{n} v_i\psi_i, \sum_{j=1}^{n} v_j\psi_j\big) = a(v,v) \geq \alpha\|v\|_{\mathcal{H}} > 0,$$

for any nonzero v. In other words, the positive definiteness of the coefficient matrix comes form the $\mathcal{H}-$ ellipticity. □

To proceed we have the following results from matrix analysis:

Proposition 1.7. Let the square matrix $\mathbf{A}$ be positive definite then

i. $\mathbf{A}^{-1}\exists$ "$\mathbf{A}$ is invertible"

ii. $\mathbf{A}\mu = b$ has a unique solution.

Proof. (i) Suppose $\mathbf{Ax=0}$ then $\mathbf{x}^T\mathbf{Ax=0}$, but $\mathbf{A}$ is positive definite, then $\mathbf{x} \equiv 0$ hence $\mathbf{A}$ has full Range and therefore $\mathbf{A}$ is invertible

(ii) Since $\mathbf{A}$ is invertible $\mathbf{A}\mu = b$ has a unique solution $\mu = \mathbf{A}^{-1}b$.

□

We can now easily establish a discrete counterpart to the Lax-Milgram lemma:

Theorem 1.8. *Existence and uniqueness of the solution to the discrete problem. There exist a unique solution vector* $(\mu_1,\ldots,\mu_M)^T$ *of the linear system* (1.13).

Proof. A is symmetric and positive definite matrix. Hence A is nonsingular. Thus the solution of the corresponding linear system of equations exists and is unique. □

1.7 Error Representation and Error Estimation

Consider the problem just described and let $\mathcal{L}$ be the differential operator

$$\mathcal{L} = -\frac{d^2}{dx^2},$$

$a(.,.)$ the bilinear form

$$a(u,v) = \int_0^1 u'v'dx, \quad u,v \in H^1(\Omega),$$

and $L(.)$, the linear functional

$$L(v) = \int_\Omega fvdx, \quad \forall v \in H^1(\Omega).$$

Then the weak and corresponding finite element formulation of this problem are, respectively: Find $u \in H^1(\Omega)$ such that

$$a(u,v) = L(v), \quad \forall v \in H_0^1(\Omega), \tag{1.14}$$

and find $u_h \in \mathcal{S}_h$ such that

$$a(u_h,v_h) = L(v_h), \quad \forall v_h \in \mathcal{V}_h. \tag{1.15}$$

Formulation (1.14) and (1.15) can be used to find an error representation formula, i.e., a way of writing down what the discretization error $u - u_h$ is. Since all $v \in \mathcal{V}_h$ are also in $H_0^1(\Omega)$, then (1.14) is valid also for $v \in \mathcal{V}_h$. Therefore we may take $v = v_h$ in (1.14); and then subtract (1.15) from the resulting equation to get

$$a(u-u_h,v_h) = 0, \quad \forall v_h \in \mathcal{V}_h. \tag{1.16}$$

or equivalently

$$\sum_K \int_K (f - \mathcal{L}u_h)v_h dx = 0, \quad \forall v_h \in \mathcal{V}_h.$$

This error representation is known as *Galerkin orthogonality* and tells us that the residual $\mathcal{R}(u_h) = f - \mathcal{L}u_h$, is orthogonal (with respect to L_2 inner product) to the space $\mathcal{V}_h$. The concept of Galerkin orthogonality will be seen to be a key result when deriving the error bounds. We proved that (section 1.3.2) the bilinear form $a(.,.)$ is coercive and continuous so that we have, for positive constants α, and γ_1

$$a(u - u_h, u - u_h) \geq \alpha \|u - u_h\|^2_{1,\Omega},$$

and

$$a(u - u_h, u - v_h) \leq \gamma_1 \|u - u_h\|^2_{1,\Omega} \|u - v_h\|^2_{1,\Omega} \quad \forall v_h \in \mathcal{V}_h.$$

Using the linearity of $a(.,.)$ with respect to the second argument we have

$$a(u - u_h, u - u_h) = a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h),$$

for any $v_h \in \mathcal{V}_h$. Note that since $v_h - u_h \in \mathcal{V}_h$, Galerkin orthogonality (1.16) tell us that the second term is zero and hence

$$a(u - u_h, u - u_h) = a(u - u_h, u - v_h). \tag{1.17}$$

Using the coercivity and continuity results we have

$$\alpha \|u - u_h\|^2_{1,\Omega} \leq a(u - u_h, u - v_h) \leq \gamma_1 \|u - u_h\|_{1,\Omega} \|u - v_h\|_{1,\Omega}.$$

Hence we have proved Céa's lemma

$$\|u - u_h\|_{1,\Omega} \leq \frac{\gamma_1}{\alpha} \min\{\|u - v_h\|_{1,\Omega} : v_h \in \mathcal{V}_h\}.$$

Thus the finite element solution u_h is the near-best fit cf [20] to u in the $H^1(\Omega)$ norm. A refined version of Céa's lemma may be proved in the case when we use the energy norm $\|.\|_a$. In this case from the (1.17) we have

$$\|u - u_h\|^2_a = a(u - u_h, u - v_h) \leq \|u - u_h\|_a \|u - v_h\|_a,$$

for any $v_h \in \mathcal{V}_h$, using the Cauchy-Schwarz inequality. Hence we have the refined version of Céa's lemma in the case of energy norm

$$\|u - u_h\|_a = \min\{\|u - v_h\|_a : v_h \in \mathcal{V}_h\}, \tag{1.18}$$

which tell us that u_h minimizes the error in the energy norm over all function in $\mathcal{V}_h$ (cf. [10]).

To make use of (1.18), we must now choose some $v \in \mathcal{V}_h$ that will yield some information about the size of error. Usually one chooses to take v as the piecewise polynomial (of the same degree as the finite element approximation) that interpolates the exact solution. Clearly this polynomial is in $\mathcal{V}_h$. Below we review some basic concepts of polynomial interpolation and derive the interpolation error estimates that play key roles in proving our convergence rates.

1.7.1 Interpolation error

The Lagrange polynomials play an important role in interpolation analysis. Given a function u defined on an interval, the problem is to find a polynomial which equals the function in n discrete separate points. There exists a unique $(n-1)$-th order polynomial P_n such that

$$P_{n-1}(x_i) = u(x_i), \qquad i = 1, \ldots, n.$$

This polynomial may conveniently be defined by use of the Lagrange interpolation polynomials defined as

$$l_i^{n-1}(x) = \frac{(x - x_1)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n)}{(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)}, \tag{1.19}$$

for $i = 1, 2, \ldots, n$, if $n > 1$, and define $l_1^0 = 1$. Since

$$l_i^{n-1}(x_j) = \begin{cases} 1, & \text{if } j = i \\ 0, & \text{if } j \neq i \end{cases}$$

we have that

$$P_{n-1}(x) = \sum_{i=1}^{n} u(x_i) l_i^{n-1}(x).$$

To emphasize the fact that we are interested in interpolating functions, we will use the notation $P_{n-1} = \pi_{n-1} u(x)$ to make explicit which function is being interpolated.

Next, we wish to be able to say something about the interpolation error. Consider an element of size h. We can let $0 \leq x \leq h$ without loss of generality. Now by the Taylor series expansion centered at x_i we have

$$u(x) = \sum_{k=0}^{n} \frac{1}{k!} u^{(k)}(x_i)(x - x_i)^k + \frac{1}{(n+1)!} u^{(n+1)}(\xi_x)(x - x_i)^{n+1},$$

for some ξ_x (depending on x) in the interval between x and ξ. If this interval is of size h, we find that, for some constant C,

$$\left| u(x) - \sum_{k=0}^{n} \frac{1}{k!} u^{(k)}(x_i)(x - x_i)^k \right| \leq C h^{n+1} \left| \frac{d^{n+1} u(\xi_x)}{dx^{n+1}} \right|.$$

The difference between the Taylor polynomial an the Lagrange interpolation polynomial is that the former approximation quickly deteriorates away from x_i.

Can we expect the interpolation polynomial to give about the same estimate? Yes: recall from elementary numerical analysis courses that

$$u(x) - \pi_n u(x) = \frac{(x - x_0)(x - x_1) \ldots (x - x_n)}{(n+1)!} \frac{d^{n+1} u(\xi_x)}{dx^{n+1}}.$$

Again we conclude that

$$|u(x) - \pi_n u(x)| \leq C h^{n+1} \left| \frac{d^{n+1} u(\xi_x)}{dx^{n+1}} \right|, \tag{1.20}$$

for some (other) constant C. The important thing to remember is that the interpolation error depends on the smoothness of the function to be approximated. If the higher derivatives are very large, raising the polynomial degree will not improve the approximation.

Let us derive in more detail the interpolation error for a piecewise linear interpolation on the interval $0 \le x \le h$. The linear interpolant can be written as

$$\pi_1 u(x) = u(0)\frac{h-x}{h} + u(h)\frac{x}{h}. \tag{1.21}$$

Choosing an arbitrary point x in the interval and using a Taylor expansion we find that

$$u(\xi) = u(x) + u'(x)(\xi - x) + \frac{1}{2}u''(\eta)(\xi - x)^2,$$

where η lies between x and ξ. Using $\xi = 0$ and $\xi = h$ we find, by use of (1.21), that

$$\begin{aligned}\pi_1 u(x) = u(x)\Big(\frac{h-x}{h} + \frac{x}{h}\Big) + \\ + u'(x)\Big(\frac{h-x}{h}(-x) + \frac{x}{h}(h-x)\Big) + \\ + \frac{1}{2}\Big(u''(\eta_1)x^2\frac{h-x}{h} + u''(\eta_2)(h-x)^2\frac{x}{h}\Big),\end{aligned}$$

leading to

$$|u(x) - \pi_1 u(x)| \le \frac{1}{2}\max_{\eta\in[0,h]}|u''(\eta)|\Big(x^2\frac{h-x}{h} + (h-x)^2\frac{x}{h}\Big) \le \frac{h^2}{8}\max_{\eta\in[0,h]}|u''(\eta)|.$$

Similarly, by differentiating (1.21) and using the Taylor expansion, we find that

$$\begin{aligned}(\pi_1 u)'(x) = u(x)\Big(-\frac{1}{h} + \frac{1}{h}\Big) + \\ + u'(x)\Big((-\frac{1}{h})(-x) + \frac{1}{h}(h-x)\Big) + \\ + \frac{1}{2}\Big(u''(\eta_1)x^2(-\frac{1}{h}) + u''(\eta_2)(h-x)^2\frac{1}{h}\Big) \\ + \frac{1}{h}\Big(u''(\eta_1)x(h-x) - u''(\eta_2)x(h-x)\Big),\end{aligned}$$

from which follows

$$|u'(x) - (\pi_1 u)'(x)| \le \frac{h}{2}\max_{y\in[0,h]}|u''(y)|.$$

These results have direct counterparts in integral norms. Below we give (without proofs) some inequalities of this nature that will be used in deriving our a priori and a posteriori error estimates.

Lemma 1.9. *Let $\pi_h u(x)$ be the piecewise linear interpolant of $u(x)$. There are interpolation constants C_i such that*

$$\|\pi_h u(x) - u(x)\|_{L_p(\Omega)} \leq C_i \|h^2 u''(x)\|_{L_p(\Omega)}, \quad 1 \leq p < \infty \tag{1.22a}$$

$$\|(\pi_h u)'(x) - u'(x)\|_{L_p(\Omega)} \leq C_i \|h u''(x)\|_{L_p(\Omega)}, \quad 1 \leq p < \infty \tag{1.22b}$$

$$\|\pi_h u(x) - u(x)\|_{L_p(\Omega)} \leq C_i \|h u'(x)\|_{L_p(\Omega)}, \quad 1 \leq p < \infty. \tag{1.22c}$$

Now, the energy norm is a weighted variant of (1.22b). As an example, we have in (1.2) that

$$\|u\|_a = \|u'\|_{L_2(\Omega)},$$

so that we may conclude that

$$\|u - \pi_h u\|_a \leq C_i h \|u''\|_{L_2(\Omega)}. \tag{1.23}$$

1.7.2 A priori error estimation in the energy norm

With the results of interpolation section we can now formulate error estimates for the finite element methods for our elliptic boundary value problem. These estimates are either in the a priori form: i.e., error estimates which are formulated in terms of the exact solution, or in the a posteriori form: that is error estimates that are formulated in terms of the computed solution and the data (right-hand side), which we shall consider, for the same model problem, in the next subsection. We have the following a priori bound for our model problem (see [10])

Lemma 1.10. *Let u and u_h be the solutions of Dirichlet problem* (1.2) *and corresponding finite element problem* (1.15)*, respectively. Then there exist an interpolation constant C_i such that*

$$\|u - u_h\|_a \leq C_i \|h\, u''\|_{L_2(\Omega)}, \quad \forall v \in \mathcal{S}_h.$$

Proof. According to the the refined version of Céa's lemma (1.18), we have

$$\|u - u_h\|_a \leq \|u - v\|_a, \quad \forall v \in \mathcal{S}_h.$$

But since $\pi_h u(x) \in \mathcal{S}_h$, then

$$\begin{aligned}\|u - u_h\|_a \leq \|u - \pi_h u\|_a &= \|u' - (\pi_h u)'\|_{L_2(\Omega)} \leq C_i \|h\, u''\|_{L_2(\Omega)} \\ &= C_i \Big(\int_0^1 h^2 \, (u''(x))^2 dx \Big)^{1/2},\end{aligned}$$

where in the last inequality above we used (1.23). □

1.7.3 A posteriori error estimation in the energy norm

Now we want to study a posteriori error estimate, where instead of the unknown value of $u(x)$, we use the known values of the approximate solution to estimate the error. This means that the error analysis performed after the computation is completed. Then we have the following a posteriori error analysis for the problem (1.2), without proof, for more details we refer to, e.g., [10].

Lemma 1.11. *There is an interpolation constant C_i, such that the error in the finite element approximation satisfies*

$$\|e(x)\|_a \leq C_i \Big(\int_0^1 h^2 \mathcal{R}^2(u_h(x)) dx \Big)^{1/2}.$$

Proof. See [10]. □

1.8 Comparison Results and Inequalities

In this part we present a comparison result and some inequalities, which will be useful in the stability and error estimates as well as in convergent analysis of our problem.

Lemma 1.12. *(**Grönwall's inequality**). Suppose that $u(t) \geq 0$ and $\varphi(t) \geq 0$ are continuous, real-valued functions defined on the interval $0 \leq t \leq T$ and $u_0 \geq 0$ is a constant. If u satisfies the inequality*

$$u(t) \leq u_0 + \int_0^t \varphi(s)u(s)ds, \qquad \forall t \in [0,T], \tag{1.24}$$

then

$$u(t) \leq u_0 \exp\Big(\int_0^t \varphi(s)ds\Big)ds, \qquad \forall t \in [0,T].$$

In particular, if $u_0 = 0$ then $u(t) \equiv 0$.

Proof. Suppose first that $u_0 > 0$. Let

$$U(t) = u_0 + \int_0^t \varphi(s)u(s)ds.$$

Then, since $u(t) \leq U(t)$, we have that

$$\dot{U} = \varphi u \leq \varphi U, \quad U(0) = u_0.$$

Since $U(t) > 0$, it follows that

$$\frac{d}{dt}\log U = \frac{\dot{U}}{U} \leq \varphi.$$

Hence

$$\log U(t) \leq \log u_0 + \int_0^t \varphi(s)ds,$$

so

$$u(t) \leq U(t) \leq u_0 \exp\Big(\int_0^t \varphi(s)\,ds\Big). \tag{1.25}$$

If the inequality (1.24) holds for $u_0 > 0$, then it also holds for all $u_0 = 0$. Taking the limit of (1.25) as $u_0 \longrightarrow 0^+$, we conclude that $u(t) \equiv 0$, which proves the result when $u_0 = 0$. □

Lemma 1.13 (Inverse estimate inequality). . *Let $\mathcal{T}_h$ be a quasi-uniform family of triangulation of $\bar{\Omega}$. Then there exist a positive constant C such that for each $v_h \in \mathcal{V}_h$*

$$\|\nabla v_h\|^2_{L_2(\Omega)} \leq Ch^{-2}\|v_h\|^2_{L_2(\Omega)}. \tag{1.26}$$

Proof. A proof can be founded in Quarteroni and et al. [65]. □

In the proofs below we shall frequently use the following Inequalities:

Lemma 1.14. *For $a, b \in \mathbb{R}$, and $\varepsilon > 0$*

$$ab \leq \frac{\varepsilon a^2}{2} + \frac{b^2}{2\varepsilon}, \tag{1.27a}$$

$$ab \leq \varepsilon a^2 + \frac{b^2}{4\varepsilon}. \tag{1.27b}$$

Proof. The proof is straightforward. □

Chapter 2

THE COUPLED PROBLEM

2.1 Continuous Model Problem

In this study we consider the coupling of two hyperbolic equations. The objective is to find the function $u : (x,t) \in \mathbb{R} \times \mathbb{R}_+ \longrightarrow u(x,t) \in \mathbb{R}$ such that $u = u(x,t)$ satisfies the following nonlinear hyperbolic system of equations:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f_R(u) = 0, \quad x > 0, \quad t > 0, \tag{2.1}$$

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f_L(u) = 0, \quad x < 0, \quad t > 0, \tag{2.2}$$

associated with an initial condition, viz

$$u(x,0) = u_0(x), \quad x \in \mathbb{R}, \tag{2.3}$$

and also a suitable "continuity" condition or "coupling" condition

$$u(0,t) = u^b(t), \quad t \geq 0, \tag{2.4}$$

at the interface $x = 0$, to be specified below, where $u_0 : \mathbb{R} \longrightarrow \mathbb{R}$ is a given function and $f_\alpha : \mathbb{R} \longrightarrow \mathbb{R}, \alpha = L, R$, denote two "smooth" functions.

This is a system of nonlinear conservation laws arising in the study of fluid problems with two different systems of equations on each side of the interface which may be fixed or moving with the flow. For instance, on one side of the interface, we assume that the flow obeys the rules of a gas dynamical system while on the other side flow is arbitrary. A more complex problem is the case of having to couple the isentropic system of gas dynamics with the Euler system of gas dynamics.

The problem (2.1)-(2.3), in general is not well-posed when the boundary data are imposed in the strong sense (2.4). In the case of nonlinear hyperbolic equations with initial and boundary condition, a first question is to understand the sense of boundary condition. In some particular cases, the boundary conditions can be founded by physical consideration, but their derivation in the general case is not obvious. The problem of finding the "correct" boundary conditions , i.e., which lead to well-posed problem, is difficult in general from both the theotrical and practical points of view. Here we review some features of these issues in details.

2.2 Initial Boundary Value Problems for Hyperbolic Equations

Numerical study of the industrial problems are, basically, consisting of simulating efficient schemes in the sense that the obtained numerical solutions satisfy, for instance, some natural constraints and which are seem to be in good agreement with the experimental data. However, in many cases, it is not so clear to understand the problem which is solved by the limit of the numerical solutions as the discretization parameter tends to 0. In this section we focus on the case where the limit equation is a hyperbolic (scalar) equation. In this case, as we mentioned earlier, the main difficulty is to understand the boundary conditions which are satisfied by the limit

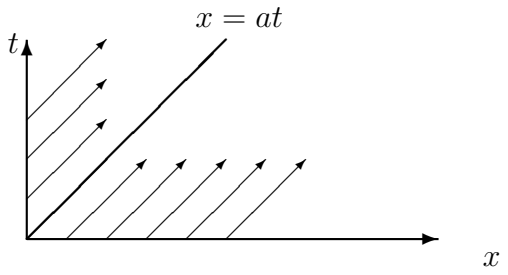

Figure 2.1: The characteristic carrying the boundary data and initial data.

of the numerical solutions. An interesting example is given by the modelization of a two phase flow in a porous medium used in reservoir simulation.

2.2.1 Linear scalar problems in one dimensional

We start with the simplest case "quarter plane problem" in one dimension, i.e., we consider the simple case of linear advection equation

$$u_t + au_x = 0,$$

where a is a constant and u is scalar.

The question is: what prescribed data on $\{x = 0, t \geq 0\}$ and on $\{x \geq 0, t = 0\}$, would guarantee a well-posed problem that we can hope to solve in the whole $\mathbb{R}_+ \times \mathbb{R}_+$? We split the cases $a > 0$ and $a < 0$ and illustrate the problem more closely:

- $a > 0$.

  In this case the characteristics are going from left to right. In particular they enter the domain (i.e., $\mathbb{R}_+ \times \mathbb{R}_+$ from the boundary $\{x = 0, t \geq 0\}$) as is shown if Figure 2.1: therefore, one needs to prescribe the solution at the boundary

$x = 0$. From the method of characteristics it is clear that if

$$\begin{cases} u(x,0) &= u_0(x), \\ u(0,t) &= g(t), \end{cases}$$

are prescribed as the initial and boundary data, then the solution of $u_t + a u_x = 0$ in $\mathbb{R}_+ \times \mathbb{R}_+$ is given by,

$$u(x,t) = \begin{cases} u_0(x-at), & \text{for} \quad x > at, \\ g(t - x/a), & \text{for} \quad x < at. \end{cases} \tag{2.5}$$

To have a smooth solution in the whole domain it is necessary that both $g(t)$ and $u_0(x)$ are smooth functions. It is also necessary that $g(t)$ and $u_0(x)$ be compatible or satisfy compatibility conditions. The most obvious necessary conditions are

$$\begin{aligned} u_0(0) &= g(0), \\ u_0'(0) &= -\frac{g'(0)}{a}, \end{aligned}$$

otherwise, the solution (2.5) is a weak solution satisfying the Rankine-Hougoniot jump condition on each side of the discontinuity $x = at$.

- $a < 0$.

  In this case the characteristics are outgoing from the interior of the domain and the information is carried from the given initial data u_0. From drawing the characteristics it is clear that in this case we are not free to prescribe the boundary data on $\{x = 0, t > 0\}$; i.e., the initial data determines boundary data (see Figure 2.2). The solution is

  $$u(x,t) = u_0(x - at), \quad x \geq 0, \quad t \geq 0,$$

  in particular

  $$u(0,t) = u_0(-at).$$

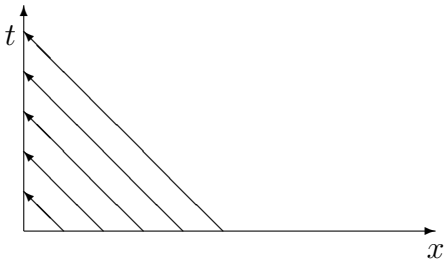

Figure 2.2: The characteristic carrying the initial data.

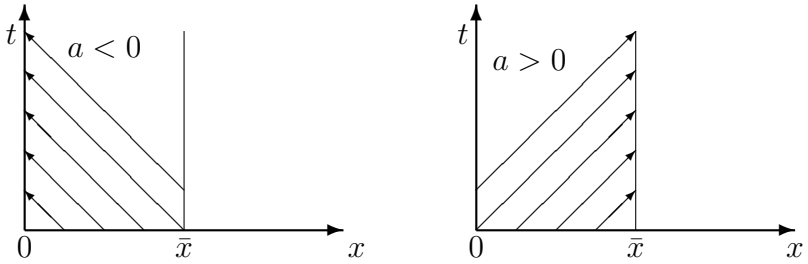

Figure 2.3: The characteristic carrying the initial data.

Note that if $a = 0$, we do not need any boundary conditions because $\frac{\partial u}{\partial x} = 0$ implies that

$$u(x,t) = u_0(x).$$

Combining both cases enables one to treat the case of a bounded interval in space, for instance the strip $\{(x,t)\big|0 \le x \le \bar{x}, t \ge 0\}$. The boundary that must be specified corresponds to incoming characteristics (see Figure 2.3)

$$\begin{cases} u(0,t) = g(t), & \text{for} \quad a > 0, \\ u(\bar{x},t) = h(t), & \text{for} \quad a < 0. \end{cases}$$

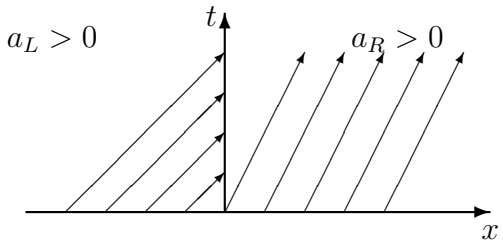

Figure 2.4: The characteristic line for $a_R > 0$ and $a_L > 0$.

2.3 The One Dimensional Coupled problem

In order to make the coupling condition more explicit, we consider the following example (see [37]):

Example 2.1. We start by considering the simplest case where

$$f_L = a_L u, \quad f_R = a_R u,$$

for some nonzero constants a_R and a_L. Due to the sign of a_L, a_R and depending on the directions of the characteristic lines we have the following cases:

1. $a_L > 0$, $a_R > 0$ or $a_L < 0, a_R < 0$ ($a_L\, a_R < 0$): we can impose the continuity condition at $x = 0$ (cf. Figure 2.4).

2. $a_L > 0,\ a_R < 0$: no continuity condition is required at $x = 0$ (cf. Figure 2.5);

3. $a_L < 0,\ a_R > 0$: we need to specify $u(0, t)$ at $x = 0$, otherwise the solution u is not defined in the fan $a_L t < x < a_R t$ (cf. Figure 2.6).

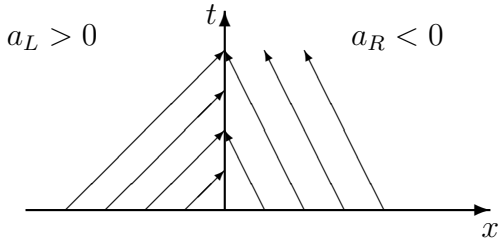

Figure 2.5: The characteristic line for $a_R < 0$ and $a_L > 0$.

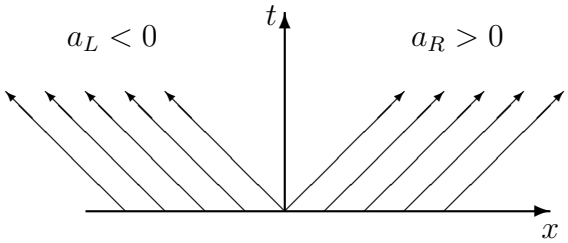

Figure 2.6: The characteristic line for $a_R > 0$ and $a_L < 0$.

2.4 A General Existence and Uniqueness Results for IBVP's

To show the well-posedness of problem (2.1)-(2.3), we consider the general nonlinear conservation laws and as a preliminary step we study the IBVP

$$\begin{cases} u_t + f(u)_x &= 0, & x \in \mathbb{R}_+, \quad \in (0,T), \\ u(x,0) &= u_0(x), & x \in \mathbb{R}_+, \\ u(0,t) &= u^b(t), & t \in (0,T). \end{cases} \tag{2.6}$$

In this section we comment (without proofs) on the existence and uniqueness results for the above IBVP and state two admissible boundary sets. For future details we refer to Godlewski and Raviart [37] and relevant papers.

Searching for an existence and uniqueness theorem for IBVPs for scalar conservation laws it is natural to apply the same method as for the Cauchy problem for

"vanishing viscosity": We thus pose the parabolic IBVP for $\varepsilon > 0$,

$$\begin{aligned} u_t^\varepsilon + f(u^\varepsilon)_x - \varepsilon u_{xx}^\varepsilon &= 0, && x \in \mathbb{R}_+, t \in (0,T), \\ u^\varepsilon(x,0) &= u_0(x), && x \in \mathbb{R}_+, \\ u^\varepsilon(0,t) &= u^b(t), && t \in (0,T). \end{aligned} \tag{2.7}$$

One can use the standard parabolic theory to show the existence of a unique solution u^ε to this problem, and one can then attempt to take the limit of u^ε as $\varepsilon \downarrow 0$. It is known that u^ε will converge to a limiting function $u(x,t)$, and this u is what we would like to define as the (hopefully) unique entropy solution of the inviscid IBVP (2.6).

Let us recall the sense of the weak entropy solution for IBVP (2.6).

Definition 2.2. A function $u \in L_\infty(\mathbb{R}_+ \times \mathbb{R}_+) \cap BV_{loc}(\mathbb{R}_+ \times \mathbb{R}_+)$ is called a weak entropy solution of the initial-boundary value problem (2.6), if for each $k \in \mathbb{R}$, and for all nonnegative test functions $\phi \in C_0^\infty(\mathbb{R}_+ \times \mathbb{R}_+)$, the following inequality holds:

$$\begin{aligned} \int_0^\infty \int_0^\infty \Big\{ |u-k|\phi_t + sgn(u-k)(f(u)-f(k))\phi_x \Big\} dxdt \\ + \int_0^\infty |u_0(x)-k|\phi(x,0)dx \\ + \int_0^\infty sgn(u^b(t)-k)(f(u(0,t))-f(k))\phi(0,t)dt \geq 0. \end{aligned} \tag{2.8}$$

Indeed, we have the following result (cf. Bardos et al. [15], see also Godlewski et al. [39] for instance):

Theorem 2.3. *Suppose that $u_0 \in BV(\mathbb{R}_+)$ and $u^b \in BV(0,T)$ for all $T > 0$. Then there exists a unique entropy solution*
$u \in L_\infty(\mathbb{R}_+ \times (0,T)) \cap BV_{loc}(\mathbb{R}_+ \times (0,T)) \cap C_0([0,T]; L_1(\mathbb{R}))$ *for all $T > 0$,*
of problem (2.6) *in the sense of* (2.8).

Let us interpret the definition of a weak entropy solution in (2.8). If the boundary data g is given along $x = 0$, and $u(t)$ denotes the "trace" u, i.e., the value we get by

taking the limit from the interior ($u(t) = \lim_{x \longrightarrow 0^+} u(x,t)$), then (2.8) says that

$$-sgn(u(t) - g(t))(f(u(t)) - f(k)) \geq 0,$$

for all k between $u(t)$ and $g(t)$, where

$$sgn(x) = \begin{cases} 1, & \text{for} \quad x > 0, \\ 0, & \text{for} \quad x = 0, \\ -1, & \text{for} \quad x < 0. \end{cases}$$

With this definition we have that the first set $\mathcal{V}(u^b(t))$ of the admissible values at the boundary due to Bardos, Leroux, Nédélec [15] (see for instance Godlewski [39]), is defined as all $u(0,t)$ such that for all $t > 0$

$$\begin{cases} f(u(0,t)) \leq f(k), & \text{for all k with} \quad u^b(t) \leq k \leq u(0,t), \\ f(u(0,t)) \geq f(k), & \text{for all k with} \quad u(0,t) \leq k \leq u^b(t). \end{cases}$$

Then the boundary condition can be written as follows:

$$u(0,t) \in \mathcal{V}(u^b(t)), \quad \forall t > 0. \tag{2.9}$$

Also a second set of *admissible boundary values*, is given as

$$\mathcal{U}(u^b) = \{v = w(0+; u^b, u_R) : u_R \in \mathbb{R}\},$$

for all $u^b \in \mathbb{R}$, where $(x,t) \longrightarrow w(x,t) = w(x/t; u_L, u_R)$ denotes the solution of the Riemann problem

$$w_t + f(w)_x = 0, \qquad x \in \mathbb{R}, t > 0,$$

with the initial data

$$u(x,0) = \begin{cases} u_L, & \text{for } x < 0, \\ u_R, & \text{for } x > 0. \end{cases}$$

Then the admissible boundary condition reads (see [25])

$$u(0,t) \in \mathcal{U}(u^b(t)), \quad t > 0. \tag{2.10}$$

By the above definition, the boundary condition for our main problem (2.1)-(2.3), can be replaced by the following coupling conditions

$$\begin{cases} u(0+,t) \in \mathcal{U}_R(u(0-,t)), \\ u(0-,t) \in \mathcal{U}_L(u(0+,t)), \end{cases}$$

where

$$\mathcal{U}_L(u^b) = \{v = w_L(0-; u_L, u^b) : u_L \in \mathbb{R}\},$$
$$\mathcal{U}_R(u^b) = \{v = w_R(0+; u^b, u_R) : u_R \in \mathbb{R}\},$$

$w_\alpha(x/t; u_L, u_R)$ denotes the solution of the Riemann problem associated with the flux functions $f_\alpha,\ \alpha = L, R$ and $u(0\pm, t)$ is the trace of the solution u at the boundary $x = 0\pm$.

Remark 2.4. *It is known [25] that these two formulation are equivalent for scalar conservation laws, namely, $\mathcal{V}(u^b(t)) = \mathcal{U}(u^b(t))$ for all $t > 0$.*

Remark 2.5. *If we suppose that $f'(v) > 0$ for all v, both conditions* (2.9) *and* (2.10) *are reduced to the boundary conditions $u(0,t) = u^b(t)$. On the other hand, if $f'(v) < 0$ for all v, these conditions are automatically satisfied and therefore we do not need to involve any constraint on the boundary value $u(0,t)$.*

Chapter 3

STREAMLINE DIFFUSION FORMULATION

In this section we introduce the *streamline diffusion finite element method* (Sd-method below) discretization of the coupled problem.

3.1 The Space-Time Discretizations and Finite Element Spaces

We start with the first equation of the above system; (2.1), and reformulate it as the following initial boundary value problem: Find a scalar function $u = u(x,t)$, where x is the space variable and t denotes the time, such that

$$\begin{cases} \dfrac{\partial u}{\partial t} + f_R'(u)\dfrac{\partial u}{\partial x} = 0, & (x,t) \in \Omega := \mathbb{R}_+ \times (0,T), \\ u(x,0) = u_0, & (x,t) \in \Omega_0 := \mathbb{R}_+ \times \{0\}, \\ u(0,t) = u^b, & (x,t) \in \Gamma := \{0\} \times (0,T), \end{cases} \tag{3.1}$$

where $(0,T)$ is a given time interval. To derive the basic stability, it is necessary to put some restrictions on $f_\alpha(\alpha = L, R)$ in the problems (1.1)-(1.3). Our basic assumptions that we called non-rarefaction assumptions are:

- $\alpha = R, \qquad f'_R(u) \le 0$ (particular $f'_R(u^b) \le 0$), $\qquad \frac{\partial}{\partial x}(f'_R(u)) \le 0.$
- $\alpha = L, \qquad f'_L(u) \ge 0$ (particular $f'_L(u^b) \ge 0$), $\qquad \frac{\partial}{\partial x}(f'_L(u)) \ge 0.$

In each case the first condition is required for existence and uniqueness of continuous problem (1.1)-(1.3), while the second condition is required for existence and uniqueness (in the discrete version of Lax-Milgram lemma) of discrete problem (see Lemma 3.1 below).

In the numerical study of the problem (3.1), because of its hyperbolic nature, the standard Galerkin finite element method does not give satisfactory convergence rates, compared to elliptic or parabolic problems. This is due to the fact that a regularizing diffusion term of the type u_{xx} is missing in the hyperbolic equations. To circumvent this phenomenon, below we shall introduce the streamline diffusion method (Sd)-method for this problem. To this approach, let $\{0 = t_0 < t_1 < ... < t_N = T\}$ be a partition of the time interval $I = (0,T)$ into the subintervals $I_n = (t_n, t_{n+1})$, with time steps $k_n = t_{n+1} - t_n$, $n = 0, 1, \ldots, N-1$ and introduce the corresponding space-time "slabs", i.e.,

$$\mathbf{S}_n = \{(x,t) : x > 0, \quad t_n < t < t_{n+1}\}, \qquad n = 0, 1, \ldots, N-1.$$

For each slab $\mathbf{S}_n$, let x_i^n be a mesh on $\mathbb{R}_+$, portioned in intervals $J_i^n = (x_{i-1}^n, x_i^n)$, with $h_i^n = x_i^n - x_{i-1}^n$. We define the global mesh function $h = h(x,t)$ by $h(x,t) = h^n(x)$ for $x \in \mathbb{R}_+$, $t \in (t_n, t_{n+1})$ and the function $k = k(t)$ by $k(t) = k_n$ for $t \in (t_n, t_{n+1})$. For $h > 0$, let T_h^n be a triangulation of the slab $\mathbf{S}_n$ into triangles K (cf. Figure 4.), satisfying as usual the minimum angle condition (see e.g., Ciarlet [20]), and indexed by the parameter h representing the maximum diameter of the triangles $K \in T_h^n$.

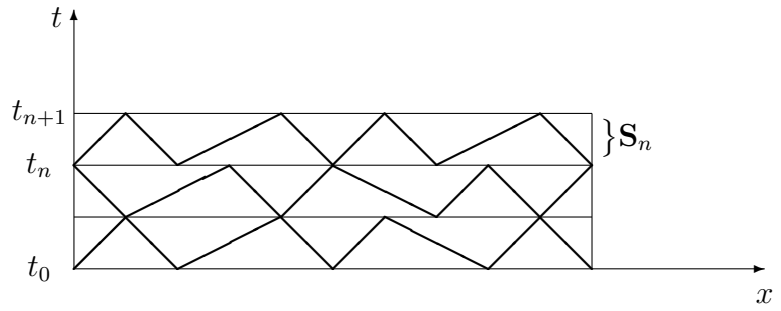

Figure 3.1: The space-time triangulation.

The triangulation of $\mathbf{S}_n$ may be chosen independently of that of $\mathbf{S}_{n-1}$, but for the sake of simplicity it must satisfy the standard quasi-uniformity conditions for finite element meshes ("uniform shape and size" condition), i.e., there exist a number $\varrho > 0$, independent of h and K such that

$$\frac{\rho_K}{h_K} \geq \varrho, \qquad \forall K \in T_h^n,$$

where the parameters h_K and ρ_K are defined as follows:

$$h_K = diam(K),$$

and

$$\rho_K = \sup\{diam(B) : B \text{ is a ball contained in } K\}.$$

This condition means that the triangles $K \in T_h^n$ are not allowed to be arbitrary thin, or equivalently, the angels of the triangles K are not allowed to be arbitrary small. The constant ϱ is a measure of the smallest angle in any $K \in T_h^n$. Let now k be a positive integer and introduce the finite element space

$$U_h^n = \{u \in H^1(\mathbf{S}_n) : u\big|_K \in P_k(K), \quad K \in T_h^n\},$$

where $P_k(K)$ denotes the set of polynomials on K of degree less than or equal to k, i.e., U_h^n is the space of all continuous piecewise polynomial functions, on $\mathbf{S}_n$, of

degree $\leq k$. Then we define the trial and test function spaces as the subspaces of U_h^n by

$$V_h^n = \{v \in U_h^n : v\big|_\Gamma = u_h^b\},$$

and

$$W_h^n = \{w \in U_h^n : w\big|_\Gamma = 0\},$$

respectively. Here u_h^b is the trace of a function in U_h^n approximating u^b on Γ. Note that we may assume that the functions $v(x,t) \in \mathcal{V}_h^n$ and $w(x,t) \in \mathcal{W}_h^n$ are vanishing for sufficiently large $|x|$.

We shall use the following notation: recall that for a given domain Q we denote by $(.,.)_Q$ the usual $L_2(Q)$ scalar product, $||.|| = \|.\|_{L_2(Q)}$ the corresponding L_2 norm, and for a positive integer s, $H^s(Q)$ will denote the usual Sobolev space of functions with square integrable derivative of order less than or equal s and with the norm $||.||_{s,Q}$. We also write $\|.\|_{\infty,Q} = \|.\|_{L_\infty(Q)}$. Further for piecewise polynomials w_i defined on the triangulation $T_h' = \{K\}$, where $T_h' \subset T_h$ and for D_i, $i = 1, 2$, being some differential operators, we use the notation

$$(D_1 w_1, D_2 w_2)_{Q'} = \sum_{K \in T_h'} (D_1 w_1, D_2 w_2)_K, \quad Q' = \bigcup_{K \in T_h} K,$$

i.e., we just sum the integrals over each element $K \in T_h'$. We also write

$$(u,v)_n = \int_{\mathbf{S}_n} uv dx dt, \qquad ||v||_n = (v,v)_n^{1/2},$$

$$< u, v >_n = \int_{\mathbb{R}_+} u(x,t_n)v(x,t_n)dx, \qquad |v|_n = < v, v >_n^{1/2},$$

$$v_+ = \lim_{s \longrightarrow 0+} v(x,t+s), \qquad v_- = \lim_{s \longrightarrow 0-} v(x,t+s).$$

3.2 Space-Time Streamline Diffusion Finite Element Method

We start with a streamline diffusion variational formulation for the continuous problem (3.1) where, roughly speaking, in order to win a small diffusive term, we have test functions of the form $v + \delta(v_t + f_R'(u)v_x)$ rather than only v: Given $u(x,0) = u_0(x)$, find $u \in H^1(\Omega)$ with $u|_\Gamma = u^b$, such that

$$\begin{aligned}\Big(u_t + f_R'(u)u_x, v + \delta(v_t + f_R'(u)v_x)\Big)_\Omega + \int_\Gamma uv\, d\sigma dt \\ = \int_\Gamma u^b v\, d\sigma dt, \qquad \forall v \in H_0^1(\Omega),\end{aligned} \tag{3.2}$$

where $d\sigma$ being surface element is assumed to be 1 in our one dimensional problem. In the above, instead of using the standard Galerkin method for the one variable (spatial or time) we used the Galerkin method simultaneously in space and time. That is, we use finite element and interpolation functions depending on time and space. Space-time Sd-method can be used to improve stabilization; however, used without care, this would lead to a very large linear system to be solved. The reason for this is that in this technique the use of continuous (in time) test and trial functions coupled all levels of time. One way to avoid this difficulty, and decrease the size of the corresponding linear system, is to work in slabs of space-time, with the help of interpolation functions that will be continuous in the spatial variables but will be discontinuous in the time variables at the common frontier of every two slabs. With the test and trial functions continuous in space and with jump discontinuities at the partition points on the time interval, this can be reformulated as follows: For $n = 0, 1, \ldots, N-1$ find $u \in H^1(\mathbf{S}_n)$, such that

$$\begin{aligned}&\Big(u_t + f_R'(u)u_x, v + \delta(v_t + f_R'(u)v_x)\Big)_n + < u_+, v_+ >_n \\ &+ \int_\Gamma uv\, d\sigma dt =< u_-, v_+ >_n + \int_\Gamma u^b v\, d\sigma dt, \qquad \forall v \in H_0^1(\mathbf{S}_n).\end{aligned} \tag{3.3}$$

The terms including $< . >$ in the above formula is a jump conditions which imposes a weakly enforced continuity condition across the slab interfaces, at t_n and is the mechanism by which information is propogated from one slab to another. For more concisely, after summing over n, we may rewrite (3.3) as follows: Find $u \in \prod_{n=0}^{N-1} H^1(\mathbf{S}_n)$, such that

$$B(u,v) = L(v), \qquad \forall v \in \prod_{n=0}^{N-1} H_0^1(\mathbf{S}_n), \tag{3.4}$$

where the bilinear form $B(.,.)$ and the linear form $L(.)$ are defined by

$$B(u,v) = \sum_{n=0}^{N-1} \Big\{ \Big(u_t + f_R'(u)u_x, v + \delta(v_t + f_R'(u)v_x) \Big)_n + < u_+ - u_-, v_+ >_n + \int_{\Gamma_n} u_+ v_+ dt \Big\},$$

and

$$L(v) =< u_0, v_+ >_0 + \int_\Gamma u^b v_+ dt.$$

We now formulate the, space-time, discrete streamline diffusion finite element method, for the problem (3.1) as follows: Find $u_h^n \in \mathcal{V}_h^n$, such that for $n = 0, 1, \ldots, N-1$

$$\begin{aligned} &\Big(u_{h,t}^n + f_R'(u_h^n)u_{h,x}^n, v_h^n + \delta(v_{h,t}^n + f_R'(u_h^n)v_{h,x}^n) \Big)_n + < u_{h,+}^n, v_{h,+}^n >_n \\ &+ \int_{\Gamma_n} u_{h,+}^n v_{h,+}^n dt =< u_{h,-}^n, v_{h,+}^n >_n + \int_{\Gamma_n} u^b v_{h,+}^n dt, \qquad \forall v_h^n \in \mathcal{W}_h^n, \end{aligned} \tag{3.5}$$

where $\delta = \bar{C}h$ with $\bar{C}$ a suitably chosen (sufficiently small, see Johnson et al. [53], [54]) positive constant, $\Gamma_n = \{0\} \times I_n$ and $u_{h,-}^0 = u_0$ is the initial data. Formally, (3.5) can be thought of as a perturbation of the standard Galerkin method corresponding to $\delta \equiv 0$. We can show that (see Nävert [63]) if $u_{h,-}(., t_n)$ is given, then (3.5) defines u_h^n uniquely in the slab $\mathbf{S}_n$ and thus u_h^n can be computed successively on all the slabs $\mathbf{S}_n$, $n = 0, 1, \ldots, N-1$, starting at $\mathbf{S}_0$, where $u_{h,-}^0 = u_0$ is given. For each n, (3.5) is equivalent to a (non) linear system of equations and thus (3.5) corresponds to an implicit scheme for (3.1). Below, in Lemma 4.1, we prove the uniqueness and hence also the existence of a solution for the (non) linear system of

equations represented by (3.5). Summing over n, (taking all the slabs together) we get the function spaces

$$\mathcal{V}_h = \prod_{n=0}^{N-1} \mathcal{V}_h^n, \qquad \mathcal{W}_h = \prod_{n=0}^{N-1} \mathcal{W}_h^n.$$

i.e., we shall seek an approximate solution $u_h \in \mathcal{V}_h$ such that for $n = 0, 1, ..., N-1$ we will have that $u_h \mid_{\mathbf{S}_n} = u_h^n$.

We emphasize that the functions in $\mathcal{V}_h$ are continuous in x and possibly discontinuous in t at discrete time levels t_n. Similar properties are valid for $w_h \in \mathcal{W}_h$.

In order to write (3.1) in a compact form suitable for analysis, and because the functions in $\mathcal{V}_h$ may be discontinuous in time, we introduce the jump terms $[v]$ across each time level by defining , for $x > 0$ and $n = 0, 1, \ldots, N-1$,

$$[v](x, t_n) = \begin{cases} v_+, & \text{if } n = 0 \\ v_+ - v_-, & \text{if } n \neq 0. \end{cases}$$

And then summing (3.5) over $n = 0, 1, \ldots, N-1$, we get the following discrete analogue of (3.3)-(3.4): Find $u_h \in \mathcal{V}_h$ such that

$$B(u_h, v) = L(v), \qquad \forall v \in \mathcal{W}_h. \tag{3.6}$$

Now subtracting (3.6) from (3.4) we get the *Galerkin orthogonality relation:*

$$B(u - u_h, v) = 0, \qquad \forall v \in \mathcal{W}_h. \tag{3.7}$$

The *Galerkin orthogonality relation:* (3.7) is very useful in our error estimates below.

Chapter 4

FINITE ELEMENT ANALYSIS FOR THE STREAMLINE DIFFUSION

Goal of the design of any numerical computational method is

- Reliability: Computational error is controlled by a predefined tolerance,
- Efficiency : Computational work is as small as possible,

and these efficiency and reliability of numerical approaches are studied in two different settings: The a priori error estimates, which are of theoretical nature and formulated in terms of the exact solution, and the a posteriori error estimates that are employed in low regularity cases and are formulated in terms of the *residual term*, i.e., computed solution and the data (right hand side) in the problem. The two approaches have different advantages and features and serve for theoretical numerical convergence analysis and computational and implementing purposes, respectively. In this section we derive the basic stability estimates for the Sd-method and also study the convergence rate of the approximate solution for (3.6) and derive a priori

error estimates. The, corresponding, a posteriori error estimates is the matter of the next section. In what follows, C and c will denote positive constants, independent of h, and not necessarily the same at each occurrence, unless otherwise explicitly stated. Normally if C and c appear in the same chain of estimates, then we shall mean $Cc = \mathcal{O}(1)$.

4.1 The Basic Stability Estimates for the Sd-Method

Below we derive the stability estimates for the Sd-method (2.2). These estimates will be of crucial importance in proving the finite element error analysis. Stability estimates would usually follow from the coercivity properties of the bilinear form $B(\cdot,\cdot)$ associated to the variational formulation for the underlying problem. For our problem we have the following stability lemma:

Lemma 4.1. *For $u \in \prod_{n=0}^{N-1} H^1(\boldsymbol{S}_n)$, and with the non-rarefaction assumptions $f_R'(u^b) \le 0$, and $\frac{\partial}{\partial x}\Big(f_R'(u)\Big) \le 0$, we have that*

$$B(u,u) \ge |||u|||^2, \tag{4.1}$$

where

$$|||u|||^2 := \frac{1}{2}\left[|u_-|_N^2 + |u_+|_0^2 + \sum_{n=1}^{N-1}|[u]|_n^2 + 2\delta\|u_t + f_R'(u)u_x\|_\Omega^2\right] + \|u_+\|_\Gamma^2.$$

Proof. Using the definition of the bilinear form B and setting $u = v$ it follows that

$$B(u,u) = (u_t,u)_\Omega + (f_R'(u)u_x,u)_\Omega + \delta\|u_t + f_R'(u)u_x\|_\Omega^2 + \|u_+\|_\Gamma^2 + \sum_{n=0}^{N-1} < [u], u_+ >_n .$$

Integrating by parts yields

$$(u_t,u)_\Omega + \sum_{n=1}^{N-1} < [u], u_+ >_n + < u_+, u_+ >_0 = \frac{1}{2}\left[|u_-|_N^2 + |u_+|_0^2 + \sum_{n=1}^{N-1}|[u]|_n^2\right].$$

Further we have that

$$(f_R'(u)u_x, u)_\Omega = -\frac{1}{2}\Big[\int_0^\infty \int_0^T \frac{\partial}{\partial x}\Big(f_R'(u)\Big)u^2\,dx + \int_0^T f_R'(u^b)(u^b)^2\Big],$$

which is obviously nonnegative by using the non-rarefaction assumptions. Thus the proof is complete. □

Note that the above lemma implies that the discrete problem (3.6) possesses a unique solution $u_h \in \mathcal{V}_h$ and that the stability of the scheme is guaranteed, (see, e.g., Lax-Milgram Lemma). In the following lemma we obtain an estimate for $\|u(t)\|_\Omega$ for all $t > 0$ as follows:

Lemma 4.2. *For any $C > 0$, we have for $u \in \prod_{n=0}^{N-1} H^1(\boldsymbol{S}_n)$ that*

$$\|u(t)\|_\Omega^2 \le \left[\sum_{n=1}^{N} |u_-|_n^2 + \frac{1}{C}\|u_t + f_R'(u)u_x\|_\Omega^2\right]\exp(Ch).$$

Proof. Using the inequality (1.27a), and c.f. Johnson and Szepessy [55] we have, for $t_n < t < t_{n+1}$, that

$$\begin{aligned}
\|u(t)\|_{\mathbb{R}_+}^2 &= |u_-|_{n+1}^2 - \int_t^{t_{n+1}} \frac{d}{ds}\|u(s)\|_{\mathbb{R}_+}^2\,ds \\
&= |u_-|_{n+1}^2 - 2\int_t^{t_{n+1}} \int_{\mathbb{R}_+} (u_t + f_R'(u)u_x)u\,dx\,ds \\
&\le |u_-|_{n+1}^2 + \frac{1}{C}\|u_t + f_R'(u)u_x\|_n^2 + C\int_t^{t_{n+1}} \|u(s)\|_{\mathbb{R}_+}^2\,ds.
\end{aligned}$$

So that by Grönwall's inequality (1.24) and for $t_n < t < t_{n+1}$ we get

$$\|u(t)\|_{\mathbb{R}_+}^2 \le \left[\frac{1}{C}\|u_t + f_R'(u)u_x\|_n^2 + |u_-|_{n+1}^2\right]\exp(Ch).$$

Integrating over $t_n < t < t_{n+1}$ and summing over $n = 0, 1, \ldots, N-1$, and using a shifting as $n - 1 \hookrightarrow n$, we obtain the desired result. □

4.2 An a Priori Error Estimate for the Sd-Method

Let us prove an a priori error estimate for the Sd-method (3.6). To obtain global error estimate; we use the standard argument for finite element and introduce the linear nodal *interpolant* $\mathcal{I}_h u \in \mathcal{V}_h$ of the exact solution u and set $\eta = u - \mathcal{I}_h u$ and $\xi = u_h - \mathcal{I}_h u$. Then we have that

$$e := u - u_h = (u - \mathcal{I}_h u) - (u_h - \mathcal{I}_h u) = \eta - \xi.$$

Recalling the *Galerkin orthogonality relation* (3.7):

$$B(e, v) = 0, \qquad \forall v \in \mathcal{V}_h \tag{4.2}$$

We can now state and prove the basic global error estimate for (3.6).

Theorem 4.3. *If $u_h \in \mathcal{V}_h$ satisfies (3.6) and the exact solution u satisfies (3.1), and further*

$$||f_R'||_{\infty,\Omega} \leq C, \tag{4.3}$$

then there is a constant C such that

$$|||u - u_h||| \leq Ch^{k+\frac{1}{2}} \|u\|_{k+1,\Omega}. \tag{4.4}$$

Proof. Using the basic stability estimate (4.1) with $u = e$ and (4.2), with $v = \xi$ we

get that

$$\begin{aligned}
|||e|||^2 \leq & B(e,e) = B(e,\eta) - B(e,\xi) = B(e,\eta) \\
= & \Big(e_t + f_R'(e)e_x, \eta + \delta(\eta_t + f_R'(e)\eta_x)\Big)_\Omega + \sum_{n=0}^{N-1} < [e], \eta_+ >_n + \int_\Gamma e_+\eta_+ dt \\
\leq & C\frac{\delta}{8}\|e_t + f_R'(e)e_x\|_\Omega^2 + 8c\delta^{-1}\|\eta\|_\Omega^2 \\
& + C\frac{\delta}{8}\|e_t + f_R'(e)e_x\|_\Omega^2 + 8c\delta\|\eta_t + f_R'(\eta)\eta_x\|_\Omega^2 \\
& + C\frac{\delta}{8}\|e_t + f_R'(e)e_x\|_\Omega^2 + 8c\delta\|\eta_x(f_R'(e) - f_R'(\eta))\|_\Omega^2 \\
& + \frac{C}{4}\sum_{n=1}^{N-1}|[e]|_n^2 + \frac{1}{4c}\sum_{n=1}^{N-1}|\eta_+|_n^2 \\
& + \frac{C}{4}|e_+|_0^2 + \frac{1}{4c}|\eta_+|_0^2 + \frac{C}{4}\|e_+\|_\Gamma^2 + \frac{1}{4c}\|\eta_+\|_\Gamma^2 \\
\leq & \frac{1}{8}|||e|||^2 + C\sum_{n=0}^{N-1}|\eta_+|_n^2 \\
& + C\left[\|\eta_+\|_\Gamma^2 + h^{-1}\|\eta\|_\Omega^2 + h\|\eta_t + f_R'(\eta)\eta_x\|_\Omega^2 + h\|\eta_x(f_R'(e) - f_R'(\eta))\|_\Omega^2\right].
\end{aligned}$$

Further we have the triangle inequality

$$\|\eta_t + f_R'(\eta)\eta_x\|_\Omega \leq \|\eta_t\|_\Omega + \|f_R'\|_{\infty,\Omega}\|\eta_x\|_\Omega,$$

and

$$\|\eta_x(f_R'(e) - f_R'(\eta))\|_\Omega \leq 2\|f_R'\|_{\infty,\Omega}\|\eta_x\|_\Omega.$$

So that by (4.3), and also using the inverse estimate inequality (1.26)

$$\|\eta_x\|_\Omega \leq Ch^{-1}\|\eta\|_\Omega,$$

and the fact that $\delta = \bar{C}h$, we get,

$$|||e|||^2 \leq C\left[\|\eta_+\|_\Gamma^2 + h^{-1}\|\eta\|_\Omega^2 + \sum_{n=0}^{N-1}|\eta_+|_n^2 + h\|\eta\|_{1,\Omega}^2\right].$$

Finally, recalling interpolation estimates it follows that (see, e.g., Ciarlet[20])

$$\left[h\|\eta_+\|_\Gamma^2 + \|\eta\|_\Omega^2 + h\sum_{n=0}^{N-1} |\eta_+|_n^2 + h^2\|\eta\|_{1,\Omega}^2 \right]^{\frac{1}{2}} \leq Ch^{k+1}\|u\|_{k+1,\Omega},$$

which proves the desired estimates. □

The basic a priori estimate (4.4) for the Sd-method should be compared with the following estimate for the standard Galerkin method

$$\|u - u_h\|_\Omega \leq Ch^k\|u\|_{k+1,\Omega},$$

for the hyperbolic problems. which, being of order $\mathcal{O}(h^k)$, is obviously not-optimal. Note that the corresponding *optimal* convergence rate for the standard Galerkin method for elliptic and hyperbolic problems is of order $\mathcal{O}(h^{k+1})$. Hence (4.4) is an improvement of the convergence order by $\mathcal{O}(h^{1/2})$, for the hyperbolic problems. It appears that the order $\mathcal{O}(h^{k+1/2})$ given in (4.4) cannot be improved, and therefore is indeed the optimal convergence rate for the hyperbolic problems, (see, Ciarlet [20] or Johnson [47] for more details).

4.3 An a Posteriori Error Estimate for the Sd-Method

In the last section we estimated the error in the standard approach, i.e., we derived a priori error estimate which bound the error in term of stability properties of the numerical method and derivatives of the exact solution. This estimate is very important since it proves the convergence of the family of discrete solution towards the solution of the equation. However it involves the regularity of the exact solution which is most often unknown, so that it cannot provide an explicit order of convergence. In other words these kind of error bounds are not useful for practical error

estimations because the derivatives of the exact solution, likewise the exact solution itself, are unknowns. The a posteriori approach instead tries to estimate the error of approximating a particular solution by using the information from computation. The foundation is a rigorous a posteriori error estimate which bound the error by computable quantities that depend on the "known computed" numerical solution rather than the unknown exact solution. The key point is that, once the discrete solution is known, the a posteriori error bound can be computed explicitly. The process of a posteriori error estimates consists of the following four ingredients:

- "Stability factor" that measure the accumulation of error; the stability factor is determined by the solution of a dual problem, obtained by linearizing the weak formulation of the differential equation around the solution to be approximated.
- Interpolation constants C_i determined only by order of the finite element method, such as angles of elements and degree of quadratures used to compute the approximated solution.
- A power of the mesh size k and h
- A quantity that measures the residual error by regularity of the approximated solution. The residual error is the remainder from substituting the approximated solution into the (weak form of the) differential equation.

In fact the a posteriori error bound for the Sd-method for hyperbolic problem have the following basic form

$$\|e\| \leq C_i S_c \|h\mathcal{R}(u_h)\|, \tag{4.5}$$

where h is the mesh size , C_i is the interpolation constant ,$\mathcal{R}(u_h)$ is the residual and $S_c = S_c(u)$ is the computable stability factors defined through a continuous

linearized dual problem and expresses the relevant stability properties of the solution u being computed. The proof of the a posteriori error estimates (4.5) includes the following steps:

- Representation of the error e in term of the residual of the finite element solution and the solution ϕ of a continuous (linearized) dual problem, introduced below.
- Use of the Galerkin orthogonality to replace ϕ by $\phi - \Phi$, where Φ is the L_2-projection of ϕ.
- Interpolation error estimate for $\phi - \Phi$ in the term of certain derivatives $D\phi$ of ϕ and the mesh size h.
- Strong stability estimates for the dual solution ϕ, estimating $D\phi$ in term of the data ψ (here, in our case, ψ_i, $i = 1, 2$, as below) of the dual problem.

In this section we consider again the following model problem: Find u such that

$$\begin{aligned} \mathcal{L}u \equiv u_t + f_R(u)_x = 0, \qquad & \text{in } \Omega, \\ u(0,.) = u^b, \qquad & \text{in } (0,T), \\ u(.,0) = u_0, \qquad & \text{in } \mathbb{R}_+, \end{aligned} \tag{4.6}$$

where f_R and Ω are defined in the preceding section and for simplicity we take $u^b \equiv 0$. Below we shall consider the following simplified version of Sd-method (3.5) with $\delta = 0$ in the problem (4.6): Find $u_h \in \mathcal{V}_h^n$, such that for $n = 0, 1, \ldots, N-1$

$$\Big(u_{h,t} + f_R(u_h)_x, v\Big)_n + < [u_h], v_+ >_n = 0, \qquad \forall v \in \mathcal{W}_h^n, \tag{4.7}$$

where $[u_h] = u_{h,+}^n - u_{h,-}^n$ and $u_{h,-}^0 = u_0$. Our aim will be to derive a posteriori error estimates for the error $e = u - u_h$, in the $L_2(L_2)$ norm, see also [57, 18], where u and u_h are the solution of the (4.6) and (4.7) respectively. An error analysis for

the case of $\delta \neq 0$, is given in [56, 71]. Further we shall first introduce the following notation:

$$\|v\|_{L_2^{\psi}(\Omega)} = (v, \psi v)_{\Omega}^{1/2},$$

where ψ is a positive weight function and recall that

$$(v, w)_{\Omega} = \sum_{n=0}^{N-1} \int_{I_n} (v, w) dt,$$

where $\|v\|_{\Omega}^2 = (v, v)_{\Omega}$ and $(.,.)$ is the inner product in $L_2(\mathbb{R}_+)$. The a posteriori error estimates contain residual of computed solution u_h defined by

$$\begin{aligned}
\mathcal{R}_0 &= u_{h,t} + f_R(u_h)_x, \\
\mathcal{R}_1 &= \frac{u_{h,+}^n - u_{h,-}^n}{k_n}, \quad \text{on} \quad \mathbf{S}_n, \\
\mathcal{R}_2 &= \frac{(\mathcal{P}_n - I)u_{h,-}^n}{k_n}, \quad \text{on} \quad \mathbf{S}_n,
\end{aligned}$$

where I is the identity operator and $\mathcal{P}_n$ is a projection operator which will be defined below. As we mentioned above, standard a posteriori error estimates for time-dependent problem as presented in [44, 57, 69, 70] and [26] typically rely on *Galerkin orthogonality*, *interpolation estimates* and *strong stability estimates* for a suitable dual problem running backward in time with a desired error functional as initial data on the right hand side as key ingredient. Of course our approach to a posteriori error analysis is slightly different of the above reference; we furthermore use the concept of *stability factor* (see, [18, 68]). Below we, especially, consider these issues in more details.

4.3.1 The dual problem

In order to obtain a representation of the error, we consider the following auxiliary problem, referred to as the linearized dual problem: Find ϕ such that

$$\begin{aligned} \mathcal{L}^*\phi \equiv -\phi_t - A^T\phi_x &= \psi_1, \qquad \text{in } \Omega, \\ \phi(0,.) &= 0, \qquad \text{in } (0,T), \\ \phi(.,T) &= \psi_2, \qquad \text{in } \mathbb{R}_+, \end{aligned} \tag{4.8}$$

where

$$A = \int_0^1 f_R'(su + (1-s)u_h)ds.$$

and $\mathcal{L}^*$ denotes the adjoint of the operator $\mathcal{L}$ defined in (4.6). Note that this problem is computed "backward", but there is a corresponding change in sign. The a posteriori error estimates is expressed in terms of a scalar product with two functions ψ_1 or ψ_2 also appearing as data in the dual problem. Depending on the choice of ψ_1 or ψ_2, we will get estimates of different norms or functionals of the error.

4.3.2 $L_2(L_2)$ a posteriori error estimates

In this subsection we will obtain an a posteriori error estimate for the error $e = u - u_h$ in the $L_2(L_2)$ norm, where u and u_h are the solutions of (4.6) and (4.7) respectively. By choosing $\psi_1 = \psi^{-1}e$ and $\psi_2 = 0$ in (4.8) we get the following linearized dual problem: Find ϕ such that

$$\mathcal{L}^*\phi \equiv -\phi_t - A^T\phi_x = \psi^{-1}e, \qquad \text{in } \Omega, \tag{4.9a}$$

$$\phi(0,t) = 0, \qquad t \in (0,T), \tag{4.9b}$$

$$\phi(x,T) = 0, \qquad x \in R_+, \tag{4.9c}$$

where A^T is defined as above and ψ is a positive weight function to be specified below. Further we suppose that $\phi(x,t) \longrightarrow 0$ as $x \longrightarrow +\infty$. Multiplying (4.9) by e

and integrating over each $\mathbf{S}_n$, integrating by parts, and summing over n, we obtain the following error representation formula:

$$\begin{aligned}\|e\|^2_{L_2^{\psi^{-1}}(\Omega)} &= (e, \psi^{-1}e)_\Omega = (e, \mathcal{L}^*\phi) \\ &= \sum_{n=0}^{N-1}(e, -\phi_t - A^T\phi_x)_n \\ &= \sum_{n=0}^{N-1}(e_t, \phi)_n - \sum_{n=0}^{N-1}\int_{\mathbb{R}_+}\Big[e\phi\Big]_{t_n}^{t_{n+1}}dx \\ &+ \sum_{n=0}^{N-1}((Ae)_x, \phi)_n.\end{aligned} \tag{4.10}$$

We rewrite the second sum as

$$\begin{aligned}J &= \sum_{n=0}^{N-1}\int_{\mathbb{R}_+}\Big[e\phi\Big]_{t_n}^{t_{n+1}}dx \\ &= \big(< e_-, \phi_- >_1 - < e_+, \phi_+ >_0 \big) + \big(< e_-, \phi_- >_2 - < e_+, \phi_+ >_1 \big) + \dots \\ &+ \big(< e_-, \phi_- >_{N-1} - < e_+, \phi_+ >_{N-2} \big) + \big(< e_-, \phi_- >_N - < e_+, \phi_+ >_{N-1} \big).\end{aligned}$$

To continue we write $\phi_-^i = (\phi_-^i - \phi_+^i + \phi_+^i), i = 1, \dots, N-1$, then we obtain

$$-J = < e_-, \phi_- >_N + < e_+, \phi_+ >_0 + \sum_{n=1}^{N-1} < [e], \phi_+ >_n + \sum_{n=1}^{N-1} < e_-, [\phi] >_n .$$

According to (4.9c), $\phi(., t_N) = 0$ and since $e_-^0 = [u^0] = 0$, we get

$$J = \sum_{n=0}^{N-1} < [u_h], \phi_+ >_n .$$

Inserting J into the (4.10) we have that

$$\begin{aligned}\|e\|^2_{L_2^{\psi^{-1}}(\Omega)} &= \sum_{n=0}^{N-1}(e_t + (Ae)_x, \phi)_n - \sum_{n=0}^{N-1} < [u_h], \phi_+ >_n \\ &= \sum_{n=0}^{N-1}(u_t + f_R(u)_x, \phi)_n \\ &- \sum_{n=0}^{N-1}(u_{h,t} + f_R(u_h)_x, \phi)_n - \sum_{n=0}^{N-1} < [u_h], \phi_+ >_n .\end{aligned}$$

So that recalling (4.6) and using the Galerkin orthogonality we obtain

$$\begin{aligned} \|e\|^2_{L_2^{\psi^{-1}}(\Omega)} &= \sum_{n=0}^{N-1}(u_{h,t} + f_R(u_h)_x, \Phi - \phi)_n \\ &+ \sum_{n=0}^{N-1} < [u_h], (\Phi - \phi)_+ >_n \equiv I + II. \end{aligned} \tag{4.11}$$

where $\Phi \in V_h$ is an interpolant of ϕ. The idea is now to estimate $\Phi - \phi$ in terms of $\psi^{-1}e$ using a strong stability estimates for the solution ϕ of the dual problem (4.9).

4.3.3 Interpolation estimates for the dual solution

We shall now choose our interpolant $\Phi \in \mathcal{V}_h$ in (4.11) to be the space-time L_2-projection of ϕ, namely if we first define the L_2-projections:

$$\mathcal{P}_n : L_2(\mathbb{R}_+) \longrightarrow \mathcal{W}_h^n,$$

$$\pi_n : L_2(\mathbf{S}_n) \longrightarrow \Pi_{0,n} = \{v \in L_2(\mathbf{S}_n) : \ v(x,\cdot) \text{ is constant on } I_n, x \in \mathbb{R}_+\}$$

in space and in space-time, respectively, by

$$\begin{aligned} \int_{\mathbb{R}_+} (\mathcal{P}_n\phi)v dx &= \int_{\mathbb{R}_+} \phi v dx, \quad \forall v \in \mathcal{W}_h^n, \\ \pi_n v|_{\mathbf{S}_n} &= \frac{1}{k_n}\int_{I_n} v(\cdot,t)dt, \quad \forall v \in \Pi_{0,n}, \end{aligned}$$

then, we can define $\Phi\big|_{\mathbf{S}_n} \in \mathcal{V}_h^n$ by letting

$$\Phi\big|_{\mathbf{S}_n} = \mathcal{P}_n\pi_n\phi = \pi_n\mathcal{P}_n\phi \in V_h^n,$$

where $\phi = \phi\big|_{\mathbf{S}_n}$. Further, if we introduce $\mathcal{P}$ and π defined by

$$\begin{aligned} (\mathcal{P}\phi)\big|_{\mathbf{S}_n} &= \mathcal{P}_n(\phi\big|_{\mathbf{S}_n}), \quad \text{and} \\ (\pi\phi)\big|_{\mathbf{S}_n} &= \pi_n(\phi\big|_{\mathbf{S}_n}), \end{aligned}$$

respectively, then we can let $\Phi \in \mathcal{V}_h$ to be

$$\Phi = \mathcal{P}\pi\phi = \pi\mathcal{P}\phi \in \mathcal{V}_h.$$

In the end of this subsection, we shall state some interpolation estimates for the projection operators $\mathcal{P}$, leaving the overall estimates of I and II to next subsection.

Lemma 4.4. *There is a constant C such that for $\mathcal{R} \in L_2(\Omega)$,*

$$|(\mathcal{R}, \phi - \mathcal{P}\phi)_\Omega| \leq C||h^2(I-\mathcal{P})\mathcal{R})||_{L_2^{\psi^{-1}}(\Omega)}||\phi_{xx}||_{L_2^{\psi}(\Omega)}. \tag{4.12}$$

Proof. See [57, 68]. □

4.3.4 The completion of the proof of a posteriori error estimates

In this subsection we state and prove a posteriori error estimate by estimating of the terms I and II in the error representation formula (4.11). To this approach we introduce the stability factors (see [18]) associated with discretization in time and space, defined by

$$\begin{aligned} S_e^t &= \frac{\|\phi_t\|_{L_2^{\psi}(\Omega)}}{\|e\|_{L_2^{\psi^{-1}}(\Omega)}}, \quad \text{and} \\ S_e^x &= \frac{\|\phi_{xx}\|_{L_2^{\psi}(\Omega)}}{\|e\|_{L_2^{\psi^{-1}}(\Omega)}}, \end{aligned} \tag{4.13}$$

respectively. We now apply the results of the previous subsections; using Cauchy-Schwarz inequality in (4.11) coupled with the interpolation estimate (4.12) and the strong stability factors (4.13), to derive the $L_2(L_2)$ a posteriori error estimates for the scheme (4.7).

Theorem 4.5. *The error $e = u - u_h$, where u is the solution of the continuous problem (4.6) and u_h that of (4.7), satisfies the following stability estimate:*

$$\begin{aligned} \|e\|_{L_2^{\psi^{-1}}(\Omega)} \leq\ & CS_e^x||h^2(I-\mathcal{P})\mathcal{R}_0)||_{L_2^{\psi^{-1}}(\Omega)} + CS_e^t||k_n\mathcal{R}_1||_{L_2^{\psi^{-1}}(\Omega)} \\ & + S_e^x||h^2\mathcal{R}_2||_{L_2^{\psi^{-1}}(\Omega)} + S_e^t||k_n\mathcal{R}_2||_{L_2^{\psi^{-1}}(\Omega)}. \end{aligned} \tag{4.14}$$

Proof. Using the notation introduce above, we may write (4.11) as

$$\|e\|^2_{L_2^{\psi^{-1}}(\Omega)} = \sum_{n=0}^{N-1}(\mathcal{R}_0, \Phi - \phi)_n + \sum_{n=0}^{N-1} < k_n \frac{[u_h]}{k_n}, (\Phi - \phi)_+ >_n = I + II.$$

Below we shall estimate the terms I and II separately. splitting the interpolation error by writing $\Phi - \phi = \Phi - \mathcal{P}\phi + \mathcal{P}\phi - \phi$ and $\Phi_n = \pi_n \mathcal{P}\phi$, we have

$$\begin{aligned} I &= \sum_{n=0}^{N-1}(\mathcal{R}_0, \Phi_n - \mathcal{P}\phi + \mathcal{P}\phi - \phi)_n \\ &= \sum_{n=0}^{N-1}(\mathcal{R}_0, \Phi_n - \mathcal{P}\phi)_n + \sum_{n=0}^{N-1}(\mathcal{R}_0, \mathcal{P}\phi - \phi)_n \\ &\leq C||h^2(I-\mathcal{P})\mathcal{R}_0)||_{L_2^{\psi^{-1}}(\Omega)}||\phi_{xx}||_{L_2^{\psi}(\Omega)}, \end{aligned}$$

where we have used the fact that $\mathcal{R}_0$ is constant in time, (making the first integral zero) and then using interpolation estimate (4.12) in the second integral. It remains to estimate the term II, to this end, we need the following notation

$$\varphi_+^n(x) = \varphi(x,t) - \int_{t_n}^t \frac{\partial}{\partial \tau}\varphi(x,\tau)d\tau dt,$$

so that

$$k_n \phi_+^n(x) = \int_{I_n} \phi(x,t)dt - \int_{I_n}\int_{t_n}^t \phi_\tau(x,\tau)d\tau dt, \tag{4.15}$$

where $\phi_\tau = \frac{\partial \phi}{\partial \tau}$ and $\phi^n = \phi(.,t_n)$.

$$\begin{aligned} II &= \sum_{n=0}^{N-1} < k_n \frac{[u_h]}{k_n}, (\Phi - \phi)_+ >_n \\ &= \sum_{n=0}^{N-1} < k_n \frac{[u_h]}{k_n}, (\Phi_n - \mathcal{P}\phi + \mathcal{P}\phi - \phi)_+ >_n \\ &= \sum_{n=0}^{N-1} < k_n \frac{[u_h]}{k_n}, (\Phi_n - \mathcal{P}\phi)_+ >_n + \sum_{n=0}^{N-1} < k_n \frac{[u_h]}{k_n}, (\mathcal{P}\phi - \phi)_+ >_n \\ &:= II_1 + II_2. \end{aligned}$$

To estimate II_1, we use (4.15) to get

$$
\begin{aligned}
II_1 &= \sum_{n=0}^{N-1} < k_n\mathcal{R}_1, (\Phi_n)_+ - \mathcal{P}\phi_+ >_n \\
&= \sum_{n=0}^{N-1} < \mathcal{R}_1, k_n\Phi_n - \mathcal{P}k_n\phi_+ >_n \\
&= \sum_{n=0}^{N-1} \Big(< \mathcal{R}_1, k_n\Phi_n - \int_{I_n} \mathcal{P}\phi(\cdot,t)dt + \int_{I_n}\int_{t_n}^{t} \mathcal{P}\phi_\tau(\cdot,\tau)d\tau dt >_n \Big) \\
&= \sum_{n=0}^{N-1} \int_{I_n}\int_{t_n}^{t} < \mathcal{R}_1, \mathcal{P}\phi_\tau(\cdot,\tau) >_n d\tau dt \\
&\le \|k_n\mathcal{R}_1\|_{L_2^{\psi^{-1}}(\Omega)} \|\mathcal{P}\phi_t\|_{L_2^{\psi}(\Omega)} \le \|k_n\mathcal{R}_1\|_{L_2^{\psi^{-1}}(\Omega)} \|\phi_t\|_{L_2^{\psi}(\Omega)}.
\end{aligned}
$$

As for the II_2-term we can write

$$
\begin{aligned}
II_2 &= \sum_{n=0}^{N-1} < k_n \frac{[u_h]}{k_n}, (\mathcal{P}\phi - \phi)_+ >_n \\
&= \sum_{n=0}^{N-1} < \frac{u_{h,+}^n - u_{h,-}^n}{k_n}, (\mathcal{P}_n - I)k_n\phi_+ >_n \\
&= \sum_{n=0}^{N-1} < \frac{\mathcal{P}_n u_{h,-}^n - u_{h,-}^n}{k_n}, (\mathcal{P}_n - I)\Big(\int_{I_n} \phi(\cdot,t)dt - \int_{I_n}\int_{t_n}^{t} \phi_\tau(\cdot,\tau)d\tau dt \Big) >_n \\
&\le \sum_{n=0}^{N-1} \int_{I_n} < \frac{(\mathcal{P}_n - I)u_{h,-}^n}{k_n}, (\mathcal{P}_n - I)\phi(\cdot,t) >_n dt \\
&+ \sum_{n=0}^{N-1} \int_{I_n}\int_{t_n}^{t} < \frac{(\mathcal{P}_n - I)u_{h,-}^n}{k_n}, (\mathcal{P}_n - I)\phi_\tau(\cdot,t) >_n d\tau dt \\
&\le \|k_n\mathcal{R}_2\|_{L_2^{\psi^{-1}}(\Omega)} \|\phi_{xx}\|_{L_2^{\psi}(\Omega)} + \|k_n\mathcal{R}_2\|_{L_2^{\psi^{-1}}(\Omega)} \|\phi_t\|_{L_2^{\psi}(\Omega)}.
\end{aligned}
$$

The a posteriori estimate now follows immediately after collecting the terms and using the definition of the stability factors (4.13). □

4.3.5 Strong stability estimates

To give the above a posteriori estimate a quantitative meaning we need to estimate the strong stability factors. In this subsection we will present some stability estimates for the model problem indicating that the corresponding stability factor is bounded by a moderate constant. More specifically, to get a quantitative estimate of the computational error $\|u - u_h\|$ in terms of S_e^t and S_e^x, it is necessary that these two quantities are bounded by some moderate constants. We now prove the following strong stability estimate for the dual problem (4.9).

Lemma 4.6. *For a positive weight function $\psi(x,t)$, the solution ϕ of the dual problem(4.9) satisfies the following estimate:*

$$\|\psi^{1/2}(\phi_t + A^T\phi_x)\|_\Omega = \|e\|_{L_2^{\psi^{-1}}(\Omega)}.$$

Proof. First, multiplying the equation (4.9a) by $-\psi(\phi_t + A^T\phi_x)$ and integrating over Ω we get

$$\begin{aligned}\int_\Omega \psi(\phi_t + A^T\phi_x)^2 dxdt &= -\int_\Omega e(\phi_t + A^T\phi_x)dxdt \\ &\le \frac{1}{2}\|\psi^{-1/2}e\|_\Omega^2 + \frac{1}{2}\|\psi^{1/2}(\phi_t + A^T\phi_x)\|_\Omega^2.\end{aligned}$$

which gives the inequality

$$\|\psi^{1/2}(\phi_t + A^T\phi_x)\|_\Omega^2 \le \|\psi^{-1/2}e\|_\Omega^2. \tag{4.16}$$

Similarly, multiplying (4.9a) by e and integrating over Ω we get

$$\begin{aligned}\int_\Omega e^2\psi^{-1} = \|\psi^{-1/2}e\|_\Omega^2 &= -\int_\Omega \psi^{1/2}(\phi_t + A^T\phi_x) \\ &\le \frac{1}{2}\|\psi^{-1/2}e\|_\Omega^2 + \frac{1}{2}\|\psi^{1/2}(\phi_t + A^T\phi_x)\|_\Omega^2.\end{aligned}$$

So that we have

$$\|\psi^{-1/2}e\|_\Omega^2 \le \|\psi^{1/2}(\phi_t + A^T\phi_x)\|_\Omega^2. \tag{4.17}$$

Then combining (4.16) and (4.17) we get the desired result. □

Also we have the following lemma:

Lemma 4.7. *There is constant $C = \sqrt{T}e^T$ such that if $\psi(x,t)$ is a positive weight function satisfying*

$$\psi_t + A^T\psi_x + A_x^T\psi \geq -\psi, \quad \text{in } \Omega, \tag{4.18}$$

then the solution ϕ of (4.9) satisfies

$$\|\psi^{1/2}\phi\|_\Omega \leq C\|e\|_{L_2^{\psi^{-1}}(\Omega)}.$$

Proof. Multiplying (4.9a) by $\psi\phi$ and integrating over $\mathbb{R}_+$ we get,

$$-(\phi_t, \psi\phi(t)) - (A^T\phi_x, \psi\phi(t)) = (e, \phi(t)).$$

Integration by part yields

$$-\frac{1}{2}\frac{d}{dt}\|\psi^{1/2}\phi(t)\|^2 + \frac{1}{2}(\psi_t, \phi^2(t)) - (A^T\phi_x, \psi\phi(t)) = (e, \phi(t)).$$

Integrating by part in space, using (4.9b) and the Cauchy-Schwarz inequality, we have that

$$\begin{aligned}-\frac{1}{2}\frac{d}{dt}\|\psi^{1/2}\phi(t)\|^2 + \frac{1}{2}(\psi_t + A^T\psi_x + A_x^T\psi, \phi^2(t)) &\leq \|\psi^{-1/2}e\|^2\|\psi^{1/2}\phi\| \\ &\leq \frac{1}{2}\|\psi^{-1/2}e\|^2 + \frac{1}{2}\|\psi^{1/2}\phi\|^2.\end{aligned}$$

Now, integrating with respect to the time variable over the interval (t,T), using (4.9c) and assumption (4.18) we get

$$\|\psi^{1/2}\phi(t)\|^2 \leq \|\psi^{-1/2}e\|_\Omega^2 + 2\int_t^T \|\psi^{1/2}\phi(t)\|^2 ds.$$

Furthermore, by the Grönwall's inequality, we have

$$\|\psi^{1/2}\phi(t)\|^2 \leq e^{2T}\|\psi^{-1/2}e\|_\Omega^2.$$

Finally, integrating with respect to time over $(0,T)$ we obtain the desired result. □

Chapter 5

A POSTERIORI ERROR ESTIMATE IN THE CASE OF ADDING ARTIFICIAL VISCOSITY

5.1 Introduction

In this chapter we continue in the framework of the Sd-method for hyperbolic problems (3.1), and add an extra artificial diffusion term of the form $-\varepsilon u_{xx}$, to our model problem, i.e., we study the following problem: Find u such that

$$\begin{aligned} \mathcal{L}_\varepsilon u \equiv u_t + f_R(u)_x - \varepsilon u_{xx} = 0, \qquad & \text{in } \Omega, \\ u(0,t) = u^b, \qquad & t \in (0,T), \\ u(x,0) = u_0, \qquad & x \in \mathbb{R}_+, \end{aligned} \tag{5.1}$$

where ε is a positive viscosity coefficient, which we typically assume to be a 'small' constant specified below. Our main goal is to derive a posteriori error estimates for the above problem based on the Sd-method. In the formulation of the Sd-method

below, if the mesh size is not small enough, the given viscosity ε will be replaced by an artificial viscosity $\hat{\varepsilon}$ depending on the computed solution u_h and the mesh size h. Therefore it is convenient and natural to split the total error $e = u - u_h$ into

$$e = u - u_h = (u - \hat{u}) + (\hat{u} - u_h) = (u - \hat{u}) + \hat{e},$$

where $\hat{u}$ satisfies (5.1) with ε replaced by $\hat{\varepsilon}$, i.e., $\hat{u}$ is the solution of the following continuous problem

$$\begin{aligned} \mathcal{L}_{\hat{\varepsilon}}\hat{u} \equiv \hat{u}_t + f_R(\hat{u})_x - \hat{\varepsilon}\hat{u}_{xx} = 0, \qquad & \text{in } \Omega, \\ \hat{u}(0,t) = u^b, \qquad & t \in (0,T), \\ \hat{u}(x,0) = u_0, \qquad & x \in \mathbb{R}_+. \end{aligned} \tag{5.2}$$

Now, the perturbation error, i.e., $u - \hat{u}$ is the difference between the solutions of two continuous problems with different viscosity ε and $\hat{\varepsilon}$, and $\hat{u} - u_h$ is the discretization error related to (5.2) with now $\hat{\varepsilon}$ considered to be given. In this case, the a posteriori error estimates (also underlying adaptive algorithm) is, in the case of space discretization only, basically as follows:

$$\|\hat{e}\|_A \leq SC_i \|\frac{h^2}{\hat{\varepsilon}}\mathcal{R}(u_h)\|_B, \tag{5.3}$$

where $\|.\|_A$ and $\|.\|_B$ are some norms, e.g., an $L_p(L_q)$ in time-space or $L_\infty(H^{-1})$ norm, h is the mesh size, and the residual $\mathcal{R}(u_h) = \mathcal{L}_{\hat{\varepsilon}}u_h$ is obtained inserting the computed finite element solution into the differential equation (5.2). Further C_i is an interpolation constant depending only on the angels of the elements of the computational mesh, and most importantly S is a stability factor which measures certain stability properties of an associated continuous linearized dual problem. An adaptive algorithm based on the a posteriori error estimate (5.3) seeks to realize the stopping criterion

$$SC_i \|\frac{h^2}{\hat{\varepsilon}}\mathcal{R}(u_h)\| \leq TOL, \tag{5.4}$$

at minimal computational cost, where TOL is a given tolerance.

5.2 The Sd-Method

The Sd-method for (5.2) can now be formulated as follows: Find $u_h^n \in \mathcal{V}_h^n$, such that for $n = 0, 1, \ldots, N-1$

$$\begin{aligned} \Big(u_{h,t}^n + f_R'(u_h^n)u_{h,x}^n, v_h^n + \delta(v_{h,t}^n + f_R'(u_h^n)v_{h,x}^n)\Big)_n + \Big(\hat{\varepsilon}u_{h,x}^n, v_{h,x}^n\Big)_n \\ + < [u_h], v_{h,+}^n >_n + \int_{\Gamma_n} u_{h,+}^n v_{h,+}^n dt = \int_{\Gamma_n} u^b v_{h,+}^n dt, \qquad \forall v_h^n \in \mathcal{W}_h^n, \end{aligned} \tag{5.5}$$

where $u_{h,-}^0 = u_0$, $[u_h] = u_{h,+}^n - u_{h,-}^n$,

$$\begin{aligned} \hat{\varepsilon} &= max(\varepsilon, C_1 h|\mathcal{R}(u_h)|/|\nabla u_h|, C_2 h^{3/2}), \\ \mathcal{R}(u_h) &= |u_{h,t} + f(u_h)_x| + \frac{|[u_h]|}{k_n}, \qquad \text{on } \mathbf{S}_n, \\ \delta &= \bar{C}h, \end{aligned}$$

with h denoting the mesh size, and C_i is a positive constant. The artificial viscosity $\hat{\varepsilon}$ acts in an implicit way to provide additional stability near to shocks ($\hat{\varepsilon} \sim C_1 h$), and less in smooth regions ($\hat{\varepsilon} \sim C_2 h^{3/2}$). In order to estimate the error $e = u - u_h$, we also need to estimate $u - \hat{u}$. To control the $u - \hat{u}$, we may adaptively refine the mesh until $\hat{\varepsilon} = \varepsilon$, giving $u = \hat{u}$, or alternatively approximate $\hat{e}$ in terms of $\varepsilon - \hat{\varepsilon}$. To approximately minimize the total number of degrees of freedom of a mesh with mesh size h satisfying (5.4), typically a simple iterative procedure is used where a new mesh-size is computed by 'equi-distribution' of element, (see [31] for parabolic and elliptic equations) contributions to the quantity $SC_i\|\frac{h^2}{\hat{\varepsilon}}\mathcal{R}(u_h)\|$ with the values of $\hat{\varepsilon}$ and $\mathcal{R}(u_h)$ taken from the previous mesh. In the a posteriori error estimates for the Sd-method (5.5), below for simplicity we assume that $u^b \equiv 0$, $\hat{\varepsilon} = \varepsilon$ is constant and that the function $h(x,t)$ is constant $h(x,t) = h$, for all x, t. Further we shall consider the following simplified version of the Sd-method with $\hat{\varepsilon} = \bar{C}h$ and $\delta = 0$: Find $u_h \in \mathcal{V}_h^n$, such that for $n = 0, 1, \ldots, N-1$

$$\Big(u_{h,t} + f_R(u_h)_x, v\Big)_n + \Big(\hat{\varepsilon}u_{h,x}, v_x\Big)_n + < [u_h], v_+ >_n = 0, \qquad \forall v \in \mathcal{W}_h^n, \tag{5.6}$$

where $[u_h] = u_{h,+} - u_{h,-}$ and $u^0_{h,-} = u_0$.

5.3 $L_2(L_2)$ a Posteriori Error Estimates

Let us now prove a basic a posteriori error estimate in the $L_2(L_2)-$norm. To prove this, we introduce the following linearized dual problem: Find φ such that

$$\mathcal{L}^*_{\hat{\varepsilon}}\varphi \equiv -\varphi_t - A^T\varphi_x - \hat{\varepsilon}\varphi_{xx} \quad = \psi^{-1}\hat{e}, \qquad \text{in } \Omega, \tag{5.7a}$$

$$\varphi(0,t) \quad = 0, \qquad t \in (0,T), \tag{5.7b}$$

$$\varphi(x,T) \quad = 0, \qquad x \in \mathbb{R}_+, \tag{5.7c}$$

where

$$A = \int_0^1 f'_R(su + (1-s)u_h)ds,$$

$\mathcal{L}^*_{\hat{\varepsilon}}$ denotes the adjoint of $\mathcal{L}_{\hat{\varepsilon}}$ defined in (5.2) and ψ is a positive weight function to be specified below. Further we suppose that $\varphi(x,t) \longrightarrow 0$ as $x \longrightarrow +\infty$. The error representation is now obtained multiplying (5.7a) by e and integrating in both space and time, together with integration by parts and using similar arguments as in (4.10):

$$\begin{aligned}
\|e\|^2_{L_2^{\psi^{-1}}(\Omega)} &= (e, \psi^{-1}e)_\Omega = (e, \mathcal{L}^*_{\hat{\varepsilon}}\varphi) \\
&= \sum_{n=0}^{N-1}(e, -\varphi_t - A^T\varphi_x - (\hat{\varepsilon}\varphi)_{xx})_n \\
&= \sum_{n=0}^{N-1}\{(e_t + (Ae)_x, \varphi)_n + (\hat{\varepsilon}e_x, \varphi_x)_n\} - \sum_{n=0}^{N-1} < [u_h], \varphi_+ >_n \\
&= \sum_{n=0}^{N-1}(u_t + f_R(u)_x - (\hat{\varepsilon}u_x)_x), \varphi)_n \\
&\quad - \sum_{n=0}^{N-1}\{(u_{h,t} + f_R(u_h)_x, \varphi)_n + (\hat{\varepsilon}u_{h,x}, \varphi_x)_n\} - \sum_{n=0}^{N-1} < [u_h], \varphi_+ >_n .
\end{aligned}$$

So that recalling (5.2) and using the Galerkin orthogonality with $\Psi \in V_h$ we have

$$\begin{aligned}\|e\|^2_{L_2^{\psi^{-1}}(\Omega)} &= \sum_{n=0}^{N-1}(u_{h,t} + f_R(u_h)_x, \Psi - \varphi)_n + \sum_{n=0}^{N-1}(\hat{\varepsilon}u_{h,x}, (\Psi-\varphi)_x)_n \\ &+ \sum_{n=0}^{N-1} < [u_h], (\Psi-\varphi)_+ >_n \equiv I + II + III,\end{aligned} \tag{5.8}$$

where $\Psi = \pi\mathcal{P}\varphi$ (see section 5.3) is assumed to be a suitable interpolant of φ. As indicated earlier, the idea is now to estimate $\Psi - \varphi$ in terms of $\psi^{-1}e$ using a strong stability estimates for the solution φ of the dual problem. To proceed we will use the following interpolation estimates, similar to Lemma 5.1:

Lemma 5.1. *There is a constant C_i such that for $\mathcal{R} \in L_2(\Omega)$*

$$|(\mathcal{R}, \varphi - \mathcal{P}\varphi)_\Omega| \le C_i||\frac{h^2}{\varepsilon}(I-\mathcal{P})\mathcal{R})||_{L_2^{\psi^{-1}}(\Omega)}||\hat{\varepsilon}\varphi_{xx}||_{L_2^{\psi}(\Omega)}, \tag{5.9a}$$

$$|(\hat{\varepsilon}u_{h,x}, (\varphi - \mathcal{P}\varphi)_x)_\Omega| \le C_i||\frac{h^2}{\hat{\varepsilon}}D_1u_h||_{L_2^{\psi^{-1}}(\Omega)}||\hat{\varepsilon}\varphi_{xx}||_{L_2^{\psi}(\Omega)}, \tag{5.9b}$$

where

$$D_1u_h(x,t) = \max\Big|\frac{[u_{h,x}(x_j^n,t)]}{h_j^n}\Big|, \qquad x \in (x_{i-1}^n, x_i^n), \quad t \in I_n,$$

and

$$[v(x_j^n)] = \lim_{\zeta \longrightarrow 0^+}(v(x_j^n + \zeta) - v(x_j^n - \zeta)).$$

Proof. See [57, 68]. □

Similarly we have also the following Lemma:

Lemma 5.2. *[44] There is a constant C_i such that for $\mathcal{R} \in L_2^{\psi^{-1}}(\Omega)$,*

$$|(\mathcal{R}, \mathcal{P}(\pi\varphi - \varphi))_\Omega| \le C_i||k_n\mathcal{R}||_{L_2^{\psi^{-1}}(\Omega)}||\varphi_t||_{L_2^{\psi}(\Omega)}. \tag{5.10}$$

Proof. Using the Cauchy-Schwarz inequality, we get

$$|(\mathcal{R}, \mathcal{P}(\pi\varphi - \varphi))_\Omega| \le \|k\mathcal{R}\|_{L_2^{\psi^{-1}}(\Omega)}\|k^{-1}(\pi(\mathcal{P}\varphi) - (\mathcal{P}\varphi))\|_{L_2^{\psi}(\Omega)}.$$

By reversing the order of integration, we have

$$\begin{aligned}\|k^{-1}(\pi(\mathcal{P}\varphi)-(\mathcal{P}\varphi))\|_{L_2^{\psi}(\Omega)} &\\ &= \Big(\sum_{n=0}^{N-1}\int_{t_n}^{t_{n+1}}\int_{\Omega} k_n^{-2}\psi^2(\pi_n(\mathcal{P}_n\varphi)-(\mathcal{P}_n\varphi))^2 dxdt\Big)^{1/2}\\ &= \Big(\int_{\Omega}\sum_{n=0}^{N-1} k_n^{-2}\|\pi_n(\mathcal{P}_n\varphi)-(\mathcal{P}_n\varphi)\|^2_{L_2^{\psi}(I_n)}dxdt\Big)^{1/2}.\end{aligned} \tag{5.11}$$

If we denote the piecewise constant interpolant of v on I_n at the point $(t_n+t_{n+1})/2$ by $\mathcal{I}_I^n v$, then using the fact that $\pi_n v$ is L_2-projection of v onto the set of piecewise constant functions, we have

$$\begin{aligned}|\pi_n(\mathcal{P}_n\varphi)-(\mathcal{P}_n\varphi)\|^2_{L_2^{\psi}(I_n)} &\\ &\le |\mathcal{I}_I^n(\mathcal{P}_n\varphi)-(\mathcal{P}_n\varphi)\|^2_{L_2^{\psi}(I_n)}.\end{aligned} \tag{5.12}$$

Further, it can be easily proved that (see cf. chapter 1, section 1.6.1)

$$\|\mathcal{I}_I^n(\mathcal{P}_n\varphi)-(\mathcal{P}_n\varphi)\|^2_{L_2^{\psi}(I_n)} \le C_i k_n\|(\mathcal{P}_n\varphi)\|_{L_2^{\psi}(I_n)}. \tag{5.13}$$

Using (5.12) and (5.13), (5.11) becomes

$$\begin{aligned}\|k^{-1}(\pi(\mathcal{P}\varphi)-(\mathcal{P}\varphi))\|_{L_2^{\psi}(\Omega)} &\\ &\le \Big(\int_{\Omega}\sum_{n=0}^{N-1} C_i^2\|(\mathcal{P}_n\varphi)_t\|^2_{L_2^{\psi}(I_n)}dx\Big)^{1/2}\\ &\le \Big(\int_{\Omega}\sum_{n=0}^{N-1} C_i^2\|\mathcal{P}_n\varphi_t\|^2_{L_2^{\psi}(I_n)}dx\Big)^{1/2}.\end{aligned} \tag{5.14}$$

If we now reverse the order of integration in (5.14) and use that

$$\|\mathcal{P}_n\varphi_t\|_{L_2(\Omega)} \le \|\varphi_t\|_{L_2(\Omega)},$$

then this proves the Lemma. □

We shall now proceed to estimate the terms I-III in the error representation formula (5.8), using the interpolation estimates in the Lemmas 5.1 and 5.2. We shall below use the "discrete second derivative" $D_2 : \mathcal{W}_h^n \longrightarrow \mathcal{W}_h^n$ defined by

$$(D_2 w, v)_n = -(w_x, v_x)_n, \quad \forall v \in \mathcal{W}_h^n,$$
$$(D_2^\varepsilon w, v)_n = -(\varepsilon w_x, v_x)_n, \quad \forall v \in \mathcal{W}_h^n.$$

For the first term we have

$$\begin{aligned} I &= \sum_{n=0}^{N-1} (u_{h,t} + f_R(u_h)_x, \Psi - \varphi)_n \\ &= \sum_{n=0}^{N-1} (\mathcal{R}_0, \Psi_n - \mathcal{P}\varphi + \mathcal{P}\varphi - \varphi)_n \\ &+ \sum_{n=0}^{N-1} (\mathcal{R}_0, \mathcal{P}\varphi - \phi)_n + \sum_{n=0}^{N-1} (\mathcal{R}_0, \mathcal{P}(\pi\varphi - \varphi))_n \equiv I_1 + I_2, \end{aligned}$$

where $\mathcal{R}_0 = u_{h,t} + f_R(u_h)_x$ and Ψ_n, $\mathcal{P}$ and π are as defined in section 5.3. By (5.9a) we have

$$|I_1| \leq C_i || \frac{h^2}{\hat{\varepsilon}} (I - \mathcal{P})\mathcal{R}_0) ||_{L_2^{\psi^{-1}}(\Omega)} || \hat{\varepsilon} \varphi_{xx} ||_{L_2^{\psi}(\Omega)},$$

similarly, using (5.10), we have

$$|I_2| \leq C_i || k_n \mathcal{R}_0 ||_{L_2^{\psi^{-1}}(\Omega)} || \varphi_t ||_{L_2^{\psi}(\Omega)}.$$

Hence,

$$|I| \leq C_i \Big(|| \frac{h^2}{\hat{\varepsilon}} (I - \mathcal{P})\mathcal{R}_0) ||_{L_2^{\psi^{-1}}(\Omega)} || \hat{\varepsilon} \varphi_{xx} ||_{L_2^{\psi}(\Omega)} + || k_n \mathcal{R}_0 ||_{L_2^{\psi^{-1}}(\Omega)} || \varphi_t ||_{L_2^{\psi}(\Omega)} \Big).$$

Next,

$$\begin{aligned} II &= \sum_{n=0}^{N-1} (\hat{\varepsilon} u_{h,x}, (\Psi - \varphi)_x)_n \\ &= \sum_{n=0}^{N-1} (\hat{\varepsilon} u_{h,x}, (\Psi_n - \mathcal{P}\varphi + \mathcal{P}\varphi - \varphi)_x)_n \\ &= \sum_{n=0}^{N-1} (\hat{\varepsilon} u_{h,x}, (\mathcal{P}\varphi - \varphi)_x)_n + \sum_{n=0}^{N-1} (\hat{\varepsilon} u_{h,x}, (\Psi_n - \mathcal{P}\varphi)_x)_n \\ &= \sum_{n=0}^{N-1} (\hat{\varepsilon} u_{h,x}, (\mathcal{P}\varphi - \varphi)_x)_n + \sum_{n=0}^{N-1} (-D_2^{\varepsilon} u_h, \mathcal{P}(\pi\varphi - \varphi))_n \equiv II_1 + II_2. \end{aligned}$$

By (5.9b) we have

$$|II_1| \leq C_i || \frac{h^2}{\hat{\varepsilon}} D_1 u_h ||_{L_2^{\psi^{-1}}(\Omega)} || \hat{\varepsilon} \varphi_{xx} ||_{L_2^{\psi}(\Omega)},$$

and by (5.10)

$$|II_2| \leq C_i || k_n D_2^{\varepsilon} u_h ||_{L_2^{\psi^{-1}}(\Omega)} || \varphi_t ||_{L_2^{\psi}(\Omega)}.$$

Thus we have that

$$|II| \leq C_i \Big(|| \frac{h^2}{\hat{\varepsilon}} D_1 u_h ||_{L_2^{\psi^{-1}}(\Omega)} || \hat{\varepsilon} \varphi_{xx} ||_{L_2^{\psi}(\Omega)} + || k_n D_2^{\varepsilon} u_h ||_{L_2^{\psi^{-1}}(\Omega)} || \varphi_t ||_{L_2^{\psi}(\Omega)} \Big).$$

Finally, for the third term III in the error representation we have

$$\begin{aligned} III &= \sum_{n=0}^{N-1} < [u_h], (\Psi - \varphi)_+ >_n \\ &= \sum_{n=0}^{N-1} < [u_h], (\mathcal{P}\varphi - \varphi)_+ >_n + \sum_{n=0}^{N-1} < [u_h], (\pi\mathcal{P}\varphi - \mathcal{P}\varphi)_+ >_n \\ &\equiv III_1 + III_2. \end{aligned}$$

Considering first III_1, we have with the L_2-projection $\mathcal{P}_n$ defined in previous chap-

ter (section 5.3)

$$\begin{aligned} III_1 &= \sum_{n=0}^{N-1} < u_{h,+} - u_{h,-}, (\mathcal{P}_n - I)\varphi_+ >_n \\ &= \sum_{n=0}^{N-1} < \mathcal{P}_n u_{h,-} - u_{h,-}, (\mathcal{P}_n - I)\varphi_+ >_n \\ &= \sum_{n=0}^{N-1} < \mathcal{R}_2(t_n)_+, k_n(\mathcal{P}_n - I)\varphi_+ >_n \end{aligned} \tag{5.15}$$

where we recall that we have defined

$$\mathcal{R}_2 \equiv \frac{(\mathcal{P}_n - I)u_{h,-}^n}{k_n}, \quad \text{on } \mathbf{S}_n.$$

Now to estimate $(I - \mathcal{P}_n)\varphi_+^n$, similarly as in (4.15) we note that

$$k_n\varphi_+^n(x) = \int_{I_n} \varphi(x,t) - \int_{I_n}\int_{t_n}^t \varphi_\tau(x,\tau)d\tau dt. \tag{5.16}$$

Inserting this representation into the right hand side of (5.15), using an estimate for $(\mathcal{P}_n - I)$ analogous to (5.9a), we get

$$\begin{aligned} |III_1| &= \Big| \sum_{n=0}^{N-1} < \mathcal{R}_2, (\mathcal{P}_n - I)\Big(\int_{I_n} \varphi(x,t)dt - \int_{I_n}\int_{t_n}^t \varphi_\tau(x,\tau)d\tau dt\Big) >_n \Big| \\ &= \Big| \sum_{n=0}^{N-1} \int_{I_n} \Big\{ < \mathcal{R}_2, (\mathcal{P}_n - I)\varphi >_n - < \mathcal{R}_2, (\mathcal{P}_n - I)\int_{t_n}^t \varphi_\tau(\cdot,\tau)d\tau >_n \Big\} dt \Big| \\ &= \Big| \sum_{n=0}^{N-1} \int_{I_n} \Big\{ < \mathcal{R}_2, (\mathcal{P}_n - I)\varphi >_n - \int_{t_n}^t < \mathcal{R}_2, (\mathcal{P}_n - I)\varphi_\tau(\cdot,\tau) >_n d\tau \Big\} dt \Big| \\ &\le \sum_{n=0}^{N-1} \Big\{ \int_{I_n} \Big| \frac{h^2}{\hat{\varepsilon}}\mathcal{R}_2\psi^{-1}\Big|_n \Big| \hat{\varepsilon}\psi\varphi_{xx}\Big|_n dt + \int_{I_n}\int_{I_n} \Big|\psi^{-1}\mathcal{R}_2\Big|_n \Big|(\mathcal{P}_n - I)\varphi_\tau\psi\Big|_n d\tau dt \Big\} \\ &\le C\Big(\|\frac{h^2}{\hat{\varepsilon}}\mathcal{R}_2\|_{L_2^{\psi^{-1}}(\Omega)} \|\hat{\varepsilon}\varphi_{xx}\|_{L_2^{\psi}(\Omega)} + \|k_n\mathcal{R}_2\|_{L_2^{\psi^{-1}}(\Omega)} \|\varphi_t\|_{L_2^{\psi}(\Omega)} \Big), \end{aligned}$$

where in the last inequality we have used that

$$\|\mathcal{P}_n\varphi_t\|_{L_2^{\psi}(\Omega)} \le \|\varphi_t\|_{L_2^{\psi}(\Omega)}.$$

Finally, for III_2 we have with $\mathcal{R}_1 \equiv (u_{h,+} - u_{h,-})/k_n)$ on $\mathbf{S}_n$,

$$
\begin{aligned}
|III_2| &= \Big| \sum_{n=0}^{N-1} < k_n \frac{[u_h]}{k_n}, (\pi\mathcal{P}\varphi - \mathcal{P}\varphi)_+ >_n \Big| \\
&= \Big| \sum_{n=0}^{N-1} < k_n\mathcal{R}_1, (\pi\mathcal{P}\varphi - \mathcal{P}\varphi)_+ >_n \Big| \\
&= \Big| \sum_{n=0}^{N-1} < \mathcal{R}_1, (\pi\mathcal{P} - \mathcal{P})k_n\varphi_+ >_n \Big| \\
&= \Big| \sum_{n=0}^{N-1} < \mathcal{R}_1, (\pi\mathcal{P} - \mathcal{P})\Big(\int_{I_n} \varphi(x,t) - \int_{I_n}\int_{t_n}^{t} \varphi_\tau(x,\tau)d\tau dt) >_n \Big)\Big| \\
&= \Big| \sum_{n=0}^{N-1} \int_{I_n} \Big\{ < \mathcal{R}_1, \mathcal{P}(\pi\varphi - \varphi) >_n - \int_{t_n}^{t} < \mathcal{R}_2, (\pi\mathcal{P} - \mathcal{P})\varphi_\tau(\cdot,\tau) >_n d\tau \Big\} dt \Big| \\
&\leq C\|k_n\mathcal{R}_1\|_{L_2^{\psi^{-1}}(\Omega)} \|\varphi_t\|_{L_2^{\psi}(\Omega)}.
\end{aligned}
$$

We have now proved the following a posteriori error estimate:

Theorem 5.3. *Suppose that the assumptions in Lemmas 5.1 and 5.2 are fulfilled. Let u_h and $\hat{u}$ be the solutions of (5.6) and (5.2) respectively. Then, there is a constant C such that:*

$$
\begin{aligned}
\|\hat{u} - u_h\|_{L_2^{\psi^{-1}}(\Omega)} \leq\ & CS_{\hat{e}}^x || \frac{h^2}{\hat{\varepsilon}}(I - \mathcal{P})\mathcal{R}_0) ||_{L_2^{\psi^{-1}}(\Omega)} + CS_{\hat{e}}^t || k_n\mathcal{R}_0 ||_{L_2^{\psi^{-1}}(\Omega)} \\
&+ CS_{\hat{e}}^x || \frac{h^2}{\hat{\varepsilon}} D_1 u_h ||_{L_2^{\psi^{-1}}(\Omega)} + CS_{\hat{e}}^t || k_n D_2^{\varepsilon} u_h ||_{L_2^{\psi^{-1}}(\Omega)} \\
&+ CS_{\hat{e}}^x || \frac{h^2}{\hat{\varepsilon}} \mathcal{R}_2 ||_{L_2^{\psi^{-1}}(\Omega)} + CS_{\hat{e}}^t || k_n\mathcal{R}_2 ||_{L_2^{\psi^{-1}}(\Omega)} \\
&+ CS_{\hat{e}}^t || k_n\mathcal{R}_1 ||_{L_2^{\psi^{-1}}(\Omega)},
\end{aligned}
\tag{5.17}
$$

where $\mathcal{R}_0$, $\mathcal{R}_1$,and $\mathcal{R}_2$ are defined as in chapter 5 and the stability factors $S_{\hat{e}}^t$, $S_{\hat{e}}^x$ are

defined by

$$S_{\hat{e}}^{t} = \frac{\|\varphi_t\|_{L_2^{\psi}(\Omega)}}{\|\hat{e}\|_{L_2^{\psi^{-1}}(\Omega)}},$$

and (5.18)

$$S_{\hat{e}}^{x} = \frac{\|\hat{\varepsilon}\varphi_{xx}\|_{L_2^{\psi}(\Omega)}}{\|\hat{e}\|_{L_2^{\psi^{-1}}(\Omega)}},$$

respectively.

5.4 Strong Stability Estimates

To give the above a posteriori error estimate a quantitative meaning, we need to determine approximations of the relevant strong stability factors $S_{\hat{e}}^{t}$ and $S_{\hat{e}}^{x}$. In this section we will consider the a posteriori error estimate (5.17) derived in the previous section based on the following dual problem

$$\mathcal{L}_{\hat{\varepsilon}}^{*}\varphi \equiv -\varphi_t - A^T\varphi_x - \hat{\varepsilon}\varphi_{xx} = \psi^{-1}e, \qquad \text{in } \Omega, \tag{5.19a}$$

$$\varphi(0,t) = 0, \qquad t \in (0,T), \tag{5.19b}$$

$$\varphi(x,T) = 0, \qquad x \in \mathbb{R}_+, \tag{5.19c}$$

assuming $\varphi(x,t) \longrightarrow 0$ for $x \longrightarrow \infty, \forall t$. We have the following main result:

Theorem 5.4. *Suppose $\hat{\varepsilon}$ is a positive constant, $\alpha \geq 0$ and $\beta \geq 0$ are constants. Suppose further that $\psi(x,t)$ is a positive weight function satisfying*

$$\psi_t + A^T\psi_x - A_x^T\psi - 2\hat{\varepsilon}\psi_x^2\psi^{-1} \geq \alpha|A_x^T|\psi + \beta\psi, \quad \text{in } \Omega, \tag{5.20}$$

then the solution φ of (5.19) satisfies

$$\begin{aligned}\|\varphi_t + A^T\varphi_x\|_{L_2^{\psi}(\Omega)}^2 + \|\hat{\varepsilon}\varphi_{xx}\|_{L_2^{\psi}(\Omega)}^2 + \|\hat{\varepsilon}^{\frac{1}{2}}\varphi_x\|_{L_\infty(L_2^{\psi}(\Omega))}^2 \\ +2\|\hat{\varepsilon}^{\frac{1}{2}}(\alpha|A_x^T| + \beta)^{\frac{1}{2}}\varphi_x\|_{L_2^{\psi}(\Omega)}^2 \leq 6\|\hat{e}\|_{L_2^{\psi^{-1}}(\Omega)}^2.\end{aligned} \tag{5.21}$$

Proof. We multiply equation (5.19a) by $-\psi(\varphi_t + A^T\varphi_x)$ and integrate over $\Omega_\tau = \mathbb{R}_+ \times (\tau, T)$ to get

$$\begin{aligned}\|\psi^{\frac{1}{2}}(\varphi_t + A^T\varphi_x)\|^2_{\Omega_\tau} + \int_{\Omega_\tau} \psi(\varphi_t + A^T\varphi_x)\hat{\varepsilon}\varphi_{xx}dxdt& \\ = -\int_{\Omega_\tau} e(\varphi_t + A^T\varphi_x)dxdt& \\ \leq \int_{\Omega_\tau} \psi^{-1}e^2dxdt + \frac{1}{4}\|\psi^{1/2}(\varphi_t + A^T\varphi_x)\|^2_{\Omega_\tau}.&\end{aligned}$$

Integrating by parts in both space and time variables in the second term on the left-hand side yields

$$\begin{aligned}&\int_{\Omega_\tau} \psi(\varphi_t + A^T\varphi_x)\hat{\varepsilon}\varphi_{xx}dxdt \\ &= \int_{\Omega_\tau} \psi\varphi_t\hat{\varepsilon}\varphi_{xx} + \int_{\Omega_\tau} \psi A^T\varphi_x\hat{\varepsilon}\varphi_{xx}dxdt \\ &= -\int_{\Omega_\tau} \hat{\varepsilon}\psi_x\varphi_x\varphi_t dxdt - \int_{\Omega_\tau} \hat{\varepsilon}\psi\varphi_x\varphi_{tx}dxdt \\ &- \frac{1}{2}\int_{\Omega_\tau} \hat{\varepsilon}\psi A^T_x\varphi^2_x dxdt - \frac{1}{2}\int_{\Omega_\tau} \hat{\varepsilon}\psi_x A^T\varphi^2_x dxdt \\ &= -\int_{\Omega_\tau} \hat{\varepsilon}\psi_x(\varphi_t + A^T\varphi_x)\varphi_x dxdt - \frac{1}{2}\int_{\Omega_\tau} (\hat{\varepsilon}\psi\varphi^2_x)_t dxdt \\ &+ \frac{1}{2}\int_{\Omega_\tau} +\hat{\varepsilon}\psi_t\varphi^2_x dxdt + \frac{1}{2}\int_{\Omega_\tau} \hat{\varepsilon}\varphi^2_x(A^T\psi_x - A^T_x\psi)dxdt.\end{aligned} \tag{5.22}$$

By substituting into above inequality we have

$$\begin{aligned}\frac{3}{4}\|\psi^{\frac{1}{2}}(\varphi_t + A^T\varphi_x)\|^2_{\Omega_\tau} + \frac{1}{2}\int_{\mathbb{R}_+} (\hat{\varepsilon}\psi\varphi^2_x)(x,\tau)dx& \\ + \frac{1}{2}\int_{\Omega_\tau} \hat{\varepsilon}\varphi^2_x(\psi_t + A^T\psi_x - A^T_x\psi)dxdt& \\ \leq \|\psi^{-\frac{1}{2}}e\|^2_{\Omega_\tau} + \int_{\Omega_\tau} \hat{\varepsilon}\psi_x\varphi_x(\varphi_t + A^T\varphi_x)dxdt& \\ \leq \|\psi^{-\frac{1}{2}}e\|^2_{\Omega_\tau} + \int_{\Omega_\tau} \hat{\varepsilon}^2\psi^2_x\varphi^2_x\psi^{-1}dxdt + \frac{1}{4}\|\psi^{\frac{1}{2}}(\varphi_t + A^T\varphi_x)\|^2_{\Omega_\tau}.&\end{aligned}$$

Collecting terms, we obtain that

$$\|\psi^{\frac{1}{2}}(\varphi_t + A^T\varphi_x)\|^2_{\Omega_\tau} + \int_{\mathbb{R}_+}(\hat{\varepsilon}\psi\varphi_x^2)(x,\tau)dx \\ + \int_{\Omega_\tau}\hat{\varepsilon}\varphi_x^2(\psi_t + A^T\psi_x - A_x^T\psi - 2\hat{\varepsilon}\psi_x^2\psi^{-1})dxdt \le 2\|\psi^{-\frac{1}{2}}e\|^2_{\Omega_\tau}.$$

We may now choose τ such that

$$\int_{\mathbb{R}_+}(\hat{\varepsilon}\psi\varphi_x^2)(x,\tau)dx = \sup_{t\in[0,T]}\int_{\mathbb{R}_+}(\hat{\varepsilon}\psi\varphi_x^2)(\cdot,t)dx = \|\hat{\varepsilon}^{\frac{1}{2}}\varphi_x\|^2_{L_\infty(L_2^\psi(\Omega_\tau)},$$

and use (5.20) to obtain

$$\|\hat{\varepsilon}^{\frac{1}{2}}\varphi_x\|^2_{L_\infty(L_2^\psi(\Omega_\tau)} + \|\hat{\varepsilon}^{\frac{1}{2}}(\alpha|A_x^T| + \beta)^{\frac{1}{2}}\varphi_x\|^2_{\Omega_\tau} \\ \le 2\|\psi^{-\frac{1}{2}}e\|^2_{\Omega_\tau} \le 2\|\psi^{-\frac{1}{2}}e\|_\Omega. \tag{5.23}$$

Choosing $\tau = 0$ yields the following inequality

$$\|\psi^{\frac{1}{2}}(\varphi_t + A^T\varphi_x)\|^2_\Omega + \|\hat{\varepsilon}^{\frac{1}{2}}(\alpha|A_x^T| + \beta)^{\frac{1}{2}}\varphi_x\psi^{\frac{1}{2}}\|^2_\Omega \\ \le 2\|\psi^{-\frac{1}{2}}e\|^2_\Omega. \tag{5.24}$$

Similarly, to obtain a bound for $\|\hat{\varepsilon}\psi^{\frac{1}{2}}\varphi_{xx}\|_\Omega$, multiply equation (5.19a) by $-\hat{\varepsilon}\psi\varphi_{xx}$ and integrating over Ω_τ to get

$$\|\hat{\varepsilon}\psi^{\frac{1}{2}}\varphi_{xx}\|^2_{\Omega_\tau} + \int_{\Omega_\tau}\hat{\varepsilon}\psi\varphi_{xx}(\varphi_t + A^T\varphi_x)dxdt \\ = -\int_{\Omega_\tau}\hat{\varepsilon}\varphi_{xx}e \le \|\psi^{-\frac{1}{2}}e\|^2_{\Omega_\tau} + \frac{1}{4}\|\hat{\varepsilon}\psi^{\frac{1}{2}}\varphi_{xx}\|^2_{\Omega_\tau}.$$

Using the integration by part for the second term in the left-hand side as in (5.22) and also using (5.20) we obtain

$$\|\hat{\varepsilon}\psi^{\frac{1}{2}}\varphi_{xx}\|^2_\Omega + \|\hat{\varepsilon}^{\frac{1}{2}}\psi^{\frac{1}{2}}(\alpha|A_x^T| + \beta)^{\frac{1}{2}}\varphi_x\|^2_\Omega \le 2\|\psi^{-\frac{1}{2}}e\|^2_\Omega. \tag{5.25}$$

Adding estimate (5.23), (5.24), (5.25) we have

$$\|\psi^{\frac{1}{2}}(\varphi_t + A^T\varphi_x)\|^2_\Omega + \|\hat{\varepsilon}\psi^{\frac{1}{2}}\varphi_{xx}\|^2_\Omega \\ + 2\|\hat{\varepsilon}^{\frac{1}{2}}(\alpha|A_x^T| + \beta)^{\frac{1}{2}}\varphi_x\psi^{\frac{1}{2}}\|^2_\Omega + \|\hat{\varepsilon}^{\frac{1}{2}}\varphi_x(\cdot,t)\|^2_{L_\infty(L_2^\psi(\Omega))} \le 6\|\psi^{-\frac{1}{2}}e\|^2_\Omega,$$

which proves the theorem. □

Corollary 5.5. *The assumption* (5.20) *of Theorem 5.1 is satisfied in the following basic cases:*

- $\psi \equiv 1$ *and* $A_x^T \leq 0$, $\beta = 0$, $\alpha = 1$, *corresponding to a shock,*
- $\psi(x,t) = (\frac{t}{T})^{1+\alpha}$, $\max A_x^T(\cdot,t) \leq \frac{1}{t}$, $\beta = 0$, $\alpha \geq 0$, *corresponding to rarefaction wave* $u(x,t) = \frac{x}{t}$, *and*
- $\psi = \exp(3\beta(t-T))$ *and* $|A_x^T| \leq \beta$, $\alpha = 1$, *corresponding to regular solution, see [57].*

Chapter 6

NUMERICAL EXPERIMENTS

6.1 Space-Time Discretization Procedure

To give numerical results obtained using the following Sd-method with $u^b \equiv 0$, that is applied to the our model problem: Find $u_h^n \in V_h^n$, such that for $n = 0, 1, \ldots, N-1$

$$\begin{aligned} &\Big(u_{h,t}^n + f_R'(u_h^n)u_{h,x}^n, v_h^n + \delta(v_{h,t}^n + f_R'(u_h^n)v_{h,x}^n)\Big)_n + \\ &< u_{h,+}^n, v_{h,+}^n >_n = < u_{h,-}^n, v_{h,+}^n >_n, \qquad \forall v_h^n \in W_h^n, \end{aligned} \tag{6.1}$$

we shall use finite element approximation on a space-time slab with the trial functions which are piecewise polynomials in space and linear in time; that is, for $(x,t) \in \mathbf{S}_n$, we let

$$u_h^n(x,t) = \sum_{i=1}^{M} \varphi_i(x)\big(\theta_1(t)\tilde{u}_i^n + \theta_2(t)u_i^{n+1}\big). \tag{6.2}$$

$\varphi_i(x) = \delta_{ij}$, $(j = 0, \ldots, M)$ is the spatial shape functions at node i; $\theta_1(t)$ and $\theta_2(t)$ are the time interpolation functions defined for the linear case as

$$\theta_1(t) = \frac{t_{n+1} - t}{t_{n+1} - t_n} = \frac{t_{n+1} - t}{k}, \text{and}$$
$$\theta_2(t) = \frac{t - t_n}{t_{n+1} - t_n} = \frac{t - t_n}{k},$$

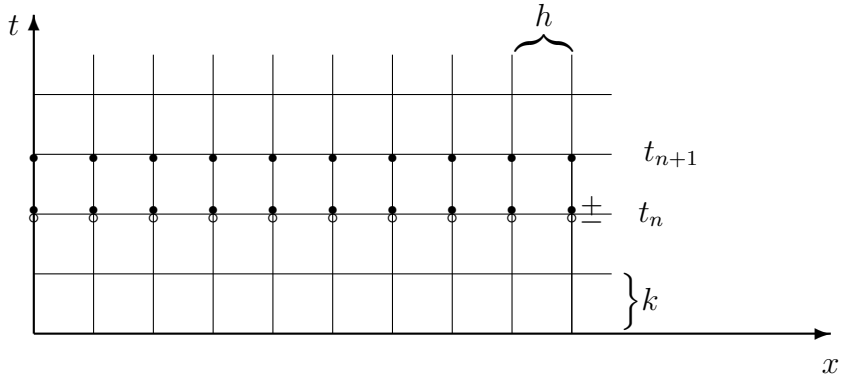

Figure 6.1: Uniform mesh

and the nodal values of u for node i at $(t_n)_+$ and $(t_{n+1})_-$ are denoted by $\tilde{u}_i^n$ and u_i^{n+1}, respectively.

The test functions v_h^n for each time slab are defined $\varphi_j(x)\theta_1(t)$ and $\varphi_j(x)\theta_2(t)$ for $j = 1, \ldots, M$. For each $n = 0, \ldots, N-1$, we have in this case that (6.1) is equivalent to the following nonlinear system of (difference) equations [24] with unknown nodal values $\tilde{u}_i^n$ and u_i^{n+1} at the points marked by solid circles in the above Figure. Here, also the known nodal values of u_i^n as data are indicated by open circle:

$$\sum_{i=1}^{M} \int_{\mathbf{S}_n} \Big[\varphi_i(x)\big(\frac{u_i^{n+1} - \tilde{u}_i^n}{k}\big) + f_R'(u^{n+1})\varphi_i'(x)\big(\theta_1(t)\tilde{u}_i^n + \theta_2(t)u_i^{n+1}\big)\Big] \times \Big[\varphi_j(x)\theta_1(t) + \delta\big((-\frac{1}{k})\varphi_j(x) + f_R'(u^{n+1})\varphi_j'(x)\theta_1(t)\big)\Big] dxdt = 0, \tag{6.3}$$

and

$$\sum_{i=1}^{M} \int_{\mathbf{S}_n} \Big[\varphi_i(x)\big(\frac{u_i^{n+1} - \tilde{u}_i^n}{k}\big) + f_R'(u^{n+1})\varphi_i'(x)\big(\theta_1(t)\tilde{u}_i^n + \theta_2(t)u_i^{n+1}\big)\Big] \times \Big[\varphi_j(x)\theta_2(t) + \delta\big((\frac{1}{k})\varphi_j(x) + f_R'(u^{n+1})\varphi_j'(x)\theta_2(t)\big)\Big] dxdt + \sum_{i=1}^{M} \int_{\mathbb{R}_+} \varphi_j(x)\varphi_i(x)\big(\tilde{u}_i^n - u_i^n\big)dx = 0, \tag{6.4}$$

for all $j = 1, \ldots, M$. Since $\theta_1(t_{n+1}) = 0$, the corresponding jump terms in (6.1) are became zero in (6.4). At each time t_n, $n = 1, \ldots, N-1$, if we compute $f'_R(u_h^n)$ at the previous time step, i.e., $f'_R(u_h^n = \sum_{i=1}^{M} \varphi_i(x)\big(\theta_1(t)\tilde{u}_i^n + \theta_2(t)u_i^{n+1}\big)) \simeq f'_R(u_i^n)$, thus we have a linear system of order $2M$ at each time step. Assuming piecewise linear φ_i functions, for $i = 1, \ldots, M$, i.e.,

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h}, & x \in [x_{i-1}, x_i], \\ \frac{x_{i+1} - x}{h}, & x \in [x_i, x_{i+1}], \\ 0 & \text{elsewhere}, \end{cases}$$

we have

$$\varphi_i'(x) = \begin{cases} \frac{1}{h}, & x \in [x_{i-1}, x_i], \\ -\frac{1}{h}, & x \in [x_i, x_{i+1}], \\ 0, & \text{elsewhere}. \end{cases}$$

Thus we can compute the entries of mass, stiffness and convection matrices, viz

$$M_{ij} = \int_{\mathbb{R}_+} \varphi_i(x)\varphi_j(x)dx = \begin{cases} \frac{2h}{3}, & j = i, \\ \frac{h}{6}, & j = i+1, i-1, \\ 0, & \text{elsewhere}, \end{cases}$$

$$N_{ij} = \int_{\mathbb{R}_+} \varphi_i'(x)\varphi_j'(x)dx = \begin{cases} \frac{2}{h}, & j = i, \\ -\frac{1}{h}, & j = i+1, i-1, \\ 0, & \text{elsewhere}, \end{cases}$$

$$K_{ij} = \int_{\mathbb{R}_+} \varphi_i(x)\varphi_j'(x)dx = \begin{cases} \frac{1}{2}, & j = i+1, \\ -\frac{1}{2}, & j = i-1, \\ 0, & \text{elsewhere}, \end{cases}$$

Further using the fact that

$$\int_{I_n} \theta_1^2(t)dt = \int_{I_n} \theta_2^2(t)dt = \frac{k}{3},$$
$$\int_{I_n} \theta_1(t)dt = \int_{I_n} \theta_2(t)dt = \frac{k}{2},$$
$$\int_{I_n} \theta_1(t)\theta_2(t)dt = \frac{k}{6},$$

we have the following equivalent forms for (6.3) and (6.4), respectively:

$$\begin{aligned}
\sum_{i=1}^{M} \Big\{ & \Big[\frac{1}{2}M_{ij} - \frac{\delta}{k}M_{ij} + \frac{\delta}{2}f_R'(u_i^n)K_{ij}\Big]u_i^{n+1} \\
& - \Big[\frac{1}{2}M_{ij} - \frac{\delta}{k}M_{ij} + \frac{\delta}{2}f_R'(u_i^n)K_{ij}\Big]\tilde{u}_i^n \\
& + \Big[\frac{k}{3}f_R'(u_i^n)(-K_{ij}) - \frac{\delta}{2}f_R'(u_i^n)(-K_{ij}) + \frac{\delta k}{3}\Big(f_R'(u_i^n)\Big)^2 N_{ij}\Big]\tilde{u}_i^n \\
& + \Big[\frac{k}{6}f_R'(u_i^n)(-K_{ij}) - \frac{\delta}{2}f_R'(u_i^n)(-K_{ij}) + \frac{\delta k}{6}\Big(f_R'(u_i^n)\Big)^2 N_{ij}\Big]u_i^{n+1} = 0,
\end{aligned} \tag{6.5}$$

and

$$\begin{aligned}
\sum_{i=1}^{M} \Big\{ & \Big[\frac{1}{2}M_{ij} + \frac{\delta}{k}M_{ij} + \frac{\delta}{2}f_R'(u_i^n)K_{ij}\Big]u_i^{n+1} \\
& - \Big[\frac{1}{2}M_{ij} + \frac{\delta}{k}M_{ij} + \frac{\delta}{2}f_R'(u_i^n)K_{ij}\Big]\tilde{u}_i^n \\
& + \Big[\frac{k}{6}f_R'(u_i^n)(-K_{ij}) + \frac{\delta}{2}f_R'(u_i^n)(-K_{ij}) + \frac{\delta k}{6}\Big(f_R'(u_i^n)\Big)^2 N_{ij}\Big]\tilde{u}_i^n \\
& + \Big[\frac{k}{3}f_R'(u_i^n)(-K_{ij}) + \frac{\delta}{2}f_R'(u_i^n)(-K_{ij}) + \frac{\delta k}{3}\Big(f_R'(u_i^n)\Big)^2 N_{ij}\Big]u_i^{n+1}\Big\} \\
& + \Big[M_{ij}\tilde{u}_i^n - M_{ij}u_i^n\Big]\Big\} = 0, \qquad \text{both for } j = 1, \ldots, M.
\end{aligned} \tag{6.6}$$

We can rewrite the above equations in the following matrix forms:

$$\begin{aligned}
\Big[(\frac{1}{2} - \frac{\delta}{k})M + (\delta - \frac{k}{6})K_f + \frac{\delta k}{6}N_f\Big]U^{n+1} \\
+\Big[(-\frac{1}{2} + \frac{\delta}{k})M - \frac{k}{3}K_f + \frac{\delta k}{3}N_f\Big]\tilde{U}^n = 0, \\
n = 0, \ldots, N-1,
\end{aligned}$$

and

$$\begin{aligned}&\Big[(\frac{1}{2}+\frac{\delta}{k})M-\frac{k}{3}K_f+\frac{\delta k}{3}N_f\Big]U^{n+1}\\+&\Big[(\frac{1}{2}-\frac{\delta}{k})M-(\delta+\frac{k}{6})K_f+\frac{\delta k}{6}N_f\Big]\tilde{U}^n=MU^n,\\&\qquad n=0,\ldots,N-1,\end{aligned}$$

where

$$\begin{aligned}U^n&=\Big[u_1^n,\ldots,u_M^n\Big]^T \quad \text{is the given data at the time level } n,\\U^{n+1}&=\Big[u_1^{n+1},\ldots,u_M^{n+1}\Big]^T,\qquad \tilde{U}^n=\Big[\tilde{u}_1^n,\ldots,\tilde{u}_M^n\Big]^T,\end{aligned}$$

and

$$\left.\begin{aligned}(K_f)_{ij}&=f_R'(u_j^n)K_{ji},\\(N_f)_{ij}&=\Big[f_R'(u_j^n)\Big]^2N_{ji}.\end{aligned}\right\}\quad \text{for}\quad i,j=1,\ldots,M.$$

Similarly, we can obtain the equation systems for left side of domain, using f_L'. Solving the resulting $4M\times 4M$ linear systems of equations we obtain the solution at each time step.

6.2 Numerical Results in the Linear Case

To show the performance of the previously describe method, we start with the linear form of our problem (2.1)-(2.3). By putting $f_\alpha(u)=a_\alpha u$ $(\alpha=L,R)$ we get the following coupling of two advection equations:

$$\begin{cases}u_t+a_Ru_x &=0, & x>0, & t>0,\\ u_t+a_Lu_x &=0, & x<0, & t>0,\\ u(x,0) &=u_0(x), & x\in[-a,a], &\\ u(-a,t) &=g(t), & & t>0,\\ u(a,t) &=h(t), & & t>0,\end{cases}$$

where $a > 0$. This problem has the explicit solution

$$u(x,t) = \begin{cases} u_0(x - a_R t), & x \in (0, a] \\ u_0(x - a_L t), & x \in [-a, 0) \end{cases}$$

for all $t > 0$ which is a weak solution (at least if u_0 is "smooth" enough).

6.2.1 Test problem 1

We, first consider the following problem

$$\begin{cases} u_t + u_x &= 0, \qquad -1 < x \leq 0, \quad t > 0, \\ u_t - u_x &= 0, \qquad 0 \leq x < 1, \quad t > 0, \\ u(x,0) &= \begin{cases} 0.5 + x & \text{if} \quad -0.5 \leq x < 0.5, \\ 0 & \text{if} \quad o.w \end{cases} \end{cases}$$

with the boundary conditions

$$u(-1,t) = u(1,t) = 0.$$

Figures 6.2 and 6.3 shows the numerical solution using streamline diffusion method for $h = 0.001$ and $k/h = 0.5$. The results are given after 2, 51, 201 and 501 time steps with $\delta = h$ and $\delta = 0$. The exact solution is represented by the solid line.

6.2.2 Test problem 2

We consider the above problem with the following discontinuous initial condition

$$u_0(x) = \begin{cases} 0, & x \leq 0.25, \\ 1, & 0.25 < x \leq 0.5, \\ 0, & x > 0.5, \end{cases}$$

with the boundary conditions $u(-1,t) = u(1,t) = 0$ and the parameters $a_L = 3$ and $a_R = -2$, over the computational domain $[-1, 1]$. Figures 6.4 and 6.5 shows the

numerical solution using streamline diffusion method for $h = 0.001$ and $k/h = 0.5$. The results are given after 2, 51, 201 and 501 time steps with $\delta = h$ and $\delta = 0$. The exact solution is represented by the solid line.

6.3 Conclusion and Observation

All the numerical experiments show that, for the considered problems, the performance of the Sd-method ($\delta = h$) is better than Galerkin method ($\delta = 0$) in terms of quality of stability. The Galerkin method for these problems is unstable, because we see that the solution is oscillating, whereas the Sd-method the solution is very close to exact solution and is not oscillating or oscillate slightly. Therefore these observation confirm the theory and show the advantages of the Sd-method.

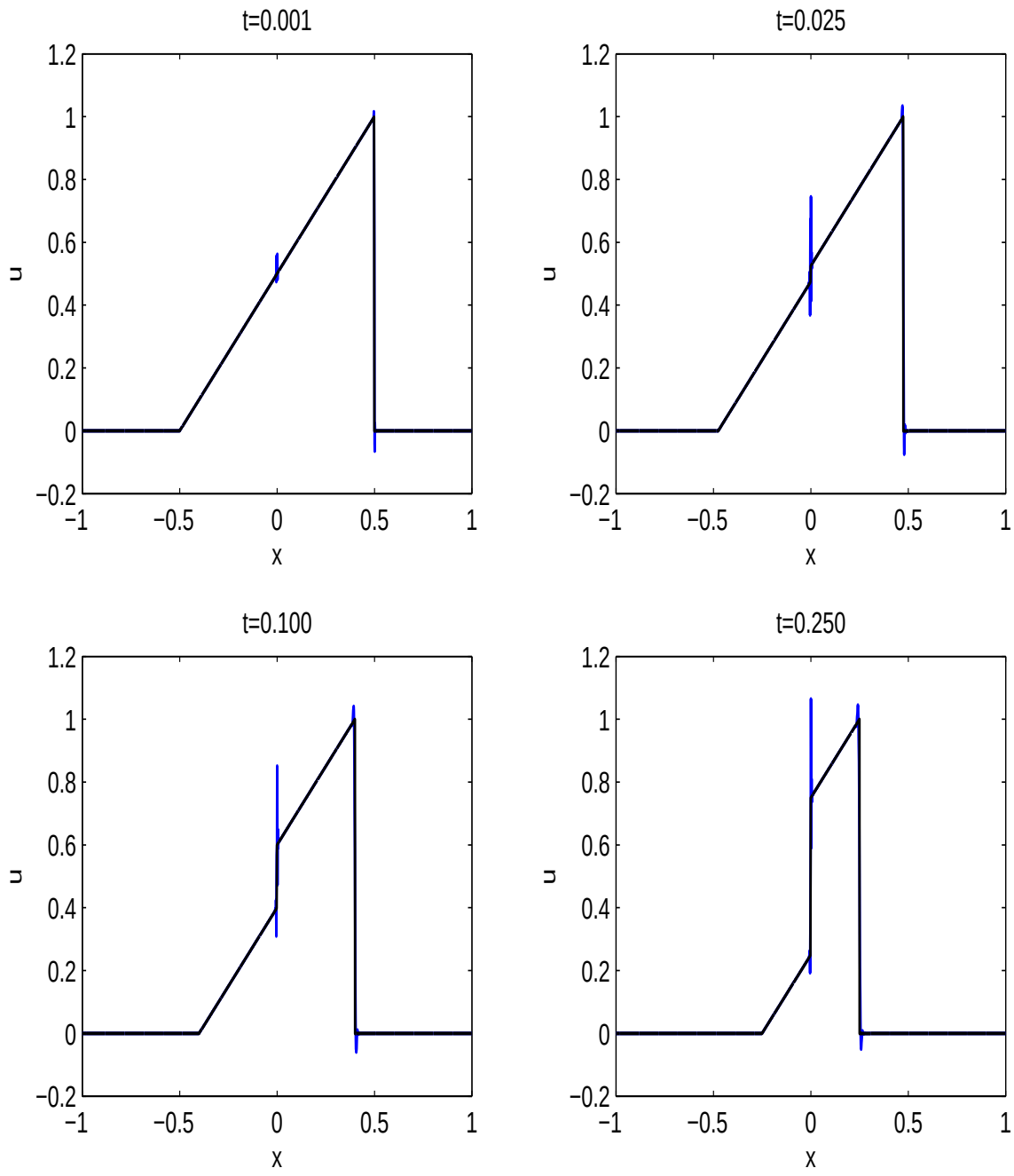

Figure 6.2: Streamline diffusion method for $h = 0.001$, $k = 0.0005$, $a_L = 1$ and $a_R = -1$. The numerical solution is given, first row: after 2 times (left), after 51 times (right) and second row: after 201 times (left), after 501 times (right). The exact solution is given by solid line.

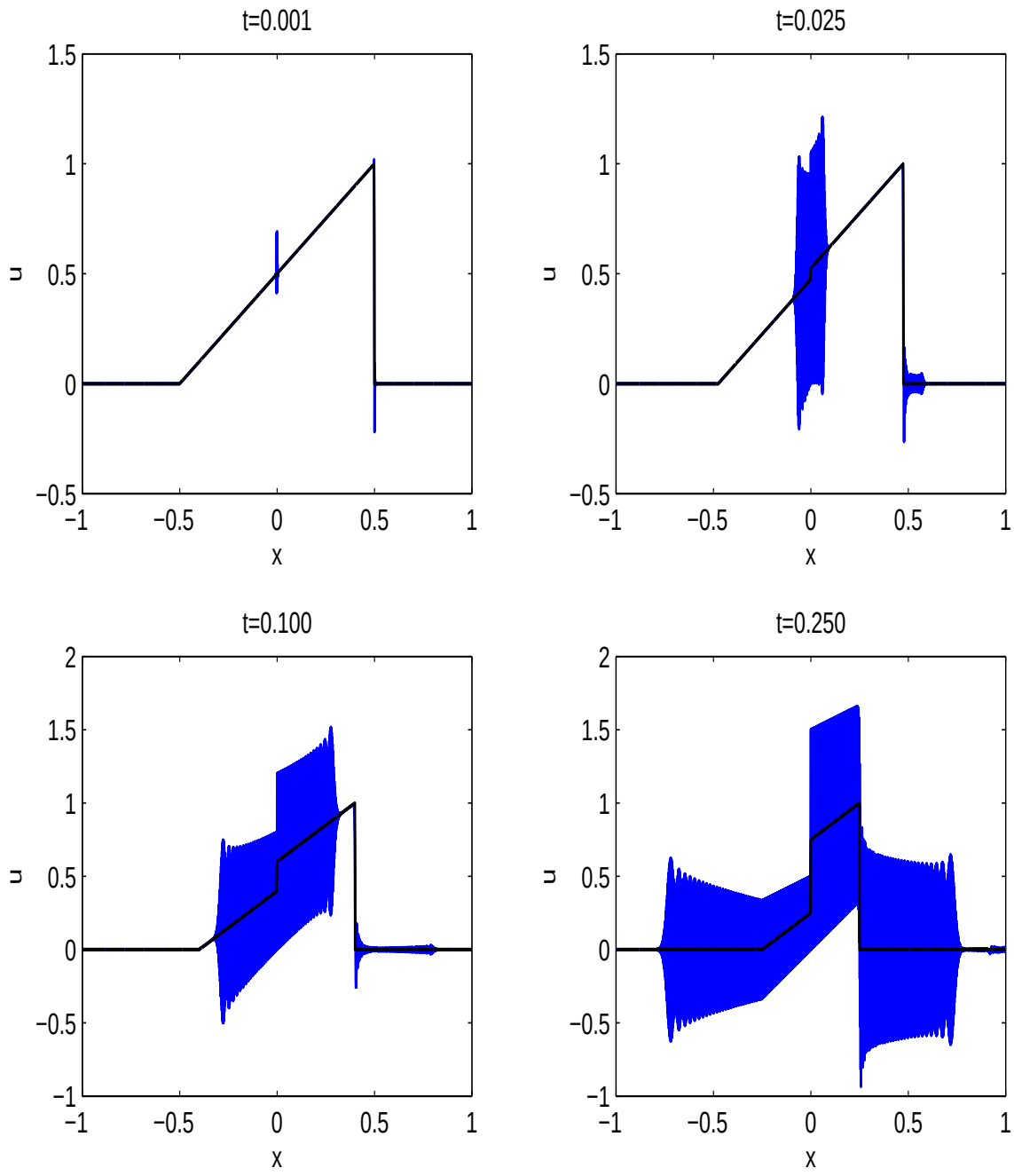

Figure 6.3: Streamline diffusion method for $h = 0.001$, $k = 0.0005$, $a_L = 1$ and $a_R = -1$. The numerical solution is given, first row: after 2 times (left), after 51 times (right) and second row: after 201 times (left), after 501 times (right). The exact solution is given by solid line.

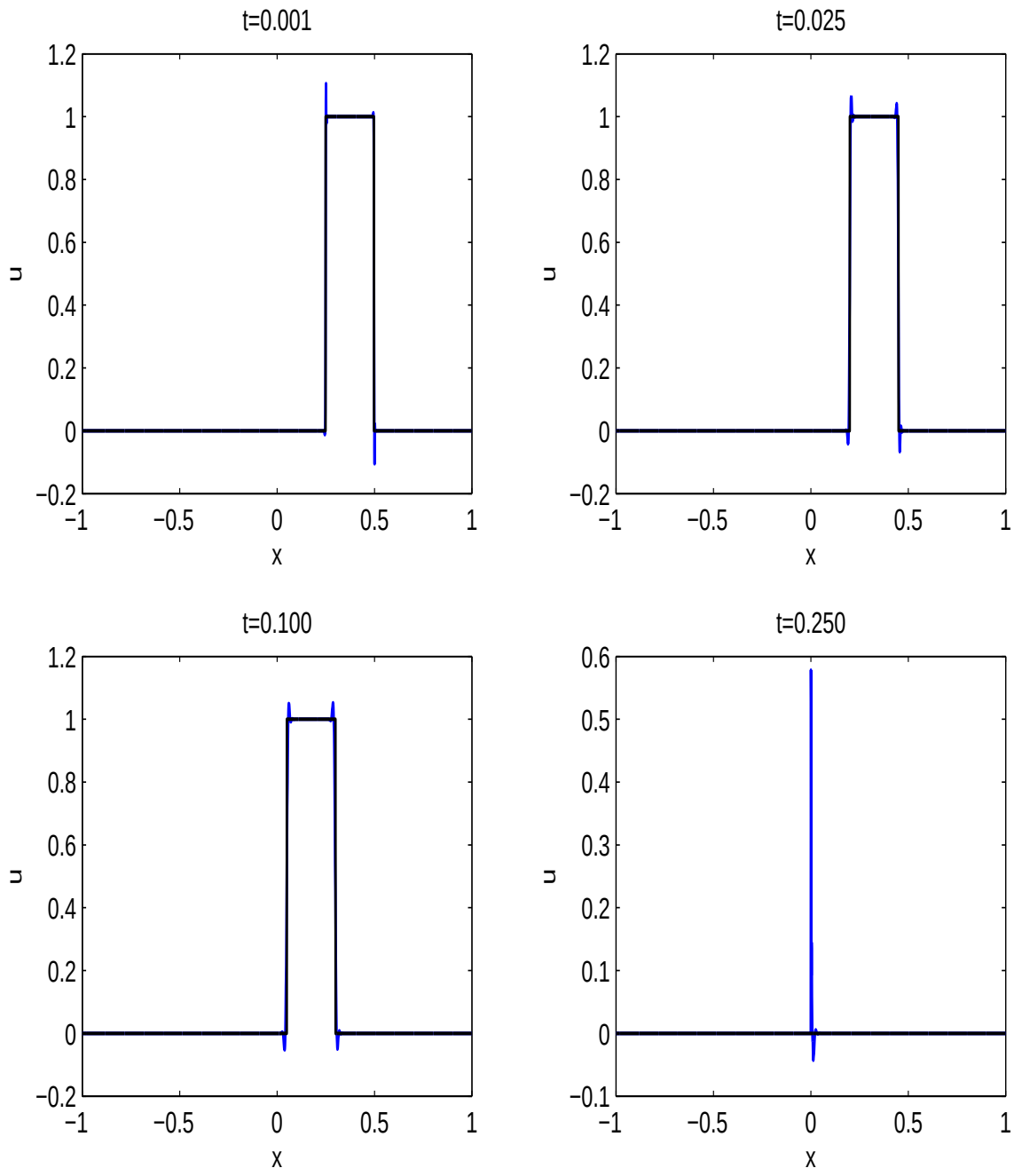

Figure 6.4: Streamline diffusion method for $h = 0.001$, $k = 0.0005$, $a_L = 3$ and $a_R = -2$. The numerical solution is given, first row: after 2 times (left), after 51 times (right) and second row: after 201 times (left), after 501 times (right). The exact solution is given by solid line.

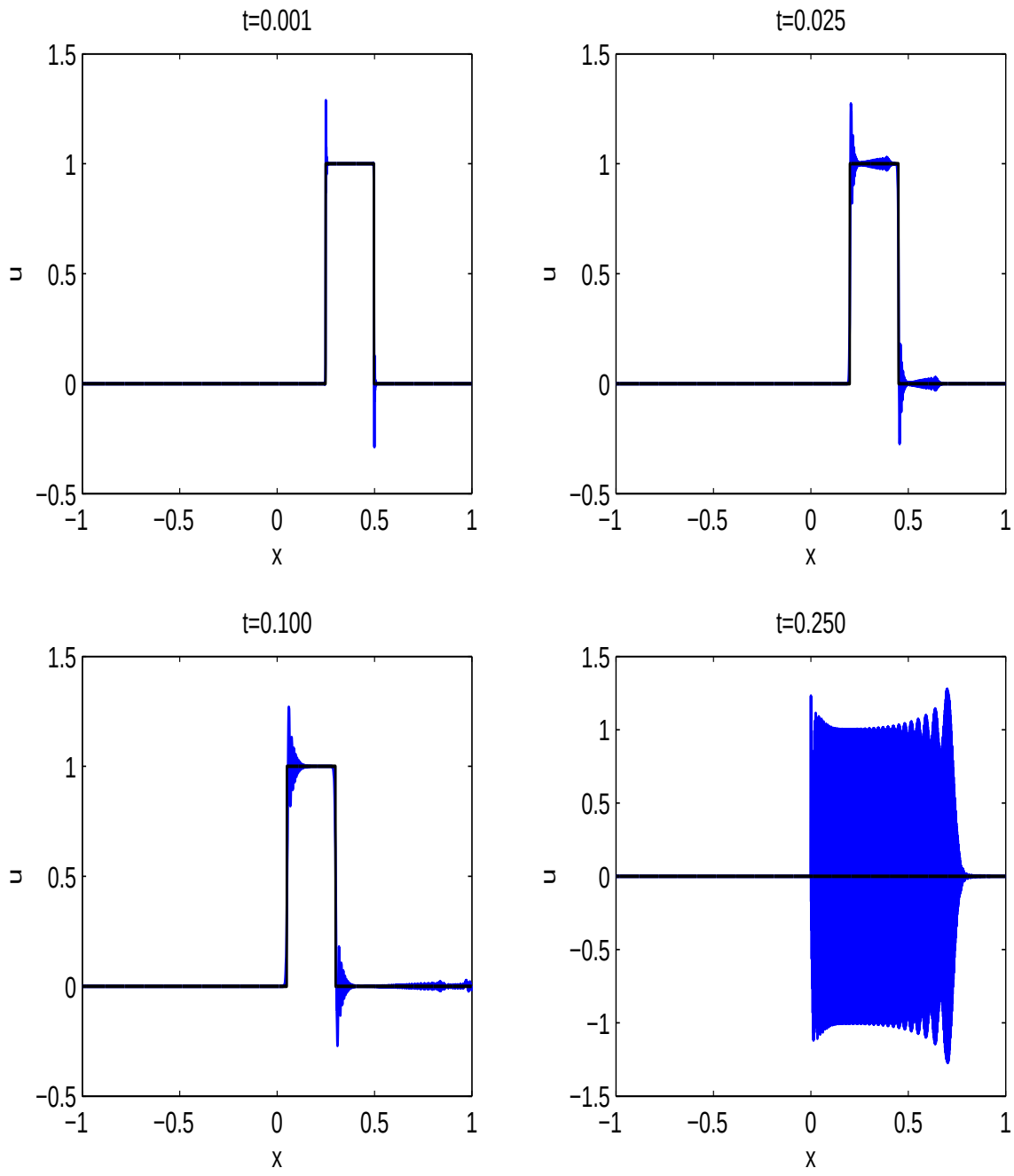

Figure 6.5: Streamline diffusion method for $h = 0.001$, $k = 0.0005$, $a_L = 3$ and $a_R = -2$. The numerical solution is given, first row: after 2 times (left), after 51 times (right) and second row: after 201 times (left), after 501 times (right). The exact solution is given by solid line.

Bibliographi

[1] R.A. Adams, **Sobolev Spaces**, Academic Press, New York 1975.

[2] M. Ainsworth and J. Tinsley Oden, **A Posteriori Error Estimation in Finite Element Analysis**. Wiley - Interscience, NY, 2000.

[3] J.H. Argyris, *Energy Theorems and Structural Analysis*, Part 1, Aircraft Eng., **26** (1954), pp. 383-394.

[4] J.H. Argyris and S. Kelsey, *Energy Theorems and Structural Analysis*, Butterworths, London, 1960.

[5] M. Asadzadeh, *Streamline Diffusion Methods for the Vlasov-Poisson Equations*, RAIRO Mathematical Modelling and Numerical Analysis , **24** (1990), pp. 177-196.

[6] M. Asadzadeh, *Streamline Diffusion Methods for the Fermi and Fokker-Planck Equations*, Transport Theory and Statistical Physics, (1997), pp. 319-340.

[7] M. Asadzadeh, *A posteriori Error Estimates for the Fokker-Planck and Fermi Pencil Beam Equations*, Math. Models and Methods in Appl. Sci., **10** (2000), pp. 737-769.

[8] M. Asadzadeh and P. Kowalczyk, *Convergence of Streamline Diffusion Methods for the Vlasov-Poisson-Fokker-Planck System*, Numer. Methods. in Partial Differential Eq., **21** (2005), pp. 472-495.

[9] M. Asadzadeh and A. Sopasakis, *Convergence of a hp streamline diffusion scheme for the Vlasov-Fokker-Planck System*, preprint (2003).

[10] M. Asadzadeh, PDE Lecture notes, `http://www.math.chalmers.se/~mohammad`

[11] M. Asadzadeh and K. Eriksson, *On adaptive finite element method for Fredholm integral equations of the second kind*, SIAM J. Numer. Anal., **31** (1994), pp. 831-855.

[12] M. Asadzadeh, *On the stability of characteristic schemes for the Fermi equation*, Appl. Comput. Math., **1** (2002), pp. 158-174.

[13] I. Babuška and W.C. Rheinboldt, *A Posteriori Error Estimate for the Finite Element Method*, SIAM J. Numer. Anal., **15** (1978), pp. 1597-1615.

[14] I. Babuška and W.C. Rheinboldt, *Adaptive Approaches and Reliability Estimates in Finite element Analysis*, Comp. Meth. Appl. Mech. Eng., **17/18** (1979), pp. 519-540.

[15] C. Bardos, A.-Y. Le Roux and J.-C. Nédéléc, *First order quasilinesr equations with boundary conditions, Comm. in Partial Differential equations*, **4** (1979), pp. 1017-1034.

[16] S.C. Brenner and L.R. Scott, **The Mathematical Theory of Finite Element Method**, Springer-Verlag, New York, 1994.

[17] F. Brezzi, C. Canuto, A. Russo: *A self-adaptive formulation for the Euler/Navier Stokes coupling*, Comp. Meth. Appl. Mech. Eng., **73** (1989), pp. 317-330.

[18] E. Burman, Adaptive Finite Element Methods for Compressible Two-Phase Flow, Math. Mod. Meth. App. Sci., **10** (2000), pp. 963-989.

[19] E. Burman, *Adaptive Finite Element Methods for Compressible Flow*, Comput. Meth. Appl. Mech. Engrg., **190** (2000), pp. 1137-1162.

[20] P.G. Ciarlet, **The Finite Element Method for Elliptic Problems**, Amesterdam, North Holland, 1987.

[21] R.W. Clough, *The Finite Element method in Plane Stress Analysis*, In: Proc. *2nd* ASCE Conf. on Electronc Comp, Pittsburgh Pa, September 1960.

[22] R.W. Clough and E.W. Wilson, *Stress Analysis of a Gravity Dam by the Finite Element Method*, In: Proc. Symp. on the use of Computers in the Civil Eng. Lab., Lisbon, Portugal, 1962. Nacional de Engenharia Civil.

[23] R. Courant, *Variational Method for the Solution of Problems of Equilibrium and Vibration*, Bulletin of the American Mathematical Society, **49** (1943), pp. 1-23.

[24] J. Donea and A. Huerta, **Finite Element Method for Flow Problems**, John Willy, 2003.

[25] F. Dubois and P.G. LeFloch, *Boundary conditions for nonlinear hyperbolic systems of conservation laws*, J. Differential Equations **71** (1988), pp. 93-122.

[26] K. Eriksson, C. Johnson, *An adaptive Finite element method for linear elliptic problems*, Math. Comput., **50** (1988), pp. 361-383.

[27] K. Eriksson, C. Johnson, *Adaptive Finite element methods for parabolic problems I: A linear model problem*, SIAM J. Numer. Anal., **28** (1991), pp. 43-77.

[28] K. Eriksson, C. Johnson, *Adaptive Finite element methods for parabolic problems II: Optimal error estimates in $L\infty L2$ and $L\infty L\infty$*, SIAM J. Numer.Anal. **32** (1995), pp. 706-740.

[29] K. Eriksson, C. Johnson, *Adaptive Finite element methods for parabolic problems IV: Non-linear problems*, SIAM J. Numer. Anal. **32** (1995), pp. 777-788.

[30] K. Eriksson, C. Johnson, *Adaptive Streamline Diffusion Finite Element Methods for Stationary Convection-Diffusion Problems*, Math. Comp. **60** (1993), pp.167-188.

[31] K. Eriksson, D. Estep, P. Hansbo and C. Johnson, *Introduction to Adaptive Methods for Differential Equations*, Acta Numerica (1995), pp. 105-158.

[32] L.C. Evans, **Partial Differential Equations**, AMS, 1998.

[33] G.B. Folland, **Introduction to Partial Differential Equations**, Princeton University Press, 1976.

[34] K.O. Friedrichs, *A Finite Difference Scheme for the Neumann and Dirichlet Problem*, Technical report University Rep. TID 4500, NYO 9760, ACE Computing and Applied Mathemaics Center, Courant Institute of Mathematical Sciences, New York, 1962.

[35] C. Führer and R. Rannacher, *An Adaptive Streamline Diffusion Finite Element Method for Hyperbolic Conservation Laws*, East-West J. Numer. Math., **5** (1997), pp. 145-162.

[36] K. Gillow, **Applications of Adaptive Finite Element Methods to Problem In Electrochemistry**, Ph.D. Thesis, Oxford University Computing Laboratory, 2000.

[37] E. Godlewski and P.-A. Raviart, *The Numerical Interface Coupling of Nonlinear Hyperbolic Systems of Conservation laws: I. The Scalar Case*, Numer. Math. **97** (2004), pp. 81-130.

[38] E. Godlewski, K.-C. Le Thanh and P.-A. Raviart, *The Numerical Interface Coupling of Nonlinear Hyperbolic Systems of Conservation laws: II. The System Case*, Preprints of the Laboratoire Jacques-Louis Lions 2004.

[39] E. Godlewski and P.-A. Raviart,**Numerical Approximation of Hyperbolic Systems of Conservation laws**, Appl. Math. Science 118, Springer-Verlag, New York 1996.

[40] P. Hansbo, Lecture note: Beyond the Elements of Finite Elements: General Principles for Solid and Fluid Mechanics Applications.

[41] P. Hansbo, *The Characteristic Streamline Diffusion Method for Convection-Diffusion Problems*, Comput. Methods Appl. Mech. Engrg., **96** (1992), pp. 239-253.

[42] P. Hansbo, *The Characteristic Streamline Diffusion Method for the Time-Dependent Incompressible Navier-Stokes Equations*, Comput. Methods Appl. Mech. Engrg., **99** (1992), pp. 171-186.

[43] P. Hansbo, *Explicit Streamline Diffusion Finite Element Methods for Compressible Euler Equations in Conservation Variables*, J. Comput. Phys., **109** (1993), pp. 274-288.

[44] P. Houston and E. Süli, *Adaptive Lagrange-Galerkin Methods for Unsteady Convection-Dominated Diffusion Problems*, Math. Comp. **70** (2001), pp. 77-106.

[45] T.J.R. Hughes and A. Brooks, *A Multidimensional upwind scheme with no Crosswind diffusion*, In: T.J.R. Hughes, eds., Finite Element methods for Convection Dominated flows, AMD-vol-**34** (1979), pp.19-35.

[46] T.J.R. Hughes and A. Brooks, *Streamline Upwind/petrov-Galerkin Formulation for Convection Dominated Flows with particular Emphasis on the Incompressible Navier-Stokes Equations*, Comput. Meth. Appl. Mech. Engrg., **32** (1982), pp. 199-259.

[47] C. Johnson, **Numerical Solution of Partial Differential Equations by the Finite Element Method**, Cambridge University press, 1987.

[48] C. Johnson, *Finite Element Methods for Convection-Diffusion Problems*, In: R. Glowinski and J.L.Lions, eds., In Computing Method in Engineering and Applied Sciences, North-Holland, Amsterdam, 1981.

[49] C. Johnson, *A new paradigm for adaptive Finite element methods*, in Proc. Mafelap 93, Brunel Univ., Wiley. Comput. Methods Appl. Mech. Engrg., **107** (1993), pp. 117-129.

[50] C. Johnson, *Discontinuous Galerkin Finite Element Methods for Second Order Hyperbolic Problems*, Comput. Meth. Appl. Mech. Engrg., **107** (1993), pp. 117-129.

[51] C. Johnson, *The Streamline Diffusion Finite Element Method for Compressible and Incompressible Fluid Flow*, Numerical Analysis 1989 (Dundee), Pitman Res. Notes Math. Ser., 228, Longman Sci. Tech., Harlow, 1990.

[52] C. Johnson and U. Nävert, *An Analysis of some Finite Element Methods for Advection-Diffusion*, In: O. Axelesson, L.S.Frank and A. Van Der Sluis, Analytical and Numerical Approaches to Asymptotics Problems in Analysis, North-Holland, Amsterdam, 1988.

[53] C. Johnson and U. Nävert and J. Pitkäranta, *Finite Element Method for Linear Hyperbolic Problems*, Comput. Methods Appl. Mech. Engrg. **45** (1984), pp. 285-312.

[54] C. Johnson and J. Saranen, *Streamline Diffusion Methods for the Incompressible Euler and Navier-Stokes Equations*, Math. Comp. **47** (1986), pp. 1-18.

[55] C. Johnson and A. Szepessy, *On the Convergence of a Finite Element Method for a Nonlinear Hyperbolic Conservation Law*, Math. Comp. **49** (1987), pp. 427-444.

[56] C. Johnson, A. Szepessy and P. Hansbo, *On the Convergence of Shock-Capturing Streamline Diffusion Finite Element Methods for Hyperbolic Conservation Law*, Math. Comp. **54** (1990), pp. 107-129.

[57] C. Johnson and A. Szepessy, *Adaptive Finite Element Methods for Conservation Laws Based on a posteriori Error Estimates*, Comm. Pure. Appl. Math. XLVIII **48** (1995), pp. 199-234.

[58] H.P. Langtangen. **Computational Partial Differential Equation. Numerical Methods and Diffpack Programming**, Texts in Computational Science and Engineering, Vol. 1 2nd ed., 1999.

[59] M.G. Larson and J. Ivarsson, *Computational a posteriori error estimation for parabolic problems*, Preprint, Chalmers Finite Element Center, Chalmers University of Technology, Sweden, 2001.

[60] M.G. Larson and A.J. Niklasson, *A Posteriori Error Estimation of Functionals in Elliptic Problems: Experiments*, Preprint 04, Chalmers Finite Element Center, Chalmers University of Technology, Sweden, 2001. to appear in M^3AS.

[61] S. Larsson and V. Thomée, **Partial Differential Equations with Numerical Method**, Springer-Verlag, 2003.

[62] P. LeFloch, *Explicit Formula for Scalar Nonlinear Conservation Laws with Boundary Conditions*, Math. Methods in the Appl. Sci., **10** (1988), pp. 265-287.

[63] U. Nävert, **A Finite Element Method for Convection-Diffusion Problems**, Ph.D. Thesis Department of Computer Science, Chalmers University of Technology, Göteborg, 1982.

[64] L.A. Oganesjan, *Numerical Solution of Plates*, In: Solution of Engineering Problems on Computers, Lensovnarchoz CBTI, Sbronik, 1663. (in Russian).

[65] A.Quarteroni and A. Valli, **Numerical Approximation of Partial Differential Equations**, Springer-Velarge, 1997

[66] R. Rannacher and G. Zhou, *Mesh Adaptation via a Predictor-Corrector Strategy in the Streamline Diffusion Method for Nonstationary Hyperbolic Systems*, In: Notes on Numerical Fluid Mechanic, **46** (1994), pp. 236-250.

[67] R. Sandboge, **Adaptive Finite Element Methods for Reactive Flow Problems**, Ph.D. Thesis, Department of Mathematics, Chalmers University of Technology, Göteborg , 1996.

[68] R. Sandboge, *Adaptive Finite Element Methods for Systems of Reaction-Diffusion Equations*, Comput. Methods Appl. Mech. Engrg. **166** (1998), pp. 309-328.

[69] E. Süli, *A Posteriori Error Analysis and Global Error Control for Adaptive Finite Element Approximations of Hyperbolic Problems*. In D. F. Griffiths and G. A. Watson, editors. Numerical Analysis. Proceedings of the 16th Biennial Conference in Numerical Analysis, Dundee, June 1995, pp. 169-190, Pitman, 1996.

[70] E. Süli and P. Houston, *Finite Element Methods for Hyperbolic Problems: A Posteriori Error Analysis and Adaptivity*, In: I. Duff and G.A. Watson, eds., State of the Art in Numerical Analysis, pp. 441-471. Oxford University Press, 1997.

[71] A. Szepessy, **Convergence of Streamline Diffusion Finite Element Method for Conservation Law**, Ph.D. Thesis, Department of Mathematics, Chalmers University of Technology, Göteborg, 1989.

[72] A. Szepessy, *Measure-Valued Solution of Scalar Conservation Laws with Boundary Conditions*, Arch. Rat. Mech. Anal., **107** (1989), pp. 181-193.

[73] M.J. Turner, R.W. Clough, H.C. Martin and L.T. Topp, *Stiffness and Deflection Analysis of Complex Structures*, J. Aeronaut. Sci., **25** (1956), pp. 805-823.

[74] M. Zlámal, *On the Finite Element Method*, Numer. Math., **12** (1968), pp. 394-409.

[75] O.C. Zienkiewicz and Y.K. Cheung, **The finite element method in Continuum and Structural Mechanics**, MCGraw Hill, New York, 1965.

[76] O.C. Zienkiewicz and R.L. Taylor, **The finite element method. Vol. 1. The basis.** Fifth edition. Butterworth-Heinemann, Oxford, 2000.

[77] O.C. Zienkiewicz and R.L. Taylor, **The finite element method. Vol. 2. Solid Mechanics.** Fifth edition. Butterworth-Heinemann, Oxford, 2000.

[78] O.C. Zienkiewicz and R.L. Taylor, **The finite element method. Vol. 3. Fluid dynamics.** Fifth edition. Butterworth-Heinemann, Oxford, 2000.

Index